Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification

Dan Lu,[1] Ming Ye,[1] and Mary C. Hill[2]

Received 16 August 2011; revised 27 June 2012; accepted 4 July 2012; published 14 September 2012.

[1] Confidence intervals based on classical regression theories augmented to include prior information and credible intervals based on Bayesian theories are conceptually different ways to quantify parametric and predictive uncertainties. Because both confidence and credible intervals are used in environmental modeling, we seek to understand their differences and similarities. This is of interest in part because calculating confidence intervals typically requires tens to thousands of model runs, while Bayesian credible intervals typically require tens of thousands to millions of model runs. Given multi-Gaussian distributed observation errors, our theoretical analysis shows that, for linear or linearized-nonlinear models, confidence and credible intervals are always numerically identical when consistent prior information is used. For nonlinear models, nonlinear confidence and credible intervals can be numerically identical if parameter confidence regions defined using the approximate likelihood method and parameter credible regions estimated using Markov chain Monte Carlo realizations are numerically identical and predictions are a smooth, monotonic function of the parameters. Both occur if intrinsic model nonlinearity is small. While the conditions of Gaussian errors and small intrinsic model nonlinearity are violated by many environmental models, heuristic tests using analytical and numerical models suggest that linear and nonlinear confidence intervals can be useful approximations of uncertainty even under significantly nonideal conditions. In the context of epistemic model error for a complex synthetic nonlinear groundwater problem, the linear and nonlinear confidence and credible intervals for individual models performed similarly enough to indicate that the computationally frugal confidence intervals can be useful in many circumstances. Experiences with these groundwater models are expected to be broadly applicable to many environmental models. We suggest that for environmental problems with lengthy execution times that make credible intervals inconvenient or prohibitive, confidence intervals can provide important insight. During model development when frequent calculation of uncertainty intervals is important to understanding the consequences of various model construction alternatives and data collection strategies, strategic use of both confidence and credible intervals can be critical.

Citation: Lu, D., M. Ye, and M. C. Hill (2012), Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, *Water Resour. Res.*, *48*, W09521, doi:10.1029/2011WR011289.

1. Introduction

[2] Environmental modeling is often used to predict effects of future anthropogenic and/or natural occurrences. Predictions are always uncertain. Epistemic and aleatory prediction uncertainty is caused by data errors and scarcity, parameter uncertainty, model structure uncertainty, and scenario uncertainty [*Morgan and Henrion*, 1990; *Neuman and Wierenga*, 2003; *Meyer et al.*, 2007; *Renard et al.*, 2011; *Clark et al.*, 2011]. Here we mostly explore the propagation of parameter uncertainty into measures of prediction uncertainty using two common measures: individual confidence intervals based on classical regression theories (*Draper and Smith* [1998] and *Hill and Tiedeman* [2007] show how prior information can be included in regression theories) and individual credible intervals (also known as probability intervals) based on Bayesian theories [*Box and Tiao*, 1992; *Casella and Berger*, 2002]. On the other hand, alternative models are also constructed in this study to allow consideration of these measures of parameter uncertainty in the context of model structure uncertainty.

[3] Confidence intervals are of interest because they generally can be calculated using tens to thousands of model runs versus the tens of thousands to millions of model runs needed to calculate credible intervals. While computational cost for evaluating Bayesian credible intervals can be reduced by using surrogate modeling such as response surface surrogates and low-fidelity physically based surrogates [*Razavi et al.*, 2012], surrogate modeling generally requires thousands of model runs to establish the response surface surrogate and the

[1]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.
[2]U.S. Geological Survey, Boulder, Colorado, USA.

Corresponding author: M. C. Hill, U.S. Geological Survey, 3215 Marine St., Boulder, CO 80303, USA. (mchill@usgs.gov)

0043-1397/12/2011WR011289

surrogate model. Considering the utility of frequently calculating uncertainty measures to evaluate the consequences of various model construction alternatives and data collection strategies [*Tiedeman et al.*, 2003, 2004; *Dausman et al.*, 2010; *Neuman et al.*, 2012; *Lu et al.*, 2012], less computationally demanding methods are appealing.

[4] While both confidence and credible intervals have been used in environmental modeling [*Schoups and Vrugt*, 2010; *Krzysztofowicz*, 2010; *Aster et al.*, 2012], it is not possible to provide a literature review and examples from such a large field. Here, we focus on groundwater modeling for the literature review below and for the complex synthetic test case considered. Since groundwater modeling and many other fields of environmental modeling share similar modeling protocols and mathematical and statistical methods for uncertainty quantification and interpretation, guidelines for using confidence and credible intervals drawn from the theoretical analyses and test cases considered in this work are expected to be generally applicable to other fields of environmental modeling.

[5] Among the pioneering groundwater works, *Cooley* [1977, 1979, 1983], *Yeh and Yoon* [1981], *Carrera and Neuman* [1986], and *Yeh* [1986] used linear confidence intervals based on observations and prior information; *Cooley and Vecchia* [1987] developed a method for estimating nonlinear confidence intervals; *Kitanidis* [1986] presented a Bayesian framework of quantifying parameter and predictive uncertainty. More recent development and applications of the intervals can be found in *Sciortino et al.* [2002], *Tiedeman et al.* [2003, 2004], *Cooley* [2004], *Montanari and Brath* [2004], *Samanta et al.* [2007], *Hill and Tiedeman* [2007], *Gallagher and Doherty* [2007], *Montanari and Grossi* [2008], *Dausman et al.* [2010], *Li and Tsai* [2009], *Wang et al.* [2009], *Parker et al.* [2010], *Schoups and Vrugt* [2010], and *Fu et al.* [2011], among others. The study of confidence and credible intervals in the context of multimodel analysis has been considered by *Neuman* [2003], *Ye et al.* [2004], *Poeter and Anderson* [2005], *Poeter and Hill* [2007], *Neuman et al.* [2012], *Lu et al.* [2012], and L. Foglia et al. (Analysis of model discrimination techniques, the case of the Maggia Valley, Southern Switzerland, submitted to *Water Resources Research*, 2012), among others.

[6] A related field is sensitivity analysis, which is commonly used to diagnose contributions to uncertainty measures [e.g., *Saltelli et al.*, 2008]. In sensitivity analysis, comparisons of linear methods related to the linear confidence intervals in this work and global methods with Monte Carlo sampling more similar to the nonlinear Bayesian methods have suggested both utility and lack of utility of the linear methods. *Foglia et al.* [2007] suggest a high level of utility; *Tang et al.* [2007] suggest little utility. Of note is that the linear measure used by *Tang et al.* [2007] (labeled CS) has dimensions of the reciprocal of the parameter values, with the predictable result that parameters with small values (such as LZPK) are rated as "most important." This contradicts their results from global methods. Multiplying by the parameter values to obtain dimensionless linear measures results in LZPK being rated as "less important" which is consistent with their results from global methods.

[7] Confidence and credible intervals are conceptually different and understanding their fundamental differences and similarities can aid appropriate use. In this study, we consider linear confidence and credible intervals for linear and linearized-nonlinear models, and nonlinear confidence and credible intervals for nonlinear models. These intervals are defined in section 2. For nonlinear models, nonlinear confidence and credible intervals account for model nonlinearity and are thus expected to be more accurate than linear intervals calculated from linearized-nonlinear models. Table 1 summarizes previous studies that compare confidence and credible intervals. The previous studies report linear confidence intervals and nonlinear credible intervals that differ by between 2% and 82%, while nonlinear confidence and credible intervals differ by between −24% and 7%. The previous studies mainly focused on comparing calculated confidence and credible intervals without a thorough discussion of underlying theories, and the large differences have not been fully understood in the environmental modeling community. In this work we seek a deeper understanding, and begin by considering the work of *Jaynes* [1976], *Box and Tiao* [1992] and *Bates and Watts* [1988].

[8] *Jaynes* [1976] presents a spirited discussion on estimation of confidence and credible intervals of distribution parameters (e.g., mean of truncated exponential distribution in his Example 5). A similar but more systematic discussion is found in *Box and Tiao* [1992]. The authors conclude that, given sufficient statistics, what we call confidence and credible intervals are "identical" [*Jaynes*, 1976, p. 199] or "numerically identical" [*Box and Tiao*, 1992, p. 86]. Sufficient statistics are completely specified by a defined set of statistics that can be calculated from the data. For example, random variable $\mathbf{X}$ follows a Gaussian distribution with unknown mean θ and variance σ^2; then the sample mean and sample variance are sufficient statistics for θ and σ^2, respectively. In this case, the confidence and credible intervals for the two parameters θ and σ^2 are identical [*Box and Tiao*, 1992, p. 61]. *Box and Tiao* [1992, p. 113] extend the discussion to linear models with multivariate Gaussian observation errors, concluding that what are herein called the parameter credible regions are "numerically identical" to the parameter confidence regions used in linear regression [*Box and Tiao*, 1992, p. 118]. As a result, confidence and credible intervals on predictions are numerically identical. This conclusion can also be found in *Bates and Watts* [1988, p. 7]. *Bates and Watts* [1988] further extend the analysis of parameter confidence and credible regions to nonlinear models with multivariate Gaussian observation errors. They conclude that, when intrinsic model nonlinearity is small, the two regions are mathematically similar and so are confidence and credible intervals on parameters and predictions (confidence and credible regions, their relation to confidence and credible intervals, and intrinsic model nonlinearity are discussed in section 2.6).

[9] Issues not considered in the references cited include the extent to which the two kinds of intervals differ from each other when intrinsic model nonlinearity is moderate and large, as is common in models of environmental systems, and resulting relative utility of the confidence intervals in these nonideal circumstances. In addition, informative prior in the Bayesian methods and inclusion of consistent prior information in regression is not considered. Prior information is of considerable utility in environmental models. This paper seeks to provide some insight into these two issues.

[10] This paper first compares confidence and credible intervals theoretically, establishing the foundations developed by *Box and Tiao* [1992] and *Bates and Watts* [1988] and extending the theory for linear models to include to

Table 1. List of Selected Recent Studies Comparing Confidence and Credible Intervals[a]

Reference and Quantity for Which the Intervals are Constructed	Model	Linear Confidence Interval (LCo)	Nonlinear Confidence Interval (NCo)	Nonlinear Credible Interval (NCr)[b]	Increase in Width Relative to NCr (%)[c]
Vrugt and Bouten [2002],	Near linear	X	N/A	X	Similar
Soil hydraulic parameters	Nonlinear	XX	N/A	X	L:82%
Gallagher and Doherty [2007], Watershed					
Parameters	Nonlinear	XXX	X	XX	L:2%; N:−24%
Predictions	Nonlinear	XXX	X	XX	L:71%; N:−7%
Finsterle and Pruess [1995], Two-phase flow Parameters	Nonlinear	X	XX	N/A	–
Christensen and Cooley [1999], Groundwater Parameters	Nonlinear	Not consistently larger or smaller		N/A	–
Predictions	Nonlinear	Not consistently larger or smaller		N/A	–
Liu et al. [2010],[d] DNAPL Parameters	Nonlinear	e	N/A	e	–
This study, Predictions	Linear simple	X	X	X	Same
	Very nonlinear simple I	XXX	XX	X	L:70%; N:13%
	Mildly nonlinear simple II	X	X	X	Same
	Nonlinear complex[f]	XX[g]	XXX	X	L: 0.1 to 33% N: 0.2 to 69%
	See Figure 5	X[h]	XX	XXX	L: −18 to −11% N: −15 to −6%

[a] The relative size of the intervals is represented by the number of Xs, with more Xs indicating larger intervals. N/A: interval types not considered. For all, consistent prior is used for the confidence and credible intervals. Except as noted, there is noninformative prior for credible intervals, and no prior for confidence intervals.

[b] Calculated using Markov-Chain Monte Carlo.

[c] L: 100 × (LCo-NCr)/NCr; N: 100 × (NCo-NCr)/NCr. Negative values mean the LCo or NCo interval is smaller than the NCr. –, results are presented graphically, so values not determined.

[d] *Liu et al.* [2010] calculated linear credible intervals using the maximum a posteriori (MAP) approach. The listing under linear confidence intervals is consistent with results of this work that show the equivalence of linear credible and confidence intervals. Same informative prior is used for linear and nonlinear credible interval.

[e] Results of the comparison between LCo and NCr were not consistently larger or smaller for different parameters.

[f] The prior is informative and prior information is included for confidence intervals.

[g] Applies for most simulated circumstances.

[h] For drawdown predicted by model INT.

include informative prior for credible intervals and prior information for confidence intervals. In particular, we prove the equivalence when the prior information used for confidence intervals is consistent with the informative prior used for credible intervals; consistency is limited to the informative prior being multi-Gaussian. For nonlinear models, we pursue a largely heuristic approach. We use simple analytical test cases to explore the conditions under which parameter confidence and credible regions and associated prediction confidence and credible intervals are numerically equivalent or nearly equivalent. We further explore computational difficulties that cause differences between the two kinds of intervals in practice. Finally, we use a complex synthetic test case to compare the differences between confidence and credible intervals in the context of differences between alternative models of a single system. The general guidelines drawn for the nonlinear models in this study are expected to be useful to practical environmental modeling.

2. Theoretical Consideration of the Uncertainty Intervals

[11] In this section, conceptual differences between confidence and credible intervals are discussed in section 2.1. In section 2.2, we discuss the situation when linear confidence and credible intervals for linear models are numerically identical. In section 2.3, we show that for linear models with multi-Gaussian observation errors, the two kinds of intervals are numerically identical without priors (noninformative priors for credible intervals and no priors for confidence intervals), and in section 2.4 with consistent priors (informative priors for credible intervals are consistent with priors defined for confidence intervals). As shown in section 2.5, the same conclusions also apply to linearized-nonlinear models. The relations between confidence and credible intervals are significantly more complicated for nonlinear models. They depend on relations between parameter confidence and credible regions defined in section 2.6. Results identify the consequences of local minima on confidence intervals and inadequate exploration of parameter space for credible intervals.

2.1. Conceptual Differences Between Confidence and Credible Intervals

[12] For a variable X, both confidence and credible intervals can be defined symbolically as

$$\mathrm{Prob}(l \le X \le u) = 1 - \alpha, \qquad (1)$$

where l and u are the lower and upper interval limits, α is significance level, and $1 - \alpha$ is confidence level. However, the definition is interpreted in different ways when estimating confidence intervals and credible intervals, rendering the two kinds of intervals conceptually different.

