Outcomes of Numerical Groundwater Model Simplification and Calibration

Ty Watson

B.Sc. (Hons)

Submitted as a requirement in full for the degree of

Doctor of Philosophy

in the

School of the Environment

Flinders University of South Australia

April 2017

Table of contents

Dedication

To Mum and Dad.

Acknowledgements

I cannot adequately express the thanks I owe to my supervisors John Doherty, Adrian Werner and Craig Simmons for what they have provided for me on countless levels: opportunity, academic wisdom and support, enthusiasm, inspiration, encouragement, tolerance, patience, approachability and friendship, to name just a few. I must also express great thanks to Steen Christensen of Aarhus University, not only for co-authorship contributions, but for mentorship and hospitality during my time in Denmark.

I am extremely grateful for the opportunity and flexibility provided by Flinders University, the financial support of an Australian Postgraduate Award and scholarship from the National Centre for Groundwater Research and Training (NCGRT). Further to financial support, I am forever grateful to the NCGRT for the many superb opportunities, including interstate and international travel for collaboration, conference presentations, workshops, training, and teaching experiences.

I also wish to thank James Craig of the University of Waterloo, as well as two anonymous reviewers for reviewing the manuscript presented as Chapter 3 of this thesis.

I must also thank my current employer Australian Groundwater Technologies (AGT) for the recent empathy and flexibility that has allowed me to complete this thesis.

I would like to thank the many great friends I have made in colleagues and visitors during my time at Flinders University, as well as during my opportunities to travel. There are far too many to name. But thank-you in particular to Matt Knowling for further inspiring me with enthusiasm for modelling and research, and for many enjoyable and beneficial discussions.

I could not have done this without the support of family and friends. Words cannot do justice to the thanks I owe to mum and dad for their infinite support. A special thank you goes also to my partner, Hannah, for bearing a huge amount in order to help me achieve this.

Without all of the above, "this book would probably be worse" (Herckenrath, 2012).

Declaration of Originality

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Ty Watson

List of Figures

List of Tables

Summary

Calibration of numerical models as a precursor to predictive uncertainty analysis is now regarded as standard practice in groundwater modelling for management and decision support. Despite computational advances facilitating increasingly complex and realistic model-based representations of natural systems, model simplifications/imperfections relative to the incomprehensible detail of reality is unavoidable.

Calibration of a simplified/imperfect model may lead to additional error in model predictions that is undetectable through standard uncertainty analysis approaches. This calibration-induced "bias" increases the risk of underestimation of potential predictive error, which defines ultimate failure of a modelling process. Assurance against modelling failure thus requires that calibration-induced predictive bias is forestalled or quantified. This thesis makes several key contributions to the knowledge base pertaining to calibration-induced predictive bias identification, and exposition of its origins, towards providing best-practice guidance for repressing its occurrence.

The first component of work is a proof of concept for the "paired model analysis" (PMA) methodology for bias identification and reduction presented by Doherty and Christensen (2011). PMA has not previously been tested for empirical consistency with theoretical expectation. PMA is applied to a highly studied synthetic example, demonstrating good agreement between PMA-quantified uncertainty with the results of established methods. The reliability of PMA in identifying calibration-induced predictive bias is systematically demonstrated, together with its capacity to reduce the consequential inflation in potential predictive error.

The second component of work builds upon the mathematical exposition of model simplification outcomes developed by Doherty and Christensen (2011). In particular it is extended to express "null-space entrainment"; a concomitant outcome of the parameter surrogacy that may occur during calibration and which is the fundamental cause of calibration-induced predictive bias. The developed linear concepts are employed in conjunction with PMA to examine the outcomes of two simplifications of a one-dimensional Richards equation-based vadose zone model. Substantial parameter surrogacy and consequential null-space entrainment is demonstrated to occur for both comparatively modest simplification (i.e., assumption of vertical

homogeneity), and more drastic simplification (i.e., replacement with a lumped parameter "bucket" model). Nonetheless, both simplified models are shown to make largely unbiased predictions of future recharge. This demonstrates that, for predictions that are similar in nature to the available calibration dataset, a model's physical basis becomes less important to its predictive performance than attainment of a "good fit".

The final component of work explores the outcomes of employing the increasingly popular pilot-point-based regularized inversion approach for calibration in a categorically heterogeneous environment. PMA is used to thoroughly examine model performance in making multiple predictions subject to several regularization weighting strategies. For some predictions, ignoring the existence of preferential flow features does not compromise the ability of the calibration and uncertainty analysis processes to substantially reduce and quantify potential predictive error. Simultaneously, calibration unavoidably inflates the potential error in other predictions beyond prior uncertainty. The results emphasize the need for prediction-specific tuning of the modelling process, to the extent that the most pragmatic approach for some predictions may be to forego calibration entirely and quantify uncertainty based on geologically realistic expressions of "expert knowledge" alone.

Chapter 1

Background and objectives

Numerical models are used globally as environmental forecasting tools. Enormous advancements in computing capabilities in recent decades have provided the capacity for increasingly complex and realistic numerical representations of natural systems (Vrugt et al., 2006; Ratto et al., 2011; Hunt and Zheng, 2012). Nonetheless, perfect model-based characterisation of subsurface environments in particular is impossible due to inevitably sparse and uncertain field information (Carrera and Neuman, 1986; Zhou et al., 2014, Anderson et al., 2015). Uncertainty in predictions made by a groundwater model is therefore inevitable, the characterization of which has long been recognised as fundamental to model use for environmental management and decision-making (e.g., Freeze et al., 1990).

The groundwater modelling community remains somewhat divided in terms of predictive uncertainty quantification philosophy (Hunt et al., 2007). The use of many stochastic model runs to explore the range of predictive possibilities within a Bayesian sampling-based framework is strongly advocated by some authors (e.g., Beven and Binley, 1992; Gómez-Hernández, 2006; Vrugt et al., 2009b). Bayesian approaches are generally recognized as providing the most reliable and robust estimates of uncertainty (e.g., Gallagher and Doherty, 2007a). However, despite the aforementioned recent advances in computing capabilities, including promise offered by cloud computing (Hunt et al., 2010; Langevin and Panday, 2012), such approaches often remain infeasible due to the requirement of a prohibitively large number of model runs (e.g., Mugunthan and Shoemaker, 2006; Mariethoz et al., 2010a; Borghi et al., 2016). Tolson and Shoemaker (2008) exemplify this, citing that the typical number of runs of their case study watershed model that would be required for a selected (informal) Bayesian approach would require 4.6 months of serial computing time (on a Pentium IV, 3 GHz computer).

Model calibration or "history-matching" as a precursor to (calibration-constrained) uncertainty analysis presents a vastly more efficient alternative to Bayesian approaches (e.g., Gallagher and Doherty, 2007a; Keating et al., 2010). The calibration process seeks an optimized set of model parameters (representing system hydraulic properties) based on which model outputs adequately reproduce real-world observations of system state (Konikow and Bredehoeft, 1992). Whilst some proponents of the Bayesian approach suggest that the notion of a single calibrated model has no place in environmental simulation, calibration now forms a standard component of defensible environmental modelling (Hunt et al., 2007; Anderson et al., 2015).

A calibrated model can never promise to provide a single accurate prediction; its role is to facilitate estimation of uncertainty bounds within which the "true" value of a prediction of future environmental behaviour can be guaranteed to lie (Doherty, 2011). The issue of groundwater model simplification is pervasive, this being attributable to three main sources:

1. Groundwater model calibration almost always constitutes an ill-posed inverse problem due to large numbers of unknown parameters (Anderson et al., 2015). Attaining a unique solution to the inverse problem is itself an implicit model simplification device as it necessitates the simplest estimated parameter field that is compatible with inevitably sparse (and "noisy") study area information (McLaughlin and Townley, 1996; Moore and Doherty, 2006; Welter et al., 2015).

2. Despite the abovementioned advancements in computing capabilities, deliberate simplification is generally required to achieve shorter execution times and numerical stability to facilitate the computationally intensive undertakings of history-matching and calibration-constrained predictive uncertainty analysis (Ratto et al., 2011; Burrows and Doherty, 2015). This also includes the necessarily subjective pre-calibration decisions faced by all modellers in terms of which model parameters and boundary conditions to fix and which to estimate through calibration (e.g., White et al., 2014). These factors may be further compounded from the practical perspective that groundwater modelling is often undertaken in a context of limited funding (Haitjema, 2011).

3. Even the most complex computer models are inherently simplified with respect to the incomprehensible complexity that defines the natural world, which inevitably includes "unknown unknowns" (Hunt and Welter, 2010; Hunt and Zheng, 2012).

"Traditional" uncertainty analysis methods seek to address the first of the above sources of model simplification and resultant uncertainty in model predictions (Refsgaard et al., 2006). An optimally formulated calibration process can theoretically provide a set of estimated parameters, and thus model predictions, that have a minimum potential for error given the available information pertaining to the system under study. This provides an ideal foundation for the critical task of quantifying this potential through post-calibration predictive uncertainty analysis. A range of techniques exist for this purpose, examples of which include linear analysis (e.g., Moore and Doherty, 2005; Gallagher and Doherty, 2007a), calibration-constrained Monte Carlo methodologies (e.g., Tonkin et al., 2007; Tonkin and Doherty, 2009; Herckenrath et al., 2011; Yoon et al., 2013) or Pareto analysis/hypothesis-testing approaches (e.g., Moore et al. 2010).

Model simplifications/imperfections pertaining to the second and third of the above sources are broadly acknowledged and have been referred to using a variety of terms, including for example model structural error (e.g., Doherty and Welter, 2010), conceptual uncertainty (e.g., Refsgaard et al., 2006), model inadequacy (e.g., Kennedy and O'Hagan, 2001), and model defects (e.g., White et al., 2014). Model imperfections present sources of potential predictive error that are supplementary to those quantified through traditional uncertainty analysis. Means of accounting for this additional error has been a topic of extensive study over many years. However, most approaches rely upon the expression of a model's imperfections as a simulator of the natural environment in the form of model-to-measurement misfit (e.g., Beven and Binley, 1992; Draper, 1995; Gupta et al., 1998; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Vrugt et al., 2005; Ye et al., 2008; Doherty and Welter, 2010; Spaaks and Bouten, 2013; Xu and Valocchi, 2015, among others).

Recent literature in particular explores additional potential predictive error induced through calibration of a simplified model, where model imperfections do not compromise its ability to achieve a "good fit" with observation data (e.g., Doherty and Christensen, 2011; White et al., 2014). This calibration-induced "bias" as defined by

Doherty and Christensen (2011) is thus undetectable through traditional approaches to uncertainty analysis. As such, it threatens underestimation of the true range of post-calibration potential predictive error. This increases the risk of "type II" statistical error, which defines the false rejection of a true hypothesis (e.g., Downes et al., 2002; Beven 2010). An example of type II statistical error in the environmental decision-support context is the occurrence of an unacceptable environmental impact despite model-based assurance that it will not occur. This defines failure of a modelling process according to Doherty and Vogwill (2016).

Avoidance of failure in environmental modelling practice requires either that bias in model predictions is accounted for, or that its occurrence is mitigated/prevented. This may be achieved through development of practical methodologies for quantifying and/or reducing the propensity for predictive error incurred though calibration of a simplified model in place of a more complex one. Doherty and Christensen (2011) present an approach that involves the pairwise use of both a complex and simplified model of the same system, allowing detection and reduction of calibration-induced predictive bias simultaneous with uncertainty quantification. These authors also present a mathematical exposition of the outcomes of calibrating a simplified model within a linear analysis framework, which is extended by White et al. (2014) to demonstrate an efficient linear approach to quantifying the effects of model imperfections upon post-calibration potential predictive error. Burrows and Doherty (2015; 2016) present novel calibration approaches also involving conjunctive complex/surrogate model usage. A reduced propensity for calibration-induced predictive bias is achieved through greater model complexity facilitated by conjunctive use of a surrogate model to increase numerical stability and reduce computational expense.

Notwithstanding the computational advantages afforded by the above methodologies relative to standalone usage of a highly complex model, their application may yet remain infeasible in many practical circumstances. Moreover, White et al. (2014) acknowledge that such methodologies, whilst exposing potential predictive error that is "invisible" to traditional uncertainty analysis methods, cannot account for the indeterminable discrepancies that exists between even the most complex of models and reality itself (this being in accordance with the third source of potential predictive error as described above). In light of this, White et al. (2014) emphasize the need for synthetic studies involving calibration of models that are subject to representative

simplifications/imperfections relative to a known "reality". The intention of this is to provide best-practice guidance through development of qualitative understanding of the causes of predictive bias, most susceptible prediction types and circumstances, and possible mitigating strategies that may be taken by a modeller. This is critical to the necessarily intuitive aspect of modelling that has been referred to by terms such as the "art" of modelling (e.g., Savenije, 2009; Doherty, 2011; White et al., 2014) and "hydrosense" (e.g., Hunt and Zheng, 2012; Simmons and Hunt, 2012; Anderson et al., 2015).

Doherty and Christensen (2011) explicate predictive bias as being caused by the surrogate roles adopted by some parameters during the calibration process, as they compensate for model imperfections in allowing a close fit between model outputs and observations to be attained. Parameter surrogacy/compensation as a potential outcome of the calibration process is widely acknowledged (e.g., Clark and Vrugt, 2006; Beven, 2006; Spaaks and Bouten, 2013; Xu and Valocchi, 2015). Doherty and Christensen (2011) explain that this is accompanied by what they refer to as "null-space entrainment". Parameters belonging to the so-called null space are those which are not informed by available calibration data. Deviation of their values from pre-calibration expected values (i.e., based on expert knowledge) should not occur if calibration is to achieve an estimated parameter set of maximum likelihood. Surrogate behaviour of simplified model parameters during calibration is accompanied by notional adjustment (i.e., "entrainment") of the null-space parameters of the more complex model that the simplified model represents. Model predictions that are sensitive to entrained null-space parameter components will be biased, thus incurring an additional (invisible) component of potential predictive error.

This thesis addresses four overarching objectives related to the pervasive issue of the outcomes of calibrating simplified groundwater models. The first two objectives concern general methodological and theoretical progression that is relevant to practical application as well as research directed at advancing understanding. Applying these methodological and theoretical concepts, the second two objectives directly address the abovementioned need for synthetic studies that explore the outcomes of calibrating simplified models, towards providing best-practice guidance and building crucial modeller intuition. More specifically, the objectives of this thesis are to:

1. Systematically validate the reliability of the "paired model analysis" methodology presented by Doherty and Christensen (2011) in performing its key functions, these being predictive bias identification/reduction and uncertainty quantification. The methodology has not previously been tested for empirical consistency with theoretical expectation.

2. Extend the Doherty and Christensen (2011) mathematical formulation of the outcomes of calibrating a simplified model. In particular, to express mathematically "null-space entrainment", which is central to the predictive outcomes of calibrating a simplified model.

3. Explore and compare the parameter and predictive outcomes of calibrating two simplified versions of a complex model built for the purpose of one-dimensional recharge simulation. A concomitant aim is to explore the occurrence of null-space entrainment through application of the theory developed in accordance with the second objective above, and examine its predictive consequences.

4. Explore the parameter and predictive outcomes of pilot-point-based regularized inversion when applied in an environment containing categorical heterogeneity. This represents a pervasive model simplification context; discrete, discontinuous features are common in the subsurface, whilst the increasingly popular use of pilot points necessitates estimation of a smooth, continuous parameter field.

The first objective is addressed in Chapter 2, which is based on a manuscript in preparation for submission. The second and third objectives are addressed in Chapter 3, which is based on a manuscript published in *Water Resources Research*. The fourth objective is addressed in Chapter 4, which is based on a manuscript in preparation for submission. Chapter 5 summarizes the conclusions of the thesis.

Chapter 2

Paired model analysis for identifying predictive bias and quantifying uncertainty: a proof-of-concept study

Abstract

This work comprises a proof of concept for the methodology presented by Doherty and Christensen (2011), herein referred to as "paired model analysis" (PMA). PMA involves conjunctive use of both a complex and simplified model in order gain efficiency and stability for quantifying predictive uncertainty whilst accounting for predictive bias induced by calibration of the latter in place of the former. PMA is yet to be tested for empirical consistency with theoretical expectation. The purpose of the present study is to verify the efficacy of PMA in performing its key functions; bias identification and uncertainty quantification. PMA is applied to an idealised synthetic example in which the model subject to calibration is structurally identical to a hypothetical "reality model", and predictions of advective transport are made following hydraulic conductivity estimation. First presented by Moore and Doherty (2005), the example is extensively studied in existing literature, providing a basis for comparison of PMA-quantified predictive uncertainty with the results of the previously applied "traditional" methods that do not account for the effects of model simplification. The structurally non-simplified example also allows analysis of the ability of PMA to detect calibration-induced predictive bias arising through other known sources in the absence of the potentially confounding influence of model structural defects. These sources are (1) overfitting with respect to measurement noise, and (2) suboptimal regularization. Results demonstrate that post-calibration uncertainty quantified through PMA is in good agreement with previous results. Calibration-induced predictive bias and the accompanying inflation in predictive error variance is shown to be expressed through PMA results where expected. Concomitantly, the results demonstrate the ability of the calibration process to simultaneously reduce the potential for error in one prediction, whilst increasing that in another, even in the idealised example of a structurally non-simplified model.

Finally, the large reduction in post-calibration potential predictive error achieved through application of PMA is demonstrated.

2.1 Introduction

Appropriate representation of prediction uncertainty has long been acknowledged as integral to environmental model-based decision making (e.g., Freeze et al., 1990). Model calibration as a precursor to calibration-constrained uncertainty analysis is now common practice (Anderson et al., 2015). This involves "history matching", whereby model parameters are adjusted such that model outputs adequately reproduce field observations (Konikow and Bredehoeft, 1992).

"Traditional" (terminology following Refsgaard et al., 2006) approaches to uncertainty analysis account for the nonuniqueness of estimated parameters in the presence of typically limited (and "noisy") field data such measurements of hydraulic head or flux (Anderson et al., 2015). That is, groundwater model calibration generally presents an ill-posed inverse problem, a unique solution to which necessitates a more parsimonious estimated parameter field than the true degree of hydraulic property detail (e.g., McLaughlin and Townley, 1996; Moore and Doherty, 2006; Welter et al., 2015). Post-calibration uncertainty analysis methods seek to quantify the resultant potential for error in model predictions made by the calibrated model, approaches to which include linear analysis techniques (e.g., Moore and Doherty, 2005; Gallagher and Doherty, 2007a), calibration-constrained Monte Carlo methodologies (e.g., Tonkin et al., 2007; Tonkin and Doherty, 2009; Herckenrath et al., 2011; Yoon et al., 2013) or Pareto analysis/hypothesis-testing methodologies (e.g., Moore et al. 2010).

Additional to the parameter uncertainty associated with solution of the inverse problem, all groundwater models are inherently simplified relative to the unknowable complexity of the natural subsurface (Hunt and Welter, 2010; Hunt and Zheng, 2012). Moreover, despite the enormous increase in computing power in recent decades, the use of large-scale physically based groundwater models remains hindered by computational limitations, particularly in a calibration context (Ratto et al., 2011). Further deliberate simplification is often required to attain the manageable run times and numerical integrity necessary for calibration and calibration-constrained uncertainty analysis (Burrows and Doherty, 2015; 2016). These are additional sources

of discrepancy between a model and reality, and thus may increase the potential for predictive error.

Model parameter and predictive error that is attributable to model simplification has been extensively studied, however most approaches rely on model inadequacies as a simulator of environmental behaviour being expressed in the form of irreducible model-to-measurement misfit. Examples of this work include Beven and Binley, 1992; Draper, 1995; Gupta et al., 1998; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Vrugt et al., 2005; Ye et al., 2008; Doherty and Welter, 2010; Spaaks and Bouten, 2013; Xu and Valocchi, 2015, among others.

Recent literature explores post-calibration predictive error in the case where model imperfections do not compromise the achievement of a "good fit" between model outputs and observation data (e.g., Doherty and Christensen, 2011; White et al., 2014). Doherty and Christensen (2011) discuss the compensatory roles played by some parameters as they compensate for model structural defects in order to achieve a close fit with calibration data. They introduce the concept of "null-space entrainment" as an inevitable accompaniment to parameter compensation. The so-called calibration null space is comprised of parameters or parameter combinations that are not informed by the available calibration dataset. Estimation of a set of parameters that have a minimum potential for error necessitates that null-space parameter components remain unperturbed from their (expert knowledge-based) pre-calibration expected values (Doherty and Christensen, 2011). Null-space parameter entrainment refers to the notional adjustment, through calibration of the simplified model, of null-space parameter components belonging to the "reality model". (Null-space entrainment is explored mathematically in Chapter 3 of the present thesis.) Predictions that are sensitive to the affected parameters thus incur an unsupported potential for wrongness, which is defined as "bias" by Doherty and Christensen (2011).

Calibration-induced predictive bias caused by the surrogate roles played by parameters in compensating for model structural error is an additional propensity for potential error that is not quantifiable through what we herein refer to as the "traditional" approach (following Refsgaard et al., 2006) to model uncertainty analysis. Moore and Doherty (2005) expound the components of predictive error variance comprising traditional uncertainty analysis, which quantifies the extent to which innate parameter uncertainty can be constrained in the presence of limited and uncertain observation

data. Calibration-induced predictive bias caused by model structural error is not unaccounted for, and therefore threatens the integrity of model predictions through increased potential for "type II" statistical error, which is defined as false rejection of a true hypothesis (e.g., Downes et al., 2002; Beven, 2010). An example of this would be the occurrence of an unwanted outcome of a proposed environmental management strategy despite model-based assurance that it is extremely unlikely. According to Doherty and Vogwill (2016), this defines failure of a modelling endeavour. As such, accounting for predictive bias (whether through prevention or quantification) is of critical importance.

Doherty and Christensen (2011) present a methodology for identifying and reducing predictive bias in a calibrated environmental model, simultaneous with predictive uncertainty quantification. The methodology, described in detail in the following section, is herein referred to as "paired model analysis" (PMA). It is proposed as an approach that may be employed in practice where a model is required to be calibrated and subsequently used to make predictions of future system behaviour. Standalone use of a highly complex model for calibration and uncertainty analysis is often thwarted by debilitating computational expense and other difficulties such as solver non-convergence and "numerical granularity" (e.g., Burrows and Doherty, 2015; 2016). PMA is designed to circumvent these issues through requiring that only a simplified model of the same system is calibrated. At the same time, it allows identification and reduction of predictive bias that may have been incurred through calibration of the simplified model in place of the more complex model.

White et al. (2014) suggest that, despite the computational benefits offered by PMA in contrast to standalone use of a complex model, the computational expense may nonetheless remain infeasible in some situations due to the requirement for repeated calibration of the simplified model. In light of this, White et al. (2014) extend the mathematical description of model defect-induced predictive error presented by Doherty and Christensen (2011). They present an efficient linear subspace-based equation for predictive error variance quantification, which includes a supplementary term that accounts for additional predictive error variance owing to calibration-induced bias.

White et al. (2014) demonstrate the utility of their methodology through application to an integrated surface-water/groundwater modelling synthetic example. They highlight

the complex and sometimes counterintuitive predictive outcomes of calibrating a simplified model, exploring different regularization schemes and objective function formulations. White et al. (2014) acknowledge that methodologies such as theirs as well as PMA are limited to identifying/quantifying calibration induced bias attributable to the discrepancy between the complex and simplified models employed in the analysis. The effect of the inevitable discrepancy between the complex model and reality itself cannot be quantified. For this reason, White et al. (2014) highlight the importance of further research in which synthetic studies are employed to characterize the predictive outcomes of calibrating models subject to representative simplifications/defects. This knowledge is also critical in the inevitable situations in which practical limitations preclude the employment of bias identification methodologies and thus necessitate standalone use of a relatively simplified model.

