AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2013WR014767

Quantifying the predictive consequences of model error with linear subspace analysis

Jeremy T. White[1], John E. Doherty[2], and Joseph D. Hughes[3]

[1]Texas Water Science Center, U.S. Geological Survey, Austin, Texas, USA, [2]Centre for Groundwater Research and Training, Flinders University, Adelaide, South Australia, Australia, [3]Florida Water Science Center, U.S. Geological Survey, Lutz, Florida, USA

Key Points:
- The expression of model error is prediction dependent
- Structural defects may be invisible during calibration
- How calibration is implemented influences the expression of model error

Correspondence to:
J. T. White,
jwhite@usgs.gov

Citation:
White, J. T., J. E. Doherty, and J. D. Hughes (2014), Quantifying the predictive consequences of model error with linear subspace analysis, *Water Resour. Res.*, *50*, 1152–1173, doi:10.1002/2013WR014767.

Received 19 SEP 2013
Accepted 3 JAN 2014
Accepted article online 9 JAN 2014
Published online 14 FEB 2014

Abstract All computer models are simplified and imperfect simulators of complex natural systems. The discrepancy arising from simplification induces bias in model predictions, which may be amplified by the process of model calibration. This paper presents a new method to identify and quantify the predictive consequences of calibrating a simplified computer model. The method is based on linear theory, and it scales efficiently to the large numbers of parameters and observations characteristic of groundwater and petroleum reservoir models. The method is applied to a range of predictions made with a synthetic integrated surface-water/groundwater model with thousands of parameters. Several different observation processing strategies and parameterization/regularization approaches are examined in detail, including use of the Karhunen-Loève parameter transformation. Predictive bias arising from model error is shown to be prediction specific and often invisible to the modeler. The amount of calibration-induced bias is influenced by several factors, including how expert knowledge is applied in the design of parameterization schemes, the number of parameters adjusted during calibration, how observations and model-generated counterparts are processed, and the level of fit with observations achieved through calibration. Failure to properly implement any of these factors in a prediction-specific manner may increase the potential for predictive bias in ways that are not visible to the calibration and uncertainty analysis process.

1. Introduction

Simplification is integral to the design and construction of any computer model. Modelers are forced to make decisions about which physical processes will be represented, how physical processes will be simulated, what spatial and temporal resolution will be used, and how boundary conditions will be represented. All of the inevitable simplifications, while necessary to form a digital representation of a natural system, result in significant discrepancies between the system states calculated by the computer model and the state observations of the system it represents. Manifestations of these discrepancies have been variously described as model uncertainty [*Draper*, 1995], model inadequacy [*Kennedy and O'Hagan*, 2001; *Gupta et al.*, 2012; *Foglia et al.*, 2013], model error [*Refsgaard et al.*, 2006; *Lin and Beck*, 2012], and model structural error [*Doherty and Welter*, 2010; *Beven*, 2005]. The discrepancies may occur in all forms of model output, including decision-making predictions of future system behavior, as well as output that corresponds to observations used for calibrating model parameters. The present study focuses on the role that model error plays in the calibration process and the implications for postcalibration model predictions.

1.1. Review of Existing Methods to Detect and Quantify the Effects of Model Error

Several methods are currently used to detect and quantify the effects of model error that are exposed during calibration. Many of the methods employ a Bayesian framework and seek to identify and statistically characterize model-to-measurement misfit generated by model error; this misfit is then used in the calculation of parameter uncertainty estimates.

In a landmark paper, *Kennedy and O'Hagan* [2001] explicitly account for model error by attributing the additional misfit generated by model error to adjustable parameters governing a stochastic process that is added to model outputs. In turn, this stochastic process prevents overfitting and underestimation of parameter uncertainty by accounting for the contribution to uncertainty made by model defects. The work of *Kennedy and O'Hagan* [2001] represents a substantial conceptual advance in the analysis and contribution of computer model error to total model parameter and predictive error, and it has inspired use by other

authors, including *Oakley* [2004], *Oakley and O'Hagan* [2002], and *Higdon et al*. [2005]. However, *O'Hagan* [2006] indicate this method is not practical in high parameter dimensions, which are common in groundwater and petroleum reservoir modeling, where many hundreds or even thousands of parameters may be needed to represent prediction-sensitive complexity and heterogeneity.

Bayesian model averaging (sometimes referred to as multimodel averaging) [*Draper*, 1995; *Raftery et al*., 1997; *Hoeting et al*., 1999; *Poeter and Hill*, 2007; *Rojas et al*., 2008; *Ye et al*., 2008] approaches the model-error problem by using an ensemble of relatively simple computer models, each of which represents an alternative conceptualization of the same system. Likelihoods are assigned to the calibrated simple models by ranking each ensemble member model according to the ability to reproduce observations of system state as well as the "complexity" of each model, usually measured by the number of adjustable parameters. Conceptual realizations that more closely reproduce observations with less parameters are assumed to be superior representations of the natural system. The uncertainty arising from model error is accounted for by propagating the ranked ensemble of alternative conceptualizations through a Bayesian inference scheme.

Multiobjective methods [*Gupta et al*., 1998; *Madsen*, 2000; *Vrugt et al*., 2003] recognize that a defective model with a single set of parameters may not necessarily be capable of reproducing all aspects of system-state observations. For example, a single set of parameters in a rainfall-runoff model may not be able to simultaneously reproduce peak flows and recession limbs of an observed hydrograph. The goal of multiobjective methods is to identify parameter sets that collectively define the Pareto frontier between many different and noncommensurate objectives, each formulated to represent a unique aspect of the observed data. The distribution of the resulting parameter sets implicitly includes some effects of model error as exposed by the model calibration process. The variability of model predictions calculated on the basis of the Pareto optimal set also includes some uncertainty arising from model inadequacies.

Generalized likelihood uncertainty estimation (GLUE) [*Beven and Binley*, 1992; *Freer et al*., 1996; *Beven et al*., 2012] uses subjective likelihood measures to account for model error. The GLUE method dichotomizes an ensemble of parameter realizations, drawn from the prior parameter distribution, into "behavioral" and "nonbehavioral" sets using subjective likelihood criteria (e.g., objective function value less than some threshold values), which are based on comparing the resulting computer model outputs of each realization with system observations. The GLUE approach recognizes that statistical characterization of model-to-measurement misfit may not be possible when model defects contribute significantly to misfit, and that a statistical characterization based on the assumed properties of measurement noise may lead to underestimation of model parameter and predictive uncertainty.

Sequential optimization and data assimilation (SODA) [*Vrugt et al*., 2005; *Spaaks and Bouten*, 2013] pairs a global optimization method with the ensemble Kalman filter [*Evensen*, 2003] to accommodate model error. Through sequential state updating, the SODA scheme uses misfit between observations and corresponding computer model outputs in excess of measurement noise to update model state variables prior to commencing the next simulation period. Irreducible misfit is attributed to model error and is explicitly accounted for in assigning parameter uncertainty.

The methods described thus far rely on model error being "visible" during calibration as misfit over and above the noise attributable to measurement error. However, our study will demonstrate that model error does not necessarily produce any misfit, thereby rendering model error invisible. During calibration, the invisible component of model error causes estimated parameters to take on values which compensate for model imperfections. The concept of "compensatory parameters" has been recognized by *Clark and Vrugt* [2006] and *Spaaks and Bouten* [2013]. When estimated parameters are assigned compensating roles, the potential for error in predictions made by a defective model may actually increase, rather than decrease, during the process of calibration. Under these circumstances, use of the methods discussed previously will fail to properly account for potential predictive bias arising from model defects and imperfections. Our study will also demonstrate that predictive error arising from calibration-induced parameter compensation is strongly prediction dependent.

Doherty and Christensen [2011] use a paired complex/simple model method to identify and quantify calibration-induced predictive bias. The complex and simple models are typically represented by high- and low-fidelity computer model variants of the same system. The high-fidelity model may include more rigorous representation of physical processes, a denser simulation grid, and/or more accurate solution schemes

than the low-fidelity model. The low-fidelity simple computer model is repeatedly calibrated against an ensemble of stochastically generated outputs from the high-fidelity computer model. Pairwise comparison of predictions made by the ensemble of simple and complex models exposes the otherwise invisible predictive bias arising from calibration-induced parameter compensation. In contrast to most of the methods discussed previously, the paired-model method can be implemented in highly parameterized contexts because the low-fidelity model must still be capable of fitting the high-fidelity model outputs to the level of measurement noise. However, the method of *Doherty and Christensen* [2011] may incur a high computational burden from repeated calibration of the low-fidelity model.

1.2. Relevance of Proposed Method

Our study employs linear subspace concepts to analyze the consequences of calibrating an imperfect model and expands on the theoretical developments of *Doherty and Christensen* [2011]. The new method requires a simple and complex model of the same system; however, in our study, the complex model is represented as additional parameters in a low-fidelity computer model, rather than requiring the construction of a high-fidelity computer model and repeated calibration of the low-fidelity computer model. This representation of the complex model greatly reduces computation burden and facilitates more efficient exploration of different calibration strategies for reducing predictive bias. Unlike some of the previous works on model error [e.g., *Kennedy and O'Hagan*, 2001; *Draper*, 1995], our linear method does not require that model error be exposed as additional model-to-measurement misfit during calibration. The linear method also scales efficiently to an arbitrarily high number of estimable parameters.

The accommodation of model defects that are not easily detectable through the calibration process is of particular interest where models are built to simulate the movement of subsurface fluids, such as groundwater modeling, petroleum reservoir modeling, and geothermal reservoir modeling. In these contexts, the complexity and variability of subsurface properties cannot be completely represented in a model. At the same time, the data set available for calibration is often limited, which results in a high-dimensional null space. As a result, calibration in these contexts facilitates parameter compensation, which allows a good fit to be found during calibration despite the fact that many nuances of subsurface geology are poorly or incorrectly represented by the model. Part of the aim of the present study is to analyze the predictive repercussions of parameter compensation in a subsurface context.

This paper is organized as follows: first, a simple groundwater flow model is used to demonstrate conceptually the predictive consequences of calibrating an imperfect model. Expressions for precalibration and postcalibration predictive error that include the contribution from model error are derived using a linear subspace framework. These expressions are subsequently used to quantitatively analyze errors associated with different types of predictions made by a much more complex synthetic model of an integrated surface-water/groundwater system. The effect that several different modeling and calibration strategies have on the manifestation of model error is explored and discussed, along with assumptions and limitations of the method.

Given the complex nature of model error, an analysis such as the one presented here can rarely be universally applied to quantify the predictive consequences of model error, especially for model error that may be invisible to the calibration process. The difficulties involved in such an analysis are part of the reason why the topic has received a comparatively small amount of attention. Our study is necessarily compromised by the fact that a "reality" complex model that encapsulates the full depth and breadth of real-world complexity cannot be incorporated in any groundwater model. It is further compromised by the linear basis for the following analysis.

However, the lack of a "reality" complex model does not invalidate the analysis presented herein. The complex model used in the following analyses attempts to include many aspects of real-world systems that are commonly omitted from groundwater modeling uncertainty analysis, yet are likely to affect model outcomes under both calibration and predictive conditions. While other model imperfections obviously are not included, the analyses attempt to address at least some of the more important ones. Additional studies can address other imperfections; the method presented herein is not restricted to the analysis of any one type of model imperfection.