[13] Take the intervals for model predictions as an example. From the frequentist point of view, a true value of the prediction exists and a confidence interval with a confidence level of, for example, 95% (α = 5%), is an interval

that is expected to include the true value "95% of the time" in repeated sampling of observations and prior information used in regression [*Jaynes*, 1976, p. 200; *Box and Tiao*, 1992, p. 86; *McClave and Sincich*, 2000, p. 282]. In other words, the interval, not the true value of the unknown prediction, is random. The interval varies with samples used to calculate the interval. If N sets of observations are sampled based on their error distribution and, for each set, the confidence interval of a prediction is evaluated, 95% of the N intervals are expected to contain the true value of the prediction. This concept was used in *Hill* [1989] and *Cooley* [1997] to test methods of estimating nonlinear confidence intervals.

[14] From the Bayesian point of view, the prediction is a random variable and a 95% credible interval is expected to include 95% of the probability distribution function (PDF) of the prediction [*Box and Tiao*, 1992, p. 86; *Casella and Berger*, 2002]. The posterior distribution summarizes the state of knowledge about the unknown prediction based on available data and prior information. The credible interval is determined via

$$\int_l^u p(g(\boldsymbol{\beta})|\mathbf{y})dg(\boldsymbol{\beta}) = 1 - \alpha, \tag{2}$$

where $\boldsymbol{\beta}$ and $g(\boldsymbol{\beta})$ are model parameters and predictions, respectively, and $p(g(\boldsymbol{\beta})|\mathbf{y})$ is the posterior distribution of $g(\boldsymbol{\beta})$ conditioned on data $\mathbf{y}$. In this study, the credible interval limits l and u are determined using the easily computed equal-tailed method as [*Casella and Berger*, 2002]

$$p(g(\boldsymbol{\beta}) \leq l|\mathbf{y}) = p(g(\boldsymbol{\beta}) \geq u|\mathbf{y}) = \alpha/2. \tag{3}$$

Other methods of estimating the credible intervals (e.g., highest posterior density (HPD) interval) are also available [*Box and Tiao*, 1992; *Chen and Shao*, 1999; *Casella and Berger*, 2002]. The HPD method is not used in this work partly because it tends to produce the smallest interval and partly simply for the convenience of reporting one set of results.

[15] In this study, we focus our discussion on relations between confidence and credible intervals within the framework of *Gaussian distributed observation errors*, which is the basis of using Gaussian likelihood functions below. The Gaussian distribution is commonly used in environmental problems and is also used in statistical books such as *Draper and Smith* [1998], *Box and Tiao* [1992], and *Hill and Tiedeman* [2007]. *Box and Tiao* [1992, p. 151] note that approximately Gaussian distributed observations are expected when the central limit theorem applies. This requires that the contributions to observation error be many independent sources, "none of which are dominant." Considering that in environmental models there are clearly many sources of error (including different sources of measurement, epistemic, and aleatory errors) and that a major goal of model development is to resolve dominant errors, the assumption of Gaussian observation errors is expected to have wide applicability.

2.2. When Linear Confidence and Credible Intervals for Linear Models are Numerically Identical

[16] For linear models, classical regression confidence intervals of *fixed but unknown* true model parameters $\boldsymbol{\beta}$ are estimated based on the distributions $p(\hat{\mathbf{b}})$ of *parameter estimates* $\hat{\mathbf{b}}$, which are functions of the observations $\mathbf{y}$. Bayesian credible intervals of *random* model parameters $\boldsymbol{\beta}$ are evaluated based on the distributions $p(\boldsymbol{\beta}|\mathbf{y})$ of *parameters* $\boldsymbol{\beta}$, which are calculated according to the Bayes' theorem,

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} = \frac{l(\boldsymbol{\beta}|\mathbf{y})p(\boldsymbol{\beta})}{p(\mathbf{y})}. \tag{4}$$

If $p(\boldsymbol{\beta}|\mathbf{y})$ is equivalent to $p(\hat{\mathbf{b}})$, the credible and confidence intervals of $\boldsymbol{\beta}$ and the intervals of linear model predictions $g(\boldsymbol{\beta})$ are numerically equivalent. *Jaynes* [1976] and *Box and Tiao* [1992] established this equivalence using the concept of sufficient statistics without prior information (noninformative prior in (4) and no prior used for the confidence interval).

[17] According to *Box and Tiao* [1992, p. 62], for a vector of observations $\mathbf{y}$ whose distribution depends on the parameter vector $\boldsymbol{\beta}$, the set of statistics $\mathbf{T}$ is said to be jointly *sufficient* for $\boldsymbol{\beta}$ if the likelihood function $l(\boldsymbol{\beta}|\mathbf{y})$ can be expressed in the form

$$l(\boldsymbol{\beta}|\mathbf{y}) \propto l_T(\boldsymbol{\beta}|\mathbf{T}) \propto p(\mathbf{T}|\boldsymbol{\beta}), \tag{5}$$

where $\mathbf{T}$ is a set of statistics calculated from $\mathbf{y}$ that fully define the likelihood function and the second relation of equation (5) uses the basic relation between likelihood and probability. If $\mathbf{y}$ is multivariate Gaussian distributed with mean simulated by a linear model with model parameters $\boldsymbol{\beta}$ and known variance, the corresponding Gaussian likelihood function $l(\boldsymbol{\beta}|\mathbf{y})$ can be fully defined by the set of sufficient statistics $\mathbf{T} = \hat{\mathbf{b}}$ so that $l(\boldsymbol{\beta}|\mathbf{y}) \propto p(\hat{\mathbf{b}}|\boldsymbol{\beta})$ [*Box and Tiao*, 1992, p. 115]. If noninformative prior is used in (4), $p(\boldsymbol{\beta}|\mathbf{y}) \propto l(\boldsymbol{\beta}|\mathbf{y})$ and thus $p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\hat{\mathbf{b}}|\boldsymbol{\beta})$ [*Box and Tiao*, 1992, p. 115]. This is the basis of the conclusion of *Jaynes* [1976, p. 199] that, "if the confidence interval is based on a sufficient statistic, … it turns out to be so nearly equal to the Bayesian interval that it is difficult to produce any appreciable difference in the numerical results; in an astonishing number of cases, they are identical. … Similarly, the shortest confidence interval for the mean of a normal distribution, whether the variance is known or unknown, …turn out to be identical with the shortest Bayesian intervals at the same level (based on a uniform prior density for location parameters and the Jeffreys prior dσ/σ for scale parameters)."

[18] Building on the theoretical foundation of *Jaynes* [1976] and *Box and Tiao* [1992], we next prove that linear confidence and credible intervals are numerically identical for the Gaussian likelihood function used widely in regression and Bayesian analysis in environmental modeling. We consider two cases, without and with informative prior information. To our knowledge, the derivation of the equivalence using informative prior information is new. Confidence intervals with prior information and credible intervals with informative prior were discussed by *Hill* [2010]. *Doherty and Hunt* [2010] and *Fienen et al.* [2010] suggested equivalence in limited circumstances. The proof of equivalence presented in this work is more theoretically rigorous and shows the equivalence to be more general than the previous works. As a corollary, we point out that linear confidence and credible intervals are also numerically identical when calculated for linearized-nonlinear models.

2.3. Equivalence of Regression and Bayesian Distributions and Intervals for Gaussian Linear Models

[19] In classical regression, a linear model without prior information on parameters is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

with

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{C}_\varepsilon), \tag{7}$$

where transformations can be applied to achieve the Gaussian distribution of (7). The covariance matrix $\mathbf{C}_\varepsilon$ is often related as $\mathbf{C}_\varepsilon = \sigma^2\boldsymbol{\omega}^{-1}$ to weighting $\boldsymbol{\omega}$ (used in regression) and a scalar σ^2 (generally unknown but can be estimated). When σ^2 is known, the distribution of estimates $\hat{\mathbf{b}}$ of the unknown true parameters $\boldsymbol{\beta}$ is multivariate Gaussian, i.e., [*Draper and Smith*, 1998, p. 94; *Seber and Lee*, 2003, p. 47],

$$\hat{\mathbf{b}} \sim N_p\left(\boldsymbol{\beta}, \left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\right) \propto \exp\left[-\frac{1}{2}\left(\hat{\mathbf{b}} - \boldsymbol{\beta}\right)^T \mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\left(\hat{\mathbf{b}} - \boldsymbol{\beta}\right)\right]. \tag{8}$$

Ideally $\mathbf{X}$ in (8) is the same as in (6), but in reality alternative models may have different $\mathbf{X}$ (more discussion see Appendix C of *Hill and Tiedeman* [2007]).

[20] For a linear prediction function $g(\boldsymbol{\beta})$ with $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$, it has

$$\begin{aligned} g(\hat{\mathbf{b}}) &\sim N\left(g(\boldsymbol{\beta}), \mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\mathbf{Z}\right) \\ &\propto \exp\left[-\frac{1}{2}\left(g(\hat{\mathbf{b}}) - g(\boldsymbol{\beta})\right)^T \left(\mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\mathbf{Z}\right)^{-1} \cdot \left(g(\hat{\mathbf{b}}) - g(\boldsymbol{\beta})\right)\right]. \end{aligned} \tag{9}$$

Correspondingly, the $(1 - \alpha) \times 100\%$ confidence interval for $g(\boldsymbol{\beta})$ is given as

$$g(\hat{\mathbf{b}}) \pm z_{1-\alpha/2}\left[\mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\mathbf{Z}\right]^{1/2}, \tag{10}$$

where $z_{1\text{-}\alpha/2}$ is a z statistic of standard normal distribution with significance level α. When σ^2 is unknown and estimated using $s^2 = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^T \boldsymbol{\omega}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)/(n-p)$, the $(1 - \alpha) \times 100\%$ confidence interval of $g(\boldsymbol{\beta})$ is

$$g(\hat{\mathbf{b}}) \pm t_{1-\alpha/2,n-p}\left[s^2\mathbf{Z}^T\left(\mathbf{X}^T\boldsymbol{\omega}\mathbf{X}\right)^{-1}\mathbf{Z}\right]^{1/2}, \tag{11}$$

where $t_{1\text{-}\alpha/2,\ n\text{-}p}$ is a t statistic with significance level α and freedom n-p, where n is the number of observations and p is the number of parameters.

[21] In the Bayesian analysis, according to (4), with the same model defined in (6) and (7), considering Jeffreys' noninformative prior defined in *Box and Tiao* [1992, pp. 41–54], it is derived in Appendix A in the auxiliary material[1] that, when σ^2 is known, the posterior distribution of $\boldsymbol{\beta}$ is multivariate Gaussian, i.e.,

$$\boldsymbol{\beta} \sim N_p\left(\hat{\mathbf{b}}, \left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \hat{\mathbf{b}}\right)^T \mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\left(\boldsymbol{\beta} - \hat{\mathbf{b}}\right)\right], \tag{12}$$

where $\hat{\mathbf{b}}$ is the maximum likelihood estimate (also the least squares estimate) of $\boldsymbol{\beta}$. Correspondingly, the posterior distribution of a linear prediction $g(\boldsymbol{\beta})$ (i.e., $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$) is multivariate Gaussian,

$$\begin{aligned} g(\boldsymbol{\beta}) &\sim N_p\left(g(\hat{\mathbf{b}}), \mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\mathbf{Z}\right) \\ &\propto \exp\left[-\frac{1}{2}\left(g(\boldsymbol{\beta}) - g(\hat{\mathbf{b}})\right)^T \left(\mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X}\right)^{-1}\mathbf{Z}\right)^{-1} \cdot \left(g(\boldsymbol{\beta}) - g(\hat{\mathbf{b}})\right)\right]. \end{aligned} \tag{13}$$

[22] Comparing equations (12) and (13) with equations (8) and (9), respectively, shows that the distributions of $\hat{\mathbf{b}}$ and $g(\hat{\mathbf{b}})$ are equivalent to those of $\boldsymbol{\beta}$ and $g(\boldsymbol{\beta})$. Therefore, the $(1 - \alpha) \times 100\%$ credible interval of $g(\boldsymbol{\beta})$ is equivalent to (10) if σ^2 is known, and (11) if σ^2 is unknown. In other words, for linear models without prior information and Gaussian distributed observation errors, the linear confidence and credible intervals are numerically identical.

2.4. Equivalence When Including Informative Prior in Bayesian Equations and Prior Information in Regression for Gaussian Linear Models

[23] The equivalence between confidence and credible intervals can also be obtained with consideration of prior information, when it is available, in Bayesian and regression analysis. Regression comes from a frequentist background, and the idea of adding prior information can sometimes seem strange. For example, *Stark and Tenorio* [2011] and *Kitanidis* [2010] suggest that only Bayesian methods include prior, while, for example, *Schweppe* [1973], *Cooley* [1983], *Hill and Tiedeman* [2007], and *Poeter et al.* [2005] have used prior within the context of regression methods for decades. The Maximum a Posteriori (MAP) method [*Carrera and Neuman*, 1986; *Oliver et al.*, 2008; *Liu et al.*, 2010] is essentially equivalent to the regression-with-prior method used in this work, but was developed from the Bayesian perspective.

[24] In the Bayesian analysis, for the model defined in (6) and (7), when prior parameter distributions are informative, a conventional form is the conjugate prior that is multivariate Gaussian with mean $\boldsymbol{\beta}_p$ and covariance matrix $\mathbf{C}_p$ [*Kitanidis*, 1986, 1997]. For this prior, it is derived in Appendix B of the auxiliary material that the $(1 - \alpha) \times 100\%$ credible interval for linear model prediction, $g(\beta) = \mathbf{Z}\boldsymbol{\beta}$, is

$$g\left(\boldsymbol{\beta}'_p\right) \pm z_{1-\alpha/2}\left[\mathbf{Z}^T\left(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X} + \mathbf{C}_p^{-1}\right)^{-1}\mathbf{Z}\right]^{1/2}, \tag{14}$$

where $\boldsymbol{\beta}'_p = (\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{X} + \mathbf{C}_p^{-1})^{-1}(\mathbf{X}^T\mathbf{C}_\varepsilon^{-1}\mathbf{y} + \mathbf{C}_p^{-1}\boldsymbol{\beta}_p)$ is mean of the Bayesian posterior distribution.