Whilst allowing efficient representation of the various contributions to potential predictive error, the White et al. (2014) linear framework in which model simplification is formulated as "included" and "omitted" parameters does not facilitate straightforward representation of all discrepancies between a simplified and a complex model. Where it is computationally tractable, PMA provides a fully nonlinear alternative to the White et al. (2014) methodology that may be employed for research purposes (PMA is utilized in this manner in Chapter 3 and Chapter 4 of the present thesis).

The PMA methodology has not previously been tested for consistency with theoretical expectation. In order to address this, the present study comprises a proof-of-concept analysis of the efficacy of PMA in performing its intended key functions, these being to (1) identify (and subsequently allow reduction of) calibration-induced predictive bias, and (2) quantify the associated post-calibration predictive uncertainty.

The present proof-of-concept study is approached by applying PMA to an idealised synthetic example wherein the model subject to calibration is free from structural simplification. That is, it has the same geometry and boundary conditions as the "reality model", as well as the same pilot-point-based parameterization mechanism (including the kriging process used to interpolate parameter values from pilot points to model cells). In other words, no sources of potential predictive error exist beyond the reaches of traditional uncertainty quantification methodologies. That is, potential predictive error arises solely through the calibration-induced parameter field

simplification that is necessary to attain uniqueness of the inverse problem as discussed above. Application of PMA in this context facilitates comparison of its results with previously published results obtained through the aforementioned traditional post-calibration uncertainty quantification methods. This provides a means of verifying the capacity of PMA to quantify post-calibration potential predictive error.

Application of PMA to an idealised non-structurally simplified model also supports a controlled examination of the ability of the methodology to identify calibration-induced predictive bias. As discussed above, this is an outcome of parameter compensation. Parameter compensation is a widely acknowledged model calibration phenomenon (e.g., Konikow and Bredehoeft, 1992; Moore and Doherty, 2005; Beven, 2006; Clark and Vrugt, 2006; Fienen et al., 2009; Doherty and Welter, 2010; Langevin and Zygnerski, 2013; Spaaks and Bouten, 2013; White et al., 2014; Xu and Valocchi, 2015). There exist other sources of compensatory parameter and thus potential calibration-induced bias that are not attributable to model structural simplifications/defects. "Overfitting" with respect to measurement noise is one such commonly cited cause of parameter error/compensation (e.g., Yeh and Yoon, 1981; Fienen et al., 2009; James et al., 2009; Doherty and Hunt, 2010). This occurs when the seeking of a close fit between model outputs and field observations leads to model parameters playing erroneous roles in order to reproduce nuances in observation data that represent measurement noise rather than physical system behaviour.

Additionally, calibration-induced parameter compensation is inevitable in the presence of "suboptimal" regularization. An optimal regularization scheme includes constraints on the correlation structure of estimated parameters such that expert knowledge pertaining to true hydraulic property variability is respected (e.g., Maurer et al., 1998; Alcolea et al., 2006). Failure to endow the calibration process with this information permits estimated parameter fields that do not respect geologically plausible spatial correlation and which are thus necessarily playing compensatory roles. These regularization concepts are further discussed below and in Chapter 3 of the present thesis. The reader is also referred to, for example, Tikhonov and Arsenin (1977); Moore and Doherty (2005; 2006); Fienen et al. (2009).

Through a synthetic example involving calibration of an idealised structurally non-simplified model, calibration-induced parameter compensation caused by overfitting and suboptimal regularization can be isolated from the influence of model structural

defects. Overfitting and suboptimal regularization are controllable sources of calibration-induced parameter compensation that provide a foundation for systematic examination of the ability of PMA to detect the resultant predictive bias.

The current proof-of-concept study seeks to verify several fundamental properties of PMA (based upon the concepts and theory presented by Doherty and Christensen (2011) in proposing the methodology) that are central to its validity of its use in practice and the interpretation of its results for research purposes. The present study aims to:

1. Confirm that PMA results indicate unbiased post-calibration model predictions following theoretically optimal calibration of a model that is structurally defect-free.

2. Confirm the ability of PMA to quantify post-calibration predictive error variance.

3. Test whether the occurrence of calibration-induced predictive bias is identified by PMA where expected (i.e., in the presence of calibration-induced parameter compensation).

4. Demonstrate the capacity of PMA to reduce calibration-induced bias and thus post-calibration predictive error variance.

This chapter is organized as follows: Section 2.2 provides a basic summary of the key theory and concepts that are relevant to the present study, including regularized inversion, predictive error variance quantification, the concept of optimal calibration, calibration-induced predictive bias, and finally the specifics of the PMA methodology. Section 2.3 details the synthetic test case. Section 2.4 provides the specific methodological steps undertaken to address the proof-of-concept aims. Results are presented in Section 2.5 and discussed progressively therein, with Section 2.6 providing a synopsis of the proof of concept results along with additional concluding remarks.

2.2 Theory and concepts

The theory summarized in subsection 2.2.1 through subsection 2.2.4 is based on well-established mathematical inversion concepts presented by, for example, Menke

(1989), Aster et al. (2005), Moore and Doherty (2005, 2006). For the sake of tractability, a linear relationship between model parameters and model outputs is assumed to apply in the following theory. This allows a model to be represented by a matrix, with model parameters representing system hydraulic properties and model outputs corresponding to observations being represented by vectors. Moreover, the values of parameter and model output vectors represent perturbations from their pre-calibration expert knowledge-based values.

2.2.1 History matching

Let the vector **k** represent the parameters employed by the model to represent system hydraulic parameters. The Jacobian matrix **X** (containing sensitivities of model outputs with respect to model parameters) represents the action of the model on **k** to produce model outputs. Observations of system state comprising the available calibration dataset are contained in **h** such that:

$$\mathbf{h} = \mathbf{Xk} + \boldsymbol{\varepsilon} \tag{2.1}$$

where the vector **ε** encapsulates measurement noise. Data assimilation or "history matching" involves adjustment of model parameters **k** in order to reduce model-to-measurement misfit, this being represented by the "measurement objective function" Φ_m defined as:

$$\Phi_m = (\mathbf{Xk} - \mathbf{h})^t \mathbf{Q}_h (\mathbf{Xk} - \mathbf{h}) \tag{2.2}$$

Here, $\mathbf{Q}_h$ is the "observation weight matrix" containing (the squares of) observation weights *q*. This matrix is ideally specified as proportional to the inverse of the covariance matrix of measurement noise C(**ε**) (e.g., James et al., 2009).

2.2.2 The null space

Unique estimation of all parameters in a complex environmental model is precluded by inevitable information deficits in available observation data (e.g., Welter et al., 2015). The concept of the null space is central to parameter nonuniqueness. By definition, a non-zero parameter set $\mathbf{k}_n$ belongs to the null space of **X** if:

$$\mathbf{0} = \mathbf{Xk}_n \tag{2.3}$$

Momentarily ignoring the presence of measurement noise **ε**, consider an estimated parameter set $\underline{\mathbf{k}}$ that fits the calibration dataset perfectly. That is:

$$\mathbf{h} = \mathbf{X}\underline{\mathbf{k}} \tag{2.4}$$

From equation (2.3) and equation (2.4) we can write:

$$\mathbf{X}(\underline{\mathbf{k}} + \mathbf{k}_n) = \mathbf{X}\underline{\mathbf{k}} = \mathbf{h} \tag{2.5}$$

thus demonstrating the nonuniqueness of $\underline{\mathbf{k}}$ due to existence of the null space.

A calibration process ideally excludes null-space parameter components from adjustment. This may be achieved through appropriate regularization as discussed below. Any perturbation of null-space parameter components from their pre-calibration expected values (based upon expert knowledge) is unsupported by the calibration dataset **h** and thus introduces asymmetry in the potential parameter error with respect to the maximum likelihood parameter set. This asymmetry necessarily induces an increased potential for error in estimated parameters and thus in predictions that are sensitive to these parameters.

2.2.3 Regularization

Calibration in most environmental modelling contexts constitutes an ill-posed inverse problem (Brunner et al., 2012), thus some form of regularization is necessary to attain uniqueness (e.g., Hunt et al., 2007). Following Moore and Doherty (2005), two alternative regularization mechanisms are employed in the present study. These are Tikhonov regularization (Tikhonov 1963a, 1963b, Tikhonov and Arsenin, 1977) and truncated singular value decomposition (SVD) (e.g., Aster et al., 2005).

2.2.3.1 Tikhonov regularization

The Tikhonov regularization theory presented herein pertains to the manner in which it is implemented by PEST (Doherty, 2016a). An ill-posed parameter estimation problem is made well-posed by supplementing **h** with a set of "regularization observations" **r**. These are expert knowledge-based preferred parameter values (or relationships) that will prevail unless information contained within the calibration dataset **h** dictates otherwise during the calibration process. A regularization objective function Φ_r is defined as:

$$\Phi_r = (\mathbf{Wk} - \mathbf{r})^t \mathbf{Q}_r (\mathbf{Wk} - \mathbf{r}) \quad (2.6)$$

where **r** is a vector containing the abovementioned "regularization observations", and $\mathbf{Q}_r$ is the regularization weight matrix. The matrix **W** defines the relationship between **r** and **k**. Where regularization observations consist of preferred parameter values, and these are defined as pre-calibration expected parameter values (as in the present study), **W** = **I** and **r** = **0**, thus equation (2.6) becomes:

$$\Phi_r = \mathbf{k}^t \mathbf{Q}_r \mathbf{k} \quad (2.7)$$

Tikhonov-regularized inversion is thus formulated as a constrained minimization problem, whereby Φ_r is minimized subject to the constraint that Φ_m of equation (2.2) is not greater than Φ_m^l, this being the target measurement objective function (referred to in PEST as the "limiting measurement objective function"). This is a user-specified threshold that defines the level of model-to-measurement misfit tolerable to be deemed as "adequate calibration". In an idealised case in which a model is free from structural defects, this is theoretically defined by a level of model-to-measurement misfit that is commensurate with measurement noise **ε**. Where the observation weight matrix is specified as the inverse of the covariance matrix of measurement noise C(**ε**), this level of fit is represented by a measurement objective function $\Phi_m = N$, where N is the number of observations comprising the calibration dataset (see Appendix A for derivation).

The Tikhonov-regularized inversion process thus involves minimization of the total objective function Φ, defined as:

$$\Phi = \Phi_m + \mu \Phi_r \quad (2.8a)$$

$$\Phi_m \leq \Phi_m^l \quad (2.8b)$$

where μ is the "regularization weight factor". This is equivalent to a Lagrange multiplier (as shown by de Groot-Hedlin and Constable, 1990) in the solution of the constrained minimization problem in which Φ_r is minimized subject to the constraint that $\Phi_m \leq \Phi_m^l$. The value of μ is determined iteratively by PEST as part of the optimization problem and thus reflects the relative weighting placed upon regularization constraints with respect to observations. Thus a decreasing value of μ

reflects an increase in the level of fit attained between observations and corresponding model outputs.

Parameters estimated through Tikhonov-regularized inversion (where equation (2.7) holds as in the present study) are given by:

$$\underline{\mathbf{k}} = (\mathbf{X}^t\mathbf{Q}_h\mathbf{X} + \mu\mathbf{Q}_r)^{-1}\mathbf{X}^t\mathbf{Q}_h\mathbf{h} \quad (2.9)$$

2.2.3.2 Truncated SVD

In the case of truncated SVD, well-posedness of the parameter estimation problem is achieved via reduction of the number of parameters estimated through the inversion process. The weighted Jacobian matrix is decomposed as follows:

$$\mathbf{X}\mathbf{Q}_h\mathbf{X} = \mathbf{V}\mathbf{S}\mathbf{V}^T \quad (2.10)$$

Here, **V** is a matrix of orthogonal unit vectors that span the parameter space of the model (i.e., eigenvectors of $\mathbf{X}\mathbf{Q}_h\mathbf{X}$). **S** is a diagonal matrix comprising singular values arranged in decreasing order. Based on **S**, partitioning occurs between the so-called "solution space" and the null space, such that:

$$\mathbf{X}\mathbf{Q}_h\mathbf{X} = \mathbf{V}_1\mathbf{S}_1\mathbf{V}_1^T + \mathbf{V}_2\mathbf{S}_2\mathbf{V}_2^T \quad (2.11)$$

$\mathbf{V}_1$ contains unit vectors that span the solution space of the inverse problem, with $\mathbf{S}_1$ containing the corresponding singular values. $\mathbf{V}_2$ contains the unit vectors that span the null space, these being associated with singular values of magnitude zero or near-zero, which are contained in $\mathbf{S}_2$.

The number of singular values at which truncation occurs to define the partitioning between $\mathbf{S}_1$ and $\mathbf{S}_2$ (and hence $\mathbf{V}_1$ and $\mathbf{V}_2$) is subjective and may be varied. The greater the number of (non-zero) singular values included in the solution space during the inversion process, the closer the level of fit sought between model outputs and observations.

Parameters estimated through truncated SVD are given by:

$$\underline{\mathbf{k}} = \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{V}^t{}_1\mathbf{X}^t\mathbf{Q}_h\mathbf{h} \quad (2.12)$$

2.2.4 Quantification of predictive error variance

Substitution of equation (2.1) into equation (2.9), and expansion of the terms of equation (2.12) and substitution of equations (2.1) and (2.10), yields a general form for parameters estimated through either Tikhonov regularized-inversion and truncated SVD:

$$\underline{\mathbf{k}} = \mathbf{R}\mathbf{k} + \mathbf{G}\boldsymbol{\varepsilon} \tag{2.13}$$

where, for Tikhonov regularization:

$$\mathbf{R} = (\mathbf{X}^t\mathbf{Q}_h\mathbf{X} + \mu\mathbf{Q}_r)^{-1}\mathbf{X}^t\mathbf{Q}_h\mathbf{X} \tag{2.14}$$

$$\mathbf{G} = (\mathbf{X}^t\mathbf{Q}_h\mathbf{X} + \mu\mathbf{Q}_r)^{-1}\mathbf{X}^t\mathbf{Q}_h \tag{2.15}$$

and for truncated SVD:

$$\mathbf{R} = \mathbf{V}_1\mathbf{V}^t_1 \tag{2.16}$$

$$\mathbf{G} = \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{V}^t{}_1\mathbf{X}^t\mathbf{Q}_h \tag{2.17}$$

Error in estimated parameters is given by:

$$\mathbf{k} - \underline{\mathbf{k}} = (\mathbf{I} - \mathbf{R})\mathbf{k} - \mathbf{G}\boldsymbol{\varepsilon} \tag{2.18}$$

where **I** is the identity matrix. The covariance matrix of post-calibration parameter error is thus given by:

$$\mathrm{C}(\mathbf{k} - \underline{\mathbf{k}}) = (\mathbf{I} - \mathbf{R})\mathrm{C}(\mathbf{k})(\mathbf{I} - \mathbf{R}) + \mathbf{G}\mathrm{C}(\boldsymbol{\varepsilon})\mathbf{G} \tag{2.19}$$

C(**k**) is the covariance matrix pertaining to the prior probability distribution of true parameters **k**.

The true value of a given (scalar) model prediction *s* is given by:

$$s = \mathbf{y}^t\mathbf{k} \tag{2.20}$$

where **y** is a vector containing sensitivities of *s* to true parameters **k**. The pre-calibration (prior) variance of prediction *s* is given by:

$$\sigma^2_s = \mathbf{y}^t\mathrm{C}(\mathbf{k})\mathbf{y} \tag{2.21}$$

The prediction made by the calibrated model is given by:

$$\underline{s} = \mathbf{y}^t\underline{\mathbf{k}} \tag{2.22}$$

Post-calibration predictive error is thus given by:

$$s - \underline{s} = \mathbf{y}^t(\mathbf{k} - \underline{\mathbf{k}}) \tag{2.23}$$

Post-calibration predictive error variance is calculable as:

$$\sigma^2_{s-\underline{s}} = \mathbf{y}^t\mathrm{C}(\mathbf{k} - \underline{\mathbf{k}})\mathbf{y} \tag{2.24}$$

Substituting equation (2.19) we therefore have:

$$\sigma^2_{s-\underline{s}} = \mathbf{y}^t(\mathbf{I} - \mathbf{R})\mathrm{C}(\mathbf{k})(\mathbf{I} - \mathbf{R})^t\mathbf{y} + \mathbf{y}^t\mathbf{G}\mathrm{C}(\boldsymbol{\varepsilon})\mathbf{G}^t\mathbf{y} \tag{2.25}$$

The first term of equation (2.25) represents the null-space contribution to error variance, this arising from the components of pre-calibration parameter uncertainty that are not informed by available observation data. The second term of equation (2.25) is the contribution to post-calibration predictive error variance owing to the presence of measurement (and structural) noise.

2.2.5 Optimal calibration

Post-calibration predictive error variance has a theoretical lower limit set by the information content inherent in available data and expert knowledge pertaining to the system under study. Subject to optimal calibration, a structurally perfect model may theoretically achieve minimum error variance status for estimated parameters and thus for model predictions that are sensitive to these parameters.

Theoretically optimal calibration as defined herein is achieved in conjunction with optimal regularization. An optimal regularization scheme is constrained to respect expert knowledge (e.g., expected spatial correlation) pertaining to the hydraulic properties of the system being modelled (e.g., Maurer et al., 1998; Alcolea et al., 2006). In the case of Tikhonov regularization, this is achieved through definition of the regularization weight matrix $\mathbf{Q}_r$ as the inverse of the true parameter covariance matrix C(**k**). For truncated SVD, this may be effected through appropriate pre-calibration parameter transformation such as the Karhunen-Loève transform. The reader is referred to Chapter 3 of the present thesis for an extended discussion of optimal pre-calibration parameter transformation.

Secondly, theoretically optimal calibration comprises an observation weight matrix $\mathbf{Q}_h$ that is proportional to the inverse of the covariance matrix of measurement noise C($\boldsymbol{\varepsilon}$). This is accompanied by appropriately setting Φ_m^l of equation (2.8b) such as to incite a level of post-calibration model-to-measurement misfit that is commensurate with measurement noise as described above.

2.2.6 Calibration-induced predictive bias

Equation (2.25) represents what we herein refer to as traditional post-calibration predictive error variance quantification. Minimization of predictive error variance is considered as the point of optimal trade-off between the null-space and solution-space contributions (i.e., the first and second terms of equation (2.25), respectively). It does not account for the additional contribution to post-calibration predictive error arising through calibration of a structurally defective model. This additional component of error is thus not visible through use of equation (2.25). Extending the linear subspace theory presented by Doherty and Christensen (2011), White et al. (2014) reformulate equation (2.25) to include a third term that accounts for this contribution. It is this additional component of post-calibration predictive error that PMA is also designed to detect (and subsequently allow reduction of).

As explained by Doherty and Christensen (2011) and explored mathematically in Chapter 3 of the present thesis, bias in estimated parameters (and thus bias in predictions that are sensitive to these parameters) occurs when the calibration process results in nominal adjustment of the null-space parameter components of the "reality model". They refer to this process as "null-space entrainment". This is an inevitable consequence of model parameters compensating for structural defects.

As explained above, null-space parameter components are by definition not informed by available observation data, thus their deviation from expert knowledge-based expected values is unsupported by the calibration dataset. Predictions that are sensitive to the affected parameters thus possess an unsupported potential for wrongness, this being defined as bias by Doherty and Christensen (2011).

Parameter compensation resulting in adjustment of null-space parameter components may also arise through other sources. One such source is overfitting with respect to measurement noise. The presence of measurement noise effectively expands the null-space, thus commanding a lesser fit between model outputs and corresponding field

observations in order to minimise error variance. Attaining a closer fit between model outputs and observations than the associated level of measurement noise results in effective null-space parameter components being included in the parameter estimation process. This is commonly recognised as parameter compensation, as it will often be expressed as model parameters attaining unrealistic values in order for the model to reproduce nuances of the calibration dataset that are attributable to measurement noise rather than physical processes (e.g., Fienen et al., 2009; Langevin and Zygnerski, 2013). In the same manner as null-space entrainment caused by calibration of a structurally defective model, predictions that are sensitive to the adjusted null-space parameter components will thus contain an unsupported propensity for predictive error, that is, calibration-induced predictive bias.

Suboptimal regularization is another source of unsupported adjustment of null-space parameters during the calibration process. As discussed above, some form of regularization is necessary in order to attain uniqueness of the inverse problem. Where the employed regularization scheme does not respect available expert knowledge (e.g., pertaining to innate hydraulic property variability and correlation), parameter compensation and thus adjustment of parameter components during the calibration process that properly belong to the "true" null space. A more comprehensive presentation of the theory and discussion of optimal regularization/parameter transformation is provided in Chapter 3 of the present thesis. Furthermore, Appendix A provides a simple mathematical representation of null-space entrainment due to suboptimal pre-calibration parameter transformation (equivalent to suboptimal regularization). These examples pertain to optimal pre-calibration parameter transformation necessary to attain optimality of regularization as effected via truncated SVD. However, Optimality of a constrained minimization (Tikhonov) regularization process is equivalently achieved through provision of Tikhonov constraints with a stochastic weighting scheme (Maurer et al., 1998).

The outcomes of parameter compensation (and accompanying null-space entrainment) for model predictive performance have been demonstrated to be highly prediction-specific (e.g., Doherty and Welter, 2010; Doherty and Christensen, 2011; White et al., 2014). In general, predictions that are very similar in type to the calibration dataset are predominantly solution-space dependent. They will therefore tend to benefit unconditionally from an improved fit between model outputs and field observations. Simultaneously, however, predictions of a different nature are likely to be null-space

dependent and thus predictive performance will be degraded by parameter compensation.

2.2.7 The paired model analysis methodology

The PMA methodology is summarized as follows:

1. A large number *n* of stochastic realizations of a complex "reality" model are generated based on expert knowledge alone. For each realization, model outputs equivalent to the available calibration dataset, as well as values for the prediction(s) of interest, are obtained through forward simulations.

2. A simplified model is developed and calibrated against the outputs generated by each of the complex model realizations. The prediction of interest is also made by each calibrated simplified model, yielding *n* complex-simple prediction pairs for each prediction of interest.

3. A scatterplot (represented schematically in Figure 2.1) of complex model prediction values (i.e., s) versus calibrated simplified model predictions (i.e., $\underline{s}$) is generated. A regression line through the s-versus-$\underline{s}$ scatterplot may then be used to identify calibration-induced predictive bias. Regression lines are calculated by Doherty and Christensen (2011) and in the present study as:

$$s = a + b\underline{s} \tag{2.26}$$

   where a and b are the regression intercept and slope, respectively. A regression line slope b of less than unity indicates the occurrence of calibration-induced predictive bias induced by nonzero null-space parameter components.

   A measure of scatter about the regression line provides a quantification of (bias-corrected) post-calibration predictive uncertainty (see Figure 2.1b). In the present study 95% prediction intervals are used for this purpose (see, for example, Draper and Smith, 1998, eq. 1.4.12 for details of prediction interval calculations).

4. Finally, the simplified model is calibrated against the available real world dataset and the prediction of interest is made using the calibrated simplified model. The s-versus-$\underline{s}$ scatterplot is subsequently used to correct for bias in the calibrated model prediction and quantify the associated uncertainty.

Figure 2.1a provides a schematic representation of predictive bias identification and correction as effected through a PMA *s*-versus-$\underline{s}$ scatterplot. Figure 2.1b depicts the subsequent quantification of post-calibration predictive uncertainty, including a representation of total post-calibration predictive uncertainty. The latter does not account for bias and is thus symmetrical about the unity line. Uncertainty bounds estimated through a given "traditional" post-calibration uncertainty analysis procedure would also be notionally symmetrical about the unity line in Figure 2.1b. However, the failure to account for calibration-induced bias would likely yield narrower uncertainty margins than the "true" margins represented in Figure 2.1b. Thus the threat of traditional post-calibration uncertainty analysis failing to capture the true prediction value within its estimated uncertainty interval following calibration of a defective model is clear.