The linearity assumption may render the outcomes of the analyses approximate. However, any analysis of the effects of model error will necessarily be at least partially approximate because as stated above, a

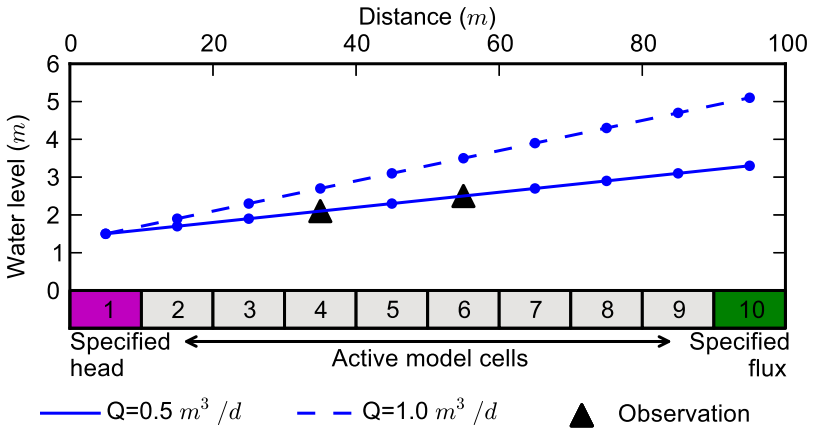

Figure 1. Simple example model domain. The "true" water level is shown for both calibration (specified flux = 0.5 m^3/d) and predictive (specified flux = 1.0 m^3/d) conditions. Simulated water levels from model cells 4 and 6 are used for calibration (black triangles).

perfect reference model can never be built, and as a result, all analyses of model error must rely on strong assumptions. While the outcomes of linear analysis may be approximate, the outcomes of such analyses are nevertheless representative enough to provide important insights into appropriate model usage. This is especially the case when an analysis attempts to compare different processing or management scenarios rather than calculate the individual outcomes of each. The validity of linear analysis in the latter context was demonstrated by *Dausman et al*. [2010] who showed that the relative ranking of different data acquisition strategies as assessed through linear analysis was independent of parameter realizations employed by a highly nonlinear variable-density model of heat and salt transports. The decision to employ a linear analysis method is further strengthened when the alternative is considered. The only known alternative to linear analysis in the present context (i.e., exploration of the effect of model imperfection in highly parameterized groundwater models) is the nonlinear paired-model method of *Doherty and Christensen* [2011]. The computational cost of such an analysis for this and most studies is infeasible. Further, the methodology of *Doherty and Christensen* [2011] does not allow explicit representation of the different contributions to the postcalibration error of model predictions with and without the use of different processing strategies that are designed to reduce the predictive effects of model error.

2. The Simple Groundwater Flow Model

A simple groundwater flow model is used to demonstrate conceptually the effects that model imperfections have on estimated parameters and on postcalibration predictions. A single forward run of the model is undertaken to generate "true" values of observations that will serve as a calibration data set. Then the model is altered to introduce a structural defect; the defective model is then calibrated against the "true" observation data set. The calibration process is repeated several times with the defective model, each time employing a different parameterization or observation processing strategy to show how these strategies may affect model predictive ability. All of the concepts in this section form the basis for a more rigorous analysis presented in the following sections.

The model is a 1-D, cell-centered finite-difference groundwater flow model with 10 cells arranged in a single row (Figure 1). Groundwater flow is simulated with U.S. Geological Survey modular finite-difference code MODFLOW-2005 [*Harbaugh*, 2005]. Each model cell is assigned a hydraulic conductivity value of 2.5 m/d. A specified flux of 0.5 m^3/d is introduced in model cell 10 at the right edge of the domain, representing inflow from upgradient; the left side of the domain is bounded by a specified head in model cell 1, which is assigned a value of 1.5 m. Two water levels calculated by the model are used as observations for subsequent calibration exercises. The observations are assumed to be free from measurement error so that effects of model error can be clearly demonstrated. Model structural error is introduced by assigning the specified head at the left side of the domain a value of 1.0 m, which is 0.5 m lower than the "true" value.

Assume expert knowledge indicates that hydraulic conductivity in the model domain is heterogeneous with an expected value of 2.5 m/d (the "true" value of hydraulic conductivity in every model cell). One

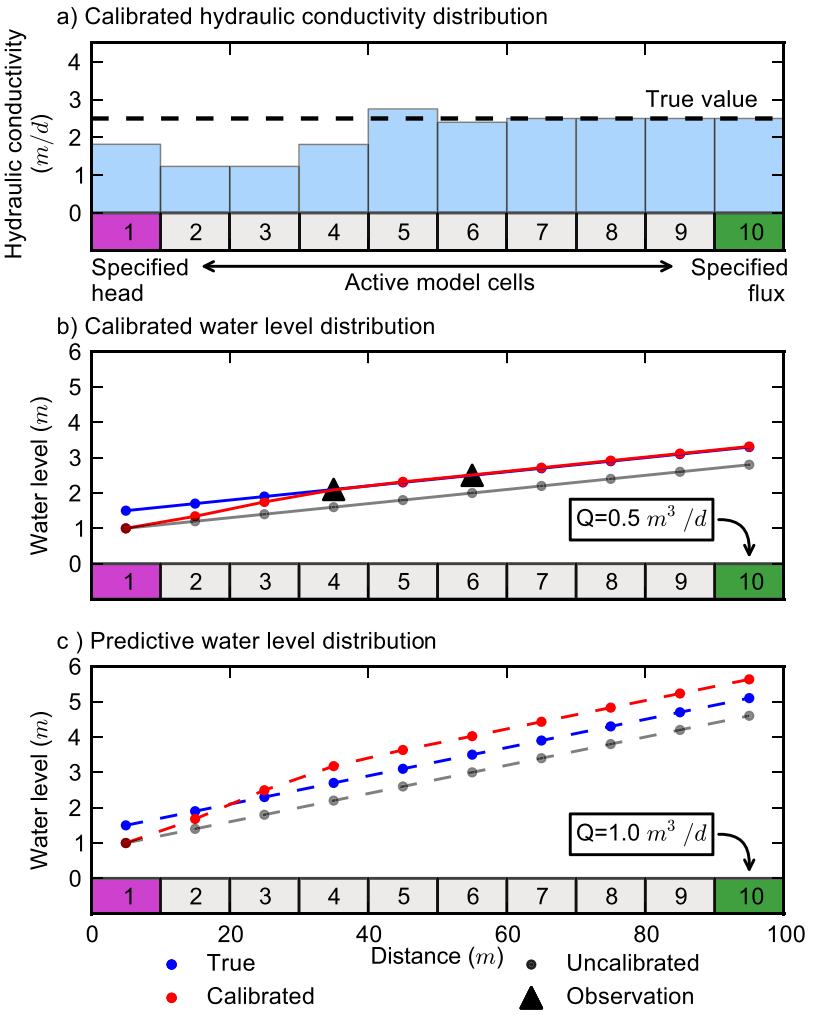

Figure 2. Results of calibration with hydraulic conductivity of all 10 model cells as independent parameters. Two water levels were used as the calibration data set. (a) Calibrated hydraulic conductivity distribution, (b) calibrated water level distribution, and (c) predictive water level distribution.

parameterization approach to accommodate expected heterogeneity is to assume that hydraulic conductivity is spatially uncorrelated and independent in all 10 model cells. Even though it is not possible to estimate each of these parameters uniquely using only two observations, the use of this number of parameters, when accompanied by appropriate subspace regularization, results in a minimum error variance solution to the inverse problem, while also allowing the calibration process the flexibility to introduce heterogeneity where required [*Aster et al.*, 2013]. Parameters or parameter combinations that are not informed by the calibration data set remain at initial values of prior maximum likelihood; these parameter combinations are not calibrated. In this example, subspace regularization, implemented with the truncated singular value decomposition (SVD) algorithm [*Aster et al.*, 2013], is used to solve the inverse problem.

Calibrating the defective model using truncated SVD results in an incorrect distribution of hydraulic conductivity (Figure 2a). However, the defective model with incorrect estimated parameter values is capable of making some reliable predictions. Water levels upgradient of the observations are correctly predicted by the calibrated, defective model, as long as the specified inflow from upgradient is unchanged (Figure 2b). This shows that the defective model is still capable of using the information in the calibration data set to improve the reliability of some predictions compared to the uncalibrated model using maximum prior likelihood parameter values. Estimating a large number of parameters has provided opportunity for parameter compensation. However, the parameter compensation "damage" is confined to the left side of the model domain (only model cells 1–5 have incorrect values). Parameters on the right side of the domain are

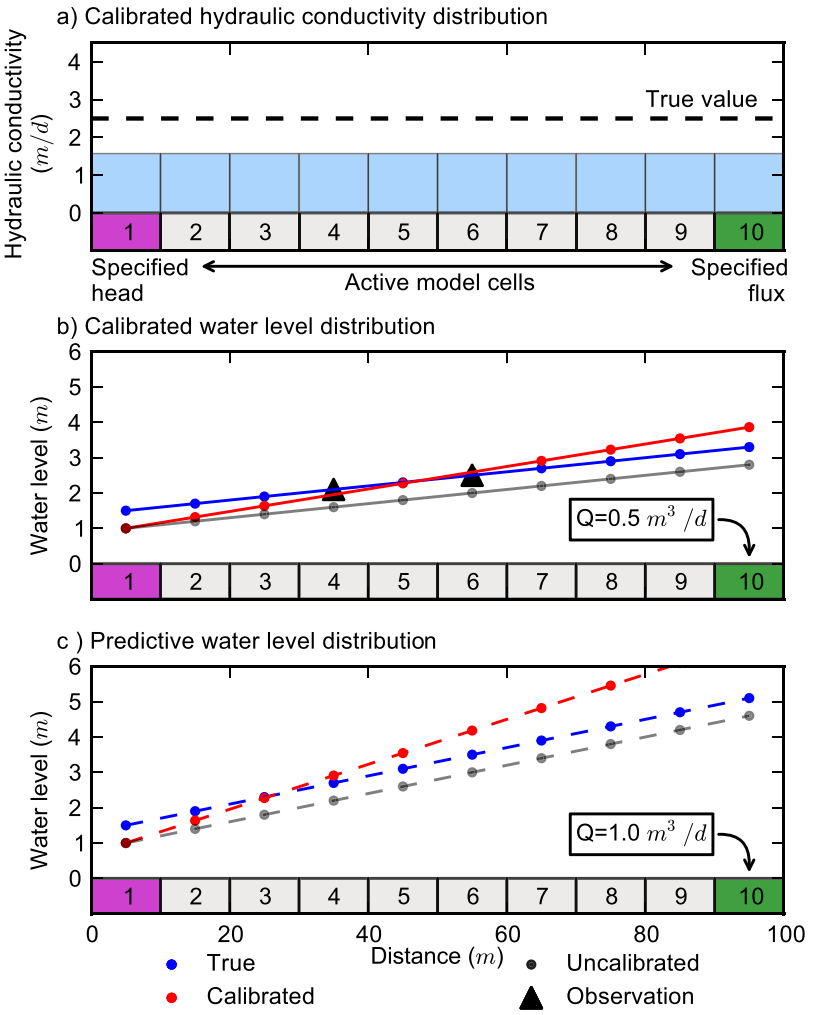

Figure 3. Results of calibration with a single hydraulic conductivity for all 10 model cells. Two water levels were used as the calibration data set. (a) Calibrated hydraulic conductivity distribution, (b) calibrated water level distribution, and (c) predictive water level distribution.

unchanged from initial, prior maximum likelihood values. However, calibration-induced bias has not been eliminated for all predictions. If during predictive model usage, the specified flux at the right side of the domain is increased to 1.0 m^3/d (Figure 2c), then water-level predictions made by the calibrated model are wrong throughout the model domain.