[25] In regression analysis, the prior information is defined as knowledge about model parameters estimated by the regression procedure [*Schweppe*, 1973, p. 104; *Cooley*, 1983; *Hill and Tiedeman*, 2007, p. 288], as might be derived from measurements or expert knowledge of transmissivity or hydraulic conductivity, recharge or discharge, specified boundary flows and heads. This knowledge is commonly included in the regression to complement the observations $\mathbf{y}$ of state variables. This is accomplished by treating the prior information as the measurement vector $\mathbf{y}_{\boldsymbol{\beta}}$ of the estimated model parameters with error vector $\boldsymbol{\varepsilon}_\beta$ and by appending $\mathbf{y}_\beta$

[1] Auxiliary materials are available in the HTML. doi:10.1029/2011WR011289.

to observation vector $\mathbf{y}$ and $\boldsymbol{\varepsilon}_\beta$ to the observation error vector $\boldsymbol{\varepsilon}$ in equation (6). For example, when the prior information has the form $\mathbf{y}_\beta = \boldsymbol{\beta} + \boldsymbol{\varepsilon}_\beta$ with the errors defined as $\boldsymbol{\varepsilon}_\beta \sim N_{npri}(\mathbf{0}, \mathbf{C}_\beta)$, it is derived in Appendix B of the auxiliary material that the $(1 - \alpha) \times 100\%$ confidence interval for linear model prediction, $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$, is equivalent to (14) when the prior information $\mathbf{y}_\beta$ is equal to the mean of Bayesian prior distribution $\boldsymbol{\beta}_p$ and the covariance of errors of the prior information $\mathbf{C}_\beta$ is equal to the covariance of the prior distribution $\mathbf{C}_p$. For a Gaussian distribution, this yields consistent priors in the two methods, even though conceptually the prior information in regression and Bayesian analysis are different. In classical regression context, the parameters $\boldsymbol{\beta}$ are taken as fixed and unknown and the randomness of the prior estimates in $\mathbf{y}_\beta$ is from the error $\boldsymbol{\varepsilon}_\beta$; in Bayesian analysis, $\boldsymbol{\beta}$ is conceptualized as being random [*Kitanidis*, 2010, among others]. For linear models with consistent priors, the linear confidence and credible intervals are numerically identical.

2.5. Equivalence of Linear Confidence and Credible Intervals for Linearized-Nonlinear Models

[26] In classical regression, the linear confidence interval is also available for a nonlinear model,

$$\mathbf{y} = f(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{15}$$

with $\boldsymbol{\varepsilon}$ defined in equation (7) when prior information is not available. To estimate a linear confidence interval, the nonlinear function $f(\mathbf{b})$ of model parameters $\mathbf{b}$ (a general vector of model parameters) is approximated by its first-order Taylor series expansion in the neighborhood of $\hat{\mathbf{b}}$, the generalized least squares estimator of $\boldsymbol{\beta}$. That is, $f(\mathbf{b}) \approx f(\hat{\mathbf{b}}) + \mathbf{X}_{\hat{\mathbf{b}}}(\mathbf{b} - \hat{\mathbf{b}})$, where $\mathbf{X}_{\hat{\mathbf{b}}} = [\partial f/\partial \mathbf{b}]_{\mathbf{b}=\hat{\mathbf{b}}}$ is the sensitivity matrix [*Seber and Wild*, 2003, p. 23–24]. If the nonlinear prediction function $g(\mathbf{b})$ is also approximated to the first-order with the prediction sensitivity vector $\mathbf{Z}_{\hat{\mathbf{b}}} = [\partial g/\partial \mathbf{b}]_{\mathbf{b}=\hat{\mathbf{b}}}$ [*Seber and Wild*, 2003, p. 192], then for σ^2 known, the $(1 - \alpha) \times 100\%$ linear confidence interval of $g(\boldsymbol{\beta})$ is similar to (10) but with $\mathbf{X}$ replaced by $\mathbf{X}_{\hat{\mathbf{b}}}$ and $\mathbf{Z}$ replaced by $\mathbf{Z}_{\hat{\mathbf{b}}}$. For σ^2 unknown, the $(1 - \alpha) \times 100\%$ linear confidence interval of $g(\boldsymbol{\beta})$ is similar to (11) but with $\mathbf{X}$ replaced by $\mathbf{X}_{\hat{\mathbf{b}}}$ and $\mathbf{Z}$ replaced by $\mathbf{Z}_{\hat{\mathbf{b}}}$.

[27] In the Bayesian analysis, the linear credible interval for the nonlinear model is evaluated after the nonlinear model is linearized. As shown in Appendix C of the auxiliary material, the derived linear credible intervals are numerically identical to the linear confidence intervals for nonlinear models. Though for a linearized model confidence and credible intervals may be numerically identical, if the linearization method is not adequate for approximating confidence and credible intervals for the parameters and predictions of a nonlinear model, they will both be equivalently wrong. Linearization tends to become progressively less adequate as the model becomes more nonlinear. More discussion on how model nonlinearity affects the above approximations is discussed by *Hill and Tiedeman* [2007, p. 393–398].

2.6. Relations Between Nonlinear Confidence and Credible Intervals for Nonlinear Models

[28] The relations between nonlinear confidence and credible intervals of nonlinear models are complicated. In classical regression, the confidence intervals of model parameters and predictions are not estimated based on the distributions of their estimates (e.g., equations (8) and (9)) as for linear models, because analytical expressions of the distributions are in general not available. Instead, the nonlinear confidence intervals are evaluated using parameter confidence regions as described below. To facilitate the discussion, we also use posterior parameter credible regions (defined below) to evaluate credible intervals. The step of determining posterior parameter credible regions is rare because credible intervals are generally calculated directly from posterior distributions of model parameters and predictions. However, a parameter credible region can be determined and is used here to relate prediction confidence and credible intervals.

[29] The nonlinear confidence and credible intervals can be numerically identical if (1) the interval limits are equivalent to the maximum and minimum values of prediction function $g(\mathbf{b})$ on the boundary of the parameter confidence and credible regions and (2) the two parameter regions are equivalent [*Vecchia and Cooley*, 1987; *Cooley* 1993a, 1993b, 1999]. The discussions below for the two conditions involve definitions of exact confidence and credible regions and approximate regions defined by the likelihood method. Based on *Vecchia and Cooley* [1987], the exact $(1-\alpha) \times 100\%$ confidence region for $\boldsymbol{\beta}$ is a region in p dimensional Euclidean space, say, R_α, that depends on $\mathbf{y}$, where $\mathbf{y}$ includes observations and possible prior information, and for which $p(\boldsymbol{\beta} \in R_\alpha) = 1 - \alpha$ holds exactly in regression theory. The exact $(1 - \alpha) \times 100\%$ posterior credible region is bounded by a contour of the posterior density function within which the posterior probability of model parameters is $(1 - \alpha) \times 100\%$ exactly [*Bates and Watts*, 1988, p. 220]. However, estimating the exact region is computationally difficult, and the approximate likelihood method is commonly used.

[30] The likelihood confidence region is defined as the set of parameter values whose corresponding objective function values, $S(\mathbf{b})$, satisfy [*Christensen and Cooley*, 1999; *Cooley*, 2004; *Hill and Tiedeman*, 2007, p. 178]

$$S(\mathbf{b}) \le S(\hat{\mathbf{b}}) \left[\frac{1}{n-p} t^2_{\alpha/2,n-p} + 1 \right]. \tag{16}$$

The confidence region defined in (16) is for calculating *individual* confidence intervals. *Scheffé-type simultaneous* confidence intervals are calculated based on the region with $S(\mathbf{b})$ satisfying

$$S(\mathbf{b}) \le S(\hat{\mathbf{b}}) \left[\frac{p}{n-p} F_{\alpha/2,p,n-p} + 1 \right], \tag{17}$$

which is equivalent to (16) if $p = 1$. The test cases in this work use individual confidence intervals calculated using (16). The parameter region defined in (16) contains the true model parameters $\boldsymbol{\beta}$ with *approximate* probability of $(1 - \alpha) \times 100\%$ for errors defined in (7) when prior information is not available and σ^2 is unknown. When intrinsic model nonlinearity is small, it contains the true model parameters with *exact* $(1 - \alpha) \times 100\%$ probability [*Donaldson and Schnabel*, 1987; *Bates and Watts*, 1988, p. 201].

[31] The likelihood credible region is also bounded by the contours of (16) and (17) as follows. When intrinsic model nonlinearity is small, consider independent Jeffrey's

noninformative priors for parameters $\boldsymbol{\beta}$ and σ, i.e., $p(\boldsymbol{\beta}) \propto C$ and $p(\sigma) \propto 1/\sigma$ (or $p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$). For errors defined in (7), the resulting likelihood function for parameters $\boldsymbol{\beta}$ and σ is

$$l(\boldsymbol{\beta}, \sigma|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}}|\boldsymbol{\omega}|^{1/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-f(\boldsymbol{\beta}))^T\boldsymbol{\omega}(\mathbf{y}-f(\boldsymbol{\beta}))\right] \tag{18}$$

where $S(\boldsymbol{\beta}) = (\mathbf{y} - f(\boldsymbol{\beta}))^T\boldsymbol{\omega}(\mathbf{y} - f(\boldsymbol{\beta}))$is the objective function, same as $S(\mathbf{b})$ used in regression. Correspondingly, the joint posterior distribution of parameters $\boldsymbol{\beta}$ and σ is

$$p(\boldsymbol{\beta}, \sigma|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma)\sigma^{-1} \propto \sigma^{-(n+1)}\exp\left(-\frac{1}{2\sigma^2}S(\boldsymbol{\beta})\right). \tag{19}$$

Integrating out parameter σ leads to the posterior distribution of model parameters $\boldsymbol{\beta}$ [*Bates and Watts*, 1988, p. 220],

$$p(\boldsymbol{\beta}|\mathbf{y}) = \int_0^\infty p(\boldsymbol{\beta}, \sigma|\mathbf{y})d\sigma \propto [S(\boldsymbol{\beta})]^{-n/2}, \tag{20}$$

and a credible region is bounded by a contour of the posterior density function or equivalently by a contour of the objective function $S(\boldsymbol{\beta})$. *Bates and Watts* [1988, p. 220] proved that the approximate $(1 - \alpha) \times 100\%$ individual posterior credible region is bounded by the $S(\mathbf{b})$ contour defined in (16).

[32] Thus, when observation errors are multivariate Gaussian distributed and the intrinsic model nonlinearity is very small, without prior information, the confidence and credible regions are equivalent and so are the nonlinear confidence and credible intervals of predictions evaluated on the regions. When consistent prior information is used for regression and Bayesian analysis in the way described in section 2.4, it is straightforward to develop the equivalence between the parameter confidence and credible regions. As a result, the confidence and credible intervals are numerically identical.

[33] Regions defined by the above approximate likelihood methods and the calculated confidence and credible intervals are accurate given the following assumptions:

[34] 1. The model accurately represents the system;

[35] 2. Model prediction $g(\mathbf{b})$ is monotonic enough that any local extreme of $g(\mathbf{b})$ within the closed parameter regions lies between the maximum and minimum values of $g(\mathbf{b})$ that occur along the boundary of the regions [*Cooley*, 1993a];

[36] 3. There is a single minimum in the objective function;

[37] 4. The residuals (defined as the difference between observed and simulated values) are multivariate Gaussian distributed to obtain a valid critical value, and

[38] 5. Model intrinsic nonlinearity is small [*Vecchia and Cooley*, 1987; *Christensen and Cooley*, 1999; *Cooley*, 2004].

[39] The assumptions can be evaluated using existing techniques such as those described in *Hill and Tiedeman* [2007] and shown below. The model intrinsic nonlinearity of assumption 5, when combined with parameter effects nonlinearity, results in total model nonlinearity [*Christensen and Cooley*, 1999]. Parameter effects nonlinearity is what can be removed after transforming model parameters in suitable ways, while intrinsic model nonlinearity results from model structure and cannot be removed by any parameter transformations. *Cooley* [2004] and *Christensen and Cooley* [2005] developed methods of measuring model total nonlinearity based on *Beale* [1960], *Linssen* [1975], and *Bates and Watts* [1980]. More discussion of model nonlinearity measures can be found in *Hill and Tiedeman* [2007, p. 142–145]. The residuals mentioned in assumption 4 are expected to have a variance-covariance matrix equal to $(\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}^T\boldsymbol{\omega}\mathbf{X})^{-1}\,\mathbf{X}^T\boldsymbol{\omega})$, so are not expected to be independent even when the elements of $\boldsymbol{\varepsilon}$ are independent. [*Cook and Weisberg*, 1982, p. 11; *Cooley and Naff*, 1990; *Hill and Tiedeman*, 2007, p. 111–113; *Aster et al.*, 2012; *Finsterle and Zhang*, 2011]. This also applies when equations (6) and (7) are augmented to include prior information.

[40] With recently developed Markov chain Monte Carlo (MCMC) techniques, the nonlinear credible intervals can be evaluated directly without relying on the concept of parameter credible region. The reason is that posterior parameter distributions can be directly simulated and that the credible intervals of predictions can be evaluated by first sorting the prediction samples from the smallest to the largest and then identifying, for example, the 2.5% and 97.5% thresholds to form 95% individual credible intervals of the predictions. However, the estimated credible interval based on the MCMC results may be inaccurate if the MCMC sampling does not correctly simulate the posterior distribution, as discussed below. Therefore, when comparing the nonlinear confidence intervals determined from a confidence region and credible intervals from MCMC methods, the assumptions of calculating accurate confidence intervals need to be examined and the MCMC sampling process needs to be checked. This is further demonstrated in two numerical examples presented in section 3.

[41] Two MCMC codes are used in this study. One is MICA developed by *Doherty* [2003]. MICA uses the Metropolis-Hastings algorithm revised by introducing a simple adaptive algorithm of the covariance matrix of the proposal distribution to obtain a reasonable acceptance rate [*Haario et al.*, 2001]. When the acceptance rate is high, all elements of the covariance matrix are simultaneously increased by a user-defined multiplier; and when the acceptance rate is low, all elements are simultaneously decreased. While MICA works well for estimating *unimodal* posterior parameter distributions, for problems with multimodal parameter distributions, it cannot sample the distribution efficiently with a single proposal distribution, because it does not have algorithms for chain jumps among multiple modes. As pointed out by *Gallagher and Doherty* [2007] and also found in this study, for problems with multiple minima the probability density functions obtained from MICA are not accurate in that the resulting density functions do not have multiple modes corresponding to the minima. To solve this problem, DREAM of *Vrugt et al.* [2008, 2009] is used. For the results presented here, the algorithm was implemented in FORTRAN and parallelized. DREAM automatically tunes the scale and orientation of the proposal distribution in the sampling process by using multiple different chains simultaneously for global exploration. It uses the current location of the chains to generate candidate points, allowing the jumps between modes of parameter distributions, and thus efficiently accommodates complex and multimodal

Table 2. Predictions and 95% Linear Confidence and Credible Intervals for Linear Simple Test Function[a]

	Prediction	Lower Limit	Upper Limit
Noninformative Prior			
Unknown σ^2, equation (11)			
Linear confidence interval	63.76	62.26	65.26
Linear credible interval (MCMC)	63.76	62.27	65.27
Known σ^2, equation (10)			
Linear confidence interval	63.76	62.37	65.16
Linear credible interval (MCMC)	63.76	62.37	65.19
Informative Prior[b]			
Linear confidence interval	63.73	62.73	64.74
Linear credible interval (MCMC)	63.73	62.72	64.75

[a] True value of the prediction is 63. MCMC results are obtained with MICA [*Doherty*, 2003] using 1,000,000 model runs. Linear confidence intervals require 10 model runs).