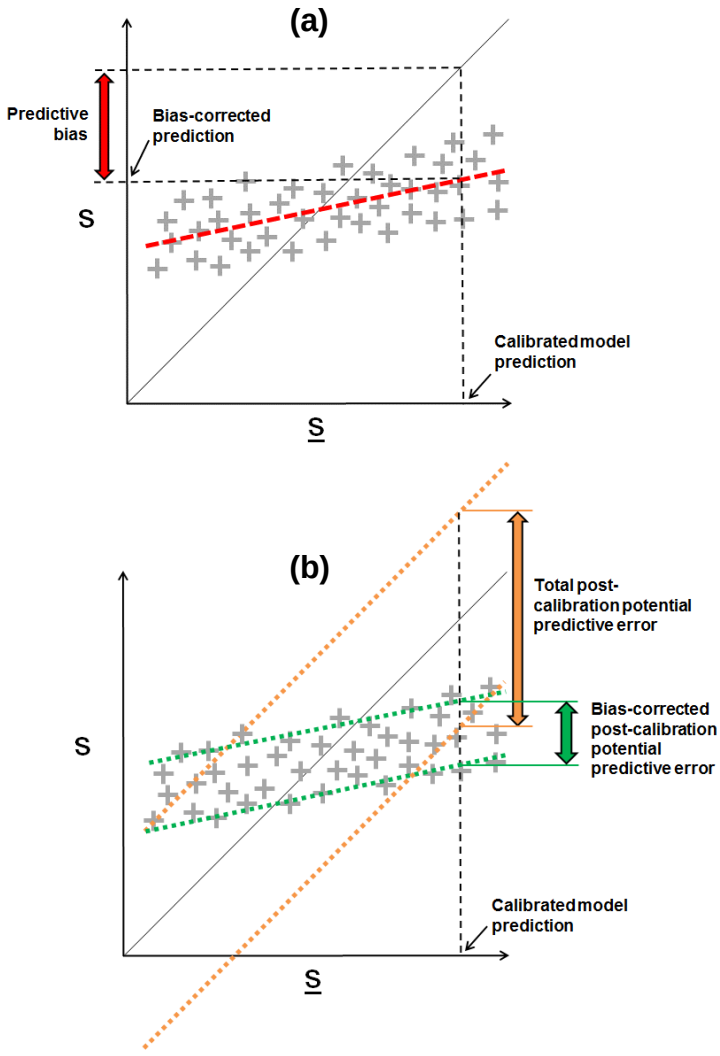

Figure 2.1. Schematic representation of a PMA *s*-versus-$\underline{s}$ scatterplot: (a) predictive bias identification/correction and (b) quantification of bias-corrected post-calibration potential predictive error (including a representation of total post-calibration potential predictive error for comparison).

2.3 Synthetic test case description

We employ a synthetic example that has been analysed extensively in multiple previous studies, including Moore and Doherty (2005, 2006), Tonkin and Doherty (2005) and Moore et al. (2010). As discussed above, this provides a foundation for comparison of the results obtained in the present study through PMA.

Figure 2.2a depicts the model domain – a single-layer confined aquifer of 10 m thickness and dimensions 800 m north-south by 500 m east-west. Water enters the system through the northern boundary as a fixed inflow of 0.1 m^3/d per metre length of boundary, and exits the system through the southern boundary which has a prescribed head of 0 m. No-flow boundaries define the western and eastern edges of the domain. Steady-state groundwater flow is simulated using MODFLOW 2000 (Harbaugh et al., 2000) with a finite-difference grid comprising 4000 10 m × 10 m cells. The movement of a particle released at the location indicated in Figure 2.2 is simulated using the ADV2 package (Anderman and Hill, 2001).

Following all previous studies that have employed the same synthetic test case, the "reality" distribution of hydraulic conductivity (K) is defined by a log exponential variogram, with a mean of 0 $\log_{10}$(m/day), a sill of 0.2 $\log_{10}$(m/day) and a range of 600 m. Figure 2.2b displays one of the stochastic $\log_{10}K$ fields as an example (accompanied by the true particle path and travel time).

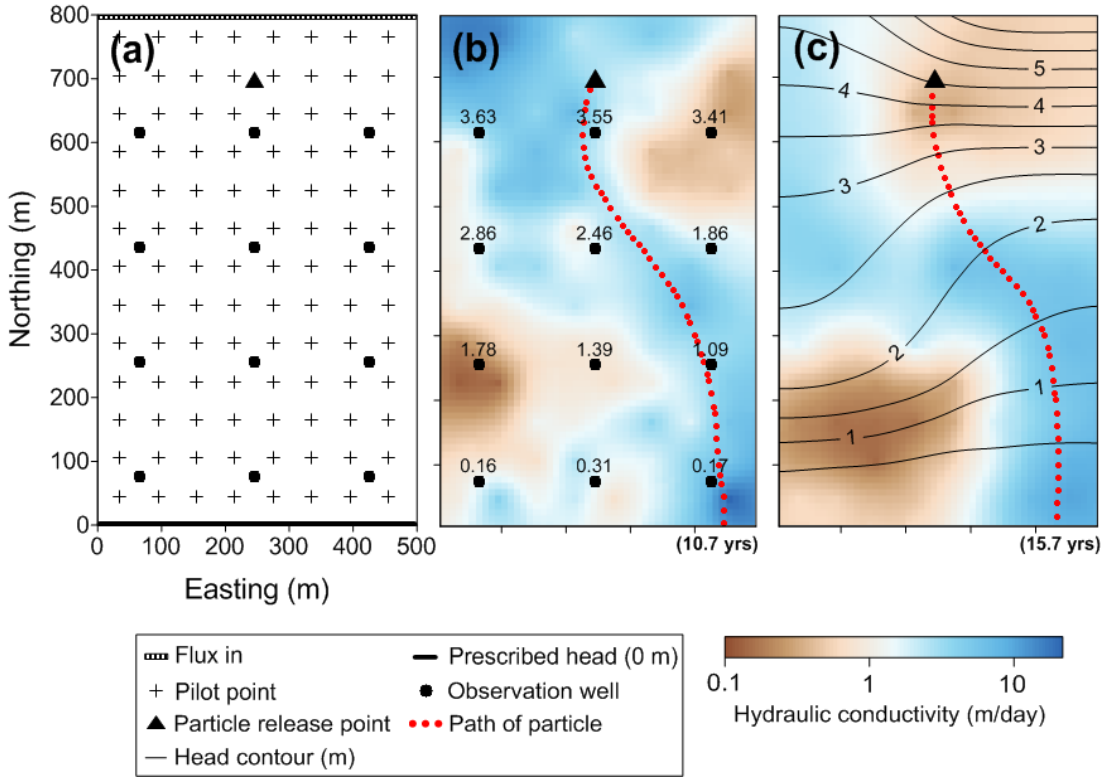

Figure 2.2. (a) Model domain and boundary conditions, including locations of pilot-points and observation wells. (b) Arbitrary stochastic "reality" $\log_{10}K$ field realization, including hydraulic head values (m) at observation wells, and true path and travel time of released particle. (c) Corresponding estimated (through stochastically weighted Tikhonov regularized based calibration for the 0.30 m standard deviation measurement noise case) $\log_{10}K$ field and hydraulic head distribution, including predicted path and travel time of released particle.

A minor difference between the present synthetic test case and previous studies is that stochastic "reality" $\log_{10}K$ fields herein are generated using pilot-points (de Marsily, 1978; de Marsily et al. 1984) rather than through direct cell-by-cell parameterization (i.e., values are generated at the 104 pilot point locations depicted in Figure 2.2 and then interpolated to model grid cells through kriging). This is to ensure complete structural equivalence between "reality" and the model for the purposes of the present study. Moore and Doherty (2005) state that the influence of kriging-induced smoothness was tested and found to be slight, to the extent that their predictive error variance analysis ignores its influence with insignificant consequences. Thus, this minor alteration in the synthetic test case does not impact upon comparison of results with those of the previous studies.

Synthetic field observation data is comprised of 12 hydraulic heads from the "reality" model, the locations of which are displayed in Figure 2.2. Independent Gaussian noise with a mean of 0 is added to each set of observations to generate the calibration

datasets. Various Gaussian noise standard deviations (0.01 m, 0.10 m and 0.30 m) are employed for different purposes of the present study – specific details of which are described in the following section).

The same 104 pilot point locations used for "reality" $\log_{10}K$ field generation are employed as adjustable parameters in the calibration process (which are equivalently interpolated to model grid cells via kriging). Tikhonov regularization and truncated SVD schemes are employed separately in present study as described above.

The Tikhonov regularization scheme involves assigning "preferred value" Tikhonov constraints (i.e., the values populating the **r** matrix of equation (2.6)) of 0 $\log_{10}$(m/day) to pilot points, this being equal to the mean of the log exponential variogram used for generation of "reality" fields. The Tikhonov scheme is stochastically weighted, whereby preferred values are accompanied by a regularization weight matrix $\mathbf{Q}_r$ (see equation 2.7) that is calculated as the inverse of the covariance matrix $C^{-1}(\mathbf{k})$ that represents the variogram upon which the "reality" $\log_{10}K$ fields are based (Maurer et al., 1998).

Figure 2.2c displays an example $\log_{10}K$ field (along with the associated hydraulic head distribution and particle fate predictions) estimated through Tikhonov-regularized inversion. This is based on calibration against "observed" steady-state hydraulic heads generated by the "reality" field of Figure 2.2b (after the addition of measurement noise of standard deviation 0.30 m).

2.4 Approach

In the context of the theoretical concepts and synthetic case study details described above, this section describes sequentially the numerical experiment and analysis components specific to addressing each of the key proof-of-concept aims outlined in the introductory text.

It should be emphasised at this point that a general comparison between the Tikhonov and truncated SVD regularization mechanisms is not an intention of this study, nor do the results facilitate such a comparison. The respective regularization schemes are formulated to emulate the numerical experiments of Moore and Doherty (2005) for the purpose of comparing PMA results with the results of these authors, this being the chief intention of the present proof of concept. The manner in which the regularization

schemes are applied (described in detail below) in the selected examples from Moore and Doherty (2005) is such the Tikhonov regularization case constitutes an example of optimal regularization, whilst the truncated SVD case represents suboptimal regularization. Parameter estimation and predictive performance differences observed herein between the two regularization mechanisms arise through this optimality/suboptimality of their application and are not inherent to the regularization mechanisms themselves. Optimality of truncated SVD as a regularization device may equivalently be achieved through appropriate pre-calibration parameter transformation as described above. As a result it would be expected to provide similar results to the present (optimal) Tikhonov scheme. Likewise, Tikhonov regularization employed in the absence of the stochastic regularization weighting scheme would represent suboptimal regularization and would be expected to produce similar results to the present untransformed truncated SVD results.

2.4.1 Optimal calibration

The first aim of the proof of concept is to confirm the theoretical notion, discussed by Doherty and Christensen (2011), that PMA yields an *s*-versus-$\underline{s}$ scatterplot with a regression line slope of unity (this indicating unbiased model predictions) for optimal calibration of a structurally non-defective model.

For this purpose an optimally weighted Tikhonov regularization scheme is employed. A very small measurement noise standard deviation of 0.01 m is applied initially, the intention being to eliminate any potential complicating effects introduced by significant measurement noise. Subsequent phases of the study involve the equivalent analysis in the presence of a greater measurement noise magnitudes (i.e., 0.10 m and 0.30 m). This allows for additional examination of this theoretical notion, which asserts that measurement noise should increase scatter about an *s*-versus-$\underline{s}$ regression line which maintains a slope of unity.

2.4.2 Quantification of predictive error variance

The second aim of the proof of concept is to confirm that the vertical scatter about the *s*-versus-$\underline{s}$ best-fit line adequately represents post-calibration predictive uncertainty. This is approached through comparison of PMA results with the uncertainty quantification results of established methods presented in existing literature.

Both Moore and Doherty (2005) and Moore et al. (2010) employ the synthetic example adopted herein to demonstrate predictive error variance quantification methodologies and concepts. Moore and Doherty (2005) apply linear subspace analysis techniques to quantify error variance in prediction of the exit location of the released particle (see Figure 2.2). They employ both Tikhonov regularization and truncated SVD schemes in separate calibration and predictive error variance analysis processes.

Moore et al. (2010) demonstrate a global optimization methodology utilising the Pareto front concept to quantify the uncertainty in prediction of the particle travel time (with regularization enforced through the same Tikhonov scheme as used by Moore and Doherty (2005)).

These previous studies differ in terms of the magnitude of artificial (independent Gaussian) measurement noise introduced to the calibration datasets – Moore and Doherty (2005) adopt a standard deviation of 0.30 m whilst Moore et al. (2010) use a standard deviation of 0.10 m. In order to facilitate comparison with each of the previous studies, three separate PMA processes are therefore undertaken, summarized in Table 2.1.

Table 2.1. Summary of PMA processes conducted for comparison of results with previous studies. σ_ε is measurement noise standard deviation.

PMA process	Prediction	σ_ε (m)	Regularization scheme
a)	Exit point	0.30	Tikhonov
b)	Exit point	0.30	Truncated SVD
c)	Log_{10} time	0.10	Tikhonov

2.4.3 Predictive error variance minimization

A number of additional PMA processes are undertaken for the purpose of extending the comparison of PMA results to include the characteristics of the error variance functions, such as the point at which their minima occur. This component of the analysis is equivalent to that conducted by Moore and Doherty (2005), thus measurement noise of standard deviation 0.30 m is used throughout.

PMA processes '(a)' and '(b)' detailed in Table 2.1 are each repeated an additional 11 times, spanning a broad range of model-to-measurement misfit (represented by the measurement objective function Φ_m). This is achieved for the Tikhonov regularization approach by varying the observation weights q comprising the diagonal of the observation weight matrix $\mathbf{Q}_h$ of equation (2.2). For calibration based on truncated

SVD, the degree of fit is controlled through the number of singular values included in the solution space (i.e., the dimensionality of the $\mathbf{V}_1$ matrix of equation (2.11)).

Whilst the analysis presented by Moore and Doherty (2005) is limited to prediction of particle exit location, the current analysis is extended to also include the travel time prediction.

2.4.4 Predictive bias identification

The two selected regularization approaches and abovementioned repetition of PMA to achieve various measurement objective functions also provides a basis for systematic examination of the ability of PMA to identify calibration-induced predictive bias.

As discussed above, overfitting and suboptimal regularization are sources of parameter compensation and thus bias in some null-space-dependent predictions. Advection-type predictions, as considered in the present study, are known to be typically sensitive to parameterization detail that cannot be inferred through calibration based on relatively sparse hydraulic head data alone (e.g. Moore and Doherty, 2005; White et al. 2014). That is, they are null-space dependent and therefore provide a suitable indicator for assessment of the ability of PMA to identify the expected calibration-induced bias arising through overfitting and suboptimal regularization, respectively.

The examination of the efficacy of PMA in identifying overfitting-induced bias is based on the specified independent Gaussian measurement noise of known standard deviation (the squares of which comprise the diagonal matrix $C(\boldsymbol{\varepsilon})$). In the case of the optimal regularization example (i.e., the Tikhonov scheme) the $\mathbf{Q}_h$ matrix of equation (2.2) is chosen as $C^{-1}(\boldsymbol{\varepsilon})$. Thus a value of 12.0 for Φ_m of equation (2.2) defines the level of model-to-measurement misfit that is commensurate with measurement noise as described above. A Φ_m value lower than 12.0 therefore represents overfitting. The further Φ_m is reduced below 12.0, the greater the induced parameter compensation in the form of inclusion of effective null-space parameter components in the parameter estimation process. It is therefore expected that PMA *s*-versus-<u>s</u> scatterplots will identify the resultant increasing degree of bias in the advective transport predictions through an increasing deviation below unity of the regression line slope *b*.

We also examine the ability of PMA to identify predictive bias arising through suboptimal regularization. As described above, optimality of the truncated SVD regularization would require pre-calibration parameter transformation based on the

covariance matrix C(**k**) that defines the innate variability of "reality" $\log_{10}K$ fields. This is not undertaken in the present truncated SVD example. For this reason, compensatory parameter behaviour is a guaranteed outcome of the parameter estimation process, through allowance of spatial variability that is not supported by the (synthetic) expert geological knowledge. The *s*-versus-$\underline{s}$ scatterplots for the truncated SVD case are thus expected to indicate more prevalent calibration-induced predictive bias than the (optimal) Tikhonov case. That is, regression line slopes that deviate below unity are expected to occur pervasively, irrespective of the attained level of fit between model outputs and observations.

2.4.5 Bias reduction and uncertainty quantification

Subsequent to identification of calibration-induced predictive bias, the proposed utility of PMA is its allowance of post-calibration bias reduction. As demonstrated by Figure 2.1a, this is achieved through *s*-versus-$\underline{s}$ scatterplots, wherein the regression line indicates the required adjustment from the calibrated model prediction to the true minimum error variance prediction. Prediction intervals based on the *s*-versus-$\underline{s}$ scatterplot regression line define the bias-corrected post-calibration prediction uncertainty. This uncertainty is theoretically be smaller than total post-calibration uncertainty due to removal of the influence of systematic error.

In order to achieve the fourth aim of the proof of concept as outlined in the introductory text, bias-corrected post-calibration predictive error variance is quantified through *s*-versus-$\underline{s}$ scatterplots (based on the standard deviation implied by the calculated 95% prediction interval). Comparison of bias-corrected predictive error variance functions with the total post-calibration predictive error variance functions enables assessment of the capacity of PMA to reduce the impact of calibration-induced predictive bias.

2.5 Results and discussion

Throughout the present study, any calibration realizations for which the target measurement objective function was not achieved (whether as a result of non-convergence of MODFLOW or due to calibration process termination criteria) are excluded from the analysis. It should be noted, however, that due to the large number of "successful" realizations relative to failed realizations in all cases, the impact of this exclusion was tested and found to be insignificant to all results.

2.5.1 Optimal calibration

Figure 2.3 displays *s*-versus-s scatterplots for the particle exit point and particle travel time predictions based on optimal calibration in the presence of a very little (σ_ε = 0.01 m) measurement noise (i.e., stochastically weighted Tikhonov regularization and $\Phi_m = \Phi_m^l = 12.0$). The associated regression statistics are provided in Table 2.2. These results are based on 995 realizations, with the target measurement objective function of 12.0 not achieved for five realizations.

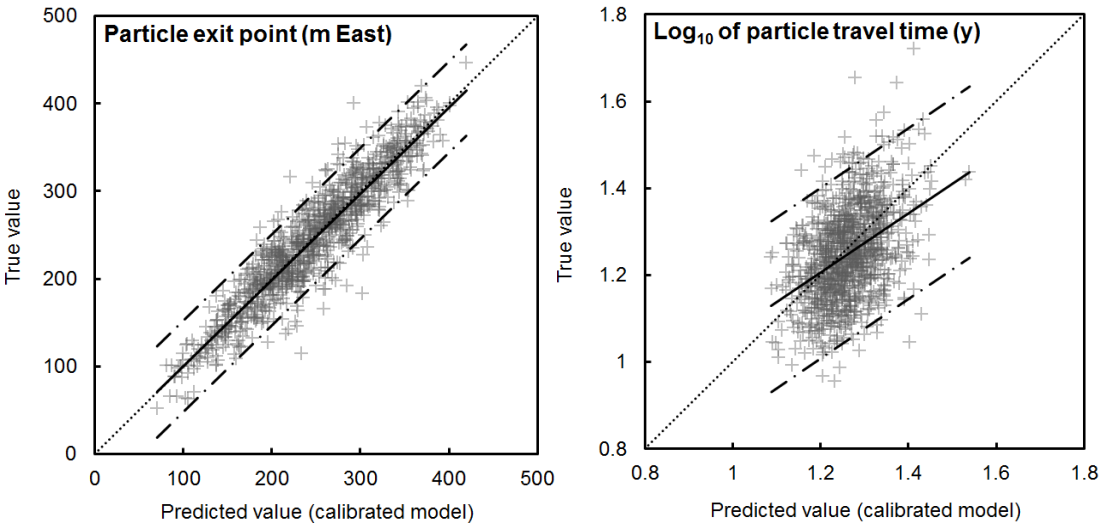

Figure 2.3. *s*-versus-s scatterplots for predictions of particle exit point and $\log_{10}$ of particle travel time based on optimal (stochastically weighted Tikhonov) regularization in the presence of independent Gaussian measurement noise of standard deviation 0.01 m. The solid line is the scatterplot regression line, the dashed lines bound the 95% prediction interval and the dotted line represents the 1:1 line.

Table 2.2. Regression coefficients and statistics pertaining to the *s*-versus-s scatterplots of Figure 2.2. *a* and *b* are the regression coefficients of equation (2.26), and r^2 is the coefficient of determination.

Prediction	*a*	*b*	r^2
Exit point	1.32	0.99	0.86
Log_{10} time	0.39	0.68	0.18

The *s*-versus-s regression line slope for the particle exit location prediction is approximately unity. That is PMA indicates unbiased prediction of particle exit location, this being consistent with theoretical expectation for a structurally non-defective model (Doherty and Christensen, 2011).

The equivalent *s*-versus-s scatterplot for the particle travel time prediction returns a regression line slope of less than unity, indicating the presence of calibration-induced bias, which is unexpected given the theoretical optimality of the calibration process.

However, this is considered to be an outcome of the poor ability of the hydraulic head-based calibration data to constrain this type of prediction, as pointed out by Moore and Doherty (2005). This is supported by the corresponding prior and posterior error distributions displayed in Figure 2.4, which demonstrate that the post-calibration potential for error in the prediction of particle travel time is barely reduced relative to the pre-calibration uncertainty (contrasting the marked reduction in uncertainty achieved for the particle exit location prediction). This is attributable to the fact that that small-scale *K* heterogeneity is a dominant control on contaminant transport rates (e.g., Eggleston and Rojstaczer, 1998; Zheng et al., 2011). Estimated fields are inevitably "blurred" and lack this detail; this being necessary for the attainment of a unique solution to the inverse problem in the presence of sparse observation data (e.g., McLaughlin and Townley, 1996; Moore and Doherty, 2006; Ulugergerli, 2011). The resultant extremely poor correlation in the particle travel time *s*-versus-$\underline{s}$ scatterplot (exemplified by the very low r^2 value in Table 2.2) limits the meaningfulness of statistics such as regression line slope.

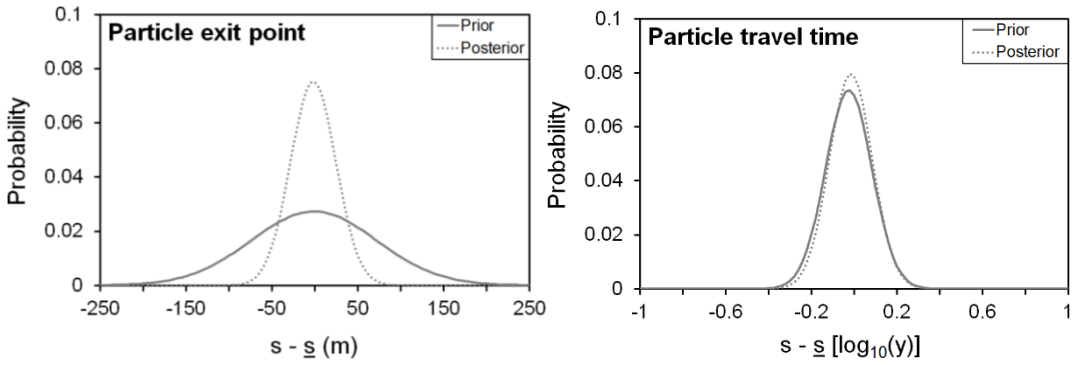

Figure 2.4. Prior and posterior predictive error ($s - \underline{s}$) distributions based on optimal (stochastically weighted Tikhonov) regularization in the presence of independent Gaussian measurement noise of standard deviation 0.01 m (corresponding to the *s*-versus-$\underline{s}$ plots of Figure 2.3). Prior distributions were calculated through unconstrained Monte Carlo analysis.