Voss [2011a, b] suggest using a parsimonious calibration approach, where only a few parameters are adjusted, and if necessary, a less-than-optimal fit with the calibration data set is accepted. This parameterization can be thought of as a strong enforcement of expert knowledge. This strategy has some intuitive appeal because estimating only a few parameters should limit the ability of parameters to compensate for model structural defects. These defects should therefore be more visible as misfit during calibration compared to the previous calibration attempt, where defects were concealed by parameter compensation. The parsimonious strategy is evaluated herein by estimating a single value of hydraulic conductivity for the entire model domain.

As expected, calibration with a single parameter exposes model error as misfit because a single hydraulic conductivity parameter cannot simultaneously reproduce both water-level observations in the defective model (Figures 3a and 3b). However, the magnitude of misfit could easily (and erroneously) be attributed to misfit arising from failure to represent hydraulic conductivity heterogeneity. Unfortunately, the estimated, domainwide hydraulic conductivity value is incorrect (1.74 m/d) and continues to compensate for model defects (the correct value is 2.5 m/d). Any prediction that is sensitive to local-scale hydraulic conductivity is

now biased as a result of the calibration process. Furthermore, calibration of the defective model with a single hydraulic conductivity value limits the ability to confine parameter compensation spatially, which, in turn negatively effects the reliability of some predictions, such as upgradient heads.

The single parameter calibration strategy can be used to demonstrate another possible (and counterintuitive) consequence of calibrating an imperfect model. In some cases, predictions made by a calibrated imperfect model may be less reliable than predictions made using an uncalibrated model parameterized using only expert knowledge. Consider Figure 3c. Setting the hydraulic conductivity for all cells in the defective model to 2.5 m/d (its correct value) produces a water level of 3.9 m in model cell 8 (0.2 m error). However, the calibrated defective model predicts a water level of 5.25 m for model cell 8 (−1.15 m error). Note this result is not limited to parsimonious calibration strategies.

Consider a case where expert knowledge suggests that hydraulic conductivity heterogeneity is spatially correlated and is best described by a $\log_{10}$ exponential variogram with sill of 1.0 and range of 300 m. The range of the exponential variogram was purposefully chosen to produce considerable correlation between adjacent model cells, while still providing flexibility in the parameterization to fit the observations. The minimum error variance solution to the inverse problem then requires use of the Karhunen-Loève (KL) transform, which recasts the inverse problem to estimate the eigencomponent coefficients of the prior parameter covariance matrix implied by this variogram (see *Watson et al*. [2013] for a more thorough description and analysis of the KL transform). Truncated SVD is again used to solve the KL-transformed inverse problem with the defective model.

The KL-transformed estimation process results in parameter compensation similar to previous attempts to calibrate the defective model (Figure 4a). Enforcement of expert knowledge with the KL transformation regionalizes compensatory parameter behavior across the model domain. The infusion of expert knowledge into the inversion process forces parameter compensation to be "suboptimal" in reducing bias for some predictions, such as upgradient heads. As a result, incorporation of expert knowledge in the inversion process may not improve a model's ability to make reliable predictions.

Doherty and Welter [2010] discuss processing of observations and model-generated counterparts to reduce the potential for calibration-induced parameter compensation. In the simple model example, if the calibration data set is replaced with a single observation that is the difference between the two observed head values, calibration of the defective model produces the correctly estimated hydraulic conductivity distribution, regardless of the parameterization approach used. Any postcalibration model predictions that depend only on hydraulic conductivity will be correct, even though the model is defective.

Using a simple groundwater model to explore the ramifications of calibrating a defective model has demonstrated several important observations related to model error:

1. Calibration of a defective model leads to errors in at least some estimated parameter values because parameters compensate for model defects.

2. Calibration-induced parameter compensation enhances the reliability of some model predictions, while degrading others.

3. For some predictions made by a defective model, parameter values based only on expert knowledge may provide more reliable predictions than parameter values inferred through calibration.

4. Estimation of many parameters facilitates parameter compensation that may reduce model-error visibility.

5. Estimation of few parameters may increase model-error visibility, but it may be difficult to separate model-error-induced misfit from parameterization-induced misfit.

6. Estimation of few parameters may reduce parameter compensation but may not necessarily eliminate it.

7. Estimation of few parameters may spread parameter compensation across large regions of the model domain.

8. Introduction of spatially correlated expert knowledge to the calibration process also regionalizes parameter compensation.

9. Some predictions will benefit from spatially localized parameter compensation.

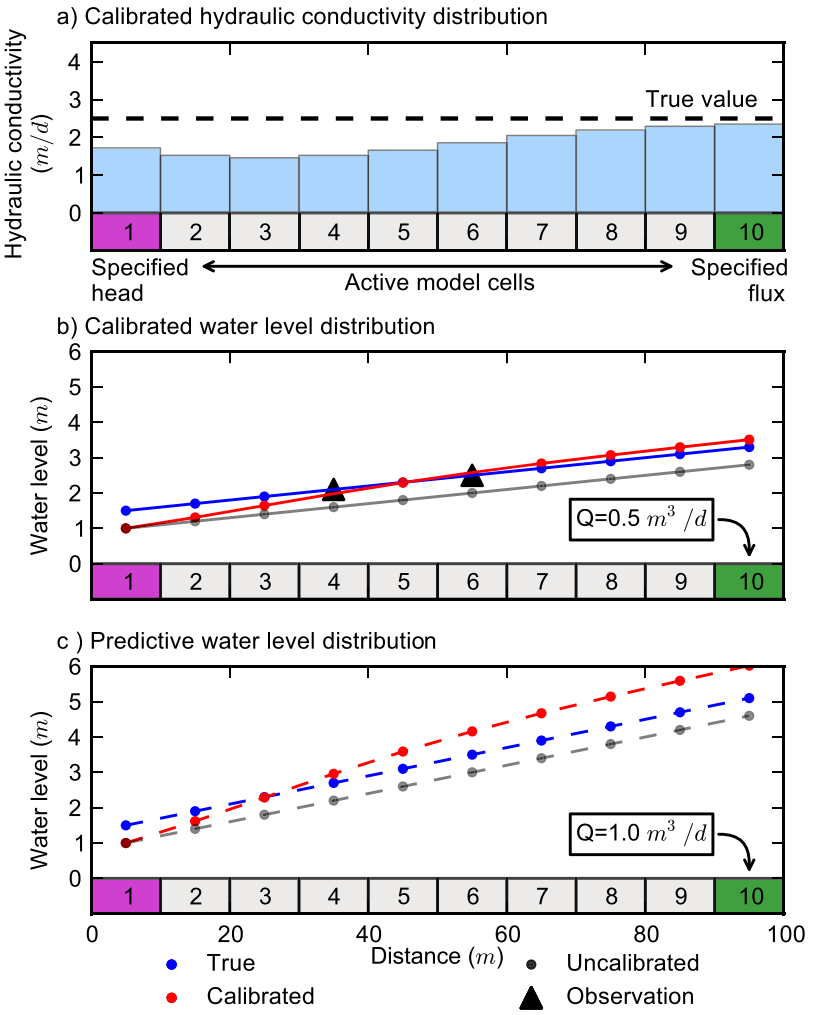

Figure 4. Results of calibrating the model using the geostatistically correlated, Karhunen-Loève-transformed parameters for hydraulic conductivity of all 10 model cells. Two water levels were used as the calibration data set. (a) Calibrated hydraulic conductivity distribution, (b) calibrated water level distribution, and (c) predictive water level distribution.

10. Formulating the inverse problem in terms of processed observations and corresponding model outputs may mitigate calibration-induced parameter compensation and related predictive bias.

The example of a simple groundwater model indicates that it may not be possible to calibrate a defective model in a way that optimizes the ability to make all predictions. A particular strategy may enhance the reliability of some predictions, while simultaneously eroding the reliability of others. All models of real systems contain defects, so calibration presents many dilemmas, some of which may be irreconcilable if a model is required to make more than one type of prediction.

3. Theoretical Basis for Linear Analysis Method

Following a similar approach to *Doherty and Christensen* [2011], consider the "true" model of reality that is later simplified to form a computer model. For tractability of the analysis, assume the model is linear. Let the matrix **Z** characterize the action of the real-world model on its real-world parameters during calibration. Then

$$\mathbf{h} = \mathbf{Z}\mathbf{k} + \epsilon, \tag{1}$$

where **h** is the vector of available observations of the natural system, **k** is the vector of real-world parameters, and ϵ is the vector of observation error. Conceptually, a natural system has an infinite level of

complexity, so **Z** has an infinite number of columns and **k** has an infinite number of rows. A computer model is a simplified representation of a complex natural system and therefore omits many parameters present in the natural system. The values of all omitted parameters cannot be known, but mathematical analysis using the omitted parameters facilitates recognition and partial quantification of the effects that simplification has on model predictive ability.

Using the construct of omitted parameters, equation (1) can be partitioned:

$$\mathbf{h} = \begin{bmatrix} \mathbf{Z}_i & \mathbf{Z}_o \end{bmatrix} \begin{bmatrix} \mathbf{k}_i \\ \mathbf{k}_o \end{bmatrix} + \epsilon = \mathbf{Z}_i\mathbf{k}_i + \mathbf{Z}_o\mathbf{k}_o + \epsilon, \tag{2}$$

where the subscripts "*i*" and "*o*" represent those parameters and corresponding processes (represented as matrices) that are, respectively, included in and omitted from the computer model. If $\mathbf{Z}_i$ represents the action of a groundwater model, then elements of $\mathbf{k}_i$ would typically include a simplified spatial representation of hydrogeologic properties, such as hydraulic conductivity and storage, as well as certain boundary condition properties, such as conductance. Elements that comprise $\mathbf{k}_o$ must therefore include much higher-density parameterization of all hydrogeologic and hydrostratigraphic properties, as well as complex spatial and temporal representations of boundary condition structures and properties. The vector $\mathbf{k}_o$ must also include correction terms for errors in the location of geological boundaries, as well as terms that account for horizontal and vertical discretization errors incurred by use of a grid or mesh in place of a continuous natural system.

3.1. Parameter Error Variance

Subjecting $\mathbf{Z}_i$ of equation (2) to the SVD [*Golub and Van Loan*, 1996] yields

$$\mathbf{Z}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i^T, \tag{3}$$

where the column vectors of $\mathbf{U}_i$ form an orthonormal basis spanning the output space of the computer model, the column vectors of $\mathbf{V}_i$ form an orthonormal basis spanning the parameter space of the simplified model, and $\mathbf{S}_i$ is a diagonal matrix of singular values, ordered from largest to smallest.