[b] Informative Prior with $\mathbf{C}_\beta = \mathbf{C}_p = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$, equation (14).

parameter posterior distributions. The test cases below illustrate the consequences of this difference between MICA and DREAM.

3. Simple Test Cases

[42] Three simple test cases with one linear and two nonlinear analytic functions are designed to demonstrate the theoretical findings above and to help better understand the similarities and differences between confidence and credible intervals. These test cases are not derived from environmental problems, but are easily designed to produce linearity or desired kinds of relevant nonlinearities. In the next section, a complicated groundwater problem is explored.

3.1. Linear Test Function

[43] The linear test function is

$$\mathbf{y} = a\mathbf{x} + m + \boldsymbol{\varepsilon}, \tag{21}$$

where the true values are $a = 2$ and $m = 3$. Twenty samples of $\mathbf{y}$ are first generated with $\mathbf{x} = \{1, 2, \ldots, 20\}$, and subsequently corrupted using one realization of white noise $\boldsymbol{\varepsilon}$, with mean zero and constant variance $\sigma^2 = 1$. Linear confidence and credible intervals, calculated for y at $x = 30$ with unknown and known σ^2, and noninformative and informative priors, are listed in Table 2, where the linear credible intervals are estimated from 500,000 MCMC parameter samples. Except for negligible numerical discrepancy, the table confirms that the linear confidence and credible intervals are numerically identical for this linear problem in all three situations. Figures 1a–1c show probability density functions (PDFs) of the two parameters and prediction based on the classical regression theory (equations (8) and (9)) and Bayesian theory (equations (12) and (13)). This simple numerical test case demonstrates the theoretical analysis in section 2.3 that confidence and credible intervals for linear models are numerically identical. The conclusions would also be expected to apply to linearized-nonlinear models.

3.2. Very Nonlinear Test Function I

[44] The nonlinear test function is

$$\mathbf{y} = \mathbf{x}/a + \sin(am\mathbf{x}) + \boldsymbol{\varepsilon} \tag{22}$$

where $a = 2$, $m = 0.1$. Twenty samples of $\mathbf{y}$ are first generated with $\mathbf{x} = \{1, 2, \ldots, 20\}$, and subsequently corrupted using one realization of white noise $\boldsymbol{\varepsilon}$, with mean zero and constant variance $\sigma^2 = 1$. This nonlinear function is designed to be nonlinear and to have local minima. For this nonlinear function, linear and nonlinear confidence intervals with no prior information and credible intervals with noninformative priors are calculated for y at $x = 30$ with known σ^2. Table 3 shows that the linear confidence and credible intervals are numerically the same but differ from the nonlinear intervals. Compared to the previous studies listed in Table 1, our results are similar to those of the nonlinear test case of *Vrugt and Bouten* [2002] in that the linear confidence intervals are larger than the nonlinear credible intervals. Our results differ from those of *Gallagher and Doherty* [2007] in that the nonlinear credible interval is smaller than the nonlinear confidence interval. Although sufficient numbers of MCMC simulations were performed, the nonlinear credible intervals of MICA and DREAM differ (Table 3).

[45] To understand better the linear intervals for nonlinear models, Figures 1d–1f were constructed to show the linear confidence and credible intervals. Similar to Figures 1a–1c, Figures 1d–1f plot the PDFs of parameters and prediction based on classical regression theory and Bayesian theory for the linearized model obtained from 500,000 MCMC samples. Again, the figures show that the classical regression (red curve) and Bayesian (blue curve) distributions are equivalent and the linear confidence (red bar) and credible (blue bar) intervals are coincident. This numerically confirms the theoretical analysis in section 2.5 for linear intervals calculated for a nonlinear problem.

[46] Figures 1d–1f also show nonlinear credible intervals. The figures show that PDFs of the linearized model (blue curves) are significantly different from those of the nonlinear model (green curves) obtained from 500,000 MCMC samples. While the PDFs of model parameters are Gaussian for the linearized model and almost Gaussian for the nonlinear model, the PDF of the model prediction for the nonlinear model (Figure 1f) is narrower and taller. This may be attributed to model nonlinearity: total and intrinsic nonlinearities of this model are 0.54 and 0.20, respectively, so that for both this model is rated as nonlinear on the four-tiered scale as shown in footnote of Table 6.

[47] Of the intervals considered, the MCMC credible intervals should provide the most accurate assessment of uncertainty. Model nonlinearity is accounted for to some degree by nonlinear confidence intervals, and Table 3 shows that for this problem the nonlinear confidence intervals are closer to the nonlinear credible intervals than are the linear intervals. The nonlinear confidence intervals are not included in Figure 1, because the underlying theory does not include definition of a PDF. Here, the nonlinear confidence intervals are evaluated on the confidence region. The plotted 95% parameter confidence region is defined using (16) and shown as the black ellipse with objective function value of 18.3 in Figure 2a. Figure 2a also shows prediction contours, and the intersection of two red prediction contours with the black ellipse define the limits (the two red dots) of the nonlinear confidence interval on the prediction with values shown in Table 3. For this nonlinear function defined in (22), the prediction is not monotonic with respect to the parameter values. While, for some orientations of the objective function contours, the shape of the prediction contours could cause the

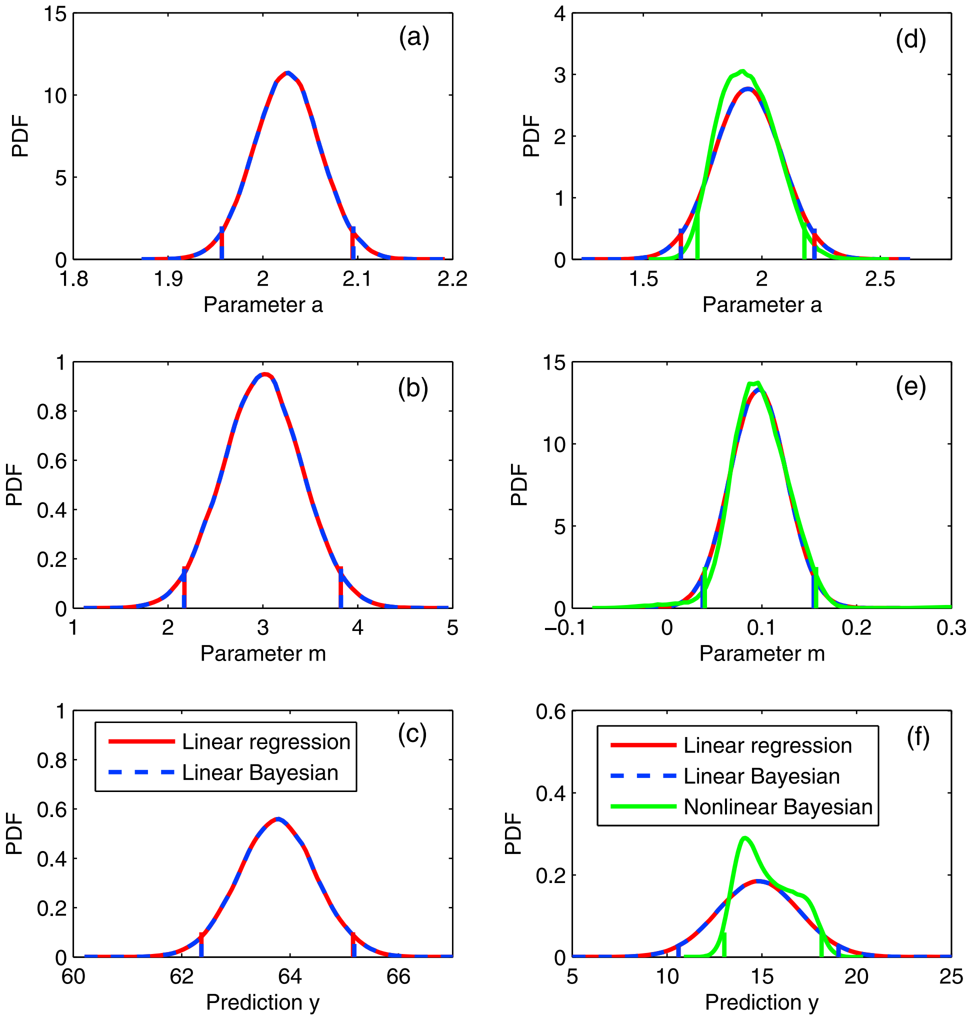

Figure 1. Probability density functions (PDFs) of parameters (*a* and *m*) and prediction (*y*) based on regression and Bayesian theories for the (a–c) linear and (d–f) very nonlinear simple test functions. Figures 1d–1f show results for linearized and nonlinear models. In each plot, the bars correspond to the 95% interval limits. The limits of the linear confidence (red) and credible (blue) intervals overlap for the linear model (Figures 1a–1c) and the linearized model (Figures 1d–1f). The PDFs and intervals are centered on estimated values.

calculation of nonlinear confidence intervals to produce nonunique results, Figure 2a clearly shows that the intersection of the objective function contour (the black ellipse) and the relevant prediction contours (the two red lines) produce unique maximum and minimum predictions (the two red dots). This analysis is consistent with the good performance of UCODE_2005 [*Poeter et al.*, 2005] in finding the nonlinear confidence interval for this problem.

[48] To better illustrate the relation between the confidence and credible intervals, the 500,000 MCMC parameter samples obtained from MICA are plotted in Figure 2a. MICA results in Figure 2a use a multivariate Gaussian proposal distribution with the covariance matrix calculated for the optimal parameters. The plotting procedure is consistent with how the credible interval is calculated, as follows. First, the MCMC parameter samples are ordered using the prediction

Table 3. Predictions and 95% Linear and Nonlinear Confidence and Credible Intervals for Very Nonlinear Simple Test Function I[a]

Type of Interval	Prediction	Lower Limit	Upper Limit	Interval Width	Change in Width Relative to NCrDr(%)
Linear confidence	14.83	10.61	19.05	8.44	65% larger
Linear credible	14.83	10.61	19.05	8.44	65% larger
Nonlinear confidence	14.83	12.80	18.42	5.62	10% larger
Nonlinear credible (MICA)	14.82	13.12	18.09	4.97	3% smaller
Nonlinear credible (DREAM) (NCrDr)	14.84	13.04	18.15	5.11	–

[a] Total and intrinsic model nonlinearities are 0.54 and 0.20, respectively, indicating this is a very nonlinear model. True value of the prediction is 17.2. Intervals are for known σ^2. Results with unknown σ^2 would be similar except that all intervals would be larger.

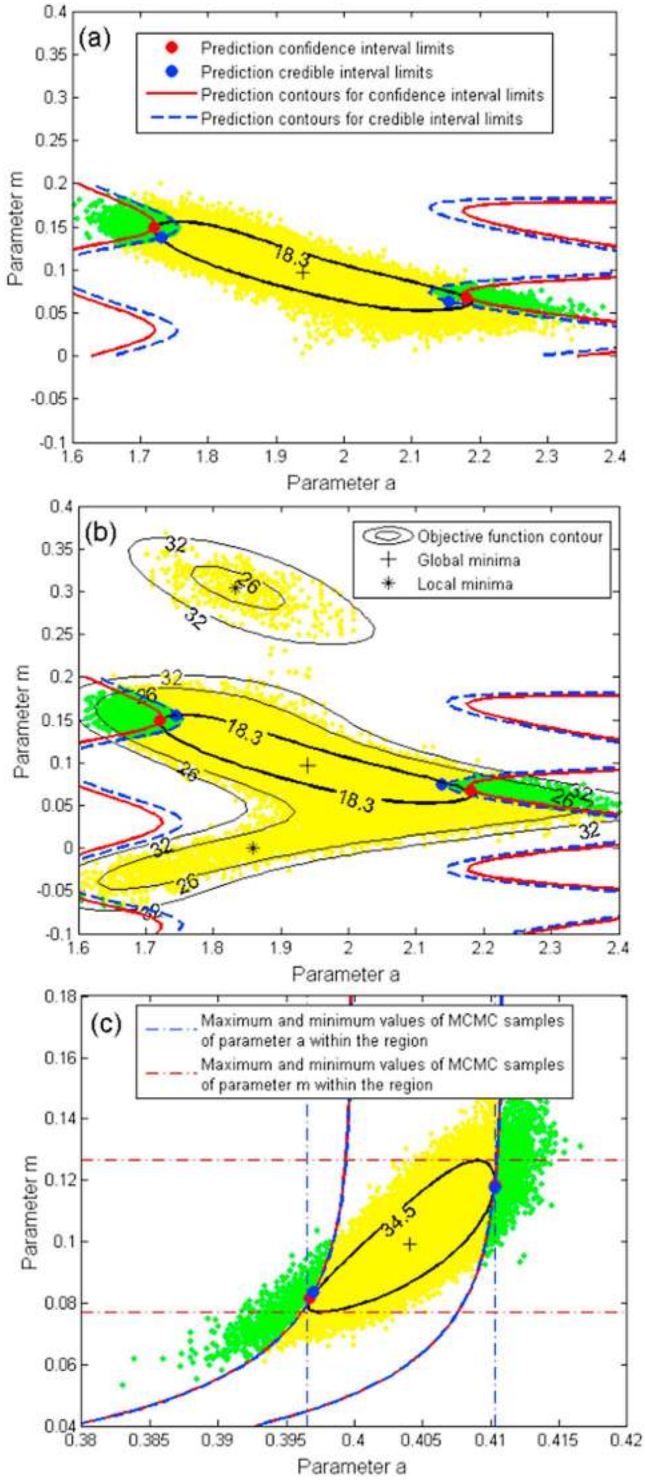

Figure 2. The 95% individual nonlinear confidence and credible intervals for (a and b) nonlinear test function I (dots in Figure 1a are samples from MICA; dots in Figure 1b are from DREAM) and (c) test function II (dots are from DREAM and similar to those from MICA). The black ellipses for an objective function value of 18.3 in (a) and (b) and of 34.5 in (c) are 95% confidence regions defined in equation (16). MCMC samples are plotted as dots: the upper 2.5% and lower 2.5% of predictions are green; the middle 95% of predictions are yellow.

values they produced, starting with the parameter set that produces the smallest value of the prediction and progressing to the parameter set that produced the largest value of the prediction. The green dots at the bottom right corner correspond to the parameter sets that produce the smallest 2.5% of the predictions, and those at the top left corner to the parameter sets that produce the largest 2.5% of the predictions. The yellow dots correspond to the parameter sets that produce the middle 95% of the predictions. The nonlinear credible interval limits equal the predictions that fall exactly at the 2.5% levels. They are identified in Figure 2a by blue dots and the values are listed in Table 3. The two blue lines are contours for the prediction values represented by the two blue dots (i.e., the nonlinear credible interval limits).