2.5.2 Quantification of predictive error variance

Figure 2.5 displays *s*-versus-$\underline{s}$ scatterplots for each of the three PMA processes summarized in Table 2.1. Table 2.3 provides the associated regression statistics. In each case the target measurement objective function of 12.0 was achieved for all 1000 realizations.

Moore and Doherty (2005) and Moore et al. (2010) present quantified post-calibration uncertainty in a variety of forms. For the purpose of direct comparison with the present

PMA results, literature-based results were converted to equivalent 95% prediction intervals and are overlayed in Figure 2.5. The horizontal position of the vertical red line corresponds to the value of the prediction made by the calibrated model in the relevant previous study, whilst the vertical span of the line represents the equivalent 95% prediction interval.

Figure 2.5 demonstrates reasonable agreement between PMA results and the equivalent results from established predictive error variance analysis methods presented in existing literature. Minor discrepancies between PMA prediction interval widths and the Moore and Doherty (2005) results (for cases '(a)' and '(b)' in Figure 2.5) are attributable to the assumption of linearity underpinning the Moore and Doherty (2005) analysis in the presence of model nonlinearity, due to which discrepancies are expected (e.g., James and Oldenburg, 1997; Christensen and Doherty, 2008; Brunner et al., 2012).

Confirmation of the influence of the linearity assumption in the present case is achieved through comparison of prior error variance values. Particle exit point prior error variance calculated through linear analysis (equation (2.21)) (results not presented for the sake of brevity) overestimates the true prior uncertainty attained through unconstrained Monte Carlo analysis (presented in Figure 2.4). This is likely a consequence of model boundary influence within the relatively small synthetic model domain. Lateral particle movement in more extreme cases is restricted by the eastern and western no-flow boundaries, thus introducing nonlinearity into model sensitivities and resulting in systematic overestimation of the uncertainty quantified through linear error variance analysis. Small discrepancies between linear and nonlinear uncertainty quantification methods are expected in general. A closer comparison is observed for prediction of travel time in Figure 2.5c as the Moore et al. (2010) result upon which comparison is based was quantified using a nonlinear method.

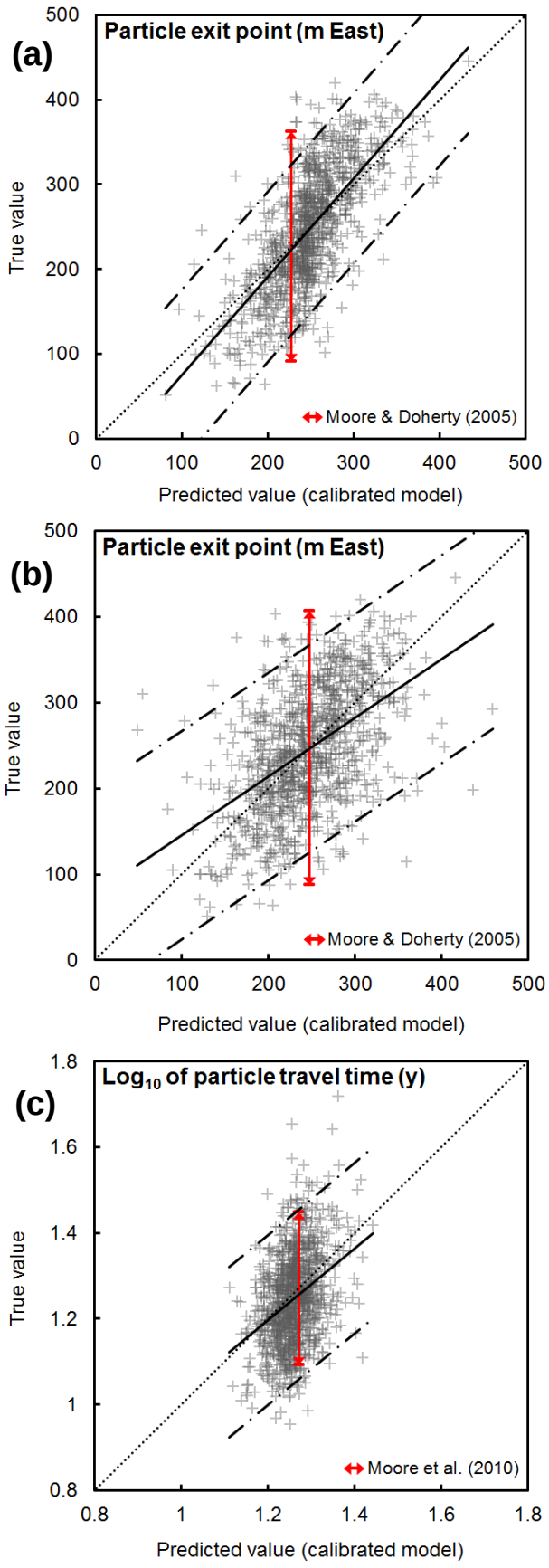

Figure 2.5. *s*-versus-*s* scatterplots for paired model analyses '(a)', '(b)' and '(c)' as per Table 2.1, overlaid by representations of results from previous studies for comparative purposes. The dotted line is the 1:1 line.

Table 2.3. Regression coefficients and statistics pertaining to the *s*-versus-$\underline{s}$ scatterplots of Figure 2.5. *a* and *b* are the regression coefficients of equation (2.26), and r^2 is the coefficient of determination.

PMA process	*a*	*b*	r^2
a)	-40.8	1.16	0.49
b)	76.8	0.68	0.26
c)	0.19	0.84	0.13

Note that the deviation from unity of the *s*-versus-$\underline{s}$ regression lines in Figure 2.5 slightly complicates the above comparison. Figure 2.1b demonstrates schematically that a deviation from unity of the regression line slope causes a difference between total post-calibration potential predictive error and that quantified through PMA *s*-versus-$\underline{s}$ scatterplots (the latter being the bias-corrected post-calibration potential predictive error in accordance with the goal of applying PMA to reduce bias). Due to the slight deviations from unity of the slopes in the Figure 2.5 plots, the comparison with previous results should strictly involve total post-calibration predictive uncertainty (i.e., the orange uncertainty margin of Figure 2.1b) instead of the displayed prediction intervals (i.e., the green uncertainty margin of Figure 2.1b). However, the difference between total and bias-corrected uncertainty was found to be negligible for the cases displayed in Figure 2.5 (in fact it slightly improves the agreement between the results), thus the comparison remains valid.

2.5.3 Predictive error variance minimization

2.5.3.1 Optimally regularized case

The *s*-versus-$\underline{s}$ scatterplots for all 12 optimally regularized (stochastically weighted Tikhonov scheme) PMA processes, in which varying degrees of model-to-measurement misfit are targeted, are displayed in Figure 2.6 (exit point prediction) and Figure 2.7 (travel time prediction). The relevant statistics are collated in Table 2.4.

In most cases the attained measurement objective function Φ_m across the ensemble of *n* paired model realizations is highly consistent as indicated by the very small values of the coefficient of variation for Φ_m (CV_{Φ_m}). This indicates average Φ_m is generally representative of the value for each individual realization. CV_{Φ_m} is greater for the very large measurement noise cases due to values below Φ_m^l being achieved for some realizations in the pre-calibration state (i.e., the initial uniform *K* of 1 m/day), resulting in immediate termination of the calibration process.

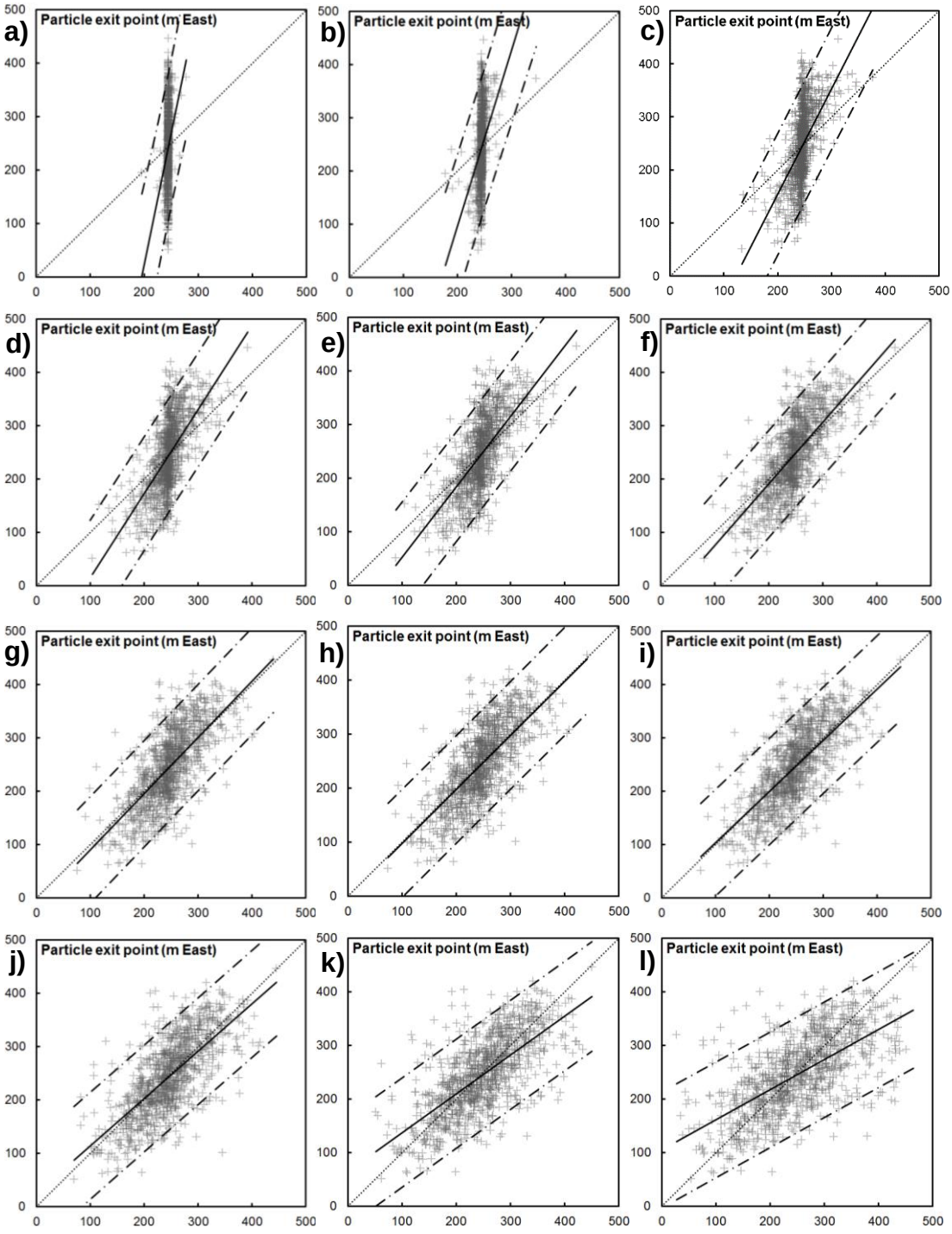

Figure 2.6. Particle exit location prediction *s*-versus-s scatterplots for Moore and Doherty (2005) Tikhonov case, for observation weights *q* of (a) 0.50, (b) 1.00, (c) 2.00, (d) 2.50, (e) 2.95, (f) 3.33, (g) 3.65, (h) 3.84, (i) 4.00, (j) 4.36, (k) 5.75 and (l) 10.0. (*q* value commensurate with σ_ε is 3.33). The solid line is the scatterplot regression line, the dashed lines bound the 95% prediction interval and the dotted line represents the 1:1 line.

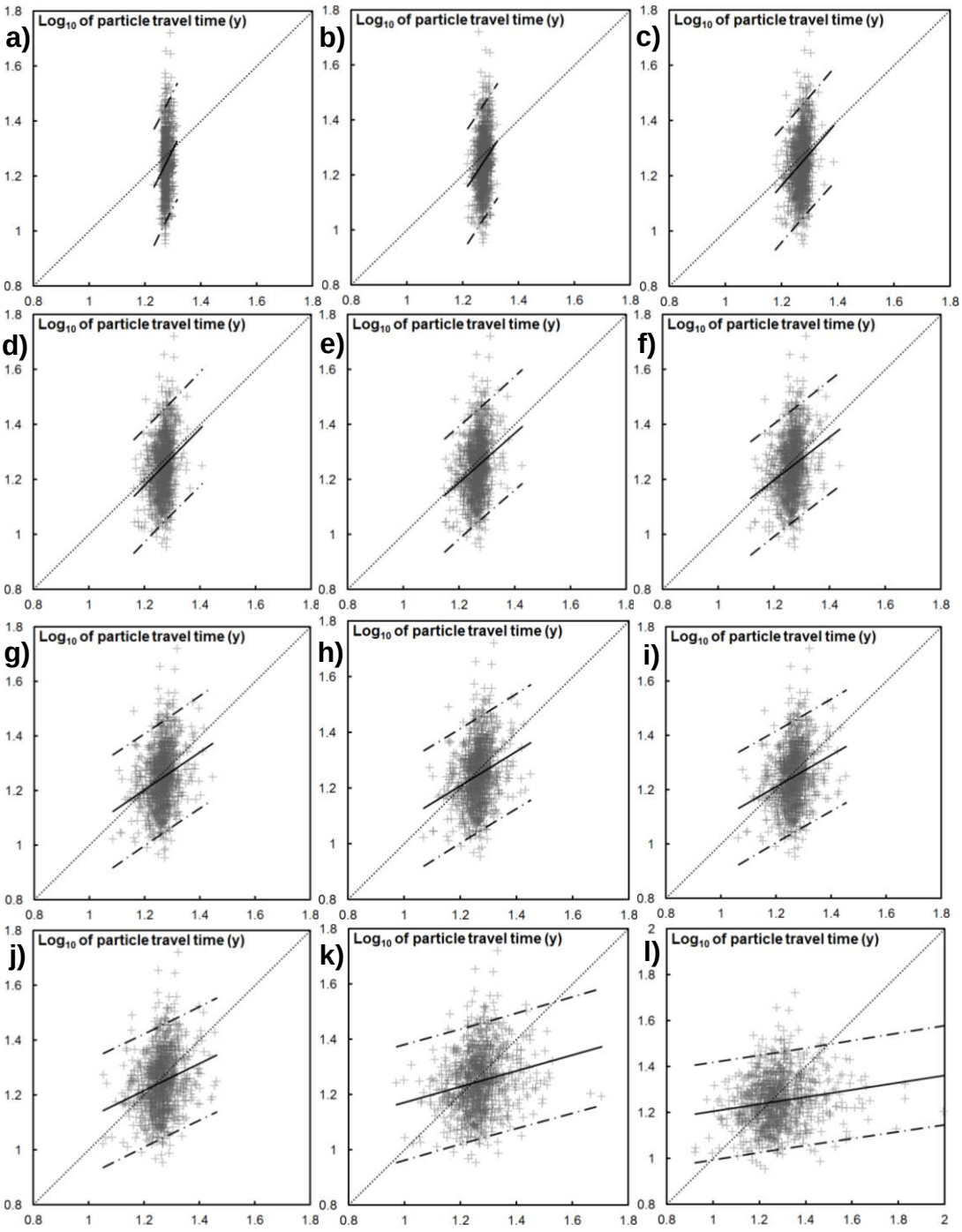

Figure 2.7. Particle travel time prediction *s*-versus-<u>s</u> scatterplots for Moore and Doherty (2005) Tikhonov case, for observation weights *q* of (a) 0.50, (b) 1.00, (c) 2.00, (d) 2.50, (e) 2.95, (f) 3.33, (g) 3.65, (h) 3.84, (i) 4.00, (j) 4.36, (k) 5.75 and (l) 10.0. (*q* value commensurate with σ_ε is 3.33). The solid line is the scatterplot regression line, the dashed lines bound the 95% prediction interval and the dotted line represents the 1:1 line.

Table 2.4. PMA statistics for Tikhonov-regularized inversion (0.30 m standard deviation measurement noise case). q is the weight applied to observations (i.e., elements of the Q_h matrix of equation (2.2)), n is the number of model-pair realizations, $\bar{\mu}$ is the average regularization weight factor, $\bar{\Phi}_m$ is the average measurement objective function[a], $\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ is average standard deviation of post-calibration model-to-measurement misfit, CV_{Φ_m} is the coefficient of variation of the measurement objective function, a and b are the s-versus-$\underline{s}$ regression coefficients of equation (2.26), and r^2 is the coefficient of determination.

						Exit point			Log_{10} travel time		
q	n	$\bar{\mu}$	$\bar{\Phi}_m$	$\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ (m)	CV_{Φ_m}	a	b	r^2	a	b	r^2
-	-	-	1283	3.11	-	-	-	-	-	-	-
0.50	1000	206	331	1.48	0.58	-970	4.96	0.04	-1.32	2.01	0.03
1.00	1000	135	118	0.93	0.26	-565	3.31	0.11	-0.71	1.54	0.05
2.00	1000	27.8	32.9	0.50	0.07	-243	1.99	0.33	-0.23	1.17	0.06
2.50	1000	13.9	21.2	0.40	0.04	-142	1.57	0.42	-0.06	1.03	0.07
2.95	1000	8.5	15.4	0.34	0.02	-79.1	1.32	0.47	0.11	0.90	0.07
3.33	1000	6.1	12.0	0.30	0.01	-40.8	1.16	0.49	0.25	0.79	0.07
3.65	1000	3.9	10.0	0.27	0.00	-15.4	1.05	0.49	0.38	0.69	0.07
3.84	1000	3.7	9.0	0.26	0.00	-2.39	1.00	0.49	0.46	0.62	0.06
4.00	1000	2.5	8.0	0.25	0.00	7.17	0.96	0.49	0.51	0.59	0.06
4.36	1000	2.3	7.0	0.23	0.00	25.4	0.89	0.49	0.62	0.49	0.06
5.75	995	1.9	4.0	0.17	0.01	64.9	0.73	0.47	0.89	0.29	0.04
10.0	951	1.7	1.3	0.10	0.00	105	0.56	0.41	1.05	0.16	0.03

[a]Average measurement objective function $\bar{\Phi}_m$ values are all presented in terms of their equivalent value for observation weights equal to the inverse of measurement noise (i.e., q = 3.33) for the purposes of comparison with truncated SVD results and Moore and Doherty (2005) results.

Based on the 12 PMA processes and following Moore and Doherty (2005), the relationship between predictive error variance and (the reciprocal of) the regularization weight factor μ is displayed as Figure 2.8. The magnitude of pre-calibration predictive error variance is represented by the horizontal dashed line. The diamond denotes the point on the error variance function at which the level of model-to-measurement misfit is commensurate with measurement noise (i.e., a Φ_m value of 12.0 as shown in Table 2.4).

Figure 2.8 indicates that for the particle exit point prediction, minimum predictive error variance does not occur at a $\bar{\Phi}_m$ value of 12.0. Rather, the minimum occurs at a higher value of $1/\mu$, that is, for a closer fit between observations and their model-generated counterparts. The minimum occurs at $\bar{\Phi}_m = 9.0$ (the case for which s-versus-$\underline{s}$ regression line slope b = 1.00), which is equivalent to a misfit standard deviation of 0.26 m (see Table 2.4). This indicates that slight overfitting is required to minimise predictive error variance for this prediction, which is consistent with the results of Moore and Doherty (2005).

Comparison of mean error variances for estimated $\log_{10}K$ values for each PMA processes reveals the same trend, with the minimum occurring for $\bar{\Phi}_m = 9.0$ (results not presented for the sake of brevity). This demonstrates that this outcome is not an artefact of the prediction and is an inherent product of the parameter estimation process itself.

Additionally, results suggest that this outcome is dependent upon the level of measurement noise. Particle exit location prediction *s*-versus-*s* scatterplot regression line slopes for the 0.01 m, 0.1 m (results not shown) and 0.3 m standard deviation measurement noise cases are 0.99, 1.12 and 1.16, respectively (for $\bar{\Phi}_m = 12.0$). In conjunction with the fact that minimum error variance corresponds to a regression line slope of unity (see Figure 2.6 and Table 2.4), this suggests an increasing degree of overfitting is required to minimise error variance in prediction of particle exit location. This observation extends the linear subspace-based results of Moore and Doherty (2005). Further examination of the influence of measurement noise upon the position of the minimum in the predictive error variance function is beyond the scope of the present study but is recommended as future work.

In terms of the particle travel time prediction, Figure 2.8 reiterates and extends the section 2.5.1 discussion. Travel time predictive error variance is barely reduced relative to its pre-calibration magnitude for any degree of calibration. Furthermore, overfitting with respect to measurement noise (discussed in depth below) inflates particle travel time predictive error variance to well beyond its pre-calibration value.

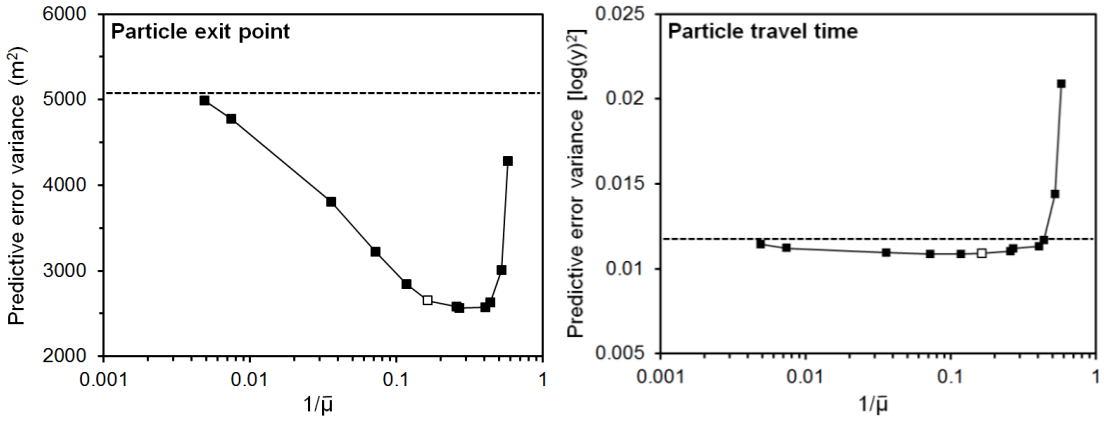

Figure 2.8. Predictive error variance versus the inverse of the average regularization weight factor ($1/\bar{\mu}$). The dashed lines represent pre-calibration predictive error variance (quantified through unconstrained Monte Carlo analysis). The hollow square in each plot represents the case for which model-to-measurement misfit is

commensurate with measurement noise (i.e., an average Φ_m of 12.0, corresponding to an average model-to-measurement misfit standard deviation of 0.30 m).

Note that average μ values in Figure 2.8 span a somewhat smaller range than the equivalent results presented by Moore and Doherty (2005). The range of μ required to achieve varying levels of fit between model outputs and observations was found to be highly case specific. Some individual realizations exhibit a larger μ range more comparable with the results of Moore and Doherty (2005), whilst a larger number of realizations exhibit smaller ranges, thus reducing the range of average μ.