The inverse problem of inferring computer model parameters on the basis of observations **h** can be solved using truncated SVD:

$$\overline{\mathbf{k}}_i = \mathbf{V}_{i1}\mathbf{S}_{i1}^{-1}\mathbf{U}_{i1}^T\mathbf{h}, \tag{4}$$

where $\overline{\mathbf{k}}_i$ is the vector of calibrated parameter values of the simplified model. The "*i*1" subscript denotes the singular elements that are associated with relatively large singular values of the computer model operator $\mathbf{Z}_i$ [*Aster et al.*, 2013; *Menke*, 1989; *Koch*, 1988].

Assuming a nondefective model, solution of the inverse problem by truncated SVD leads to an estimated parameter field that is necessarily simpler than the true field because it excludes true-field null-space parameter components. The calibration process therefore estimates an orthogonal projection of true parameters onto a small "solution subspace" of parameters space, which is controlled by the information in the observations. If the model is free from defects, then exclusion of null-space components from estimation eliminates parameter bias and approaches a minimum error variance solution to the inverse problem.

Substitution of equation (2) into equation (4) and using the orthonormal properties of $\mathbf{U}_i$ yields

$$\overline{\mathbf{k}}_i = \mathbf{R}\mathbf{k}_i + \mathbf{G}\epsilon + \mathbf{G}\mathbf{h}_o \tag{5}$$

$$\mathbf{R} = \mathbf{V}_{i1}\mathbf{V}_{i1}^T \tag{6}$$

$$\mathbf{G} = \mathbf{V}_{i1}\mathbf{S}_{i1}^{-1}\mathbf{U}_{i1}^T \tag{7}$$

$$\mathbf{h}_o = \mathbf{Z}_o\mathbf{k}_o. \tag{8}$$

Errors in the estimated parameters can be calculated by rearranging equation (5) as

$$\mathbf{k}_i - \overline{\mathbf{k}}_i = (\mathbf{I} - \mathbf{R})\mathbf{k}_i - \mathbf{G}\epsilon - \mathbf{G}\mathbf{Z}_o\mathbf{k}_o, \tag{9}$$

where $\mathbf{I}$ is the identity matrix. The postcalibration covariance matrix of parameter error can be calculated from equation (9) using basic matrix theory for propagation of covariance [*Koch*, 1988]:

$$\Sigma_{\overline{\mathbf{k}}_i - \mathbf{k}_i} = (\mathbf{I} - \mathbf{R})\Sigma_{\mathbf{k}_i}(\mathbf{I} - \mathbf{R})^T + \mathbf{G}\Sigma_\epsilon\mathbf{G}^T + \mathbf{G}\mathbf{Z}_o\Sigma_{\mathbf{k}_o}\mathbf{Z}_o^T\mathbf{G}^T, \tag{10}$$

where $\Sigma_{\mathbf{k}_i}$ and $\Sigma_{\mathbf{k}_o}$ are the precalibration parameter covariance matrices described by the prior probability distribution of the computer model parameters $\mathbf{k}_i$ and the omitted parameters $\mathbf{k}_o$, respectively. The three terms of equation (10) are herein referred to as the null-space, solution-space, and model-error contribution to parameter error variance, respectively. The matrix Σ_ϵ is the covariance matrix of measurement noise, which is often assumed to be diagonal.

Moore and Doherty [2005] analyze equation (10) applied to model calibration in the absence of structural defects (without the third term). The first term on the right side of equation (10) describes the null-space contribution to postcalibration parameter error variance, which arises from the inability to estimate certain parameters and parameter combinations because of an information deficit in the calibration data set. The second term is the "solution-space" contribution to postcalibration parameter error variance, which expresses the potential for error in estimated parameters that is inherited from measurement error. If there is no noise associated with the observations $\mathbf{h}$, and if the model has no structural defects, the second and third terms are not present in equation (10). Under these conditions, truncation can take place where singular values become zero. However, consideration of measurement noise forces truncation to take place at numerically nonzero singular values because the second term rises rapidly with decreasing singular values due to the $\mathbf{S}_{i1}^{-1}$ matrix. Truncation at nonzero singular values effectively increases the dimensionality of the null-space and limits the ability of the calibrated model to fit observations.

The quantity $\mathbf{Z}_o\mathbf{k}_o$ in the third term of equation (9) has been referred to as "structural noise" by *Doherty and Christensen* [2011] and exposes model structural defects to the calibration process. If $\mathbf{Z}_o\mathbf{k}_o$ is orthogonal to $\mathbf{U}_1^T$, structural noise will be visible as misfit during calibration. Conversely, components of $\mathbf{Z}_o\mathbf{k}_o$ that are nonorthogonal to $\mathbf{U}_1^T$ are treated as information by the calibration process and result in compensatory parameter values.

The $\mathbf{S}_{i1}^{-1}$ matrix (included in $\mathbf{G}$, see equation (7)) in the third term of equation (9) may indicate that model structural defects should be accommodated in the same way as measurement noise by using fewer singular values for calibration, which further decreases the dimensionality of the solution space and with it, the level of fit possible with observations. Implementing this approach to model error requires specification of an appropriate level of fit for calibration, which is controlled by the number of singular components used for calibration. However, in practice, implementing this strategy is difficult. *Doherty and Welter* [2010] show that the covariance matrix associated with structural noise is typically singular, which precludes the use of an error-based weighting scheme in those contexts where model-to-measurement misfit is dominated by structural noise (as it often is). Without an ability to stochastically characterize structural noise and then accordingly adopt an error-based weighting scheme, an appropriate truncation point that best accommodates model error cannot be found. We will demonstrate that some predictions do in fact benefit from calibrating with fewer singular components, while the quality of other predictions may be damaged.

To reduce the number of terms in equation (10), as well as the following equations, we assume no correlation between $\mathbf{k}_i$ and $\mathbf{k}_o$; this maintains simplicity of the analysis and clarifies the resulting conclusions.

3.2. Predictive Error Variance

Following equation (2), the "true" value (scalar) of a prediction made by the model of reality is

$$s = \mathbf{y}_i^T\mathbf{k}_i + \mathbf{y}_o^T\mathbf{k}_o, \tag{11}$$

where $\mathbf{y}_i$ and $\mathbf{y}_o$ are vectors of sensitivities of the prediction to parameters included in and omitted from

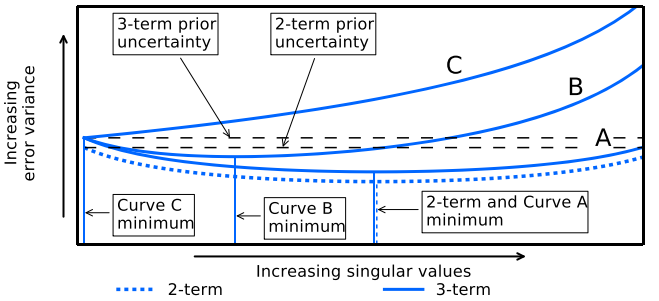

Figure 5. A conceptual plot of different expressions of equation (14). Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Curves A–C show some possible modalities of equation (14). Increasing the number of singular values used in the solution effectively increases the fit achieved by calibration, which effects the predictive error variance.

the computer model, respectively. Equation (11) demonstrates that the true value of the prediction potentially depends on both the $\mathbf{k}_i$ and $\mathbf{k}_o$ parameter vectors.

The value of the same prediction made with the calibrated computer model is

$$\bar{s} = \mathbf{y}_i^T \bar{\mathbf{k}}_i, \tag{12}$$

where $\bar{\mathbf{k}}_i$ is the vector of calibrated parameter values, computed using equation (5). Postcalibration predictive error is then

$$s - \bar{s} = \mathbf{y}_i^T \bar{\mathbf{k}}_i - (\mathbf{y}_i^T \mathbf{k}_i + \mathbf{y}_o^T \mathbf{k}_o). \tag{13}$$

Using equation (13) and standard matrix formulas for propagation of covariance, the potential for predictive error (i.e., the variance of model predictive error) can be calculated as

$$\sigma^2_{s-\bar{s}} = \mathbf{y}_i^T (\mathbf{I} - \mathbf{R}) \Sigma_{\mathbf{k}_i} (\mathbf{I} - \mathbf{R})^T \mathbf{y}_i + \mathbf{y}_i^T \mathbf{G} \Sigma_\epsilon \mathbf{G}^T \mathbf{y}_i + \mathbf{p} \Sigma_{\mathbf{k}_o} \mathbf{p}^T, \tag{14}$$

$$\mathbf{p} = \mathbf{y}_i^T \mathbf{V}_{i1} \mathbf{S}_{i1}^{-1} \mathbf{U}_{i1}^T \mathbf{Z}_o - \mathbf{y}_o^T. \tag{15}$$

Similar to parameter error variance (equation (10)), the first and second terms of equation (14) represent the null-space and solution-space contributions to predictive error variance. If model parameters are KL transformed prior to estimation, and if observation weighting is measurement-error based, then *Moore and Doherty* [2005] show that the first term of (14) falls monotonically with increasing singular value, while the second term rises monotonically. The sum of the first two terms has a minimum that marks the optimal number of singular components to use in calibrating a defect-free computer model. The predictive error variance associated with this minimum is slightly above the posterior uncertainty of the prediction inferred through Bayesian analysis. The reduction in predictive error variance between the precalibration value (at zero singular values) and the minimum value is a measure of the reduction in predictive uncertainty accrued by calibration (Figure 5).

The third term of equation (14) describes the contribution that model defects make to predictive error variance. Some possible modalities of the third term are depicted schematically in Figure 5 (curves A–C). Unlike the first two terms of equation (14), the third term is not monotonic and the effect of the third term of equation (14) is prediction specific. For some predictions, it will have no effect in spite of considerable calibration-induced parameter compensation (curve A), which occurs when the first two terms of equation (15) are equal. These types of predictions tend to be similar in character to the calibration data set [*Doherty and Welter*, 2010]. Conversely, for predictions that are highly sensitive to compensatory parameters, error variance may not ever fall with increasing singular value (e.g., Figure 5, curve C). In this case, even a minor adjustment of model parameters as part of calibration may undermine the predictive ability of the computer model.

For most predictions, the total error variance curve will have a defined minimum (e.g., Figure 5, curve B). However, the minimum of this curve may occur at a singular value that is smaller (or larger) than would have been computed under the assumption of a defect-free model $\mathbf{Z}_i$. The lateral shift in the minimum implies a different level of fit with observation data is needed to achieve the full reduction in error variance. However, the situation is made more complex by the nonmonotonic behavior of the third term in equation (14). The magnitude of the third term is influenced by several factors, including the alignment of the parameter solution matrix $(\mathbf{V}_{i1} \mathbf{S}_{i1}^{-1} \mathbf{U}_{i1}^T)$ with $\mathbf{Z}_o$, the sensitivity of the prediction to the omitted parameters $(\mathbf{y}_o)$, and the prior uncertainty of the omitted parameters $(\Sigma_{\mathbf{k}_o})$.

4. Application of the Linear Method to a Synthetic Model and Evaluation of its Effectiveness

The linear analysis method is applied to several types of predictions made by a synthetic model to evaluate the effectiveness of different calibration strategies in reducing predictive error variance. The synthetic model was constructed to represent many processes and attributes common to real-world integrated surface-water/groundwater modeling. It is anticipated that some of the outcomes presented herein can guide improvements to real-world model parameterization and calibration practices.