[49] Figure 2a shows that the nonlinear confidence interval is larger than the credible interval (the distance between the two red lines is larger than that between the two blue lines), but numerically they should be identical if all the assumptions are satisfied to calculate the confidence interval and the posterior distribution is correctly sampled by MCMC in the calculation of the credible interval. To better understand the discrepancy, Figure 2b was constructed similarly to Figure 2a but using results from DREAM. Figure 2b indicates that there are two local minima (black stars) of the objective function in addition to the one detected by nonlinear regression and MICA (black plus) and the parameter values corresponding to the minima differ from each other substantially, as shown in Table 4. The nonlinear confidence intervals defined by (16) are evaluated on the objective function contours centered on a minimum. For the three different minima, the objective function contours defined by (16) are different and so are the nonlinear confidence intervals, as shown in Table 4.

[50] Most likely, local minima affect estimation of confidence and credible intervals centered at the global minimum more as the values of the local minima approach the value of the global minimum. In Figure 2b the objective function contours defined by (16) centered at the local minima are larger than that of 18.3 centered at the global minimum, and the effect on calculated confidence intervals centered at the global minimum is modest (10% larger as shown in Table 3 and 4). MICA only samples the region surrounding the global objective function minimum (black plus in Figure 2a). DREAM successfully identifies the multimodal target parameter posterior distribution as shown in Figure 2b. Figure 2b indicates that, because of the local minima, the percent of samples covered by the 95% parameter confidence region as shown in Figure 2a is actually less than 95%. In other words, because (16) uses the statistics corresponding to a 95% confidence level to define the confidence region, the calculated nonlinear confidence interval on the prediction is larger than what it should be. It appears that this would be typical in the presence of multimodal posterior PDFs for confidence intervals constructed using the global minimum. Table 3 indicates that the more accurate nonlinear credible interval from DREAM is a little smaller than the nonlinear confidence interval but larger than the incorrect nonlinear credible interval from MICA produced using a single proposal distribution.

3.3. Mildly Nonlinear Test Function II

[51] Nonlinear test function I gives different nonlinear confidence and credible intervals because of unsatisfied

Table 4. Optimal Parameter Values, Minimum Objective Function Values, and Corresponding Prediction Uncertainty Intervals for Very Nonlinear Simple Test Function I[a]

Type of minima	Parameter Value *a*	Parameter Value *m*	Objective Function	Prediction	Lower Limit	Upper Limit	Interval Width	Change in Width Relative to NCrDr (%)[c]
Local 1	1.86	0.00046	23.1	16.16	13.89[b]	18.40[b]	4.51[b]	12% smaller
Local 2	1.83	0.305	24.9	15.50	14.07[b]	17.55[b]	3.48[b]	32% smaller
Global	1.94	0.096	14.7	14.83	12.80[b]	18.42[b]	5.62[b]	10% larger
Global[a]	1.94	0.098	14.7	14.84	13.04[c]	18.15[c]	5.11[c]	–

[a]The first two rows correspond to two local minima marked as asterisks in Figure 2b; the third row corresponds to global minimum marked as "plus" in Figure 2a; the last row corresponds to the global minimum marked as "plus" in Figure 2b.

[b]Nonlinear confidence intervals.

[c]Nonlinear credible interval (DREAM) (NCrDr).

assumptions for the calculation of the confidence interval: there are important multiple minima in the objective function and the intrinsic model nonlinearity is high. Here, we used another nonlinear test function to evaluate whether the calculated nonlinear confidence and credible intervals are closer if the assumptions are satisfied to a greater degree: there are no competing local minima and the intrinsic nonlinearity is smaller. This nonlinear test function, adopted from *Draper and Smith* [1998, p. 475], is

$$\mathbf{y} = a + (0.49 - a)e^{-m(\mathbf{x}-8)} + \boldsymbol{\varepsilon} \tag{23}$$

where $a = 0.4$, $m = 0.1$. Forty samples of **y** are first generated with $\mathbf{x} = \{8, 9, \ldots, 47\}$, and subsequently corrupted using one realization of white noise $\boldsymbol{\varepsilon}$, with mean zero and constant variance $\sigma^2 = 0.01$. For this nonlinear function, nonlinear confidence intervals and credible intervals are calculated for y at $x = 50$.

[52] As expected, the nonlinear confidence and credible intervals are almost the same (Table 5). Figure 2c explores their relations by plotting the confidence interval calculated based on equation (16) and the credible interval determined from DREAM parameter samples based on the equal-tailed method. The coincidence of the red and blue curves confirms the similarity between nonlinear confidence and credible intervals. The reason is that the parameter confidence region is identical to its credible region. As shown in Figure 2c, the 95% individual confidence region is delineated by the black objective function contour with value of 34.5 calculated using equation (16). This contour corresponds to the 95% individual parameter credible region because the probability of each individual parameter within or on the boundary of the region is 95%. This conclusion is drawn in the following procedure for the two parameters *a* and *m*. Based on the 95% individual confidence region (the black ellipse), the minimum and maximum values of parameter *a* are located and represented by the two vertical dotted dash blue lines; the parameter samples between the two lines is 95.02% of all the MCMC samples. Following the same procedure, the minimum and maximum values of parameter *m* are represented by the two horizontal dotted dash red lines and 95.01% of all the MCMC samples are enclosed by the two lines. We can conclude that the 95% confidence and credible regions are numerically identical. Based on the 95% individual parameter confidence region (the black contour in Figure 2c), the 95% individual prediction confidence interval can be evaluated. According to the discussion in section 2.6, these confidence interval limits are the minimum and maximum prediction values on the confidence region, which is equivalent to the credible region. This equivalence means that these two predictions define both 95% individual confidence and credible intervals. The credible intervals produced in this manner include about 95.02% of all the MCMC samples, which again verifies the equivalence of the 95% confidence and credible intervals. This suggests that for this mildly nonlinear problem with small intrinsic nonlinearity, the 95% parameter credible region is identical to its 95% confidence region and the 95% credible interval of prediction evaluated on the credible region is identical to its 95% confidence interval evaluated on the confidence region.

[53] The correspondence of the regions and intervals between classical regression and Bayesian analysis is not an accident for the following reasons. (1) The problem is well conditioned: the ratio of the largest and smallest singular values of $(\mathbf{X}^{\mathrm{T}}\boldsymbol{\omega}^{1/2})$ is about 6.4, which corresponds to the composite scaled sensitivities values being substantial and similar in value and parameter correlation being small [*Hill*, 2010]. In Figure 2c, this is clear because the contour is not elongated parallel to either axis or at an angle to the axes. (2) There is one minimum. (3) The single mode parameter posterior probability distribution can be easily sampled by MCMC techniques and MICA and DREAM produce similar results. (4) The prediction function is monotonic and the maximum and minimum values are on the boundary of the region. (5) The intrinsic model nonlinearity is very small, 0.002, suggesting that the approximate likelihood region is almost correct. Consequently, in this case, equation (16) can be used to calculate an accurate individual nonlinear confidence interval and MCMC samples can accurately represent the posterior distribution of parameters and an accurate individual nonlinear credible interval.

4. Complex Test Case of Synthetic Groundwater Models

[54] The complex test case of synthetic groundwater models is more directly relevant to environmental problems than the simple test cases and is designed to extend our

Table 5. Predictions and 95% Nonlinear Confidence and Credible Intervals for Mildly Nonlinear Simple Test Function II[a]

Type of Interval	Prediction	Lower Limit	Upper Limit
Nonlinear confidence	0.406	0.3998	0.4109
Nonlinear credible	0.406	0.3997	0.4109

[a]Intrinsic model nonlinearity is 0.002. True value of the prediction is 0.401.

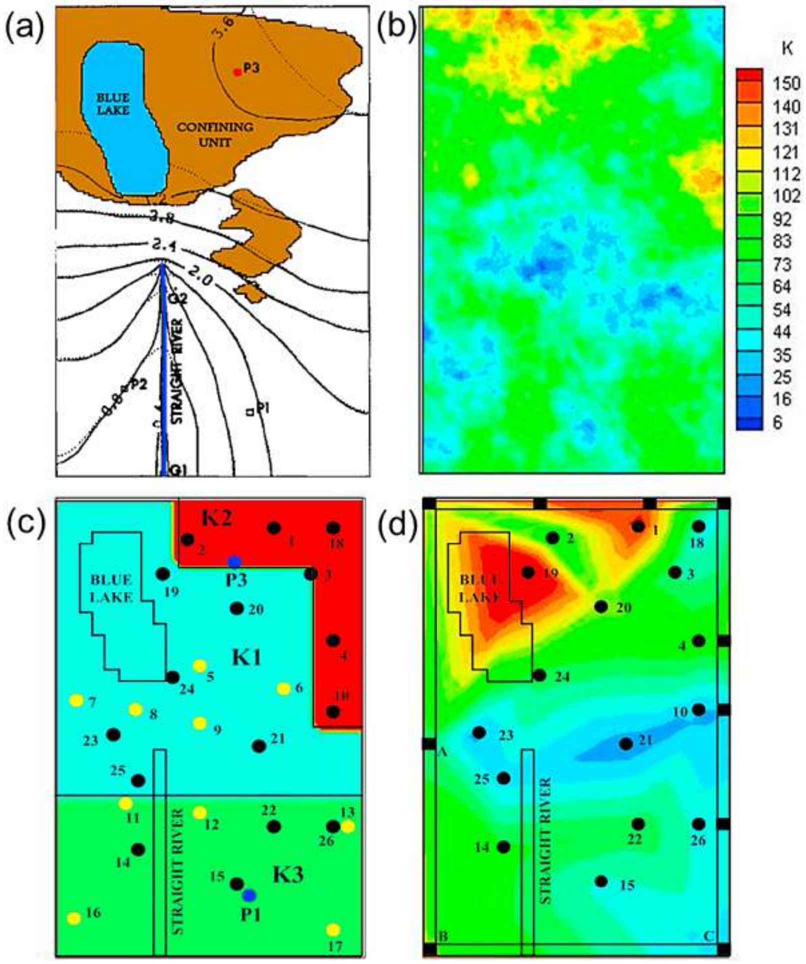

Figure 3. (a) Modeling domain with the confining unit, blue Lake, and Straight River, (b) the true horizontal hydraulic conductivity field (K field) with values from 6 to 150 m/d, and (c–d) the calibrated K fields for the three-zone model (3Z) with K2 up to 317 m/d and the interpolation model (INT) with highest value 180 m/d. (Figure 3c) Configuration of the three zones of K field for 3Z. (Figure 3d) Dots are nodes of linear triangular basis functions constructed to interpolate K field for INT. The hydraulic conductivities are changed in the regression at the sixteen numbered inner dots and at the three squares labeled A, B and C; the hydraulic conductivities of the other seven squares without labels are set. In Figure 3c dots represent 27 wells where hydraulic heads are observed at the top and bottom of the system. Black dots (same as those in Figure 3d) also have hydraulic conductivity measurements and yellow dots only have heads observations; pumping is taken at wells P1 and P3 (blue dots), and drawdown is predicted at well P3.

understanding of the confidence and credible intervals to more practical issues in the context of model uncertainty. We consider several alternative models with different levels of model error, and evaluate the measures of uncertainty in the context of true prediction values and the assumptions required to obtain accurate confidence and credible intervals.

4.1. Model Description, Sensitivity Analysis, and Calibration

[55] This test case includes simulations using a synthetic true model and three alternative calibrated models developed based on data from the true model; all the models are steady state. The true model is described in *Hill et al.* [1998], where it was used to study nonlinear regression methods. The three-dimensional modeling domain is an undeveloped alluvial valley with top water table boundary (subject to areal recharge) and impermeable bottom and lateral boundaries. As shown in Figure 3a, the system has a clay confining unit, a lake (Blue Lake), and a river (Straight River). In the true and all calibrated modes, the river is modeled as a head-dependent boundary. All models use the true distribution of the confining layer shown in Figure 3a because *Hill et al.* [1998] showed that results were insensitive to reasonable variations. Each model layer is characterized by heterogeneous horizontal hydraulic conductivity with values ranging from 6 to 150 m/d in the true model (Figure 3b) and 27 to 317 m/d in the calibrated models (Figures 3c and 3d). No vertical variation in aquifer hydraulic conductivity is considered in any model.

[56] Development scenario A of *Hill et al.* [1998] considered in this study includes pumpage at wells P1 and P3 (locations shown in Figure 3c). Model predictions of interest are drawdown of the water table at well P3 and percent decrease of streamflow at the gauge site G2 (Figure 3a).

[57] The three alternative models represent three different ways of parameterizing the heterogeneous horizontal

hydraulic conductivity. The first model (HO) is the simplest, treating the domain as homogeneous. The second model (3Z) has three zones shown in Figure 3c, which produced the best overall model calibration among many other zone configurations considered by *Hill et al.* [1998]. In the third model (INT), the field of horizontal hydraulic conductivity is parameterized by interpolation using linear triangular basis functions constructed on the points in Figure 3d. Model HO is developed in this study; models 3Z and INT are modified from those of *Hill et al.* [1998].

[58] The alternative models differ from the true model in several ways. The most important difference is that, instead of modeling the lake as a head-dependent boundary as in the true model and *Hill et al.* [1998], the volume of the lake is simulated as high hydraulic conductivity cells to avoid the modeler intervention described by *Hill et al.* [1998]. The true model has five layers; the calibrated models all have three as follows. In the north, the bottom of layer 1 coincides with the bottom of the lake; in the south, the bottom of layer 1 is deep enough to ensure that no cell goes dry during the simulation. The bottom of layer 2 coincides with the top of the confining unit which is simulated as vertical leakance between model layers 2 and 3.

[59] The defined parameters common to all the models are net lake recharge (LAKERCH), areal recharge rate (RCH), leakance of the confining unit (KV), and vertical anisotropy (VANI). The calibrated parameters specific to the individual models are hydraulic conductivities and the use of one (for models HO and 3Z) or three (for model INT) parameters for the conductance of the riverbed.