2.5.3.2 Suboptimally regularized case

The *s*-versus-$\underline{s}$ scatterplots pertaining to all 12 PMA processes undertaken for the truncated SVD case are displayed in Figure 2.9 and Figure 2.10 for the exit point and travel time predictions, respectively. Table 2.5 details the corresponding statistics.

In contrast to the Tikhonov-regularized case, a significant degree of variability in Φ_m between realizations occurs in each PMA process (i.e., Φ_m for a given number of pre-truncation singular values is highly case-dependent). This is indicated by the large values of CV_{Φ_m} in Table 2.5 relative to the equivalent Tikhonov-case values in Table 2.4. The effect of this degree of variability about the mean was tested through reproduction of results subject to Φ_m-based filtering (results not presented in the interests of brevity). An observed immunity of key *s*-versus-$\underline{s}$ regression characteristics to this filtering suggests that the large inter-realization Φ_m variability does not impact upon the interpretation of results in the present study.

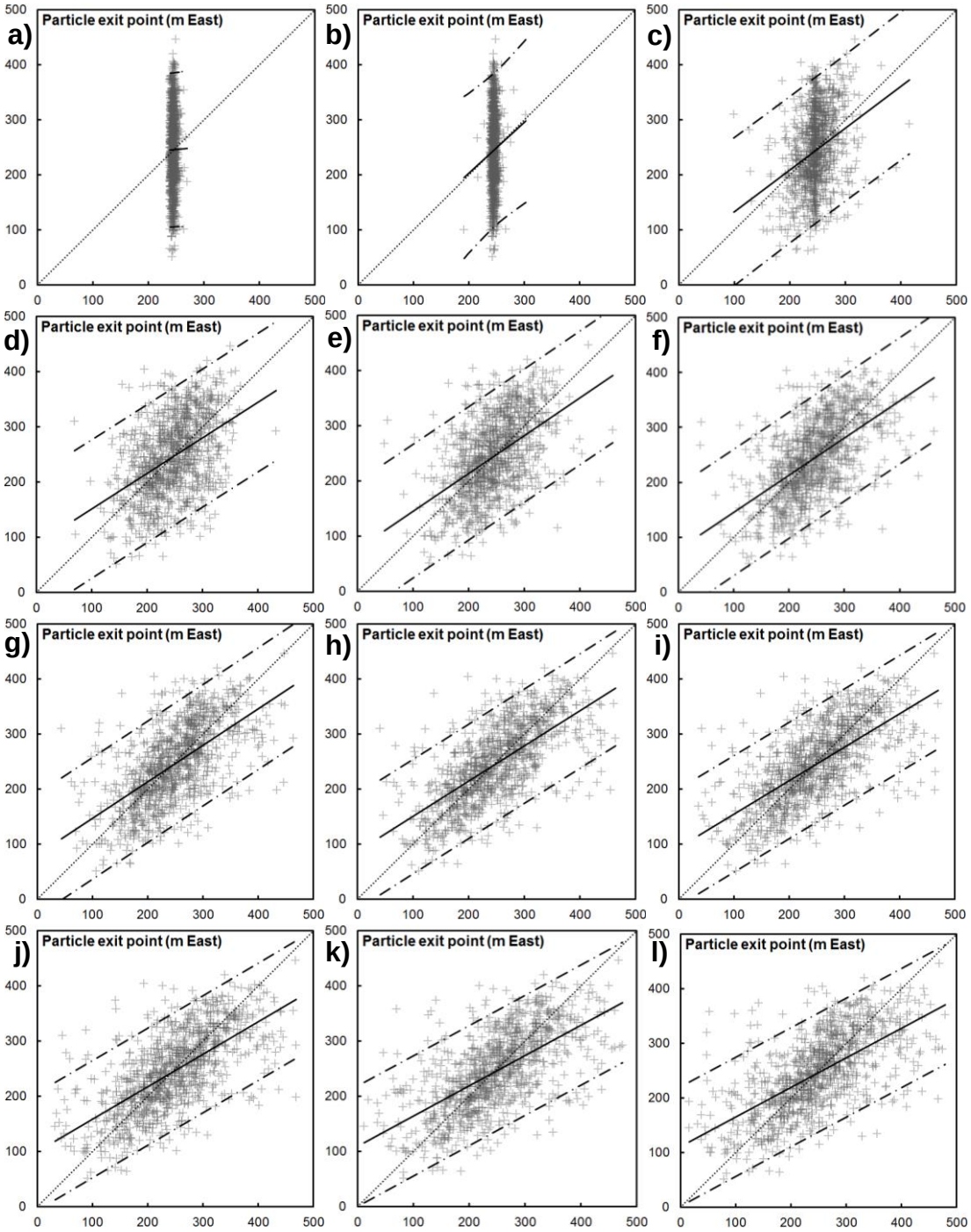

Figure 2.9. Particle exit point prediction *s*-versus-s scatterplots for the suboptimally regularized (untransformed truncated SVD) case, for (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, (f) 6, (g) 7, (h) 8, (i) 9, (j) 10, (k) 11 and (l) 12 pre-truncation singular values. The solid line is the scatterplot regression line, the dashed lines bound the 95% prediction interval and the dotted line represents the 1:1 line.

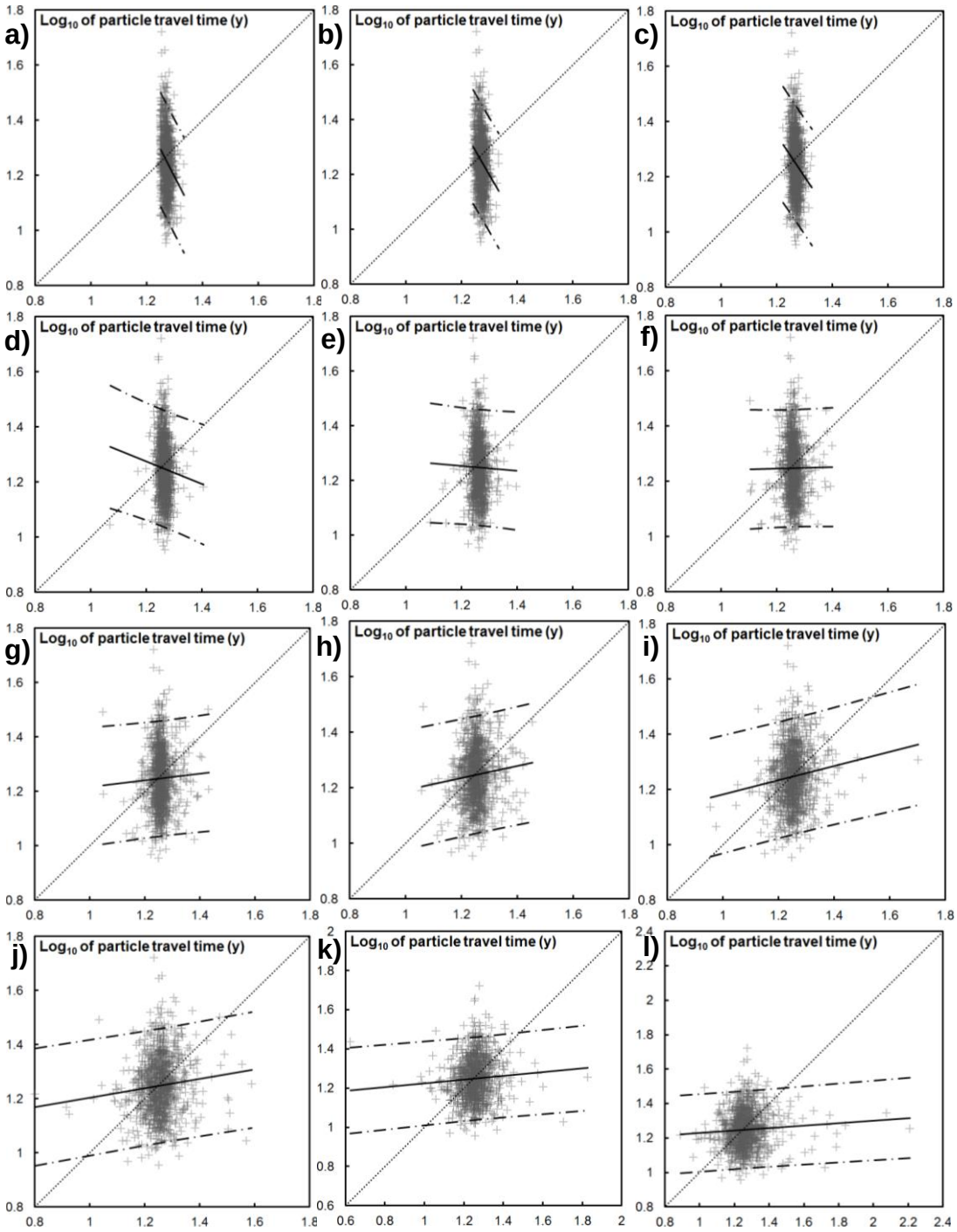

Figure 2.10. Particle travel time prediction *s*-versus-*s* scatterplots for the suboptimally regularized (untransformed truncated SVD) case, for (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, (f) 6, (g) 7, (h) 8, (i) 9, (j) 10, (k) 11 and (l) 12 pre-truncation singular values. The solid line is the scatterplot regression line, the dashed lines bound the 95% prediction interval and the dotted line represents the 1:1 line.

Table 2.5. PMA statistics for truncated SVD-based inversion (0.30 m standard deviation measurement noise case). SVs denotes the number of pre-truncation singular values employed in the calibration process (i.e., the number of singular values assigned to the solution space, represented by S_1 of equation (2.10)), *n* is the number of model-pair realizations, $\bar{\Phi}_m$ is the average measurement objective function, $\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ is average standard deviation of post-calibration model-to-measurement misfit, CV_{Φ_m} is the coefficient of variation of the measurement objective function, *a* and *b* are the *s*-versus-$\underline{s}$ regression coefficients of equation (2.26), and r^2 is the coefficient of determination.

					Exit point			Log_{10} travel time		
SVs	*n*	$\bar{\Phi}_m$	$\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ (m)	CV_{Φ_m}	*a*	*b*	r^2	*a*	*b*	r^2
0	-	1283	3.11	-	-	-	-	-	-	-
1	1000	131	0.76	3.35	222	0.10	0.00	3.75	-1.97	0.05
2	1000	38.1	0.48	1.46	19.1	0.92	0.00	3.47	-1.75	0.05
3	1000	24.2	0.38	1.69	57.3	0.76	0.10	3.13	-1.49	0.03
4	1000	17.0	0.33	1.18	86.8	0.64	0.20	1.76	-0.41	0.00
5	1000	12.4	0.28	1.07	76.8	0.68	0.26	1.37	-0.09	0.00
6	1000	9.05	0.24	1.05	77.3	0.68	0.33	1.21	0.03	0.00
7	1000	6.30	0.20	1.12	80.4	0.66	0.38	1.10	0.12	0.00
8	1000	3.88	0.16	0.99	86.0	0.64	0.45	0.97	0.22	0.01
9	997	2.51	0.12	1.22	94.3	0.61	0.43	0.93	0.26	0.01
10	996	1.39	0.09	1.38	99.8	0.59	0.42	1.03	0.17	0.01
11	985	0.62	0.05	2.39	109	0.55	0.39	1.13	0.10	0.00
12	966	0.13	0.01	3.42	111	0.54	0.40	1.16	0.07	0.00

Table 2.5 shows that use of five pre-truncation singular values results in the average Φ_m value closest to 12.0 (i.e., 12.4, which corresponds to an average model-to-measurement misfit standard deviation of 0.28 m). This is consistent with the results presented by Moore and Doherty (2005), which also demonstrate that the use of five singular values results in the Φ_m value closest to 12.0 (i.e., 11.19, corresponding to an average model-to-measurement misfit standard deviation of 0.29 m). Also consistent with Moore and Doherty (2005) (as well as the above Tikhonov regularisation-based results), Figure 2.11 demonstrates that the error variance for the prediction of particle exit location is further reduced with the use of additional pre-truncation singular values beyond five (i.e., overfitting is again required to minimise predictive error variance).

The predictive error variance function for particle exit location in Figure 2.11 differs slightly from the results presented by Moore and Doherty (2005). Figure 2.11 displays a distinct predictive error variance minimum at eight singular values, beyond which a significant increase is observed, which is not present in the results of Moore and Doherty (2005). Figure 2.11 in fact suggest that the combination of suboptimal regularization and overfitting (i.e., use of 11 or 12 pre-truncation singular values) almost entirely erodes the gains achieved through calibration such that post-calibration exit location error variance is nearly equal to its pre-calibration value.

In terms of the particle travel time prediction, Figure 2.11 demonstrates that, accompanied by suboptimal regularization (SVD in the absence of appropriate pre-calibration parameter transformation in this case), calibration employing any number of singular values results in inflation of predictive error variance beyond its pre-calibration magnitude. Thus, even for a model that is "well-calibrated" from a history matching point of view (and for any degree of history matching for that matter), the calibration process improves the ability of the model to predict particle exit point, whilst simultaneously degrading its ability to predict travel time relative to the uncalibrated model. Previous authors demonstrate counterintuitive prediction-specificity in the outcomes of calibration as a result of model structural simplifications and defects (e.g., Christensen and Doherty, 2008; White et al., 2014). The present results demonstrate this for an idealised case wherein the model and "reality" are structurally equivalent, thus this outcome is solely a consequence of the (suboptimal) regularization mechanism employed to attain $\log_{10}K$ field uniqueness.

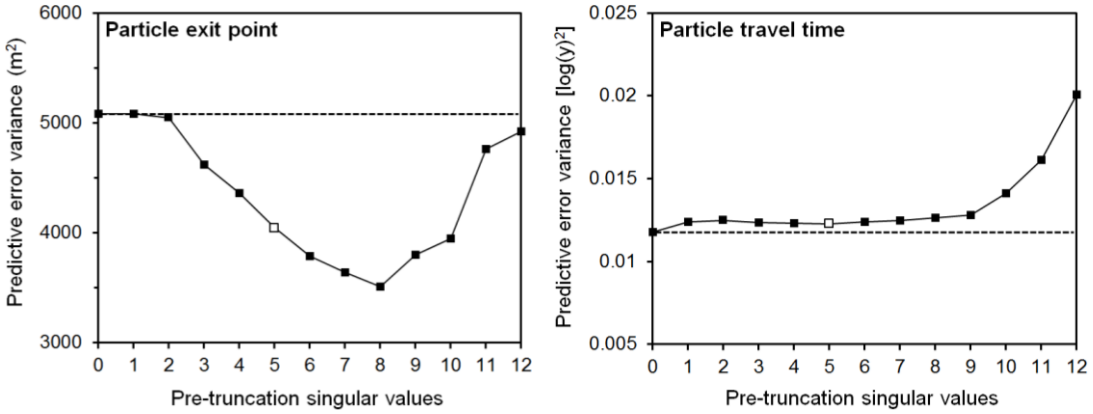

Figure 2.11. Predictive error variance versus number of singular values employed in the truncated SVD inversion process. The dashed lines represent pre-calibration predictive error variance (quantified through unconstrained Monte Carlo analysis and representing the use of zero pre-truncation singular values). The hollow square in each plot represents the case for which the average standard deviation of model-to-measurement misfit is most commensurate with measurement noise (i.e., an average Φ_m of 12.4, corresponding to an average model-to-measurement misfit standard deviation of 0.28 m).

2.5.4 Identification of predictive bias

Figure 2.12 provides an illustrative example to support the following discussion of calibration-induced parameter compensation and consequential predictive bias. Figure 2.12a provides a representative example of a "reality" $\log_{10}K$ field. This is accompanied by three post-calibration $\log_{10}K$ fields (each pertaining to the 0.30 m

standard deviation measurement noise case). Figure 2.12b represents optimal calibration and regularization (i.e., stochastically weighted Tikhonov scheme and post-calibration model-to-measurement misfit that is commensurate with measurement noise ($\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ = 0.30 m)); Figure 2.12c is the same case in which overfitting has occurred ($\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ = 0.10 m); Figure 2.12d represents suboptimally regularized (untransformed truncated SVD) calibration using five pre-truncation singular values.

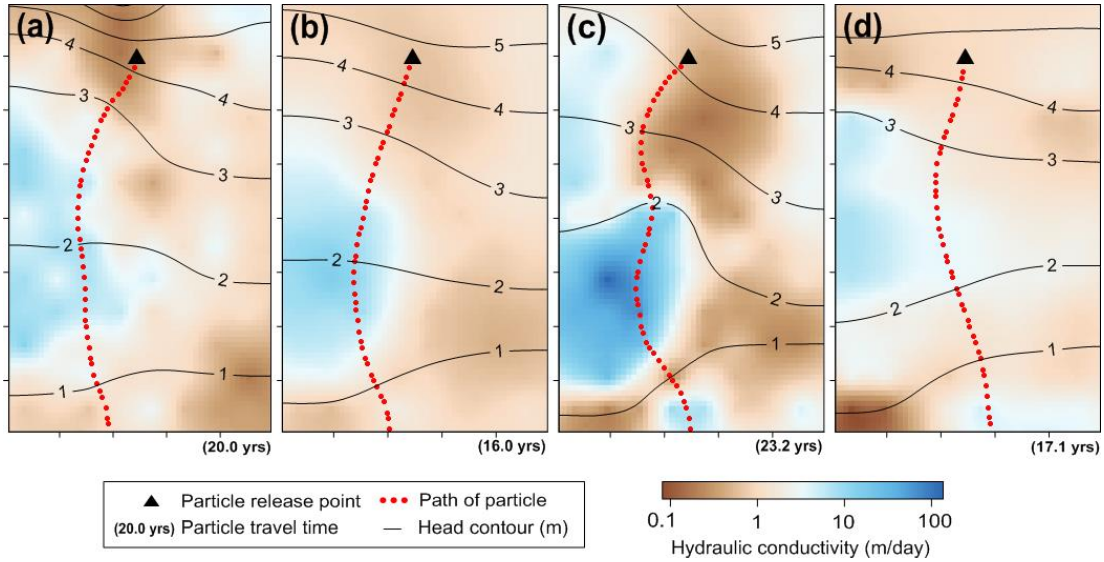

Figure 2.12. (a) Arbitrary "reality" $\log_{10}K$ field realization accompanied by three example post-calibration $\log_{10}K$ fields yielded through different PMA processes (all for the σ_ε = 0.30 m case): (b) Tikhonov-regularized calibration to level commensurate with measurement noise ($\bar{\Phi}_m$ = 12.0; $\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ = 0.30 m); (c) Tikhonov-regularized case including substantial overfitting ($\bar{\Phi}_m$ = 1.3; $\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ = 0.10 m); and (d) calibration effected through truncated SVD employing five pre-truncation singular values ($\bar{\Phi}_m$ = 12.4; $\bar{\sigma}_{\mathbf{X\underline{k}-h}}$ = 0.28 m).

2.5.4.1 Overfitting-induced bias

Figure 2.12c exemplifies the classic compensatory parameter behaviour induced by overfitting. It is clear that the model has been forced to introduce unrealistic parameter variability in order to closely fit the measurement dataset (i.e., a model-to-measurement misfit standard deviation of 0.10 m) contaminated by measurement noise of standard deviation 0.30 m. The range of $\log_{10}K$ variability in the estimated field is substantially inflated relative to the "reality" field (i.e., Figure 2.12a), and indeed violates "geological plausibility" defined by the variogram used to generate the ensemble of "reality" fields.

Simultaneously, the field of Figure 2.12c includes substantially more spatial detail than the "optimally calibrated" equivalent field in Figure 2.12b. This small-scale detail comprises parameter components that belong to the artificially expanded null space owing to the presence of a 0.30 m standard deviation measurement noise. As explained above, the calibration dataset does not support estimation of these null-space parameter components. However, attaining a model-to-measurement misfit standard deviation of 0.10 m has resulted in their adjustment and thus an expected increase in error potential (bias) in predictions that are sensitive to them.

This expectation is confirmed through inspection of Table 2.4 (and Figure 2.6l). PMA for this case yields *s*-versus-$\underline{s}$ regression line slopes of substantially less than unity for both predictions (0.56 and 0.16 for exit location and travel time, respectively). Thus PMA clearly identifies substantial calibration induced predictive bias caused by significant overfitting with respect to measurement noise.

The suite of *s*-versus-$\underline{s}$ scatterplots for both predictions based the optimally regularized case demonstrate a monotonic reduction in regression line slope as model-to-measurement misfit is reduced (see Figure 2.6, Figure 2.7, and/or Table 2.4). This monotonicity indicates that the PMA process is inherently consistent, and thus supports its reliability as an identifier of the presence of calibration-induced predictive bias (or lack thereof).

2.5.4.2 Suboptimal regularization-induced bias

PMA results pertaining to the suboptimally regularized (untransformed truncated SVD) case reveal that *s*-versus-$\underline{s}$ regression line slopes of less than unity are yielded for both predictions based on calibration employing any number of pre-truncation singular (see Figure 2.9, Figure 2.10 and/or Table 2.5). That is, consistent with theoretical expectation, PMA results indicate pervasive calibration-induced predictive bias resulting from any attempt at calibration accompanied by suboptimal regularization. In fact Figure 2.10 includes negative regression line slopes, indicating an inversely proportional relationship between model-predicted and "true" travel times up to and including the use of five pre-truncation singular values (for which $\overline{\Phi}_m$ is approximately commensurate with measurement noise).

As described above, the failure to undertake appropriate pre-calibration parameter transformation in accordance with geological plausibility effectively hardwires

parameter compensation into the calibration process. Any degree of parameter adjustment through the calibration process thus includes unsupported adjustment of "reality model" null-space parameter components (the definition of which is based on C(**k**); this being discussed in detail in Chapter 3 of the present thesis). The inevitable outcome is calibration-induced predictive bias in null-space dependent predictions such as advective transport.

This is demonstrated visually by Figure 2.12. Figure 2.12d is the parameter field estimated in the presence of suboptimal regularization employing five pre-truncation singular values (based on the "reality" field in 2.12a with 0.30 m standard deviation measurement noise added). The measurement objective function associated with Figure 2.12d is 13.1, thus it is an example for which model-to-measurement misfit is approximately commensurate with measurement noise, but with a very slight degree of underfitting. Despite this, the field of Figure 2.12d displays a visibly higher degree of variability than its optimally regularized counterpart in Figure 2.12b. Whilst Figure 2.12b clearly captures the key features of Figure 2.12a and resembles a smoothed version of "reality", Figure 2.12b exhibits some erroneous small-scale variations, some of which are clearly influenced by the locations of observation wells (see Figure 2.2a).

A comprehensive analysis of the covariance structures of the various estimated parameter field ensembles to support the above commentary based on visual inspection of the example parameter fields in Figure 2.12 is beyond the scope of the present study.

As a consequence of the pervasive bias caused by suboptimal regularization, predictive error variance is inflated relative to optimal regularization for most values of $\overline{\Phi}_m$. Figure 2.13 displays predictive error variance versus $\overline{\Phi}_m$ for both regularization approaches to facilitate direct comparison. This highlights the erosion of the benefits of data assimilation caused by suboptimal regularization. For prediction of travel time, the regularization optimality (or lack thereof) determines whether predictive performance is improved or degraded through calibration (relative to pre-calibration predictive error variance).

White et al. (2014) discuss the notion that inclusion of spatially correlated expert knowledge may cause parameter compensation to spread across larger regions of the model domain. This may explain the rise of the optimally regularized case error variance above that of the suboptimally regularized case for low values of $\overline{\Phi}_m$ (see

Figure 2.13). That is, the rapidly increasing degree of parameter compensation induced by an increasing degree of overfitting is more spatially expansive in the optimally regularized case, thus having a greater influence on predictions than the more localised parameter compensation facilitated by suboptimal regularization.