As has already been discussed, use of linear analysis comes with advantages and disadvantages. One disadvantage is that it relies on sensitivities that may vary with the actual values of model parameters. Hence, the outcomes of linear analyses, such as those discussed below, can only be approximate. This does not, however, invalidate its ability to provide insight and guidance.

In the examples presented below, linear analysis is used as a basis for comparing the effectiveness of different calibration strategies. The outcomes of the analysis are often very effective in demonstrating the superiority of one strategy over another in optimizing the reliability of a particular type of prediction in the synthetic example. While the analysis is synthetic and linear, it demonstrates that accommodation of model defects cannot be ignored in real-world modeling practice. In any modeling context, it is likely that a strategy can be found that mitigates the problems arising from model error in prediction-sensitive ways, and that part of the task of modeling in any given context must be to develop those strategies.

4.1. Model Construction

A synthetic integrated surface-water/groundwater model was constructed to simulate an aquifer system with two aquifers separated by a semiconfining unit. Simulation is performed with MODFLOW-2005 [*Harbaugh*, 2005], which uses a cell-centered finite-difference approximation to solve the groundwater flow equation. The upper aquifer unit intersects a dynamic surface-water system, simulated by the surface water routing (SWR1) process [*Hughes et al.*, 2012]. A specified flux, representing mountain-front recharge is specified along the northern edge of the model and a head-dependent flux is along the southern edge. Uniform recharge is applied to the top of the domain. Several extraction wells are located in the upper and lower aquifer units (Figures 6 and 7). The model was designed to allow groundwater to discharge from the model through (1) the surface-water system, (2) extraction wells, or (3) the southern head-dependent flux (MODFLOW GHB type) boundary.

The model is vertically discretized into 20 layers, each of which is approximately 5 m thick. The upper aquifer, the semiconfining unit, and the lower aquifer are comprised of 6, 2, and 12 layers, respectively (Figure 7). The model is horizontally discretized into 300 $\times$ 300 square 100 m^2 cells. The active model domain narrows in the east-west direction with depth, conceptually representing a sediment-filled valley overlying impermeable bedrock. Model properties were assigned by hydrogeologic unit. A uniform hydraulic conductivity of 100.0 m/d was specified for the upper and lower units; a uniform value of 1.5 m/d was specified for the middle unit. Specific storage and specific yield values of 0.001 m^{-1} and 0.15, respectively, were assigned to all units.

The initial groundwater level was specified to be 0.5 m below land surface. The synthetic calibration data set was generated using data from a 150 day period simulated using daily stress periods. Predictions were made for a 120 day period immediately following the calibration period. Daily mountain-front recharge, areal recharge, and extraction well rates were generated stochastically as first-order Markov processes. The minimum value from each generated series was used as forcing in the prediction period to represent drought conditions. Extraction well 7 only pumps during the prediction period. The construction and design of the model represent a common modeling application that might be used by regulators to determine maximum sustainable pumping rates during drought conditions.

4.2. Observations

The calibration data set is composed of model simulated water levels from locations A–C (Figure 6). At each location, daily water levels in both the upper and lower aquifer units were "measured," yielding 900 observations. Observation location A was purposefully placed 200 m from extraction well 3, which is a typical regulatory requirement. Each water-level observation was assigned a weight of 100.0, commensurate with an

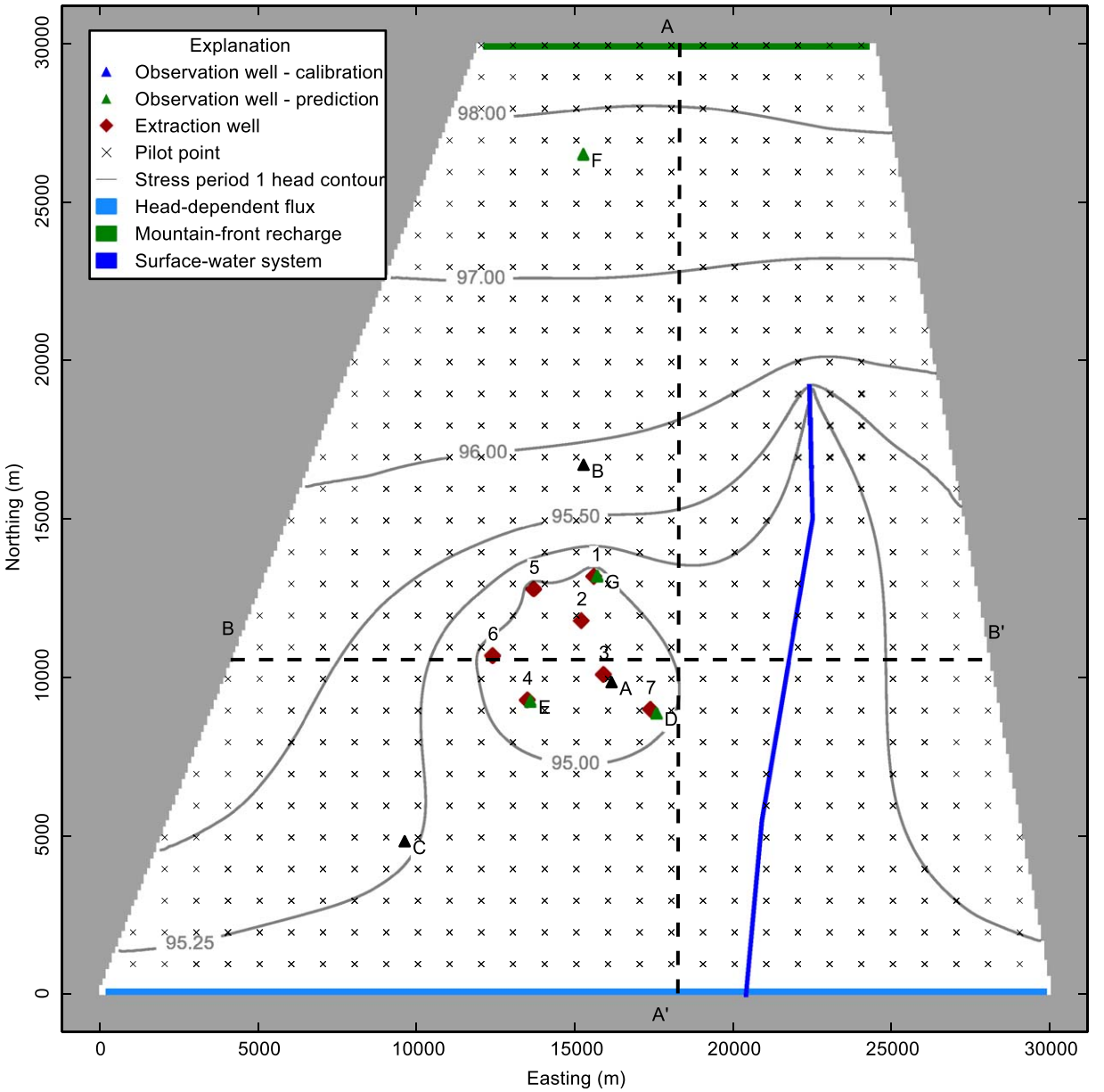

Figure 6. Plan view of the synthetic model domain. The traces of cross sections A-A′ and B-B′ are shown.

assumed measurement error standard deviation of 0.01 m (these weights are squared in the formulation of the overall objective function).

A data processing strategy was applied to observations and model-generated counterparts to evaluate the potential to mitigate calibration-induced predictive bias. Each observation time series was processed using its difference from the first value in the series, effectively shifting each series to a relative datum, which *Doherty and Welter* [2010] suggest may reduce parameter compensation. The relative-datum processing alleviates the need of the calibration process to reproduce absolute heads while maintaining replication of temporal head variability. The ability of a model to simulate absolute heads can be easily compromised by incorrect definition of boundary conditions or coarse grid resolution. By focusing on reproducing relative head variability, the information in the observation data set related to local storage and hydraulic conductivity properties is largely preserved. A transformation matrix representing the relative-datum processing

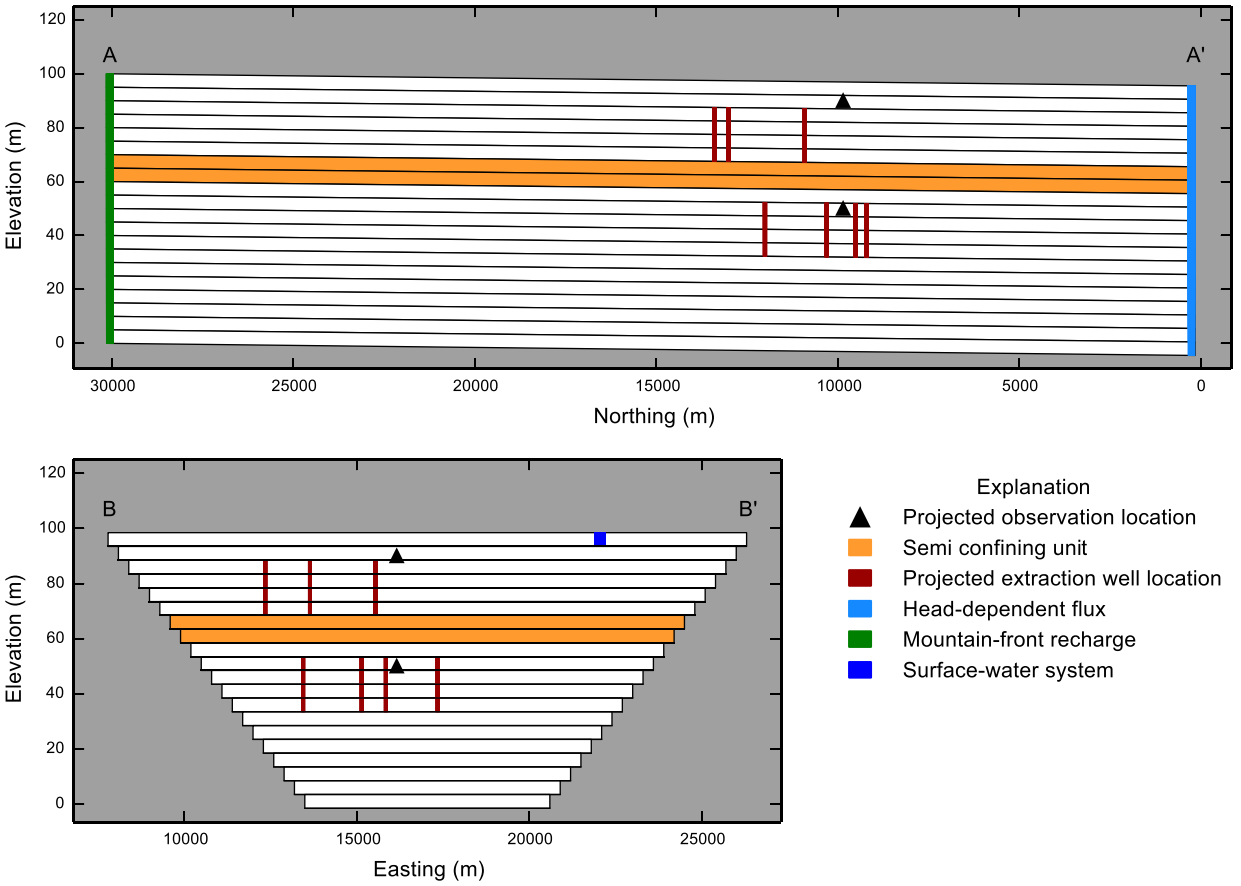

Figure 7. Cross sections A-A′ and B-B′ through synthetic model domain. Cross-section locations are shown in Figure 6. Projected observation location A is shown.

operation was constructed and used to propagate measurement noise from native observations to processed observations, yielding a nondiagonal covariance matrix of measurement noise. The covariance matrix was then used in the second term of equation (14) to establish error-based observation weighting.