[60] Calibration data include 54 observations of hydraulic head from 27 wells (two heads from each well in layers 1 and 3, Figure 3c) and one observation of lake stage. Observations of streamflow gain (groundwater discharge to the stream) are also used in the calibration. To investigate the effects of more streamflow data on prediction accuracy and uncertainty, two steamflow gain data sets are used: two observations from gauges G1 and G2 (Case I) or eighteen observations from each cell of the river (Case II). In Case I, the observation of G2 and the difference between G1 and G2 are used in the calibration. The level of detail provided by Case II is generally not realistic in practice but here the value of such data is evaluated. Prior information on LAKERCH is used in all three models. For model INT, prior information also includes sixteen measurements of hydraulic conductivity from the wells shown in Figure 3d.

[61] Errors added to the observed heads and the lake stage are based on typical values for the data type; the weights used in model calibration are calculated solely based on the added measurement errors and do not account for model error. The observations of hydraulic head and lake stage have white noise with variance of 0.01 m^2, the inverse of which is the weights. For measurements of net lake recharge and hydraulic conductivity used as prior information, coefficients of variation of 50% and 20%, respectively, are used to define the added white noise and the weights. No measurement errors are added to the created steamflow data observations. Realistic coefficients of variation are used to determine the weighting: 10% and 20% in cases I and II, respectively. A larger value is used in Case II because it is expected that such detailed flows are likely to be measured with larger percent errors. Because no errors are added, any discrepancies are due only to model error.

[62] Estimated parameters are selected based on the composite scaled sensitivity (CSS) and parameter correlation coefficient (PCC) obtained from a sensitivity study [*Hill and Tiedeman*, 2007, p. 50–54]. CSS values identify parameters that possibly can be reasonably estimated using the calibration data. Sensitive parameters that are highly correlated with one or more other parameters, as indicated using PCC, possibly cannot be estimated uniquely with the available data. If present, selected highly correlated parameters generally need to be removed from the set of estimated parameters. Take the 3Z model as an example: Figures D1a and D1c in Appendix D of the auxiliary material show that parameter RCH has the highest CSS values and VANI and KV have the smallest. This is also true for the HO and INT models. In this work, using heads alone would result in all parameters except VANI to be completely correlated (absolute value of PCC equal to 1.00) so that the parameters would be interdependent and nonunique. This is because all hydraulic conductivities and boundary flows would have been estimated with no flow specified or observed. Adding the measured streamflow gains and losses provide sufficient information that all the parameters can be uniquely estimated: the largest absolute value of PCC was 0.93, which is too much smaller than 1.00 to cause nonunique parameter estimates.

[63] Model calibration is conducted using the modified Gauss-Newton method in UCODE_2005 to minimize the following objective function

$$S(\mathbf{b}) = \sum_{i=1}^{NH} \omega_{h_i}\left(y_{h_i} - \hat{y}_{h_i}(\mathbf{b})\right)^2 + \sum_{j=1}^{NQ} \omega_{q_j}\left(y_{q_j} - \hat{y}_{q_j}(\mathbf{b})\right)^2 + \sum_{k=1}^{NPR} \omega_{p_k}\left(y_{p_k} - \hat{y}_{p_k}(\mathbf{b})\right)^2 \qquad (24)$$

where $\mathbf{b}$ is a vector of NP parameters; NH and NQ equal the number of hydraulic-head and streamflow observations, respectively; NPR is the number of prior information values; y_{h_i} and $\hat{y}_{h_i}(\mathbf{b})$ represent the observed and simulated hydraulic head, respectively; ω_{h_i} represents the weight for the head observation; y_{q_j} and $\hat{y}_{q_j}(\mathbf{b})$ represent the observed and simulated streamflow, respectively; ω_{q_j} represents the weight for the steamflow; y_{p_k} and $\hat{y}_{p_k}(\mathbf{b})$ represent the prior parameter value and corresponding estimate of the parameter in the regression, respectively; and ω_{p_k} represents the weight for the prior estimate.

4.2. Simulating the Confidence and Credible Intervals

[64] The MCMC simulation of credible intervals is conducted using DREAM. The prior distributions of parameters used in the Bayesian analysis are consistent with the prior information used in the classical regression. For parameters without prior information in the regression, uniform prior distributions with large bounds are used in MCMC to create noninformative priors; for parameters with prior information in the regression, Gaussian prior distributions are applied in MCMC with the mean and variance set to equal y_{p_k} and $\omega_{p_k}^{-1}$ of (24). Details of the parameter prior distribution can be found in Appendix E of the auxiliary material.

[65] Ten chains are used in the MCMC simulation, each of which generated 20,000 samples after the chains converge. The convergence is monitored by the factor R, as

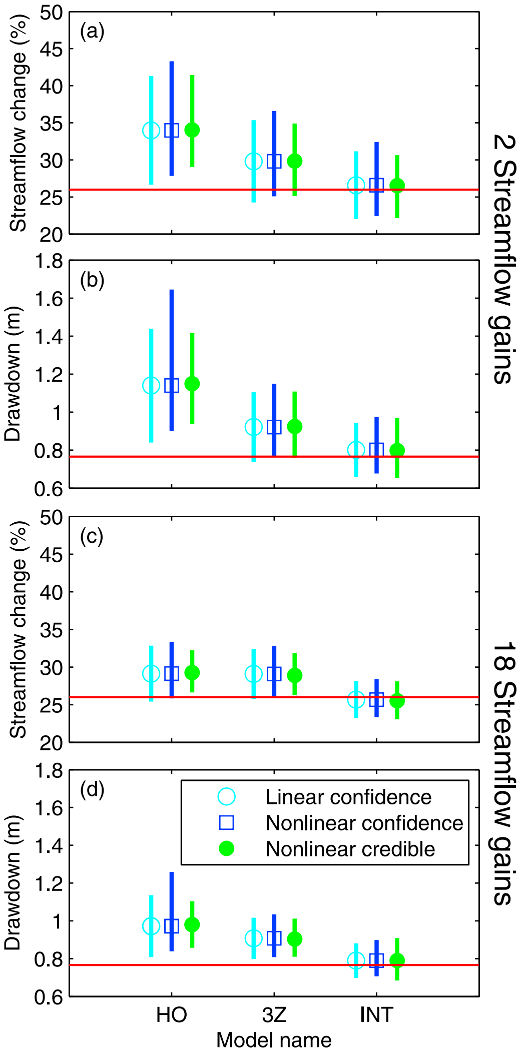

Figure 4. Linear and nonlinear confidence intervals and nonlinear credible intervals (≤106, ≤1594, and 420,000 model runs, respectively) for two predictions: (a and c) streamflow change at gauge site G2 and (b and d) drawdown of the water table at well P3. Here 2 or 18 observations of streamflow gain are used. The horizontal lines represent true values of the predictions. The nonlinear credible intervals are calculated using DREAM.

defined and suggested by *Gelman et al.* [1995]. The largest R for the first 2000 iterations (total 20,000 parameter samples of 10 chains) is 1.05, which suggests convergence to the posterior probability distribution because it is below the critical value 1.2). Model predictions are calculated using the 200,000 samples, and for each prediction the 95% equal-tailed credible interval is estimated subsequently.

[66] Linear measures of parameter uncertainty required 2 × NP + 1 forward model runs without pumping, where NP = 6, 8, or 26 for models HO, 3Z, and INT, respectively. The same number of prediction model runs with pumping is required to propagate the parameter uncertainty into prediction uncertainly (i.e., based on equation (11)). Thus, to calculate linear confidence intervals on predictions, 26, 34, and 106 model runs are needed for the HO, 3Z, and INT models, respectively.

[67] Nonlinear confidence intervals on predictions required an optimization process for each interval limit. In this work, the two nonlinear intervals for each of the HO, 3Z, and INT models required about 303, 412, and 1594 model runs, respectively.

[68] Nonlinear credible intervals require about 420,000 model runs without and with pumping for each model. Any number of predictions can be simulated with the MCMC results, so that if many predictions are considered MCMC can become competitive with nonlinear confidence intervals [*Shi et al.*, 2012]. However, the number of predictions is rarely that large.

4.3. Linear and Nonlinear Confidence and Credible Intervals

[69] The 95% linear and nonlinear confidence and credible intervals calculated for two predictions and either two or eighteen streamflow gain observations are shown in Figure 4. We have already shown and demonstrated (Table 1 and Figures 1a–1c) that linear confidence and credible intervals are numerically identical (as discussed in sections 2.3 and 2.4), so only linear confidence intervals are plotted.

[70] Figure 4 shows that, for models 3Z and INT, the three intervals for both predictions in both cases are very close to each other relative to the distance from the prediction to the true value. For HO, the credible intervals are smaller than both the linear and nonlinear confidence intervals, especially for predicted drawdown (Figure 5). The reasons are multiple. First, HO is the least accurate model. Second, the intrinsic nonlinearity of HO is the largest as shown in Table 6 and Figure 5. Third, the prediction function of drawdown is not monotonic with respect to parameters LAKERCH and RCH, though the insensitivity of these parameters to drawdown and the slight nonmonotonicity suggest the last effect may be small.

[71] The results presented in this work cannot be used to test the significance level of the intervals because only a few samples are considered. However, true values located far outside the intervals can be used to suggest that the intervals are too small. The intervals for HO in Figures 4a, 4b, and 4d exclude the true values by enough that the accuracy of the prediction significance level is drawn into question. For model 3Z, the true value of drawdown is excluded from both the confidence and credible intervals after including the more detailed observations, suggesting that model 3Z may not produce reliable measures of prediction uncertainty. These examples are consistent with the idea that simple models tend to underestimate prediction uncertainty.

[72] Using the more detailed observations of streamflow gain reduces interval width of predictions for all models due to reduced parameter uncertainty. In model 3Z, for example, the 95% linear confidence interval for parameter K2 is reduced from 134–741 m/d to 136–557 m/d by adding the more detailed observations. Using the more detailed observations of streamflow gain also improves predictions accuracy slightly, as evidenced by the simulated values in Figures 4c and 4d being slightly closer to the true values than those in Figures 4a and 4b.

4.4. Evaluation

[73] In this section, the reasons that the true values are excluded from the confidence and credible intervals of models HO and 3Z are investigated by examining

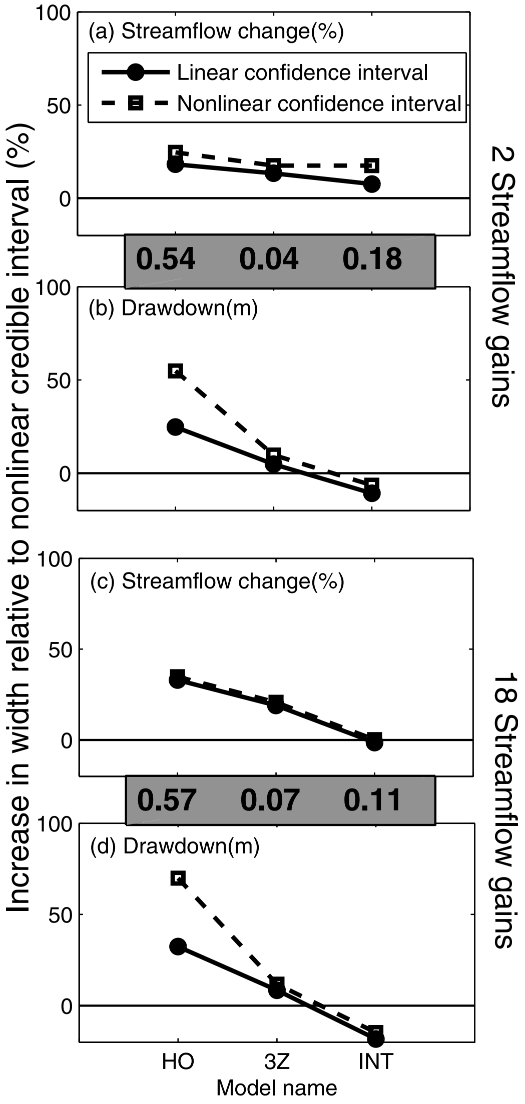

Figure 5. Relative increase in width of linear and nonlinear confidence intervals to nonlinear credible intervals for streamflow change and drawdown with (a–b) 2 and (c–d) 18 observations of streamflow gain. The measures of intrinsic nonlinearity from Table 6 are inserted in the shaded boxes for easy comparison.

assumptions required for deriving accurate intervals. Derivation of accurate linear and nonlinear confidence intervals assumes that true errors are Gaussian with zero mean and that model errors are either insignificant relative to measurement errors (i.e., the models are nearly error-free representations of the system) or the model errors are zero-mean and random (i.e., the models have no distinct biases), and en masse produce a Gaussian contribution to errors. The Gaussian assumptions are needed to construct confidence intervals (though not for the regression to optimize parameter values) but in theory are not needed for credible intervals because non-Gaussian likelihood functions and prior distributions are allowed in Bayesian analysis. For consistent comparison, this work uses a Gaussian likelihood function and, when applicable, a Gaussian prior distribution. Thus, this work does not investigate situations in which non-Gaussian errors are accounted for by credible intervals, but only considers the ability of credible intervals to correctly account for model nonlinearity, including resulting local minima. Using the Gaussian structure for non-Gaussian errors may lead to less accurate model simulations [*Sun et al.*, 2009] and overestimation of predictive uncertainty [*Schoups and Vrugt*, 2010]. The assumptions about Gaussian distributed errors and the size and characteristics of model errors are examined for both the confidence and credible intervals.

[74] However, the focus on Gaussian residual distributions is expected to be of wide interest because they are apparently often applicable and commonly used [*Renard et al.*, 2011]. For non-Gaussian residuals, as models of a given system evolve, bias in residuals may diminish and residual structures may become Gaussian after certain transforms (e.g., the Box-Cox transformation). In this sense, the Gaussian assumption of residuals may become increasingly realistic. On the other hand, our complex example suggests that model errors are similarly debilitating for both confidence and credible intervals, and dominate over the differences between confidence and credible intervals. Therefore, faith that somehow credible intervals are more reliable in the face of model error seems somewhat questionable based on the results shown here.