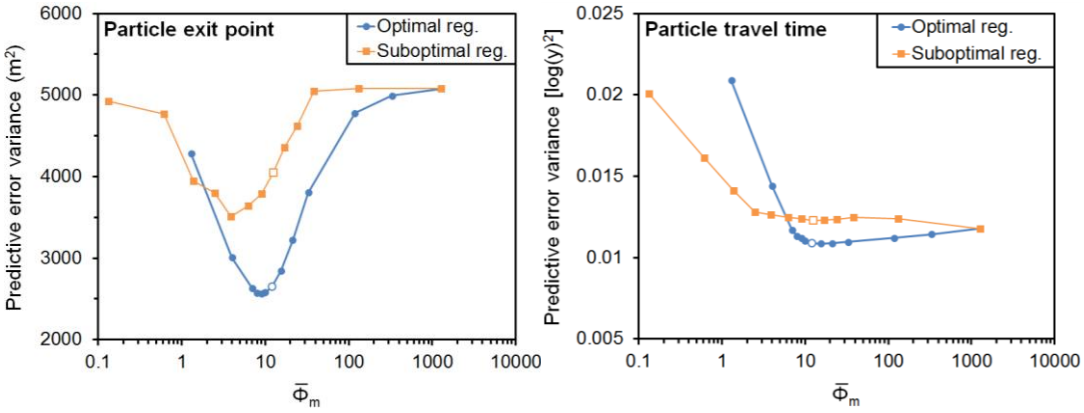

Figure 2.13. Predictive error variance versus average measurement objective function $\bar{\Phi}_m$ based on both optimal regularization (stochastically weighted Tikhonov scheme) and suboptimal regularization (truncated SVD in the absence of appropriate parameter transformation). The hollow markers represent the cases for which model-to-measurement misfit is most commensurate with measurement noise (see Table 2.4 and Table 2.5 for details).

2.5.5 Bias-corrected post-calibration uncertainty

Figure 2.14 provides plots of predictive error variance versus average Φ_m equivalent to Figure 2.13, but for predictive error variance quantified through the PMA *s*-versus-*s* scatterplots (i.e., the 95% prediction interval for the true minimum error variance prediction that has been corrected for calibration-induced bias, as represented by Figure 2.1).

The predictive error variance functions of Figure 2.13 are also displayed in Figure 2.14 as dashed lines to facilitate direct comparison. Figure 2.14 thus clearly illustrates 1) the extent of the inflation in predictive error variance attributable to calibration-induced parameter surrogacy and 2) the capacity of PMA to reduce the component of predictive error variance caused by predictive bias. This is particularly evident for low values of average Φ_m where bias is most severe due to substantial overfitting. For low values of average Φ_m, bias-corrected predictive error variance is relatively comparable regardless of whether regularization is optimal or suboptimal, and regardless of the degree of overfitting. This indicates that bias-corrected post-calibration predictive

error variance quantified through PMA *s*-versus-s scatterplots is almost entirely void of the deleterious effects of calibration-induced predictive bias.

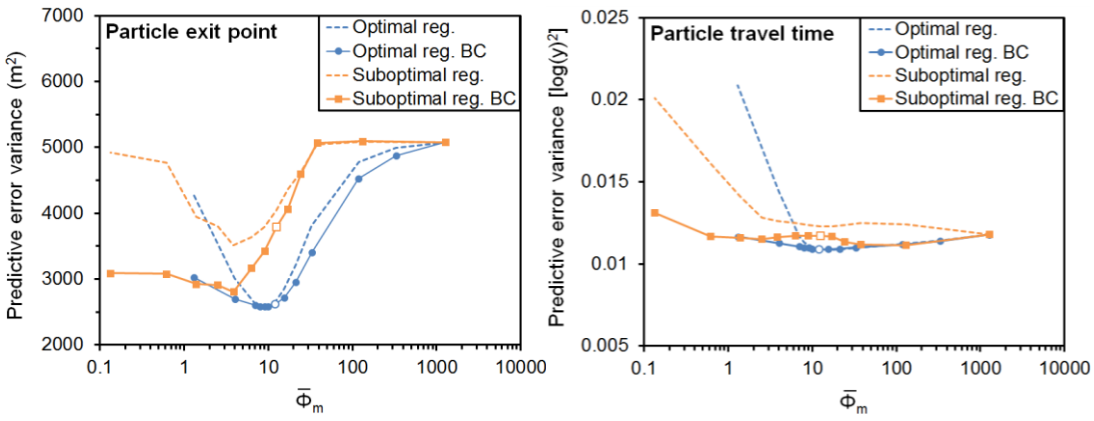

Figure 2.14. Bias-corrected (denoted as 'BC') predictive error variance functions quantified through PMA *s*-versus-s scatterplots (total predictive error variance functions from Figure 2.13 are shown as dashed lines for comparative purposes). The hollow markers represent the cases for which model-to-measurement misfit is most commensurate with measurement noise (see Table 2.4 and Table 2.5 for details).

2.5.6 Robustness of *s*-versus-s metrics

Doherty and Christensen (2011) employed 483 paired model realizations for the calculation of *s*-versus-s statistics in their study. An additional 500 realizations were plotted independently by these authors for testing. They observed qualitative agreement between the independent *s*-versus-s plots constructed using both sets of realizations. Quantitative 95% prediction interval testing by these authors returned acceptable results, albeit highlighting a degree of nonstationary scatter attributed to model nonlinearity.

In the present study an arbitrarily large sample size of 1000 complex-simple model pair realizations was employed for PMA. In order to test the suitability of this sample size a convergence test was performed to verify the integrity of key PMA metrics. This involved repetition of PMA for an incrementally increasing sample size (including a total of 23 PMA repetitions using sample sizes ranging from 10 realizations to the full set of 1000 realizations). For completeness, a convergence test was performed for both Tikhonov (optimally regularized) and truncated SVD (suboptimally regularized) approaches. For this purpose, the q = 3.33 and 5 pre-truncation singular values cases were adopted, respectively. The sample size-dependence of several PMA-derived metrics was analyzed, these being *s*-versus-s regression line slope, predictive error variance and bias-corrected predictive error variance.

Figure 2.15a displays *s*-versus-s regression line slope versus complex-simple model pair sample size. Figure 2.15b displays PMA-derived predictive error variance (both total and bias-corrected) versus sample size. Figure 2.15a and Figure 2.15b indicate acceptable stabilization of the values of all tested metrics, most significantly after the sample size reaches approximately 500.

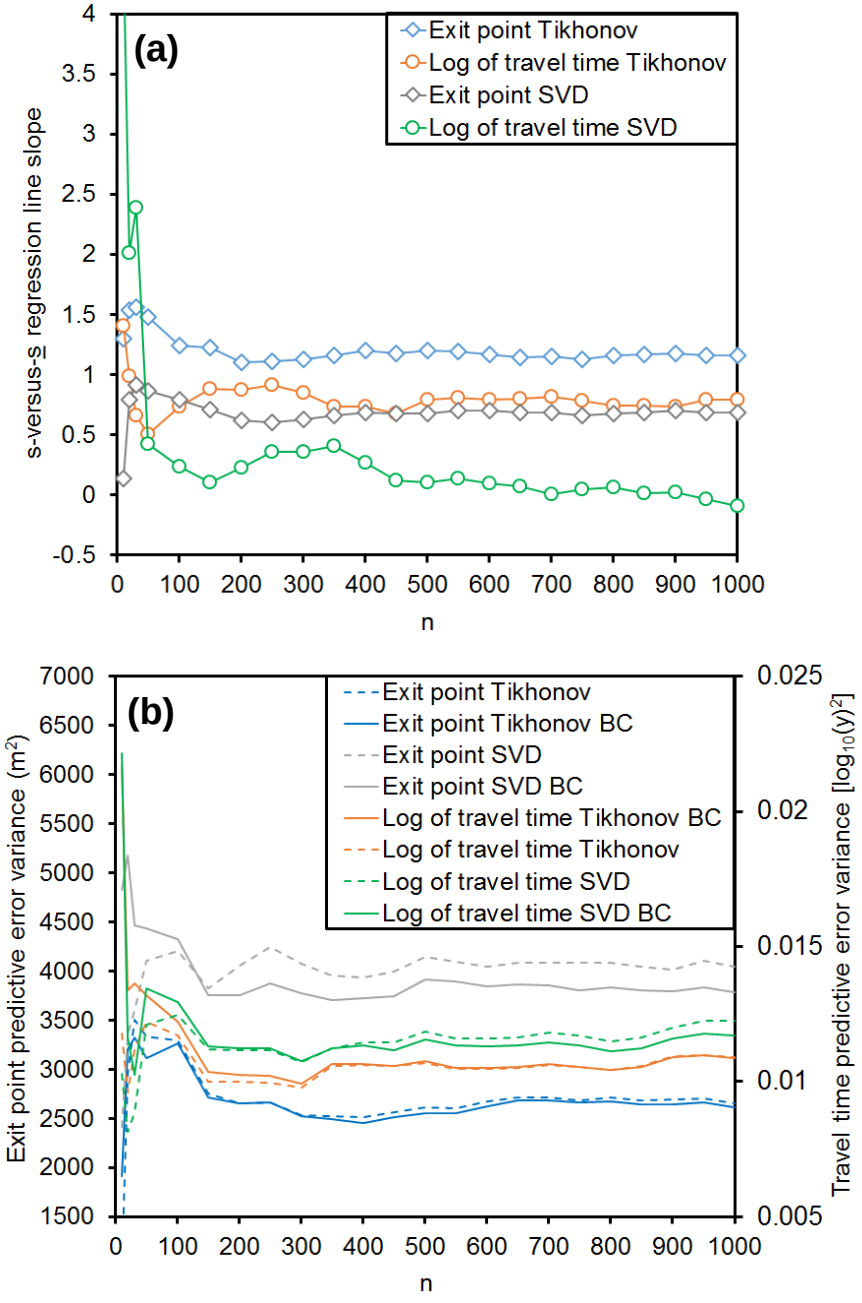

Figure 2.15. Convergence (with respect to *s*-versus-s sample size *n*) test for PMA metrics including (a) *s*-versus-s regression line slope and (b) total and bias-corrected (BC) predictive error variance represented by the dashed and solid lines, respectively. Tikhonov example pertains to observation weights $q = 3.33$ and truncated SVD example pertains to 5 pre-truncation singular values. Sample-size increments in (b) are equivalent to those in (a); markers in (b) not displayed to reduce clutter.

2.6 Conclusions

The present study comprises a systematic proof of concept for the PMA methodology, which was first presented by Doherty and Christensen (2011) and has not previously been tested for empirical consistency with theoretical expectations. For this purpose, PMA is applied to a hypothetical synthetic example, studied extensively in previous literature, in which the model to be calibrated and "reality" are structurally identical. This facilitates comparison of PMA results with well-established "traditional" uncertainty analysis results, as well as examination of its ability to identify calibration-induced bias in the absence of the complicating influence of structural simplifications.

Reflecting the key aims of the current proof of concept outlined in the introductory text, the outcomes of the present study are as follows:

1. Subject to theoretically optimal regularization and calibration, PMA indicates unbiased post-calibration prediction of particle exit location, as theoretically expected. PMA applied to the prediction of particle travel time indicates a small amount of calibration-induced bias. This is attributed to the inability of the hydraulic head dataset to constrain this prediction due to its high sensitivity to small-scale heterogeneity.

2. Post-calibration predictive error variance quantified through PMA is demonstrated to be in good agreement with equivalent results from previous studies attained via well-established "traditional" uncertainty analysis methods. Discrepancies, where they occur, are within a range that is acceptably attributable to the linearity assumption upon which some of the previous results are based.

3. PMA identifies the occurrence of calibration-induced predictive bias where expected in the present synthetic example. Known sources of compensatory parameter behaviour (accompanied by unsupported adjustment of null-space parameter components), namely overfitting with respect to measurement noise, and suboptimal regularization is clearly elucidated by PMA. In the case of overfitting with respect to measurement noise, predictive bias is proportional to the degree of overfitting as expected. In the case of suboptimal regularization, predictive bias is pervasive as expected, due to parameter surrogacy being inherent to the formulation of the inverse problem.

4. PMA is demonstrated to allow extensive retroactive mitigation of the deleterious effects of calibration-induced predictive bias. The vast majority of post-calibration predictive error variance inflation associated with bias is removed through quantification of bias-corrected predictive error variance using *s*-versus-*s* scatterplots.

The present study also yields some additional findings that are not directly related to the main aims of the proof of concept but are nonetheless pertinent, as well as some more general insights. These are summarized as follows:

- PMA applied in the presence of optimal regularization yields a monotonic change in regression line slope with measurement objective function. This monotonicity strengthens the validation of PMA as it demonstrates inherent consistency and thus reliability as a bias identification (and reduction) tool.

- PMA results based on theoretically optimal calibration suggest that as the level of measurement noise contaminating the calibration dataset is increased, an increasing degree of overfitting is required to minimise the particle exit location prediction error variance. This extends the single observation of this phenomenon identified by Moore and Doherty (2005) via their linear subspace analysis. Post-calibration parameter (i.e., $\log_{10}K$) error variance was found to produce the same trend, indicating that this is not a prediction-related artefact but inherent to the parameter estimation process itself. Further research is recommended to examine this phenomenon.

- The present results demonstrate that, even through use of hypothetical structurally perfect model (which is unattainable in reality), a poorly forged calibration process (for example, in the form of an absence of appropriate pre-calibration parameter transformation) has the ability to instil a greater potential for predictive error in a "well-calibrated" model than if the model had not been calibrated at all.

Chapter 3

Parameter and predictive outcomes of model simplification

Note: this chapter is based on the following paper:

Watson, T. A., J. E. Doherty, and S. Christensen 2013. Parameter and predictive outcomes of model simplification. Water. Resources Research 49, doi:10.1002/wrcr.20145.

Abstract

Simplification is an unavoidable aspect of model usage. Even complex, physically based models are simplifications of reality. More profound simplification is required to construct the "lumped parameter" models of semi-physical basis that are often employed for simulation of large-scale processes operative over one or many watersheds. Simplification can lead to model predictive error beyond that which would be expected on the basis of study-area information deficits alone. Building on a recently developed mathematical description of the model simplification process, this work employs linear subspace methods to analyse in detail the nature and ramifications of that process when applied to a one-dimensional, Richards equation-based unsaturated zone model used to predict recharge to a groundwater system. Two simplified versions of this model are examined. The first achieves simplification through assuming vertical parameter uniformity. The second achieves simplification through use of a lumped parameter model in place of the Richards equation-based model. Relationships between parameters employed by the complex model and those used by each of the simplified models are analysed. The nature of predictive errors incurred through simplification is explored. Also explored is the ability of the calibration process to decrease the propensity for model error in making some predictions, while increasing the propensity for model error in the making of others – an outcome that may be considered counter-intuitive from a Bayesian perspective, but which is a natural consequence of suboptimal simplification.

3.1 Introduction

The issue of simplification (as well as the closely related issues of model reduction and parameter upscaling) is central to environmental simulation. This is especially the case where modelling is carried out for the purpose of environmental management. In these circumstances a model is required to make one or a number of predictions on which decisions may be based. The extent to which the process of model simplification induces errors in predictions required of the model must be assessed so that decision-makers and stakeholders can thereby be aware of the credibility of such predictions.

The need for a proper understanding of model simplification arises first and foremost from the fact that all models are simplifications of reality. Hence they are imperfect simulators of the systems that they purport to represent. In addition to this, considerable simplification is often required for a model to be calibrated, for calibration uniqueness can only be attained at the cost of parameter simplification. Ideally, such simplification should achieve a status of minimized error variance for estimated parameters and for predictions which depend on them. Theoretically, this can be achieved through implementation of various types of mathematical regularization; see, for example, Tikhonov and Arsenin (1977), Menke (1984), Aster et al. (2005), and Moore and Doherty (2005; 2006). Following calibration, calibration-constrained Monte-Carlo methodologies such as those described by Tonkin and Doherty (2009), Herckenrath et al. (2011), or hypothesis-testing methodologies such as that described by Moore et al. (2010) can be employed for analysis of the potential for error in predictions made by the simplified model. Ideally, analysis of the potential for errors in predictions made by a simplified/calibrated model is (almost) equivalent to analysis of the inherent uncertainty of these predictions given the information available for the system under study.

Theoretically, uncertainty analysis without the need for simplification as a precursor to that analysis can be undertaken in a Bayesian framework under the assumption that a model's inadequacies as a simulator of real-world environmental processes are small enough to be ignored. Examples of such analysis include the work of Harmon and Challenor (1997), Kuczera and Parent (1998), Campbell et al., (1999), Campbell and Bates (2001), Makowski et al. (2002), Qian et al. (2003), Kanso et al. (2003), Vrugt et al. (2009a) and references cited within these studies. Kennedy and O'Hagan (2001) extended Bayesian analysis to include the contributions made by simplification-

induced model-to-measurement misfit to inferred posterior parameter uncertainty. Their analysis, however, was applied to parameter spaces of relatively low dimension where contributions to predictive uncertainty incurred by the existence of inestimable parameters, and/or inestimable combinations of parameters, are small or non-existent.

In many cases of model design and usage, simplification is not carried out in such a mathematically controlled manner as that which is implemented through regularized inversion. Consequently, a mathematical description of the simplification process is rarely available. It is therefore difficult to account for the contribution that such simplification makes to the potential for error in predictions made by the simplified model.

Recognition of the need for simplification dates back as far as modelling itself. Meisel and Collins (1973) discuss the need for model simplification in order to achieve (among other benefits) computational savings in an optimization context. More recently, Ratto et al. (2011) highlight that, despite the enormous advances in computing power over recent decades, computational limitations still remain a major barrier to use of large-scale, process-based simulation models in a decision-making context. Razavi et al. (2012) provide a review of the growing number of documented incidences of the use of simplified or surrogate models in place of complex, physically based models in studies that demand computation of model outcomes on the basis of many different sets of what they call "explanatory variables", the nature of these depending on the nature of the study being undertaken.

In response to the challenges posed by the need for model simplification, the recent literature documents a wide range of approaches to reducing the computational expense of simulating natural and man-made systems. Strategies include model emulation (or "metamodelling") (e.g., Kennedy and O'Hagan, 2001; Oakley and O'Hagan, 2002; Sivakumar, 2008; Young and Ratto, 2009, 2011; Castelletti et al., 2011; Stone, 2011), model "reduction" (e.g., Vermeulen et al., 2004, 2005, 2006; Cheng et al., 2011), and parameter upscaling (e.g., Farmer, 2002; Pachepsky et al., 2006; Gerritsen and Lambers, 2008; Mondal et al., 2010). Meanwhile, less formal simplification strategies involving parameter and/or process lumping have been applied as a matter of course in model design and deployment over many years. See, for example, Lewis and Walker (2002), Dripps and Bradbury (2007), Zhu and Sun (2009), Francés et al. (2010), Martínez-Santos and Andreu (2010), Andreu et al. (2011)

and Touhami et al., (2012), all of whom modelled recharge to regional groundwater systems, this being the context of the example model discussed in the present study.

While considerable effort has been devoted to seeking simplification strategies that reduce the computational burden of environmental simulation, few studies have explored the effects of simplification on a model's predictive performance. Deleterious repercussions of simplification can include the introduction of predictive bias, and a loss of ability to quantify the full range of uncertainty associated with a prediction of interest; the latter is a fundamental requirement of model usage in decision support (Freeze et al., 1990). Such studies are difficult to undertake, for they often require that a simplified model be paired with a more complex one, with the latter providing metrics by which the former's performance can be judged. Nevertheless, this approach was taken by Aanonsen (2008) and Scheidt et al. (2011) in the petroleum context, by Vrugt et al. (2004) and Schoups and Hopmans (2006) in the vadose zone context, and by Doherty and Christensen (2011) in the groundwater context. In most modelling contexts however, while the imperfect nature of model-based simulation is recognized, little or no attempt is generally made to quantify the effects of model imperfections on model predictive performance, for time and resources typically permit no such investigation.

The present study seeks to improve our understanding of the effects of model simplification by undertaking such an investigation. It uses as its starting point theory and techniques developed by Doherty and Christensen (2011), who provided a generalized mathematical characterization of the model simplification process. By characterizing simplification as the omission from a model of parameters and processes that prevail in the real world, they were able to apply subspace concepts in their analysis. They then characterized simplification induced model predictive error as arising from one or more of the following sources.

- Failure to represent parameter/process detail to which historical measurements of system state comprising the calibration dataset are sensitive.

- Failure to represent parameter/process detail to which predictions of interest are sensitive.

- The compensatory roles that parameters of a defective model are forced to play during the calibration process, and then continue to play when the model is used to make predictions.

The first of these represents a failure of the model calibration process to extract as much information from the calibration dataset as is available in that dataset. Because the (over-)simplified model provides no receptacles for such information, the post-calibration propensity for error of some model predictions may be higher than it needs to be, given the available dataset. The gap between the information content of the calibration dataset and the receptacles that the model provides to hold that information is expressed as simplification-induced model-to-measurement misfit; this is commonly referred to as "structural noise". Ideally, stochastic characterization of such noise would allow the effects of simplification to be at least partially included in the quantification of model predictive error. Methods such as those described by Kennedy and O'Hagan (2001), Cooley (2004), Cooley and Christensen (2006) and Cui et al. (2011) could be used for this purpose. However, Doherty and Welter (2010) point out that such analysis is likely to be hampered by the fact that the covariance matrix of structural noise is generally singular.

The second source of error identified by Doherty and Christensen (2011) represents a failure on the part of a simplified model to represent the so-called "null space" contribution to predictive uncertainty. This source of uncertainty arises from a sensitivity of model predictions to parameters and/or parameter combinations that are not inferable through the model calibration process. That is, it arises from simplifications that do not degrade a model's ability to replicate the past, but may compromise its ability to represent the future. In general, the magnitude of this term increases with the extent to which predictions of interest are different, or occur under different conditions, from those employed for model calibration.

The third of the above sources of error can promote predictive bias. Doherty and Christensen (2011) show that adjustment of parameters of an imperfect model to achieve a good fit between model outputs and members of the calibration dataset requires that some parameters assume roles that they were not necessarily designed to play. At the same time, null-space parameter components are unwittingly adjusted away from their pre-calibration expected values, a process that Doherty and Christensen (2011) refer to as "null-space entrainment". Predictions which are

sensitive to thus-adjusted null-space parameter components become biased as a result. Under certain circumstances this bias can dominate predictive error, engendering greater propensity for error in a model that has been calibrated than in a model that has not been calibrated at all. In contrast, if a prediction is entirely dependent on parameter combinations that are informed by the calibration dataset (i.e., so-called "solution space" parameter combinations), its propensity for predictive error may be significantly reduced by the model calibration process, regardless of model defects and regardless of the compensatory roles played by some model parameters during the calibration process and the degree of null-space entrainment endured by others. In general, this applies to predictions that are comprised of model outputs which are very similar in type and location to those used for model calibration.

This study extends the work of Doherty and Christensen (2011) in examining the theory and ramifications of model simplification in contexts where a model must be calibrated before being used in a predictive capacity. With some slight modification of Doherty and Christensen's (2011) original theory, simplification is viewed in the present study as parameter transformation and decomposition. The requirements of optimal transformation/decomposition are outlined, and the repercussions of suboptimal transformation/decomposition are described. The theory and concepts discussed herein are then illustrated using a relatively complex model of water movement through a heterogeneous soil profile built for the purpose of groundwater recharge estimation, together with two simplified versions of this same model.