4.3. Parameterization

An effort was made to represent many uncertainties associated with simulation of a natural hydrologic system by specifying 6288 parameters. Separate groups of pilot points [*Doherty*, 2003], each placed on a regular 500 m grid, were used to parameterize hydraulic conductivity, specific storage, specific yield, and layer thickness for each of the three hydrogeologic units (Figure 6). Spatially varying recharge multiplier pilot points were used to perturb the spatially uniform input recharge time series; these represent the sensitivity of distributed recharge. The leakance coefficient for each of the 52 SWR1 reach groups was also included as model parameter, as were head-dependent flux stage and conductance. Mountain-front recharge rate and extraction rates assigned to each of the six calibration-period extraction wells were also perturbed as monthly varying parameters; treating extraction rates as parameters acknowledges imperfect knowledge of historical pumping rates (a common occurrence in some aquifers). Extraction well 7 is active only during the prediction period and has an assumed known rate.

4.4. Calculation of Error Variance

Use of equation (14) requires that model parameters be partitioned into the elements of $\mathbf{k}_i$ (estimable computer model parameter) and $\mathbf{k}_o$ (omitted parameters). For this analysis, 2310 parameters were assigned to $\mathbf{k}_o$, effectively eliminating these parameters from the calibration process while acknowledging the fact that they are uncertain and therefore likely to be specified with incorrect values during construction of the computer model. Omitted parameters are comprised of layer thickness pilot points, GHB stage parameters,

Table 1. Parameter Prior Uncertainty

Model Element	Assignment[a]	Log Standard Deviation
Hydraulic conductivity	$\mathbf{k}_i$	1.0
Mountain-front recharge	$\mathbf{k}_i$	0.0328
Specific storage	$\mathbf{k}_i$	0.425
SWR1 reach leakance	$\mathbf{k}_i$	0.044
Specific yield	$\mathbf{k}_i$	0.325
Extraction rate	$\mathbf{k}_o$	0.0328
GHB conductance	$\mathbf{k}_i$	0.25
GHB stage	$\mathbf{k}_o$	0.000571
Layer thickness	$\mathbf{k}_o$	0.119
Recharge multipliers	$\mathbf{k}_o$	0.0218

[a]For use in equation (14).

historical extraction well rate parameters, and recharge pilot point multiplier parameters. These omitted parameters represent sources of model error in most groundwater modeling applications because they are typically assumed to be known and are not adjusted during calibration. Omitted parameters can also be considered representative of other typical sources of model error. For example, spatial recharge multiplier parameters can also be considered representative of model error arising from missing unsaturated zone processes and/or surface-water processes that increase recharge as a result of localized ponding. Hydrostratigraphic unit thickness parameters can serve as proxies for model error arising from vertical discretization. Uncertainty in historical extractions rates may also represent, to some extent, horizontal discretization errors near extraction wells since uncertainty in extraction rates may serve as a proxy for incorrect propagation of drawdown near extraction wells.

Prior uncertainties associated with all parameters are shown in Table 1. For distributed properties, represented by pilot point parameters, a covariance matrix was constructed using an exponential variogram with a range of 1500 m and a sill of 1.0. This covariance matrix was assigned to each set of distributed parameters and scaled by the standard deviation listed in Table 1. Note that even though the synthetic model was assigned a maximum prior likelihood spatially uniform values for hydraulic properties, use of the linearity assumption (sensitivities are independent of parameter values), combined with covariance matrices constructed in this manner, can be used to represent the prior uncertainty of a heterogeneous distribution as described by the variogram; a similar strategy was followed by *Dausman et al.* [2010]. Parameter covariances are used in equation (14); however, sensitivities employed in this equation (required to fill the $\mathbf{Z}$ matrix and $\mathbf{y}$ vector) are computed using spatially uniform parameter values that are centered on the prior probability distribution. The combined prior parameter covariance matrix for all parameters is assembled as a block-diagonal matrix by combining all of the individual scaled pilot point covariance matrices, together with the standard deviations listed in Table 1, into a single matrix. This matrix was subsequently used for the application of the KL transform.

The model was run a total of 6289 times to calculate the finite-difference perturbation sensitivities of model-generated outputs with respect to each parameter; outputs include the system-state observations used for calibration as well as several predictions. The sensitivities resulting for these simulations were used to populate the $\mathbf{Z}_i$ and $\mathbf{Z}_o$ matrices and the $\mathbf{y}_i$ and $\mathbf{y}_o$ vectors.

Predictions were generated from model outputs during simulation of the 120 day prediction stress period (i.e., stress period number 151) and include groundwater levels, groundwater-surface-water exchange volume, and advective travel time through the groundwater system. Travel time predictions were calculated with the particle tracking code MODPATH [*Pollock*, 2012]. A single particle is released at the top center of the active domain in layer 1 at the start of the 120 day prediction period; the total distance the particle travels is recorded as a prediction.

5. Results

Application of equation (14) to several types of predictions made by the synthetic model reveals a large potential for prediction degradation by calibration-induced bias arising from model defects. Easily

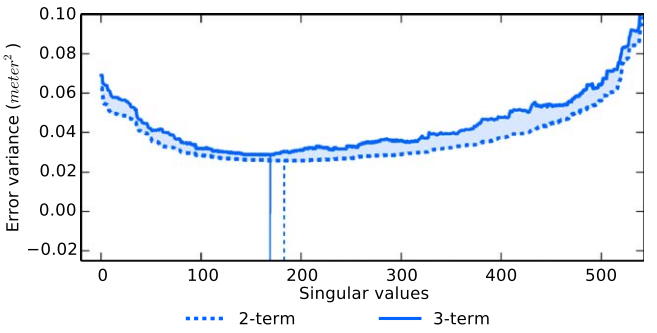

Figure 8. Predictive error variance function for water level at ungauged location G. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). The addition of the third (model-error) term does not substantially increase error variance.

implemented strategies can defend against this bias if used properly. However, improper use of these strategies may exacerbate predictive bias.

For the predictions considered herein, curves of error variance are plotted against the number of singular values used for calibration, exposing the relation between the location of minimum predictive error variance and the level of calibration. Two sets of curves are presented for each prediction. The "two-term" results employ only the first two terms of equation (14), which is predictive error variance under the assumption of no model defects. The "three-term" results include all of the terms of equation (14), which is the predictive error variance including the contribution of model error.

5.1. Omitting Observations

The error variance function for predicting water levels at location G was calculated using only observations from locations B and C for calibration, because excluding observations from location A may be presumed to minimize the potential for predictive bias since location A is near an extraction well whose pumping rate is uncertain. However, even without an apparent source of structural noise, the third term of equation (14) is nonzero (Figure 8). The difference between the two-term and three-term prior uncertainty (i.e., the intersection of the error variance functions with the *y* axis) is less than 0.01 m^2, indicating that predicting head at this location is only slightly sensitive to the omitted parameters $\mathbf{k}_o$. However, inclusion of the third term does shift the error variance minimum from near 180 singular values to about 165 singular values, implying that for this prediction, seeking a poorer fit is appropriate in the presence of model error. The three-term minimum of error variance is also slightly higher than its two-term counterpart, indicating model defects have reduced the reliability of the calibrated model to make this water-level prediction.

The same process was applied to predicting the volumetric groundwater discharge to surface water during the 120 day prediction period; this type of prediction is a key objective for integrated surface-water/groundwater models. Only unprocessed observations from locations B and C were used to calibrate the model prior to making the prediction under drought conditions. However, in contrast to the water-level prediction, model error plays a much larger role in the predictive error variance function (Figure 9). The model-error (third) term dominates the other two terms of equation (14) for singular values less than 400. The two-term minimum of error variance 1.5×10^6 $(m^3)^2$ $(volume^2)$ occurs near 180 singular values. However, the three-term minimum of error variance is 2.5×10^6 $(m^3)^2$, occurring near 450 singular values. For this prediction, the intuitive strategy of seeking a poorer fit to the calibration data set as a mechanism to reduce the negative effects of model error is not effective, and results in a potential for predictive error that exceeds the prior uncertainty if less than 200 singular values are used for calibration. A more appropriate calibration strategy for this prediction is to seek a good fit so that more parameters can adopt compensating roles, which in turn has a beneficial effect in reducing error variance.

Equation (14) also can be used to compare the relative benefit of different strategies to mitigate the potentially degrading effects of model error on postcalibration predictions. A direct comparison of error variance functions for the predicted water level at location A (at the end of the prediction period) was computed for two cases. In the first case, observations from location A (computed during the calibration period) are included in the calibration data set; the second case excludes them. Calibration with observations from location A, despite their likely contamination by structural noise, significantly reduces the error variance for predicting water level at the same location (Figure 10). Results indicate that in addition to structural noise, observations at this location must also contain significant information pertaining to prediction-sensitive

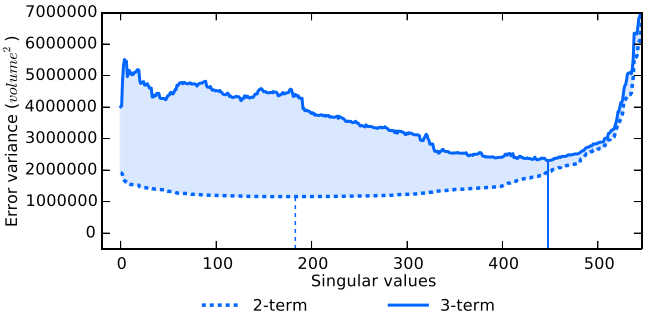

Figure 9. Predictive error variance function for volumetric groundwater contribution to surface water. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). The third (model-error) term increases error variance and shifts the minimum from near 180 singular values to near 450 singular values. For reference, the initial forward model run predicts an exchange volume of 25,000 m^3.

local hydraulic properties such that the compensatory roles induced by contaminated observations have a beneficial effect for this prediction. Furthermore, the profound error variance drop at four singular values for the three-term equation indicates that the information content of observations from this site informs a single parameter solution-space eigencomponent (column of $\mathbf{V}_{i1}$ of equation (14)). Successful use of a low-dimensional solution space suggests that a parsimonious parameterization scheme may be sufficient to minimize error variance for this prediction.