[75] To investigate whether true errors are Gaussian or equivalently weighted true errors are standard normal, the weighted-residuals are evaluated. Weighted-residuals tend to be correlated even when true errors are independent, but

Table 6. Total Nonlinearity, Intrinsic Nonlinearity, Correlation Coefficient and Regression Statistics and Their Critical Values, in Parentheses, of the Three Alternative Groundwater Models

	Case I: 2 Streamflow Gains			Case II: 18 Streamflow Gains		
Statistic	HO	3Z	INT	HO	3Z	INT
Total nonlinearity[a]	2.44	0.81	13.85	2.79	0.91	2.27
Intrinsic nonlinearity[a]	0.54	0.04	0.18	0.57	0.07	0.11
R^2_N[b]	0.989 (0.96)	0.986 (0.96)	0.989 (0.97)	0.987 (0.97)	0.988 (0.97)	0.982 (0.97)
s[c]	1.49 (1.25–1.84)	1.27 (1.06–1.57)	1.05 (0.88–1.30)	1.33 (1.14–1.60)	1.17 (1.0–1.4)	0.95 (0.81–1.14)

[a] Calculated using equation (7.15) and (7.16) of *Hill and Tiedeman* [2007] by UCODE_2005. Critical values for both measures: >1.0 highly nonlinear; 0.09–1.0 nonlinear; 0.01–0.09 moderately nonlinear; <0.01 effectively linear.

[b] Calculated using equation (6.18) of *Hill and Tiedeman* [2007] by UCODE_2005. R^2_N normality test for weighted-residuals evaluated for the observations and the prior information. Critical values are in parentheses. Larger R^2_N values indicate normally distributed weighted-residuals.

[c] Confidence intervals for s are shown in parentheses and are calculated using equation (6.2) of *Hill and Tiedeman* [2007] by UCODE_2005. Confidence intervals that are entirely above 1.0 indicate model fit that is significantly worse than would be consistent with the weighting. If the weighting is error-based, as in this study, such large values of s indicate important model error and possibly model bias.

normality of the weighted-residuals suggests normality of the true errors. Also, it is the normality of the weighted-residuals that is important for (16) used to calculate nonlinear confidence intervals. The statistical variable, R_N^2 is the correlation coefficient between the weighted-residuals (ordered from smallest to largest) and the normal order statistics [*Brockwell and Davis*, 1987, p. 304; *Hill and Tiedeman*, 2007, p. 110], and can be used to test if the weighted-residuals are normally distributed and also satisfy the more demanding condition of independence. R_N^2 provides a more powerful test than the common Kolmogorov-Smirnov statistic. Values of R_N^2 close to 1.0 indicate that the weighted-residuals are independent and normally distributed. Its values and corresponding critical values at significance level of 0.05 are listed in Table 6 for the three models. Because the R_N^2 values are all larger than the critical values, it is concluded that the assumption of error normality is satisfied for all models in the synthetic groundwater test case.

[76] To investigate whether model errors are significant relative to measurement errors, *Hill and Tiedeman* [2007, p. 303], *Aster et al.* [2012], and others suggest using the standard error of regression, $s = \left[S\left(\hat{\mathbf{b}}\right)/(NH + NQ + NPR - NP)\right]^{1/2}$, where $S\left(\hat{\mathbf{b}}\right)$ is from minimizing equation (24). If the model fit is consistent with assigned weighting then the calculated standard error of regression s is close to 1.0. Larger values of s indicate either additional errors, commonly epistemic and aleatory errors, not accounted for in the weighting or that the weights do not correctly reflect the intended errors [*Hill and Tiedeman*, 2007]. In this synthetic study the measurement error is known and incorporated properly in the weights for all observations except flows, but no model error is included. For flows, the lack of added error and use of coefficients of variation of 10% and 20% mean that even values of 1.0 suggest some model error. Thus, s values larger than 1.0 indicate model errors. The s values and their corresponding 95% confidence intervals listed in Table 6 suggest significant model errors for models HO and 3Z. The model errors of HO and 3Z mainly result from oversimplified conceptualizations of the horizontal hydraulic conductivity (Figure 3).

[77] Other investigations of model error consider plots of weighted-residuals and optimized parameter values. Figures D1b and D1d in Appendix D of the auxiliary material plot weighted-residuals against weighted simulated values from model 3Z for the case of using two and eighteen observations of streamflow gain for calibration, respectively. While the calibration results are acceptable, the weighted-residuals of model 3Z are not fully random in that they are mostly positive for weighted simulated values between 26 and 32. This may indicate systematic model error in model 3Z. Though results for HO and INT are not shown, model HO has worse overall fit than 3Z and model INT has the best fit with weighted-residuals randomly distributed. Both models HO and 3Z suffer from biased parameter estimates. For example, for model 3Z the estimated value of K2 equals 317 m/d, more than two times the largest hydraulic conductivity of the true model (Figure 3); the estimated values of riverbed conductance for models HO and 3Z, which equal 312.94 and 303.29, respectively, are far from the true value of 244 m^2/d per meter of stream, however this is not clear evidence of model bias because streambed conductances depend on grid size [*Mehl and Hill*, 2010]. In this synthetic problem, we know that the bias is mainly caused by oversimplification of horizontal hydraulic conductivity. No such problems occur for the INT model, for which the hydraulic conductivity parameterization is most closely coordinated with the pattern of the true field (Figures 3b and 3d).

[78] The groundwater examples suggest that both intrinsic model nonlinearity and model error affect the differences between nonlinear confidence and credible intervals, but their relation is not simple. Table 6 and Figure 5 show that the largest differences between the confidence and credible intervals occur for model HO, which has the largest intrinsic model nonlinearity and worst model fit. However, results for models 3Z and INT do not support this relation clearly. For example, the smaller standard error of regression s in Table 6, the distribution of weighted-residuals, and more reasonable parameter values indicate that INT has less model error than 3Z, but some of the confidence intervals are closer to the credible intervals for 3Z than for INT (Figure 5d). This is probably because the intrinsic model nonlinearity of INT (0.18 and 0.11 for case I and II) is larger than that of 3Z (0.04 and 0.07) (Figure 5). For both, difficulties associated with model error remain.

5. Discussion

[79] This work is focused on understanding differences and similarities between the confidence and credible intervals from theoretical and heuristic perspectives. We limit our discussions and conclusions to the problems that the observation errors are multivariate Gaussian distributed, which is commonly useful in environmental modeling. In this study, linear and nonlinear confidence and credible intervals are considered for both linear and nonlinear models; results are summarized and compared to other published work in Table 1.

[80] For *linear* confidence and credible intervals of *linear* models, analytical expressions were used to show that, when no prior information is used in the regression and the non-informative prior parameter distribution is used in the Bayesian calculation, the two kinds of intervals are numerically identical. If prior information used in the regression is consistent with the prior distribution used in the Bayesian calculation, the confidence intervals are equivalent to the credible intervals; otherwise, the two kinds of intervals differ.

[81] For *linear* confidence and credible intervals of *linearized-nonlinear* models, we show that the two kinds of intervals are again numerically identical. However, both confidence and credible intervals are approximate, and their accuracy depends on model nonlinearity and model error.

[82] For *nonlinear* confidence and credible intervals of *nonlinear* models, analytical expressions are not available and heuristic investigations were pursued. Confidence intervals were calculated using the approximate likelihood methods of *Christensen and Cooley* [1999] and *Cooley* [2004], while credible intervals were calculated using the MCMC techniques of *Gallagher and Doherty* [2007] and *Vrugt et al.* [2008, 2009].

[83] Three numerical experiments are conducted to explore the differences between linear and nonlinear confidence and credible intervals for nonlinear models. The experiments with simple nonlinear functions indicate that the intervals are similar for consistent priors, as long as the

assumptions used to calculate the confidence intervals are satisfied and the credible intervals are correctly simulated by MCMC techniques. Therefore, in practice, the expected advantage produced by calculating computationally demanding nonlinear credible intervals can be evaluated by considering the underlying assumptions. In the very nonlinear test function I, the nonlinear credible interval is smaller than the linear and nonlinear confidence intervals because the intrinsic model nonlinearity is large and multiple minima exist in the objective function. But in the mildly nonlinear test function II, when all the assumptions are satisfied, the confidence and credible parameters regions are numerically coincident, and correspondingly the nonlinear confidence and credible intervals of predictions are numerically identical. Even in the very nonlinear test case I, the linear and nonlinear confidence intervals constructed using the global minima were close enough to the credible intervals to suggest potential utility given the small number of model runs required by the confidence intervals (10 and 232 for the linear and nonlinear confidence intervals, respectively; 1,000,000 for the credible intervals).

[84] The numerical experiment of groundwater modeling introduces the complication of model error and resulting prediction bias on the performance of confidence and credible intervals. The confidence and credible intervals are generally closer to each other when model error is small and the intrinsic model nonlinearity is not very large, despite some exception for the 3Z and INT models (Figure 5). For model HO, which has greater model error and nonlinearity, the difference between confidence and credible intervals is the largest. However, the more important result is that for model HO all of the intervals perform similarly and not very well in that they excluded the true value by a wide margin. While more experience is needed to determine common performance, this example and theoretical considerations suggest poor performance of all intervals is likely when model bias dominates. This suggests that it may be useful to calculate the less computationally demanding confidence intervals early in model development, and calculate the computationally demanding credible intervals as the model becomes a better representation of the system. An advantage of this approach is that it allows more routine calculation of uncertainty intervals and associated measures of, for example, the importance of observations and parameters to predictions [*Tonkin et al.*, 2007; *Dausman et al.*, 2010]. This results in greater understanding of simulated dynamics and increased model transparency.

[85] We suggest that model nonlinearity can be measured using the intrinsic model nonlinearity statistic [*Bates and Watts*, 1980; *Cooley*, 2004]. A difficulty with this statistic is that it considers only the nonlinearity of observations with respect to the parameters, a difficulty discussed by *Hill and Tiedeman* [2007, p. 189–193] and *Cooley* [2004]. In this work, the predictions considered are similar to the observations used and intrinsic model nonlinearity is expected to provide information about the linearity of the predictions as well. When predictions differ substantially from observations due to differences between calibration and prediction conditions or because different quantities are considered, it may not be as useful a measure of nonlinearity.

[86] The results of very nonlinear simple test case I suggests that the match between nonlinear confidence and credible intervals is affected strongly by the existence of local minima. Of concern are how common local minima are in practice, what types of models are likely to have local minima, and how local minima can be detected, as also noted, for example, by *Kavetski et al.* [2006] in the context of surface hydrologic modeling. When models include many kinds of observations and are characterized by few parameters or many parameters with substantial prior information or regularization, parameter sensitivities are likely to be larger, absolute values of parameter correlation coefficients are likely to be further from 1.00, the inverse problem is more likely to be well posed, and local minima are likely to be less common. Thus, decisions made regarding model data and construction are critical to how similar confidence and credible intervals are likely to be in a given problem. *Kavetski and Clark* [2010] also discuss the importance of model numerical methods to the presence of local minima.

6. Conclusions

[87] The results presented in this work suggest that for *linear* models, confidence and credible intervals are mathematically and numerically identical when either (1) parameter prior information is absent for the confidence intervals and noninformative for the credible intervals or (2) any prior information defined for the confidence intervals is consistent with informative prior defined for the credible intervals. Because prior information for the confidence intervals is defined using only means, variances, and covariances, consistency means that the prior information in the credible intervals needs to be Gaussian.

[88] For *nonlinear* models, confidence and credible intervals can be numerically identical when all the assumptions are satisfied in the calculation of confidence intervals and credible intervals are calculated precisely. In practice, problem complexity leads to difficulties for both types of intervals. This work shows how competing local minima degrade the accuracy of confidence intervals. MCMC methods used to calculate credible intervals can suffer from numerical imprecision caused by inadequate sampling of the parameter space; long execution times and large parameter dimensions of many practical models can make it difficult or impossible to conduct enough model runs for sufficiently sampling. The accuracies of both confidence and credible intervals suffer from model errors. In practice nonlinear confidence intervals often differ from nonlinear credible intervals and the differences do not in themselves indicate which uncertainty interval should be more trusted. It is often assumed that the difficulties cited here for confidence intervals are more debilitating than those cited for credible intervals. This work suggests that the situation is likely to be more nuanced, and that both confidence and credible intervals can be important to uncertainty evaluation of environmental models.

[89] The computational expense of credible intervals (for the complex problem considered in this work, about 500,000 model runs versus about 100 and 1500 model runs for linear and nonlinear confidence intervals, respectively) is likely to influence the choice of the uncertainty analysis method used. Whether one uses confidence intervals or credible intervals, the results of this work indicate that the difference between alternative models can be more critical than the differences between confidence and credible intervals in many practical circumstances.

[90] Suggestions for users include the following. (1) Report all uncertainty intervals with an evaluation of model error. Model error can be evaluated using the sum of squared weighted-residuals when the weighting is based on observation error (including epistemic error), as well as plots of weighted-residuals and evaluation of estimated parameter values. For nonlinear models, linear intervals and nonlinear confidence intervals need to be reported with an evaluation of model nonlinearity. (2) Nonlinear confidence intervals can be estimated accurately using the approximate likelihood method, if the assumptions are approximately satisfied. The intervals should be reported with an examination of the assumptions. (3) Nonlinear credible intervals calculated from MCMC techniques can be estimated accurately if the parameter space is adequately explored. MCMC techniques face their own challenges for high-dimension problems with multimodal distributions because it is difficult to adequately search high-dimensional parameter spaces. (4) It appears likely that less computationally demanding confidence intervals are adequate early in model development when model error is likely to adversely affect all types of intervals. This provides more opportunity to take advantage of the rich set of related methods for identifying observations and parameters important to predictions. Use of the computationally demanding credible intervals can be reserved for when the model becomes a better representation of the system.

[91] **Acknowledgments.** The authors thank David Draper and Bruno Mendes for inspiring discussions, John Doherty and Jasper Vrugt for providing the MICA and DREAM codes, respectively, used for the MCMC simulation, Dmitri Kavetski for helping to verify the MCMC results, and Alberto Montanari and three anonymous reviewers for comments on the manuscript. This work was supported in part by NSF-EAR grant 0911074, DOE-SBR grant DE-SC0002687 and DOE Early Career Award DE-SC0008272.

References

Aster, R. C., B. Borchers, and C. H. Thurber (2012), *Parameter Estimation and Inverse Problems*, 2nd ed., 360 pp., Elsevier, Amsterdam.

Bates, D. M., and D. G. Watts (1980), Relative curvature measures of nonlinearity, *J. R. Stat. Soc. B*, *42*(1), 1–25.

Bates, D. M., and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications*, 365 pp., John Wiley and Sons, New York.

Beale, E. M. L. (1960), Confidence regions in non-linear estimation, *J. R. Stat. Soc., B*, *22*(1), 41–76.

Box, E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, 588 pp., Wiley, New York.