It is salient to point out that while Vrugt et al. (2004) and Schoups and Hopman (2004) also addressed the issue of model simplification in the vadose zone context, the present study differs from these previous studies in that its particular focus is on the potential for error in simplified model parameters, and in predictions that are sensitive to them, that is incurred through the act of calibrating the simplified model. As such, it forms a useful complement to this previous work. It is also salient to point out that a significant difference between the example used in this study and that employed by Doherty and Christensen (2011) is that calibration of the simplified models used in the present example constitutes a well-posed inverse problem. Furthermore, one of the simplified models is over-simplified, as the fit between outputs of this model and measurements comprising the calibration dataset are contaminated by structural noise. The effects of such "over-simplification" on model predictive performance are examined. In a further development of the theory presented by Doherty and Christensen (2011), the

relationships that parameters of a simplified model have with parameters of a partnered complex model (and, by inference, to the hydraulic properties of reality itself) are examined. The degree of null-space entrainment engendered through adjustment of simplified model parameters and its effects on model predictions are also examined through linear analysis.

This chapter is organized as follows. Section 3.2 provides a brief review of the theory of simplification presented by Doherty and Christensen (2011). This theory is then extended to include the issue of optimal parameter transformation and the role of expert knowledge in seeking such optimality. Following that, transformations are developed through which the relationships between simplified model parameters and complex model parameters can be better understood. In section 3.3 we introduce a synthetic Richards equation-based model and two simplifications of it. In section 3.4 the paired model methodology of Doherty and Christensen (2011), in conjunction with theory presented in section 3.2, are applied to this suite of models as they are employed to make predictions of groundwater recharge. Section 3.5 presents a discussion of the outcomes of these analyses. Section 3.6 draws conclusions that are salient not only to the models that are discussed in this study, but to environmental models in general.

3.2 Concepts and theory

3.2.1 Introduction

The theoretical analysis of simplification presented herein rests on subspace concepts, whereby a simplified model is viewed as the outcome of a parameter transformation and decomposition process. One advantage of adopting such an approach is that, as we shall discuss, optimality of parameter transformation and decomposition (and hence of simplification) can, at least in principle, be defined. The success or otherwise of any particular simplification strategy can then be assessed according to this metric. It is important to point out that our analysis is not intended to constitute a mechanism for simplification and/or parameter upscaling that one would necessarily use in a real world context. It does, however, provide a means to understand and assess the outcomes of model simplification implemented in whatever way a modeller chooses.

Our analysis assumes that the relationship between model outputs and parameters employed by a model is linear, and hence can be represented as a matrix. This

assumption is violated by most models; however it allows the use of subspace methods in our analysis. This, in turn, exposes outcomes of the simplification process which would be otherwise difficult or impossible to explore. These outcomes are not diminished by model nonlinearity; rather they are made more complex. Given that the intentions of our study are to expose and explore the general nature of these outcomes rather than their details in any specific modelling context, our analysis is not invalidated by the nonlinear nature of most models. Nevertheless, some numerical experiments were carried out to address this issue, and to thereby ensure the integrity of the conclusions drawn in the examples section of this chapter; details are provided in section 4.6.

The following subsections provide a brief review of aspects of linear analysis that are salient to our analysis of model simplification.

3.2.2 Linearization concepts

3.2.2.1 General

As stated above, to facilitate the application of subspace concepts and theory, a linear relationship between environmental process outputs and parameters pertaining to those processes is assumed. Reality, and any model that simulates it, are thus represented as matrices operating on parameters; the latter representing properties of a system. For ease of analysis we consider reality to be a very complex model (herein referred to as the "reality model"), and numerical simulators of reality to be simpler models that strive to provide the same outputs under the same conditions. In the text that follows we therefore make repeated reference to a "reality model" as our starting point for examining the effects of simplification, the latter being a necessary accompaniment of any attempt to simulate reality. Though a "reality model" does not actually exist (for only reality itself exists) we have retained this terminology in the following text in preference to the term "complex model" to depict the starting point for our analysis in order to reinforce the concept that all models, even the most complex, are gross simplifications of reality. Hence any model, no matter how complex, is subject to the same phenomena that we discuss below when parameters of that model are adjusted in order to ensure that its outputs better match the observed behaviour of the real world.

For convenience in the analysis that follows, parameter values are formulated as perturbations from their expert knowledge-based expected values; model outputs are

treated in the same manner. Thus parameter values of zero give rise to model output values of zero. Adoption of this protocol reduces the complexity of the following equations. It also sets the mathematical context for optimal usage of subspace concepts in the model calibration process, namely that parameters (or combinations of parameters) should be assigned values of zero (and hence be informed by expert knowledge alone) unless information within the dataset supports estimation of these parameters (or parameter combinations).

In accordance with the protocols just described, let **k** (a vector) denote the hydraulic properties of a real world system, or equivalently the parameters used by a "reality model" which simulates that system perfectly. Let **Z** represent the action of that model under calibration conditions, and let **h** represent the calibration dataset. The latter is contaminated by measurement noise ε so that:

$$\mathbf{h} = \mathbf{Z}\mathbf{k} + \boldsymbol{\varepsilon} \tag{3.1}$$

3.2.2.2 The null space

The matrix **Z** of reality will normally have many more columns than rows, for reality is heterogeneous and complex, and its parameters are many. Unique estimation of these parameters from a calibration dataset is not possible. A parallel concept to that of parameter nonuniqueness is that of the null space. By definition, a non-zero parameter set $\mathbf{k}_n$ belongs to the null space of **Z** if:

$$\mathbf{0} = \mathbf{Z}\mathbf{k}_n \tag{3.2}$$

Suppose that a parameter set $\underline{\mathbf{k}}$ can be found that fits the calibration dataset perfectly. Then:

$$\mathbf{h} = \mathbf{Z}\underline{\mathbf{k}} \tag{3.3}$$

By adding equation (3.2) to equation (3.3) the nonuniqueness of $\underline{\mathbf{k}}$ in the face of the existence of a null space is demonstrated.

Matrices that have more rows than columns can also possess a null space. However, a matrix with more columns than rows will surely possess a null space. In many modelling contexts the purpose of model simplification is to reduce the number of parameters employed by an existing model so that the null space is eliminated, thereby promulgating uniqueness of its calibration. However, calibration uniqueness can also

be achieved mathematically (and optimally, as will be explained) through the process of singular value decomposition (SVD).

3.2.2.3 Singular value decomposition

Through SVD, any matrix **Z** can be decomposed as:

$$\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}^t \tag{3.4}$$

where **U** and **V** are orthonormal square matrices whose columns are unit vectors which span the output and parameter spaces of **Z** respectively. **S** is a matrix with diagonal elements, referred to as "singular values", ordered from highest to lowest and all of which are positive or zero. Partitioning of **S** on the basis of zero and non-zero singular values leads to concordant partitioning of **U** and **V**. Applying subscripts 1 and 2 to the partitions that correspond to non-zero and zero singular values respectively, equation (3.4) becomes:

$$\mathbf{Z} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}^t_1 + \mathbf{U}_2\mathbf{S}_2\mathbf{V}^t_2 = \mathbf{U}_1\mathbf{S}_1\mathbf{V}^t_1 \tag{3.5}$$

In practice, singular values that are close to zero, in addition to those that are exactly zero, are relegated to $\mathbf{S}_2$ in order to prevent "overfitting" (whereby a model is forced to reproduce characteristics of the calibration dataset that are more likely to represent measurement error rather than true system behaviour). The solution to the inverse problem of model calibration found through SVD is given by (see, for example, Aster et al., 2005):

$$\underline{\mathbf{k}} = \mathbf{V}_1\mathbf{S}^{-1}_1\mathbf{U}^t_1\mathbf{h} \tag{3.6}$$

The fact that this amounts to a form of parameter simplification is demonstrated by pre-multiplying both sides of equation (3.6) by $\mathbf{V}^t_1$ to yield:

$$\underline{\boldsymbol{\alpha}} = \mathbf{V}^t_1\underline{\mathbf{k}} = \mathbf{S}^{-1}_1\mathbf{U}^t_1\mathbf{h} = \mathbf{S}^{-1}_1\boldsymbol{\varphi} \tag{3.7a}$$

where:

$$\boldsymbol{\varphi} = \mathbf{U}^t_1\mathbf{h} \tag{3.7b}$$

$\underline{\boldsymbol{\alpha}}$ is a vector comprising estimates of the scalar projections of the real world parameter set **k** onto each of the orthogonal unit vectors $\mathbf{v}_{1i}$ comprising the columns of $\mathbf{V}_1$. Collectively these orthogonal unit vectors span the parameter solution space; this is

the orthogonal complement of the null space – orthogonal because the projection of one of these subspaces onto the other is zero or, in more colloquial terms, because there is no overlap between them. The smaller is the dimensionality of the solution space, the fewer of these scalar projections are estimated, for the dimensionality of the solution space is defined as the number of columns comprising the $\mathbf{V}_1$ matrix. Meanwhile, parameter projections onto the $\mathbf{v}_{2i}$ vectors which span the null space are not estimated. These projections therefore retain their pre-calibration values of zero.

Projections of parameters onto the $\mathbf{v}_i$ vectors can be considered to be linear combinations of the original parameter set **k**. The ratios in which these parameters are combined are given by the elements of each $\mathbf{v}_i$ vector. The calibration process thus effectively provides estimates for multipliers pertaining to some of these combinations (i.e., parameter combinations belonging to the solution space), while multipliers for other parameter combinations are assigned a value of zero as the calibration dataset provides insufficient information for their estimation. At the same time, because of the diagonal status of $\mathbf{S}^{-1}{}_1$, each element of $\underline{\boldsymbol{\alpha}}$ is calculated directly from its corresponding element of **φ** through multiplication by the corresponding element of $\mathbf{S}^{-1}{}_1$. The i'th element of **φ** is the scalar projection of the observation dataset onto the i'th column of **U**, the latter being denoted as $\mathbf{u}_i$. In a similar fashion to $\mathbf{v}_i$ for parameters, each $\mathbf{u}_i$ contains coefficients that combine observations in a linear manner. Equation (3.7) thus states that the *i*'th combination of observations expressed by $\mathbf{u}_i$ is uniquely and entirely informative of the *i*'th combination of parameters expressed by $\mathbf{v}_i$. The reader is referred to texts such as Aster et al. (2005) for further details.

3.2.2.4 Optimal calibration

Let *s* (a scalar) be a prediction made by the reality model. Let the vector **y** denote the sensitivity of this prediction to parameters **k** of the reality model. Then:

$$s = \mathbf{y}^t\mathbf{k} \tag{3.8a}$$

When the prediction is made using the calibrated model, it is calculated as:

$$\underline{s} = \mathbf{y}^t\underline{\mathbf{k}} \tag{3.8b}$$

It can be shown (see Moore and Doherty, 2005) that the error variance of the prediction made by the calibrated model is:

$$\sigma^2_{\underline{s}\text{-}s} = \mathbf{y}^t\mathbf{V}_2\mathbf{V}^t{}_2C(\mathbf{k})\mathbf{V}_2\mathbf{V}^t{}_2\mathbf{y} + \mathbf{y}^t\mathbf{V}_1\mathbf{S}^{-1}{}_1C(\boldsymbol{\varepsilon})\mathbf{S}^{-1}{}_1\mathbf{V}^t{}_1\mathbf{y} \quad (3.9)$$

where C(**ε**) is the covariance matrix of measurement noise. C(**k**) is the covariance matrix associated with the prior probability distribution of parameters **k**. As such it is an encapsulation of expert knowledge.

Let us suppose that C(**k**) and C(**ε**) can be expressed as follows:

$$C(\mathbf{k}) = \sigma^2_k\mathbf{I} \quad (3.10a)$$

$$C(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon\mathbf{I} \quad (3.10b)$$

The first of these equations states that, in terms of expert knowledge, all parameters are independently and equally variable, with no statistical correlation between them. The second states that the errors associated with measurements comprising the calibration dataset are also independent and of equal magnitude for all measurements. If the conditions given by equation (3.10) are met then equation (3.9) becomes:

$$\sigma^2_{\underline{s}\text{-}s} = \sigma^2_k\mathbf{y}^t\mathbf{V}_2\mathbf{V}^t{}_2\mathbf{y} + \sigma^2_\varepsilon\mathbf{y}^t\mathbf{V}_1\mathbf{S}^{-2}{}_1\mathbf{V}^t{}_1\mathbf{y} \quad (3.11)$$

The first term on the right of equation (3.11) falls monotonically as the number of singular values that are assigned to the solution space increases, whilst the second term rises monotonically at the same time. Meanwhile, the sum of these terms falls from its pre-calibration value (equal to pre-calibration predictive uncertainty) if no singular values are retained, achieves a minimum value at some number of positive singular values, and then rises, approaching infinity as the magnitude of singular values comprising the diagonal elements of $\mathbf{S}_1$ approaches zero. The minimum value of the predictive error variance curve defines the optimum number of singular values to employ in calibrating the model.

Optimality of calibration can also be viewed from a parameter, as well as from a predictive, point of view. As the dimensionality of the solution space is increased, the error variance of the α_i scalars comprising the elements of the $\underline{\boldsymbol{\alpha}}$ vector of equation (3.7) can be computed using a slight modification of equation (3.11). If the error variance of an α_i is greater after calibration than before calibration (where its propensity for error is based on expert knowledge alone), it should not be estimated, and the corresponding $\mathbf{v}_i$ vector should not be included in the solution space. In other words, parameters and parameter combinations which are not estimated through the

calibration process should be assigned values based on expert knowledge alone, for this endows such parameters and parameter combinations with less potential for error than that which they would accrue through the calibration process.

Satisfaction of equations (3.10a) and (3.10b) is important for achieving optimality of calibration through minimization of predictive error variance through selection of the appropriate number of singular values to employ in the calibration process. Where equation (3.10a) in particular is not met, and especially where C(**k**) has off-diagonal elements, it is easy to find cases where a graph of $\sigma^2_{\underline{s}\text{-}s}$ versus number of singular values rises before it falls. Note also that the number of singular values used in estimation of model parameters can be taken as a measure of calibration-induced simplification. The use of a small number of singular values implies a small dimensionality of the solution space, and hence a high degree of simplification.

3.2.2.5 Optimal parameter transformation (the Karhunen-Loève transform)

Rarely will expert knowledge be such that equation (3.10a) automatically holds. However, conceptually at least, it can be achieved through appropriate parameter transformation.

Let the matrices **F** and **E** (the former being orthonormal and the latter being diagonal) be defined through the following equation in which the necessarily positive definite symmetric matrix C(**k**) is subject to SVD:

$$C(\mathbf{k}) = \mathbf{F}\mathbf{E}\mathbf{F}^t \tag{3.12}$$

Now let the vector **m** be defined as:

$$\mathbf{m} = \mathbf{E}^{-1/2}\mathbf{F}^t\mathbf{k} \tag{3.13a}$$

so that, by pre-multiplication of both sides of equation (3.13a) by $\mathbf{E}^{1/2}$ and **F**:

$$\mathbf{k} = \mathbf{F}\mathbf{E}^{1/2}\mathbf{m} \tag{3.13b}$$

Using standard matrix relationships for propagation of covariance it is easily shown that:

$$C(\mathbf{m}) = \mathbf{I} \tag{3.14}$$

Comparing equation (3.14) with the condition for optimal calibration expressed by equation (3.10a), it follows that parameter estimation should take place in **m**-space rather than **k**-space if it is to achieve a minimum error variance status for estimated parameters and for predictions which depend on them. From equations (3.1) and (3.13b):

$$\mathbf{h} = \mathbf{Zk} + \boldsymbol{\varepsilon} = \mathbf{ZFE}^{1/2}\mathbf{m} + \boldsymbol{\varepsilon} = \mathbf{Ym} + \boldsymbol{\varepsilon} \tag{3.15}$$

where:

$$\mathbf{Y} = \mathbf{ZFE}^{1/2} \tag{3.16}$$

3.2.3 Simplification and subspaces

3.2.3.1 Simplification strategies

Strategies through which complex and heterogeneous natural systems are represented in a numerical model are often based on notions of averaging and/or fixing. In a groundwater model, for example, many facies may be simulated as a single layer; parameters assigned to that single layer are hydraulic properties vertically averaged over those facies. Similarly, horizontal spatial hydraulic property averaging is required in order to assign parameters to the (possibly large) cells or elements used by a regional numerical model.

The process of model structure simplification has much in common with the process of parameter simplification that is often undertaken prior to model calibration in order to achieve well-posedness of the resulting inverse problem. The latter involves the fixing of some parameters at expert knowledge-informed values and the amalgamation of others so that average properties, rather than parameterization detail, are subject to estimation. As discussed above, optimality of parameter simplification required for model calibration can be achieved through SVD following appropriate parameter transformation. This too can be viewed as the process of fixing certain parameters and parameter combinations at "known" values (these being parameters and parameter combinations which lie entirely within the null space) while estimating a limited number of "averaged" parameters. The "averaging coefficients" (i.e., the elements of the $\mathbf{v}_i$ vectors comprising the columns of the $\mathbf{V}_1$ matrix of equation (3.5)) are defined in a manner that guarantees orthogonality to null-space parameter components and

achieves a minimum error variance status for averaged parameters thus estimated, and for predictions which depend on them.

3.2.3.2 Optimal model simplification

Doherty and Christensen (2011) addressed the concept of optimality of model simplification through analysing model simplification in linear subspace terms. They showed that simplification can be viewed as a kind of parameter decomposition, this resulting in a set of actual and/or notional parameters which are "included" in the simplified model, together with a complimentary set of parameters which are "omitted" from this model. Under the assumption that the simplified model must be calibrated as part of its field deployment, they demonstrated that a necessary condition for achieving optimality of model simplification is that the parameter space decomposition implied by simplification be an orthogonal decomposition, and that the outcomes of this decomposition process resemble, as much as possible, that implied by transformation according to equations (3.13a) and (3.13b), followed by SVD of the resulting **m** parameter space.

Doherty and Christensen (2011) characterized optimal model simplification as that which leads to predictions of minimized error variance. (Note that this error variance may be far from zero, as it is bounded from below by the innate uncertainty associated with model predictions given all available information pertaining to the system of interest.) They showed that where simplification is not in accordance with the transformation and decomposition process described above, a consequence may be inadvertent adjustment of parameter components that properly belong to the null space as the simplified model is calibrated. Predictions that are sensitive to thus entrained null-space parameter components will be biased and will therefore fail to achieve minimum error variance status. Such simplification is therefore suboptimal. Simultaneously with null-space entrainment certain model parameters and/or parameter combinations assume surrogate roles as they compensate for model defects while allowing model outputs to fit the calibration dataset; this further contributes to potential predictive bias. (A simple mathematical demonstration of calibration-induced null-space parameter entrainment incurred through failure to comply with optimality of parameter transformation as described by equations (3.13a) and (3.13b) is provided in Appendix A.)

Doherty and Christensen (2011) further showed that the adverse effects of a suboptimal simplification strategy are prediction-specific. Where a prediction is similar in nature to data comprising the calibration dataset (and therefore is sensitive solely to parameter combinations occupying the solution space), optimality of simplification as far as that prediction is concerned requires only that a model be capable of replicating historical system behaviour well; a recognizably physical basis for its parameters is of secondary importance. Alternatively, where a model is required to make predictions under different conditions, or of a different type, to the observations which comprise the calibration dataset (as is often the case), its design must be such that these different conditions can indeed be simulated, and that its ability to make such predictions with minimized error variance is enhanced, rather than eroded, by the process of model calibration. The alignment of calibration and simplification subspaces discussed above is important in achieving this.

3.2.3.3 Paired model analysis

If model simplification approaches optimality, then all predictions made by the model after it has been calibrated are of minimized error variance regardless of their degree of solution and null space dependence. However, lack of representation of real world null-space parameter components in the simplified model may preclude the possibility of exploring the error variance associated with its predictions, and hence of quantifying their uncertainties. Doherty and Christensen (2011) propose a methodology for model predictive uncertainty analysis that involves conjunctive use of a simplified and complex model in order to overcome this problem. At the same time, this methodology allows identification of, and correction for, calibration-induced predictive bias. Although requiring construction of a complex model for use in conjunction with the simplified model, a benefit of this approach is that the complex model does not require calibration – an undertaking which may be hampered by long run times and/or numerical instability of the model. Moreover, use of a physically based complex model provides the means through which expert knowledge pertaining to a particular study site can be best expressed.

The methodology is as follows.

- Generate many different expert knowledge-based stochastic realizations of a complex model and its parameters, with these realizations including those

aspects of the system that are likely to contribute most to the uncertainty of predictions of interest. Obtain a suite of such predictions from the stochastic model realizations. Let each such prediction be referred to as s. Additionally, compute complex model outputs that correspond to observations comprising the available calibration dataset.

- For each realization of the complex model, calibrate a simplified model against those complex model outputs that correspond to members of the calibration dataset. Then make the prediction of interest using each calibrated simplified model. Let these predictions be referred to as $\underline{s}$.

- Produce a scatterplot of s-versus-$\underline{s}$. A regression (best-fit) line through the scatterplot can be used to correct simplified model predictions for simplification and calibration-induced predictive bias. Meanwhile, scatter about the line of best fit quantifies predictive uncertainty (this often being dominated by the sensitivity of a prediction to null-space parameter components that are not represented in the simplified model).

- Calibrate the simplified model against the real world dataset. On the basis of information available from the s-versus-$\underline{s}$ plot, correct this prediction for bias and quantify its uncertainty.

The described paired model analysis, whilst presented by Doherty and Christensen (2011) as a practical methodology for predictive uncertainty quantification and bias correction, also serves as a metric by which the success, in terms of predictive performance, of a given simplification approach can be judged. It is employed in the present study for this purpose.

3.2.4 Relationships between complex and simplified model parameters

In accordance with the approach taken by Doherty and Christensen (2011) we express the effects of simplification as the omission of parameters, and the processes that operate on them, from a complex "reality model" in order to derive the actual numerical model that we use in place of reality. Thus any model that we use to simulate reality can be viewed as possessing a suite of "visible" parameters, together with a set of implied "invisible" parameters. The latter specify corrections that should be made to any aspect of the physical, numerical and/or parameter structure of the model that

would allow that model to become a perfect replica of the real world and the processes that are operative therein. This is, of course, a simplistic notion; indeed any analysis of simplification will itself be a simplification. However, as will be demonstrated below, this conceptualization of simplification allows us to gain some important insights into the effects of the simplification process that would not otherwise be so clearly visible.