Model predictions of advective water movement (including particle locations and times) are often sensitive to null-space parameter components, which is an outcome of the reliance on parameterization details that cannot be inferred from observation data used during calibration [*Moore and Doherty*, 2005]. To explore null-space-dependent predictions, equation (14) was applied to the particle travel distance prediction described previously. Comparisons of error variance functions for the particle travel distance using unprocessed observations were computed for cases that include and omit observations from location A (Figure 11). In contrast to the previous prediction, the structural noise accompanying observations from location A significantly increases the potential error for predicting particle travel distance. Excluding observations from location A increases the two-term error variance, which is consistent with these observations informing prediction-sensitive parameters. However, consideration of the three-term results shows that excluding these observations lowers the minimum error variance. For this prediction, parameter compensation induced by calibrating a defective model with observations from location A compromises the ability to make reliable predictions.

5.2. Observation Processing

Equation (14) was used to examine the error variance associated with predicted water level at location D for two alternative calibration strategies. The first strategy uses only unprocessed observations; the second uses only observations processed with the relative-datum processing operation. Observations from locations A–C were included in both calibration data sets. The two-term error function for this prediction is insensitive to the data processing, indicating this type of processing has not removed any prediction-related information from the calibration data set (Figure 12). However, differences in the three-term error functions are marked. Processed observations yield a much lower predictive error variance, indicating observation preprocessing has removed some aspects of the model-error signal that induce parameter compensation. Removing these components, in turn, reduces compensation-induced bias for this prediction (Figure 12).

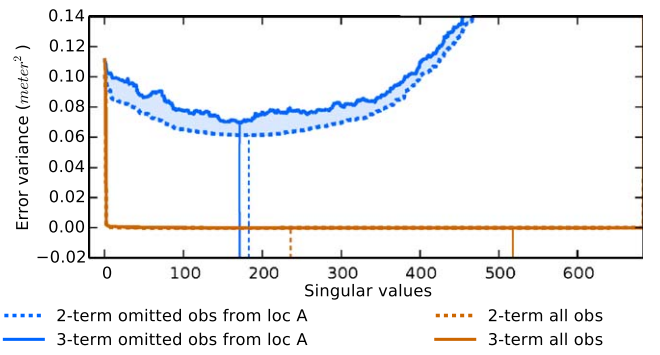

Figure 10. Predictive error variance function for water level at location A. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Excluding observations from location A increases predictive error variance.

5.3. KL Transformation

Equation (14) was employed to evaluate use of the KL transform in model calibration and subsequent

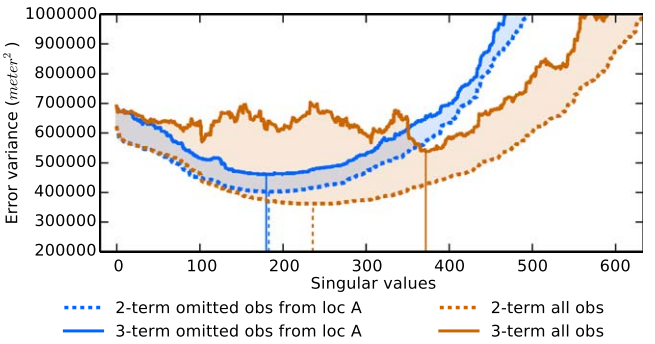

Figure 11. Predictive error variance function for particle travel distance. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Excluding observations from location A decreases error variance. For reference, the initial forward model run predicts a travel distance of 11,400 m.

prediction of water level at location F. Unprocessed observations from locations A–C were used for two model calibration exercises. In one calibration, parameters were KL transformed prior to calibration, while in the other they were not. KL transformation was accomplished using the prior parameter covariance matrix described previously. For some predictions, inclusion of expert knowledge in calibration through the KL transformation is essential to eliminate additional model error incurred through improper adjustment of parameters (Figure 13). Without KL transformation, any attempt to calibrate the model raises the potential for predictive error. For other predictions, like water level at locations D (Figure 14) and E (Figure 15), use of the KL transform prevents localized parameter compensation, which increases the potential for predictive error. The calibration-induced compensatory roles assigned to the primary eigencomponents of the prior parameter covariance matrix spread the effects of model error over larger regions of the model domain, resulting in degraded predictive ability.

6. Discussion

A new method was developed to assess the predictive consequences of calibrating a simplified computer model for the purpose of extracting information from a limited number of system-state observations. The linear basis for the method has advantages and disadvantages. Real-world systems can be (highly) nonlinear; therefore, analysis based on a linear assumption may only be approximate. However, this does not preclude the ability of linear subspace analysis to provide useful information, as well as powerful insights into the strengths and weakness of different approaches to model calibration. Use of linear subspace methods facilitates analysis of systems that have thousands of unknowns. The synthetic analysis described herein used 6288 parameters and required completing 6289 model simulations, which can easily be completed in parallel. Given the complex nature of groundwater and petroleum reservoir systems, failure to represent a high level of system complexity in a computer model may invalidate the usefulness of the model.

Another benefit of the method is that the equations that emerge from linear analysis do not rely on particular parameter values, or on the values of any particular observed value. Only sensitivities and the prior parameter covariance matrix are required by equation (14). It follows that analyses based on equation (14) do not require that the model be calibrated or that observations used for calibration even exist. As a result, linear analyses can be used to assess many different options for formulation of an inverse problem prior to actual calibration, including design of a model, parameterization, and processing of observations and model-generated counterparts.

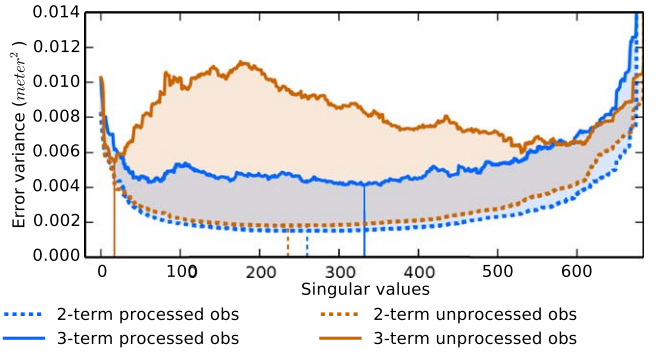

Figure 12. Predictive error variance function for water level at location D. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Observations were processed to a relative-datum to reduce error variance.

Given the complex relations between computer models and corresponding natural systems, all methods for analysis of model error must necessarily rely on

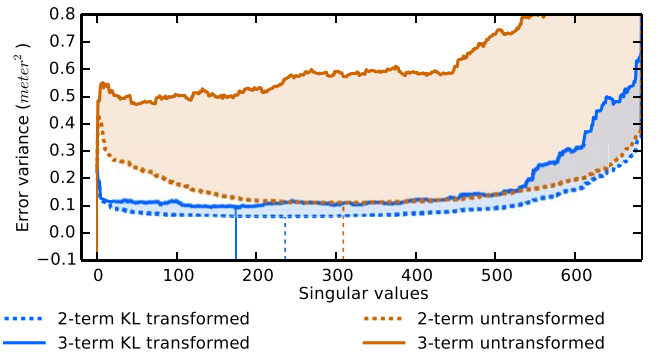

Figure 13. Predictive error variance function for water level at location F. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Use of the KL transform reduces error variance.

simplifying assumptions to form the foundation for a tractable analysis. Use of equation (14) assumes availability of the model of "reality" [$\mathbf{Z}_i$ $\mathbf{Z}_o$] that is simplified to form the computer model $\mathbf{Z}_i$. In real-world modeling practice, even the most complex and dense parameterization schemes cannot represent the full suite of omitted parameters needed to fill $\mathbf{k}_o$. Incomplete knowledge of all omitted parameters makes the effects of model error calculated with our new method approximate. However, in contrast to most existing methods, use of equation (14) can detect the invisible component of model error and therefore does not require that model error generate any misfit during calibration [e.g., *Kennedy and O'Hagan*, 2001]. In highly parameterized settings, where parameter compensation is more likely to occur, identifying the invisible component of structural noise is critical for estimating the uncertainty of some predictions.

Application of equation (14) to predictions made with an integrated surface-water/groundwater model reveals a variety of outcomes that are strongly prediction dependent. For predictions that are highly sensitive to null-space parameter combinations of $\mathbf{Z}_i$, postcalibration error variance may be reduced very little from its prior uncertainty because calibration does little to constrain the prediction-sensitive parameters [*Moore and Doherty*, 2005]. It follows that calibration also induces little bias in these types of predictions. In contrast, for predictions that are in general very similar in location and character to observations used for calibration, history matching with a defective model can be very effective in reducing the potential for predictive error by extracting information from the observations, even though calibration may induce prediction-sensitive parameters to assume compensating roles. For these predictions, the model can function effectively as a "black box" since its main design and calibration criterion are the minimizations of model-to-measurement misfit. In fact, for these types of predictions, reducing model-to-measurement misfit may be more important for minimizing error variance than the application of expert knowledge. This philosophy underpins the design of model emulators [*Young and Leedal*, 2013]. For these types of predictions, the physical basis of a process model can be safely abandoned.

Most model predictions lie somewhere between these two extremes; most models are built to evaluate future stresses that may be somewhat different from historical stresses. Therefore, matching model outputs to system-state observations only informs some of the prediction-sensitive parameter combinations, while other parameter combinations remain uninformed and are in the null space of the real-world model operator [$\mathbf{Z}_i$ $\mathbf{Z}_o$] of equation (2). *Doherty and Christensen* [2011] indicate that as a result of suboptimal simplification, these types of predictions are most vulnerable to calibration-induced bias.

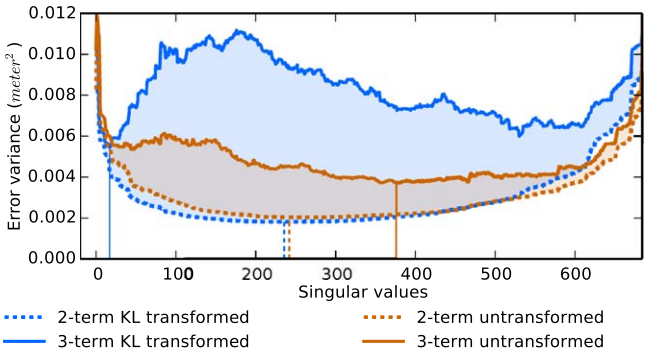

Figure 14. Predictive error variance function for water level at location D. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Use of the KL transform increases error variance.

We have shown how the potential invisibility of bias associated with these types of predictions makes robust estimation of posterior uncertainty difficult. If not properly accounted for, invisible calibration-induced bias may invalidate

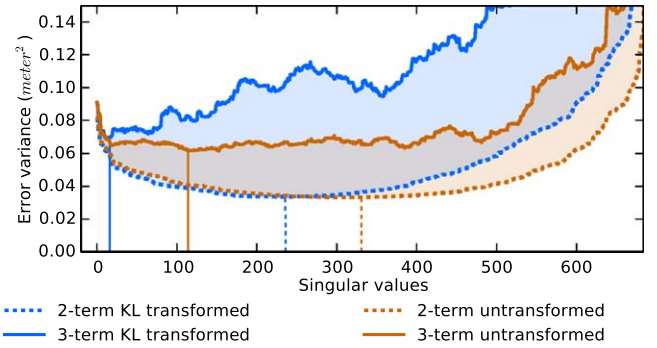

Figure 15. Predictive error variance function for water level at location E. Vertical lines mark curve minima. Two-term results are calculated using only the first two terms of equation (14); three-term results are calculated using all the terms of equation (14). Use of the KL transform increases error variance.

posterior uncertainty estimates (for example, see Figure 9). The analyses presented above are possible because the omitted parameter space ($\mathbf{k}_o$) is completely known in the synthetic, albeit complex model. In real-world modeling analyses, the full dimension of $\mathbf{k}_o$ cannot be known, which means that robust quantification of uncertainty for some types of predictions is nearly impossible. However, using the method presented herein, the potential for calibration-induced predictive bias arising from commonly recognized sources of model error can be estimated.