Brockwell, P. J., and R. A. Davis (1987), *Time Series, Theory and Methods*, 600 pp., Springer, New York.

Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210, doi:10.1029/WR022i002p00199.

Casella, G., and R. L. Berger (2002), *Statistical Inference*, 2nd ed., 660 pp., Duxbury, Belmont, Calif.

Chen, M., and Q. Shao (1999), Monte Carlo estimation of Bayesian credible and HPD intervals, *J. Comput. Graph. Statist.*, *8*(1), 69–92.

Christensen, S., and R. L. Cooley (1999), Evaluation of confidence intervals for a steady-state leaky aquifer model, *Adv. Water Resour.*, *22*(8), 807–817, doi:10.1016/S0309-1708(98)00055-4.

Christensen, S., and R. L. Cooley (2005), User guide to the UNC process and three utility programs for computation of nonlinear confidence and prediction intervals using MODFLOW-2000, *U.S. Geol. Surv. Tech. Methods Rep.*, *2004-1349*.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Cook, R. D., and S. Weisberg (1982), *Residuals and Influence in Regression*, 230 pp., Chapman and Hall, New York.

Cooley, R. L. (1977), A method of estimating parameters and assessing reliability for models of steady-state groundwater flow: 1. Theory and numerical properties, *Water Resour. Res.*, *13*(2), 318–324, doi:10.1029/WR013i002p00318.

Cooley, R. L. (1979), A method of estimating parameters and assessing reliability for models of steady state groundwater flow: 2. Application of statistical analysis, *Water Resour. Res.*, *15*(3), 603–617, doi:10.1029/WR015i003p00603.

Cooley, R. L. (1983), Incorporation of prior information on parameters into nonlinear regression groundwater flow models: 2. Applications, *Water Resour. Res.*, *19*(3), 662–676, doi:10.1029/WR019i003p00662.

Cooley, R. L. (1993a), Exact Scheffé-type confidence intervals for output from groundwater flow models: 1. Use of hydrogeologic information, *Water Resour. Res.*, *29*(1), 17–33, doi:10.1029/92WR01863.

Cooley, R. L. (1993b), Exact Scheffé-type confidence intervals for output from groundwater flow models 2. Combined use of hydrogeologic information and calibration data, *Water Resour. Res.*, *29*(1), 35–50, doi:10.1029/92WR01864.

Cooley, R. L. (1997), Confidence intervals for ground water models using linearization, likelihood, and bootstrap method, *Ground Water*, *35*(5), 869–880, doi:10.1111/j.1745-6584.1997.tb00155.x.

Cooley, R. L. (1999), Practical Scheffe-type credibility intervals for variables of a groundwater model, *Water Resour. Res.*, *35*(1), 113–126, doi:10.1029/98WR02819.

Cooley, R. L. (2004), A theory for modeling groundwater flow in heterogeneous media, *U.S. Geol. Surv. Prof. Pap.*, *1679*.

Cooley, R. L., and R. L. Naff (1990), Regression modeling of ground-water flow, in *U.S. Geological Survey Techniques of Water-Resources Investigations*, Book 3, Chap. B4, 232 pp., U.S. Geol. Surv., Washington, D.C.

Cooley, R. L., and A. V. Vecchia (1987), Calculation of nonlinear confidence and prediction intervals for groundwater flow models, *Water Resour. Bull.*, *23*(4), 581–599, doi:10.1111/j.1752-1688.1987.tb00834.x.

Dausman, A. M., J. Doherty, C. D. Langevin, and M. C. Sukop (2010), Quantifying data worth toward reducing predictive uncertainty, *Ground Water*, *48*(5), 729–740, doi:10.1111/j.1745-6584.2010.00679.x.

Doherty, J. (2003), *MICA: Model-independent Markov Chain Monte Carlo Analysis*, Watermark Numer. Comput., Brisbane, Australia.

Doherty, J., and R. J. Hunt (2010), Response to comment on "Two statistics for evaluating parameter identifiability and error reduction," *J. Hydrol.*, *380*, 489–496, doi:10.1016/j.jhydrol.2009.10.012.

Donaldson, J. R., and R. B. Schnabel (1987), Computational experience with confidence regions and confidence intervals for nonlinear least squares, *Technometrics*, *29*, 67–82.

Draper, N. R., and H. Smith (1998), *Applied Regression Analysis*, 3rd ed., John Wiley, New York.

Fienen, M. N., J. E. Doherty, R. J. Hunt, and H. W. Reeves (2010), Using prediction uncertainty analysis to design hydrologic monitoring networks: Example applications from the great lakes water availability pilot project, *U.S. Geol. Surv. Sci. Invest. Rep.*, *2010-5159*.

Finsterle, S., and K. Pruess (1995), Solving the estimation-identification problem in two-phase flow modeling, *Water Resour. Res.*, *31*(4), 913–924, doi:10.1029/94WR03038.

Finsterle, S., and Y. Zhang (2011), Error handling strategies in multiphase inverse modeling, *Comput. Geosci.*, *37*, 724–730, doi:10.1016/j.cageo.2010.11.009.

Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross-validation and other methods, *Ground Water*, *45*(5), 627–641, doi:10.1111/j.1745-6584.2007.00341.x.

Fu, G., D. Butler, S.-T. Khu, and S. Sun (2011), Imprecise probabilistic evaluation of sewer flooding in urban drainage systems using random set theory, *Water Resour. Res.*, *47*, W02534, doi:10.1029/2009WR008944.

Gallagher, M., and J. Doherty (2007), Parameter estimation and uncertainty analysis for a watershed model, *Environ. Model. Softw.*, *22*, 1000–1020, doi:10.1016/j.envsoft.2006.06.007.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, 696 pp., Chapman and Hall, London.

Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*(2), 223–242, doi:10.2307/3318737.

Hill, M. C. (1989), Analysis of accuracy of approximate, simultaneous, nonlinear confidence intervals on hydraulic heads in analytical and numerical test cases, *Water Resour. Res.*, *25*(2), 177–190, doi:10.1029/WR025i002p00177.

Hill, M. C. (2010), Comment on "Two statistics for evaluating parameter identifiability and error reduction" by John Doherty and Randall J. Hunt, *J. Hydrol.*, *380*, 481–488, doi:10.1016/j.jhydrol.2009.10.011.

Hill, M. C., and C. R. Tiedeman (2007), *Effective Calibration of Ground Water Models, With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, 480 pp., John Wiley and Sons, New York.

Hill, M. C., R. L. Cooley, and D. W. Pollock (1998), A controlled experiment in ground water flow model calibration, *Ground Water*, *36*(3), 520–535, doi:10.1111/j.1745-6584.1998.tb02824.x.

Jaynes, E. T. (1976), Confidence intervals vs Bayesian intervals, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, edited by W. L. Harper and C. A. Hooker, pp. 175–257, D. Reidel Publ., Dordrecht, Netherlands.

Kavetski, D., and M. P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008896.

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Calibration of conceptual hydrological models revisited: 2. Improving optimization and analysis, *J. Hydrol.*, *320*, 187–201, doi:10.1016/j.jhydrol.2005.07.013.

Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resour. Res.*, *22*(4), 499–507, doi:10.1029/WR022i004p00499.

Kitanidis, P. K. (1997), *Introduction to Geostatistics, Applications in Hydrology*, 272 pp., Cambridge Univ. Press, New York, doi:10.1017/CBO9780511626166.

Kitanidis, P. K. (2010), Bayesian and geostatistical approaches to inverse problems, in *Large-scale inverse problems and quantification of uncertainty*, edited by L. Biegler et al., pp. 71–86, Wiley, New York, doi:10.1002/9780470685853.ch4.

Krzysztofowicz, R. (2010), Decision criteria, data fusion and prediction calibration: A Bayesian approach, *Hydrol. Sci. J.*, *55*(6), 1033–1050, doi:10.1080/02626667.2010.505894.

Li, X., and F. T.-C. Tsai (2009), Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod, *Water Resour. Res.*, *45*, W09403, doi:10.1029/2008WR007488.

Linssen, H. N. (1975), Nonlinearity measures: A case study, *Stat. Neerl.*, *29*, 93–99, doi:10.1111/j.1467-9574.1975.tb00253.x.

Liu, X., M. A. Cardiff, and P. K. Kitanidis (2010), Parameter estimation in nonlinear environmental problems, *Stochastic Environ. Res. Risk Assess.*, *24*, 1003–1022, doi:10.1007/s00477-010-0395-y.

Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, *35*, 69–82, doi:10.1016/j.advwatres.2011.10.007.

McClave, J. T., and T. Sincich (2000), *Statistics*, 8th ed., Prentice Hall, Upper Saddle River, N.J.

Mehl, S. W., and M. C. Hill (2010), Grid-size dependence of Cauchy boundary conditions used to simulate stream-aquifer interactions, *Adv. Water Resour.*, *33*, 430–442, doi:10.1016/j.advwatres.2010.01.008.

Meyer, P. D., M. Ye, M. L. Rockhold, S. P. Neuman, and K. J. Cantrell (2007), *Combined Estimation of Hydrogeologic Conceptual Model, Parameter, and Scenario Uncertainty, NUREG/CR-6940, PNNL-16396*, U.S. Nucl. Regul. Comm., Washington, D.C.

Montanari, A., and A. Brath (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *40*, W01106, doi:10.1029/2003WR002540.

Montanari, A., and G. Grossi (2008), Estimating the uncertainty of hydrological forecasts: A statistical approach, *Water Resour. Res.*, *44*, W00B08, doi:10.1029/2008WR006897.

Morgan, M. G., and M. Henrion (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge Univ. Press, Cambridge, U.K., doi:10.1017/CBO9780511840609.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305, doi:10.1007/s00477-003-0151-7.

Neuman, S. P., and P. J. Wierenga (2003), *A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites, NUREG/CR-6805*, U.S. Nucl. Regul. Comm., Washington, D.C.

Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, doi:10.1016/j.advwatres.2011.02.007, in press.

Oliver, D. S., A. C. Reynolds, and N. Liu (2008), *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, 392 pp., Cambridge Univ. Press, New York.

Parker, J. C., U. Kim, M. Widdowson, P. Kitanidis, and R. Gentry (2010), Effects of model formulation and calibration data on uncertainty in dense nonaqueous phase liquids source dissolution predictions, *Water Resour. Res.*, *46*, W12517, doi:10.1029/2010WR009361.

Poeter, E. P., and D. A. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, *43*(4), 597–605, doi:10.1111/j.1745-6584.2005.0061.x.

Poeter, E. P., and M. C. Hill (2007), MMA, A computer code for multimodel analysis, *U.S. Geol. Surv. Tech. Methods*, *6–E3*.

Poeter, E. P., M. C. Hill, E. R. Banta, S. W. Mehl, and S. Christensen (2005), UCODE_2005 and six other computer codes for universal sensitivity analysis, inverse modeling, and uncertainty evaluation, *U.S. Geol. Surv. Tech. Methods*, *6–A11*.

Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water Resour. Res.*, *48*, W07401, doi:10.1029/2011WR011527.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008), *Global Sensitivity Analysis, The Primer*, John Wiley and Sons, New York.

Samanta, S., D. S. Mackay, M. K. Clayton, E. L. Kruger, and B. E. Ewers (2007), Bayesian analysis for uncertainty estimation of a canopy transpiration model, *Water Resour. Res.*, *43*, W04424, doi:10.1029/2006WR005028.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Schweppe, F. C. (1973), *Uncertainty Dynamic Systems*, Prentice Hall, Englewood Cliffs, N.J.

Sciortino, A., T. C. Harmon, and W. W.-G. Yeh (2002), Experimental design and model parameter estimation for locating a dissolving dense nonaqueous phase liquid pool in groundwater, *Water Resour. Res.*, *38*(5), 1057, doi:10.1029/2000WR000134.

Seber, G. A. F., and A. J. Lee (2003), *Linear Regression Analysis*, 549 pp., Wiley, New York, doi:10.1002/9780471722199.

Seber, G. A. F., and C. J. Wild (2003), *Nonlinear Regression*, 768 pp., Wiley, New York.

Shi, X., M. Ye, S. Finsterle, and J. Wu (2012), Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling, *Vadose Zone J.*, doi:10.2136/vzj2011.0147, in press.

Stark, P. B., and L. Tenorio (2011), A primer of frequentists and Bayesian inverence in inverse problems, in *Large-scale Inverse Problems and Quantification of Uncertainty*, edited by L. Biegler et al., pp. 9–32, Wiley, New York.

Sun, A. Y., A. P. Morris, and S. Mohanty (2009), Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques, *Water Resour. Res.*, *45*, W07424, doi:10.1029/2008WR007443.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2007), Comparing sensitivity analysis methods to advance lumped watershed model identification and calibration, *Hydrol. Earth Syst. Sci.*, *11*, 793–817, doi:10.5194/hess-11-793-2007.

Tiedeman, C. R., M. C. Hill, F. A. D'Agnese, and C. C. Faunt (2003), Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system, *Water Resour. Res.*, *39*(1), 1010, doi:10.1029/2001WR001255.

Tiedeman, C. R., D. M. Ely, M. C. Hill, and G. M. O'Brien (2004), A method for evaluating the importance of system state observations to model predictions, with application to the Death Valley regional groundwater flow system, *Water Resour. Res.*, *40*, W12411, doi:10.1029/2004WR003313.

Tonkin, M., C. R. Tiedeman, D. M. Ely, and M. C. Hill (2007), OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics, *U.S. Geol. Surv. Tech. Methods*, *6–E2*, 115 pp. [Available at http://water.usgs.gov/software/OPR-PPR.html.]

Vecchia, A. V., and R. L. Cooley (1987), Simultaneous confidence and prediction intervals for nonlinear regression models with application to a groundwater flow model, *Water Resour. Res.*, *23*(7), 1237–1250, doi:10.1029/WR023i007p01237.

Vrugt, J. A., and W. Bouten (2002), Validity of first-order approximations to describe parameter uncertainty in soil hydraulic models, *Soil Sci. Soc. Am. J.*, *66*, 1740–1751, doi:10.2136/sssaj2002.1740.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling:

Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 271–288.

Wang, Q. J., D. E. Robertson, and F. H. S. Chiew (2009), A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resour. Res.*, *45*, W05407, doi:10.1029/2008WR007355.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

Yeh, W. W.-G. (1986), Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(2), 95–108, doi:10.1029/WR022i002p00095.

Yeh, W. W.-G., and Y. S. Yoon (1981), Aquifer parameter identification with optimum dimension in parameterization, *Water Resour. Res.*, *17*(3), 664–672, doi:10.1029/WR017i003p00664.