A number of strategies can be applied as part of calibration to identify model structural defects as well as reduce the contribution that model structural defects make to the error variance of some predictions, including

1. Seeking a prediction-specific level of fit between observations and model outputs.

2. Processing of observation data and model-generated counterparts in ways that "filter out" expressions of model error before calibration.

3. Use of manual and/or mathematical parameterization and regularization techniques that allow spatial and/or temporal containment of parameter compensation.

The use of equation (14) in the above analyses has shown that the application of each of these three strategies is prediction dependent. This is in direct contrast to the common practice of using a single, calibrated model for assessing posterior predictive uncertainty of many different types of predictions. Calibration-induced predictive bias arising from model error is shown to severely compromise this approach to model usage. Calibration instead should be implemented differently for different predictions. Equation (14) can be used as a guide in implementing these prediction-specific calibration strategies. The same model may need to be repeatedly calibrated with different parameterization, different regularization, and/or different objective function formulations to achieve the maximum reduction of potential error for a suite of different predictions.

Parameterizing uncertain aspects of a computer model is an inherently subjective process. In groundwater modeling settings, a parsimonious parameterization has been recommended by *Voss* [2011a, b], *Hill and Tiedeman* [2006], and others. Purposefully limiting the number of estimable parameters may not degrade the reliability of model predictions that strongly resemble the historical system responses. For other predictions that are sensitive to null-space parameter combinations, parsimonious parameterization introduces another unnecessary layer of model imperfection that may contribute significantly to postcalibration error variance. Parameterization-induced error can be minimized by adopting a highly parameterized approach to model calibration and uncertainty analysis. Highly parameterized inversion is not suggested to achieve a better fit with observations because a well-fit imperfect model can substantially increase the potential bias for some predictions. Instead, high parameter dimensionality provides mathematical inversion and associated regularization techniques with maximum flexibility in separating the solution space of $\mathbf{Z}_i$ from null space, which, in theory, should reduce the misalignment of these subspaces with the subspaces of the reality model [$\mathbf{Z}_i$ $\mathbf{Z}_o$]. Additionally, predictions with a high null-space dependency require the use of many parameters to effectively assess posterior predictive uncertainty.

7. Conclusion

The present paper expands the theory originally presented by *Doherty and Christensen* [2011] and investigates the ramifications of using structurally defective models to predict the response of complex

environmental systems to future stresses. The focus is on systems that involve subsurface flow because these systems are normally accompanied by a high level of spatial heterogeneity, in lithologies, boundaries, and hydraulic properties.

The theory and examples provided demonstrate that the reliability of predictions made by a defective model may or may not be seriously compromised by a model error. For some model predictions, defects can be "calibrated out." These predictions tend to be similar to the calibration data set. However, the reliability of other predictions made by the same calibrated model may be seriously compromised by the calibration process. Unfortunately, the model which is used to make predictions cannot be used to analyze the effect of model error on those predictions. However, through necessarily synthetic studies, such as the present one, an analogue of a natural system can be used to infer the extent of possible calibration-induced bias, as well as strategies that may mitigate these effects for common types of model predictions used to inform decision making.

The present study demonstrates that construction, calibration, and predictive usage of a defective model is considerably more complicated than for a presumed defect-free model. As shown, strategies that protect the reliability of a prediction from the worst effects of model error may actually degrade the reliability of other predictions. This is in contrast to the concept of a single "calibrated model" that is used for making a suite of predictions. This study suggests that many aspects of the model construction and calibration process must be designed for a single type of model prediction. If another type of prediction is needed, another construction and calibration strategy may be necessary to optimize the ability to make that prediction. Specific aspects of the calibration process that may need prediction-specific adjustment include formulation of the objective function, the level of fit sought with the calibration data set, and formulation of an appropriate regularization strategy.

The present study also demonstrates that the link between a well-calibrated model and reliable predictions may be broken. For some predictions, achieving a good fit during the calibration process results in a large reduction in error variance, regardless of the compensatory roles that parameters assume from the calibration process. For other predictions, any adjustment of model parameters to improve the fit with the calibration data set induces bias that can severely compromise predictive reliability.

Given the contrast between common modeling practice and the outcome of the present study, it is hoped that methods such as the one presented herein continue to be applied to synthetic studies that resemble real-world contexts. Such studies will provide modelers with the qualitative knowledge needed for the "art" of environmental modeling. The outcomes of these synthetic studies will provide the environmental modeling community with greater confidence for making subjective model construction, calibration, and application decisions that are an inherent part of the modeling process.

Acknowledgments
The authors would like to acknowledge Kim Haag for editing the original manuscript. The authors would also like to acknowledge Steffen Mehl and three anonymous reviewers whose comments and suggestions improved the quality of this manuscript.

References

Aster, R., B. Borchers, and C. Thurber (2013), *Parameter Estimation and Inverse Problems*, Academic, Waltham, Mass.
Beven, K. (2005), On the concept of model structural error, *Water Sci. Technol.*, *52*(6), 167–175.
Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.
Beven, K., P. Smith, I. Westerberg, and J. Freer (2012), Comment on "Pursuing the method of multiple working hypotheses for hydrological modeling" by P. Clark et al., *Water Resour. Res.*, *48*, W11801, doi:10.1029/2012WR012282.
Clark, M. P., and J. A. Vrugt (2006), Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters, *Geophys. Res. Lett.*, *33*, L06406, doi:10.1029/2005GL025604.
Dausman, A., J. Doherty, C. Langevin, and M. Sukop (2010), Quantifying data worth toward reducing predictive uncertainty, *Ground Water*, *48*(5), 729–740.
Doherty, J. (2003), Ground water model calibration using pilot points and regularization, *Ground Water*, *41*(2), 170–177.
Doherty, J., and S. Christensen (2011), Use of paired simple and complex models to reduce predictive bias and quantify uncertainty, *Water Resour. Res.*, *47*, W12534, doi:10.1029/2011WR010763.
Doherty, J., and D. Welter (2010), A short exploration of structural noise, *Water Resour. Res.*, *46*, W05525, doi:10.1029/2009WR008377.
Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc., Ser. B*, *57*(1), 45–97.
Evensen, G. (2003), The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dyn.*, *53*(4), 343–367, doi:10.1007/s10236-003-0036-9.
Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resour. Res.*, *49*, 260–282, doi:10.1029/2011WR011779.
Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*(7), 2161–2173.
Golub, G., and C. Van Loan (1996), *Matrix Computations, Johns Hopkins Stud. in the Math. Sci.*, Johns Hopkins Univ. Press, Baltimore, Md.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763.
Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.
Harbaugh, A. W. (2005), MODFLOW-2005, the U.S. Geological Survey modular ground-water model: The ground-water flow process [electronic], *U.S. Geol. Surv. Tech. and Methods 6-A16*, U.S. Dep. of the Inter., U.S. Geol. Surv., Reston, Va.
Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne (2005), Combining field data and computer simulations for calibration and prediction, *SIAM J. Sci. Comput.*, *26*(2), 448–466, doi:10.1137/S1064827503426693.
Hill, M., and C. Tiedeman (2006), *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley, Hoboken, N. J.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.
Hughes, J., C. Langevin, K. Chartier, and J. White (2012), Documentation of the surface-water routing (SWR1) process for modeling surface-water flow with the U.S. Geological Survey Modular Ground-Water Model (MODFLOW2005), *U.S. Geol. Surv. Tech. and Methods 6-A40*, U.S. Dep. of the Inter., U.S. Geol. Surv., Reston, Va.
Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*(3), 425–450.
Koch, K.-R. (1988), *Parameter Estimation and Hypothesis Testing in Linear Models*, Springer, New York.
Lin, Z., and M. B. Beck (2012), Accounting for structural error and uncertainty in a model: An approach based on model parameters as stochastic processes, *Environ. Model. Software*, *27–28*, 97–111, doi:10.1016/j.envsoft.2011.08.015.
Madsen, H. (2000), Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, *J. Hydrol.*, *235*(34), 276–288, doi:10.1016/S0022-1694(00)00279-1.
Menke, W. (1989), *Geophysical Data Analysis: Discrete Inverse Theory*, 2nd ed., Academic, San Diego, Calif.
Moore, C., and J. Doherty (2005), Role of the calibration process in reducing model predictive error, *Water Resour. Res.*, *41*, W05020, doi:10.1029/2004WR003501.
O'Hagan, A. (2006), Bayesian analysis of computer code outputs: A tutorial, *Reliab. Eng. Syst. Saf.*, *91*(10–11), 1290–1300, doi:10.1016/j.ress.2005.11.025.
Oakley, J. (2004), Estimating percentiles of uncertain computer code outputs, *J. R. Stat. Soc., Ser. C*, *53*(1), 83–93, doi:10.1046/j.0035-9254.2003.05044.x.
Oakley, J., and A. O'Hagan (2002), Bayesian inference for the uncertainty distribution of computer model outputs, *Biometrika*, *89*(4), 769–784.
Poeter, E. P., and M. C. Hill (2007), MMA, a computer code for multi-model analysis: Constructed using the JUPITER API [electronic], *U.S. Geol. Surv. Tech. and Methods 6-E3*, U.S. Dep. of the Inter., U.S. Geol. Surv., Boulder, Colo.
Pollock, D. W. (2012), User guide for MODPATH version 6—A particle-tracking model for MODFLOW, *U.S. Geol. Surv. Tech. and Methods 6-A41*, U.S. Dep. of the Inter., U.S. Geol. Surv., Reston, Va.
Raftery, A. E., D. Madigan, and J. A. Hoeting (1997), Bayesian model averaging for linear regression models, *J. Am. Stat. Assoc.*, *92*(437), 179–191, doi:10.1080/01621459.1997.10473615.
Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.
Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*, W12418, doi:10.1029/2008WR006908.
Spaaks, J. H., and W. Bouten (2013), Resolving structural errors in a spatially distributed hydrologic model, *Hydrol. Earth Syst. Sci. Discuss.*, *10*(2), 1819–1858, doi:10.5194/hessd-10-1819-2013.
Voss, C. (2011a), Editor's message: Groundwater modeling fantasies. Part 1: Adrift in the details, *Hydrogeol. J.*, *19*(7), 1281–1284.
Voss, C. (2011b), Editor's message: Groundwater modeling fantasies. Part 2: Down to earth, *Hydrogeol. J.*, *19*(8), 1455–1458.
Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.
Vrugt, J. A., C. G. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.
Watson, T. A., J. E. Doherty, and S. Christensen (2013), Parameter and predictive outcomes of model simplification, *Water Resour. Res.*, *49*, 3952–3977, doi:10.1002/wrcr.20145.
Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.
Young, P. C., and D. Leedal (2013), *Data-Based Mechanistic Modelling and the Emulation of Large Environmental System Models*, pp. 111–131, John Wiley, Hoboken, N. J.