Let the reduced number of parameters employed by a simplified model of the study site described by **Z** be represented by the vector **p**. Let the model which acts on this reduced set of parameters be designated as **X**. Then, under calibration conditions, from equation (3.1):

$$\mathbf{h} = \mathbf{Xp} + (\mathbf{Zk} - \mathbf{Xp}) + \boldsymbol{\varepsilon} \tag{3.17}$$

The term (**Zk** – **Xp**) can be viewed as structural error (or sometimes "structural noise" when its presence becomes apparent through attempts to calibrate the model) as it represents simplification-induced model-to-measurement misfit. As discussed above, analysis of this can prove difficult for a number of reasons. Furthermore, if a simplified model fits the calibration dataset well, this term may be very small (or even non-existent). Hence it is often more fruitful to examine the ramifications of simplification on the calibration process from the point of view of its effect on parameters rather than its effects on model outputs; see, for example, Vrugt (2005), Kavetski et al. (2006a, 2006b) and Kuczera et al. (2006). Following Doherty and Christensen (2011) we write:

$$\mathbf{Zk} - \mathbf{Xp} = \mathbf{Z}_o\mathbf{k}_o \tag{3.18}$$

where the subscript "o" stands for "omitted". $\mathbf{k}_o$ represents parameters omitted from the "reality" model in building the simplified model, while $\mathbf{Z}_o$ represents the omitted processes which operate on them. Equation (3.17) therefore becomes:

$$\mathbf{h} = \mathbf{Xp} + \mathbf{Z}_o\mathbf{k}_o + \boldsymbol{\varepsilon} \tag{3.19}$$

The simplified model parameter set **p** can be characterized as being derived from the complex model parameter set **k** through a decomposition operation. If this is denoted by the matrix **L** then:

$$\mathbf{p} = \mathbf{Lk} \tag{3.20}$$

so that from equation (3.17):

$$\mathbf{h} = \mathbf{XLk} + (\mathbf{Z} - \mathbf{XL})\mathbf{k} + \boldsymbol{\varepsilon} \quad (3.21)$$

To simplify the following analysis we will assume that the model simplification process is such that **X** is of full rank, and that estimation of **p** from the calibration dataset therefore constitutes a well-posed inverse problem. If we further assume for simplicity that measurement noise is zero, and then provide all observations with the same weight, a value of **p** can be obtained for any **k** through calibrating the simplified model against a calibration dataset generated by the complex model. Thus:

$$\mathbf{p} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{h} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Zk} = \mathbf{Lk} \quad (3.22)$$

Equation (3.22) provides the relationship between **p**-space, the parameter space of the simplified representation of the real world that is the **X** model, and **k**-space, the parameter space of the reality model (which we characterize as the **Z** model). Through use of this relationship, the composition of any simplified model parameter in terms of reality model parameters can be established.

While simplified models are often abstractions of reality, their designers often state that their parameters can be informed by expert knowledge; in fact they are often built specifically with this in mind. These considerations apply particularly to the lumped parameter soil moisture store models that form the basis of many regional rainfall/runoff/recharge simulators; indeed such a model is examined later in this study. Equation (3.22) can be used to examine whether any particular simplified model parameter does indeed perform the function that it was designed to perform. Furthermore, the extent to which expert knowledge should be respected in terms of the degree of simplified model parameter variability allowed during calibration can be judged by computing the prior covariance matrix C(**p**) of the simplified model parameter set. From equation (3.22), this can be calculated from that of the real-world parameter set using standard relationships for propagation of covariance as:

$$\mathrm{C}(\mathbf{p}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z}\mathrm{C}(\mathbf{k})\mathbf{Z}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \quad (3.23)$$

We note that this equation is similar to that derived by Gallagher and Doherty (2007b).

Cooley and Christensen (2006) discuss the special case where model simplification is undertaken through assuming spatial parameter uniformity in place of heterogeneity while retaining all other computational aspects of the complex model in the simplification process. (This corresponds to one of the examples presented herein.)

They show that in this case the difference between estimated properties **p** and spatially averaged properties of the complex model, denoted as $\mathbf{p}_*$, is given by:

$$\mathbf{p} - \mathbf{p}_* = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z}(\mathbf{I} - \boldsymbol{\gamma}(\boldsymbol{\gamma}^t\boldsymbol{\gamma})\boldsymbol{\gamma}^t)\mathbf{k} \tag{3.24a}$$

The matrix $\boldsymbol{\gamma}$ is defined by the equation:

$$E(\mathbf{k}) = \boldsymbol{\gamma}\underline{\mathbf{p}} \tag{3.24b}$$

where:

$$\underline{\mathbf{p}} = E(\mathbf{p}) \tag{3.24c}$$

with $E(\,)$ being the expected value operator. They show for a linear model that over many realizations:·

$$E(\mathbf{p} - \mathbf{p}_*) = E(\mathbf{p} - \underline{\mathbf{p}}) = \mathbf{0} \tag{3.25}$$

However for any one realization **p** and $\mathbf{p}_*$ will generally differ, as equation (3.24) shows. Cooley and Christensen (2006) also show that the discrepancy between **p** and $\mathbf{p}_*$ can be reduced, though not eliminated, by employing an empirically determined weighting matrix instead of measurement weights in estimating **p**. This empirically determined weighting matrix is an estimate of the inverse of the total error covariance matrix, where the total error is the sum of observation error and structural error caused by the model's neglect of spatial heterogeneity.

3.2.5. Back-transformation to complex model parameter space

To conform with nomenclature introduced above we will continue to employ **k** to represent a "reality" parameter set, or the parameter set employed by an equivalent complex model; **Z** represents the action of that model. We will assume, however, for the sake of simplicity in development of the theory, that equation (3.10a) holds (this may be either automatically or through appropriate parameter transformation) so that optimality of simplified parameterization can be achieved through SVD of **Z**.

As decomposition of **k** to **p** involves parameter reduction, it is not possible to find a unique back-transformation from **p** to **k**. However, it is possible to find a unique transformation from **p** to the solution space of **Z**. This follows from the fact that, for a given **h**, **p** is unique and the projection of **k** onto its solution space is unique. This transformation is accomplished by seeking that $\underline{\mathbf{k}}$ (where the underscore signifies

solution space occupancy) which provides the same **h** as **Xp**. We will refer to this back-transformed parameter set as $\underline{\mathbf{k}}_p$ and define the transformation as **N**. Thus, from equation (3.6):

$$\underline{\mathbf{k}}_p = \mathbf{Np} = \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{Xp} \tag{3.26}$$

where $\mathbf{V}_1$, $\mathbf{S}_1$ and $\mathbf{U}_1$ are defined through SVD of **Z**. Using the specification for **L** provided by equation (3.22), equation (3.21) can be expanded as:

$$\mathbf{h} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Zk} + (\mathbf{Z} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z})\mathbf{k} + \boldsymbol{\varepsilon} \tag{3.27}$$

We now introduce the identity (Aster et al., 2005):

$$\mathbf{V}_1\mathbf{V}^t{}_1 + \mathbf{V}_2\mathbf{V}^t{}_2 = \mathbf{I} \tag{3.28}$$

where $\mathbf{V}_1$ and $\mathbf{V}_2$ are defined through equations (3.4) and (3.5), with partitioning taking place according to minimization of $\sigma^2_{\underline{s}\text{-}s}$ through equation (3.11). Hence, from equation (3.1):

$$\mathbf{h} = \mathbf{Zk} + \boldsymbol{\varepsilon} = \mathbf{ZV}_1\mathbf{V}^t{}_1\mathbf{k} + \mathbf{ZV}_2\mathbf{V}^t{}_2\mathbf{k} + \boldsymbol{\varepsilon} = \mathbf{Z}\underline{\mathbf{k}}_i + \mathbf{Zk}_n + \boldsymbol{\varepsilon} \tag{3.29}$$

where we define $\underline{\mathbf{k}}_i$ as the "ideal" value of the calibrated parameter set. As such, it is the projection of the reality parameter vector **k** onto a solution space whose dimensions are those determined by minimization of predictive error variance of equation (3.11). It is thus calculable as:

$$\underline{\mathbf{k}}_i = \mathbf{V}_1\mathbf{V}^t{}_1\mathbf{k} \tag{3.30}$$

Meanwhile $\mathbf{k}_n$ is orthogonal to $\underline{\mathbf{k}}_i$. This includes null space components of **k**, as well as components of **k** that are relegated to the null space as they are not worth estimating because their potential for error will be greater after calibration than before calibration. From equations (3.5) and (3.30) and the orthonormality of the vectors comprising **V**:

$$\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{Zk} = \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{Z}\underline{\mathbf{k}}_i = \underline{\mathbf{k}}_i \tag{3.31}$$

If both sides of equation (3.29) are now pre-multiplied by $\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1$ and (3.31) is substituted into the right side we then obtain:

$$\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{h} = \underline{\mathbf{k}}_i + \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{ZV}_2\mathbf{V}^t{}_2\mathbf{k} + \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\boldsymbol{\varepsilon} \tag{3.32}$$

Through SVD of **Z** and using the orthonormality of **V** and **U**, it is easy to show that the second term on the right of equation (3.32) is zero. Thus:

$$\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{h} = \underline{\mathbf{k}}_i + \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\boldsymbol{\varepsilon} \tag{3.33}$$

If both sides of equation (3.27) are now pre-multiplied by $\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1$, we obtain, with the help of equations (3.22) and (3.26):

$$\mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{h} = \underline{\mathbf{k}}_p + \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{Z}\mathbf{k} + \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\boldsymbol{\varepsilon} \tag{3.34}$$

The matrix $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is an orthonormal matrix spanning the range space of the simplified model **X**. $(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)$ spans the orthogonal complement of this. If this includes only system outputs calculated on the basis of parameter combinations corresponding to singular values that are below the singular value cutoff threshold, orthogonality of these outputs to $\mathbf{U}_1$ guarantees that the second term on the right of equation (3.34) is zero. However if it includes any outputs that have a non-zero projection onto the $\mathbf{U}_1$ subspace, this term will not be zero. In other words, the second term of equation (3.34) describes structural noise created by an inability of the simplified model **X** to fit those aspects of the system response (encapsulated in $\mathbf{U}_1$) that are considered to be worth fitting from a parameter estimation point of view. Ideally, the design of a simplified model should be such as to reduce this term to as close to zero as possible. To the extent that this is accomplished, a comparison of equation (3.34) with equation (3.33) reveals that $\underline{\mathbf{k}}_p$ approaches $\underline{\mathbf{k}}_i$. Calibration of the simplified model thus achieves an effective real-world parameter set that has the same projection onto the real-world solution space as would have been achieved if the reality model itself were calibrated in an ideal manner.

From equations (3.22) and (3.26) the relationship of $\underline{\mathbf{k}}_p$ to real-world parameters **k** is calculable as:

$$\underline{\mathbf{k}}_p = \mathbf{V}_1\mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z}\mathbf{k} \tag{3.35}$$

Equation (3.35) describes only the solution space projection of the effective real-world parameter set achieved through calibration of the simplified model. Let the vector $\underline{\boldsymbol{\beta}}$ contain the scalar projections of $\underline{\mathbf{k}}_p$ into each of the axes of parameter space defined through SVD of the reality model matrix **Z**. Then:

$$\underline{\boldsymbol{\beta}} = \mathbf{V}^t{}_1\underline{\mathbf{k}}_p = \mathbf{S}^{-1}{}_1\mathbf{U}^t{}_1\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z}\mathbf{k} \tag{3.36}$$

Meanwhile, the true (reality model) set of solution space projections $\boldsymbol{\beta}$ can be calculated as:

$$\boldsymbol{\beta} = \mathbf{V}^t_1\mathbf{k} \tag{3.37}$$

Errors in these projections are therefore calculable as:

$$\underline{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{S}^{-1}_1\mathbf{U}^t_1\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z} - \mathbf{V}^t_1)\mathbf{k} \tag{3.38}$$

Given the assumption that equation (3.10a) holds and thus the covariance matrix of $\mathbf{k}$ (i.e., C($\mathbf{k}$)) is $\mathbf{I}$, the covariance matrix of solution-space parameter projection error can be calculated from equation (3.38) as:

$$C(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{S}^{-1}_1\mathbf{U}^t_1\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z} - \mathbf{V}^t_1)(\mathbf{S}^{-1}_1\mathbf{U}^t_1\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z} - \mathbf{V}^t_1)^t \tag{3.39}$$

If a simplified model is capable of fitting a noise-free calibration dataset perfectly, C($\underline{\boldsymbol{\beta}} - \boldsymbol{\beta}$) is $\mathbf{0}$, indicating achievement of correct real-world solution-space parameter projections through calibration of the simplified model. However, through unavoidable, simultaneous adjustment of real-world null-space parameter components as simplified model parameters are adjusted, these correct real-world solution-space parameter projections may be accompanied by non-zero (and hence biased) real-world null-space parameter projections. The propensity for this to occur can be calculated using an appropriately modified version of equation (3.39) as:

$$C(\underline{\boldsymbol{\beta}}_n) = (\mathbf{S}^{-1}_2\mathbf{U}^t_2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z})(\mathbf{S}^{-1}_2\mathbf{U}^t_2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z})^t \tag{3.40}$$

where

$$\underline{\boldsymbol{\beta}}_n = \mathbf{V}^t_2\underline{\mathbf{k}}_p = \mathbf{S}^{-1}_2\mathbf{U}^t_2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Z}\mathbf{k} \tag{3.41}$$

Equivalent to equation (3.37), the true set of null-space parameter projections $\boldsymbol{\beta}_n$ is given by:

$$\boldsymbol{\beta}_n = \mathbf{V}^t_2\mathbf{k} \tag{3.42}$$

However, these can never be known.

The presence of non-zero elements of C($\underline{\boldsymbol{\beta}}_n$) represents suboptimality of the model simplification process, as it implies that parameter components that are not inferable from the calibration dataset have been interjected into the effective parameter set of

the reality model through calibration of the simplified model (i.e., null-space entrainment). Though such components may indeed be present in the real-world parameter set **k**, their calibrated values must be zero if minimum error variance status of all model predictions is to be achieved (rather than just those that are solution space-dependent). The $\mathbf{S}^{-1}{}_2$ term of equation (3.40) suggests that these unwanted terms may grow large as singular values get small. The error variance of some simplified model predictions may grow very large accordingly.

3.3 Synthetic case study – description

The theory and concepts developed above are now applied to a synthetic one-dimensional vadose zone example in which a model is to be built and calibrated for the purpose of transient groundwater recharge estimation. A "complex" model was developed, together with two different simplifications of this model. The latter differ in their degree of simplification, with one involving only parameter simplification and the other involving substantial process and structural simplification. These are now described.

3.3.1 Complex model

The complex model was constructed using HYDRUS-1D (Šimůnek et al., 2009). HYDRUS-1D simulates variably saturated flow in porous media by solving the one-dimensional Richards equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} K\left(\frac{\partial h}{\partial z} + 1\right) - S \tag{3.43}$$

where θ [L^3L^{-3}] is the volumetric water content, t [T] is time, h [L] is the pressure head, z [L] is the vertical coordinate, S [$L^3L^{-3}T^{-1}$] is the sink term and K [LT^{-1}] is the unsaturated hydraulic conductivity function. The latter is defined using the Mualem-van Genuchten (MVG) model for unsaturated soil hydraulic properties (van Genuchten, 1980), implemented using the following equations.

$$K(S_e) = K_s S_e^l [1 - \left(1 - S_e^{1/m}\right)^m]^2 \tag{3.44}$$

where:

$$S_e(h) = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \begin{cases} [1 + |\alpha h|^n]^{-m} & h < h_s \\ 1 & h \geq h_s \end{cases} \quad (3.45)$$

and:

$$m = 1 - 1/n \quad (3.46)$$

In equations (3.45) and (3.46), S_e [-] is the effective water content, θ_r [L^3L^{-3}] and θ_s [L^3L^{-3}] are the residual and saturated water contents respectively, α [L^{-1}] is the inverse of the air-entry pressure head h_s [L], n [-] is a pore-size distribution index, K_s [LT^{-1}] is the saturated hydraulic conductivity and l [-] is a pore-connectivity parameter.

The complex HYDRUS-1D model used in our study simulates water movement within a 10-layer vertical soil column of 500 cm depth. The top of the column is defined by an atmospheric boundary condition. This switches between prescribed head and prescribed flux depending on the pressure head at the soil surface; meanwhile any excess water on the soil surface is immediately removed as surface runoff. A seepage face condition comprises the lower boundary. Root water uptake is simulated using the Feddes (1978) model of water uptake reduction. This model enforces cessation of root water uptake below wilting point and close to saturation, with a linear transition to a constant, optimal uptake between these two extremes. Daily time series of precipitation and potential evapotranspiration over a period of 522 weeks spanning 1st January 1990 to 31st December 1999 serve as inputs to the model. These time series were measured at the WMO 06072 weather observation station in Ødum, Denmark. The first 285 weeks of these data were used during the calibration and predictive phases of model deployment. The entirety of this dataset was employed during a 522 week model warm-up period; see below. Only transpiration, with no evaporation, is assumed to occur in the simulated soil column.

One thousand stochastic realizations of soil column hydraulic properties were generated based on synthetic expert knowledge encapsulated in prior parameter probability distributions. Two levels of variability were employed in assigning values to different parts of the 1-D model domain. First, for each soil column, random values were generated for θ_s and θ_r based on normal distributions, and for K_s, α and n based on log-normal distributions. The means μ and standard deviations σ_1 of these distributions appear in the first two columns of Table 1. A random set of each

parameter type was then generated and assigned to the 10 layers comprising the model domain, each layer being of 50 cm thickness. (Log-)normal distributions were once again employed, with the mean of each distribution being the previously generated random value for each parameter type; standard deviations σ_2 for these secondary distributions appear in the third column of Table 1. This stochastic parameter generation process was repeated for each of the 1000 realizations. A single value of 100 cm for root depth was employed in all realizations. (Lack of numerical differentiability of model outputs with respect to this parameter precluded its inclusion in the linear analysis documented below.)

For all parameter set realizations, HYDRUS-1D was run in order to generate a calibration dataset, as well as three different predictions. Unfortunately, HYDRUS-1D did not converge for two of the 1000 parameter set realizations; thus the nonlinear analysis presented below is based on 998 realizations. In all cases the model was run for a 522 week warm-up period driven by the precipitation and potential evapotranspiration time series described above. The calibration dataset was assumed to consist of observations of total weekly drainage through the bottom boundary of the column for the next 230 weeks (i.e., weeks 523 to 752 of the simulation). Predictions were then made over weeks 753 to 807 of the simulation. The three different predictions considered in this study are (1) the total recharge summed over all of the 55 weeks comprising the prediction period, (2) the maximum recharge occurring during any 4-week interval within the prediction period, and (3) the maximum recharge occurring during any 1-week interval within the prediction period.

For ease of reference, the above model is referred to as "complex HYDRUS" hereafter.

Table 3.1. Statistical parameters used in generation of stochastic realizations of soil hydraulic properties employed by the HYDRUS-1D complex model. μ and σ_1 are the mean and standard deviation, respectively, for the first level of random parameter value generation, while σ_2 represents the standard deviation defining inter-layer parameter variability within one particular soil column.

Parameter	μ	σ_1	σ_2
Log(K_s) [cm/day]	2.03	0.5	0.1
θ_s [-]	0.41	0.05	0.01
θ_r [-]	0.065	0.02	0.004
Log(α) [-]	-1.12	0.5	0.1
Log(n) [-]	0.28	0.1	0.02

3.3.2 Simplified models

As mentioned above, two simplified models were employed in conjunction with complex HYDRUS. Both of these were driven by the same precipitation and evapotranspiration time series as complex HYDRUS. In obtaining the first simplified model, herein referred to as "simplified HYDRUS", only parameterization simplification was undertaken; the 10 layer heterogeneous soil column of complex HYDRUS was simply replaced by a column that is homogeneous in all parameters.

A lumped parameter "bucket" recharge model (herein referred to as LUMPREM, i.e., "LUMped Parameter REcharge Model") was employed as the second simplified model. Like the two HYDRUS models, LUMPREM works on a daily time step. Evapotranspirational losses E [L] from the soil moisture store are calculated using the equation:

$$E = f E_p \frac{1 - e^{-\gamma v'}}{1 - 2e^{-\gamma} + e^{-\gamma v'}} \tag{3.47}$$

where f [-] is a crop factor, E_p [L] is potential evapotranspiration, γ [-] is a shape parameter, and v' [-] is the relative volume of water in the bucket, i.e., V/V_{max}, where V [L^3] is the current volume of water in the bucket and V_{max} [L^3] is the total bucket volume.

Water lost as recharge to the groundwater domain (i.e., R [LT^{-1}]) is calculated as:

$$R = K_s [v']^l \left[1 - \left(1 - [v']^{1/m}\right)^m\right]^2 \tag{3.48}$$

where m [-] is a shape parameter. An additional parameter *rdelay* [T] defines the delay between water draining from the soil moisture store and the same water appearing as recharge to the groundwater system.

In conducting the numerical experiments discussed below, LUMPREM was run over the same time periods as were the HYDRUS models (including the 522 week warm-up period), and generated the equivalent calibration and predictive outputs.

3.3.3 Calibration and prediction

For each of the 998 realizations of the 50 stochastic parameters comprising the complex HYDRUS parameter set, both of the simplified HYDRUS and LUMPREM models were calibrated against the 230 weekly recharge observations comprising the calibration dataset as generated by the complex HYDRUS model. Equal weights were assigned to all recharge observations. For each of the simplified HYDRUS and LUMPREM models, five parameters were estimated using this calibration dataset. For simplified HYDRUS, these were θ_s, θ_r and the logs of K_s, α and n, whilst for LUMPREM, the logs of V_{max}, *rdelay*, K_s, m and f were estimated. (The logarithms, rather than native values, of most parameters were estimated in order to improve model linearity and also to provide a degree of parameter normalization; this being implicit in the estimation of the logs of parameters.)

Calibration of each of these simplified models against the 230 week recharge dataset constitutes a well-posed inverse problem. No random noise was added to the complex HYDRUS outputs in generating each calibration dataset; hence failure to achieve a perfect fit during the calibration of each simplified model is an outcome of model simplification alone. Calibration of LUMPREM was effected using the Gauss-Marquardt-Levenberg parameter estimation method through PEST (Doherty, 2016a), while calibration of the simplified HYDRUS model was accomplished using the adaptation of the CMA-ES (Covariance Matrix Adaptation Evolution Strategy) algorithm of Hansen and Ostermeier (2001) and Hansen et al. (2003) available through the PEST suite. While use of CMA-ES requires more model runs than that of PEST-implemented Gauss-Marquardt-Levenberg parameter estimation, it affords better protection against model output numerical granularity and entrapment in local optima.

Estimation of a set of LUMPREM parameters required approximately 20 s on a 3.07 GHz Intel Core i7 CPU. In contrast, calibration of a simplified HYDRUS model required several hours. It is also pertinent to note that attempts were made to calibrate the complex HYDRUS model (using regularized inversion to estimate its 50 adjustable parameters) against complex HYDRUS-generated observation datasets. However, this could not be achieved because HYDRUS numerical instability resulted in frequent model run failures during these attempted calibration processes. This exemplifies the difficulties that are often encountered in calibrating complex models, and therefore

illustrates the attractiveness of using a relatively simple model in the calibration process.

As stated above, the simplified HYDRUS and LUMPREM models were calibrated against all of the 998 datasets generated by the complex HYDRUS model, each calculated using a different stochastic parameter field. The outcomes of these 1996 calibration exercises were 998 simplified HYDRUS and LUMPREM counterparts to the 998 complex HYDRUS model realizations. The same predictions as those that were made with the complex HYDRUS models were then made using each of the calibrated simplified models.

3.3.4 Calculation of sensitivities

The linear analysis discussed below employs sensitivities of the 230 weekly recharge model outputs used for calibration purposes and the three outputs used for predictive purposes, to the adjustable parameters of the complex HYDRUS, simplified HYDRUS, and LUMPREM models. Complex HYDRUS and simplified HYDRUS sensitivities were evaluated using finite differences, with all parameters slightly perturbed from their expected values (see Table 3.1). For LUMPREM, a parameter set obtained through calibration of this model against a dataset generated using the expected values of complex HYDRUS model parameters was employed in finite difference sensitivity calculation.

3.4. Results

Outcomes of all analyses are now presented. They are discussed as they are presented. A more general discussion follows in section 3.5.

3.4.1 Quality of calibration

Figure 3.1 shows fits attained between simplified and complex model outputs for one particular realization of the complex HYDRUS parameter set. These are representative of the average fits of both simplified HYDRUS and LUMPREM across the 998 complex HYDRUS realizations. This figure demonstrates that weekly recharge values generated using a complex HYDRUS parameter field can be matched in nearly every detail by an appropriately parameterized simplified HYDRUS model. Given that the latter has only five parameters and that the former employs 50 parameters, if an attempt were made to calibrate a complex HYDRUS model against a calibration dataset of this

type, the dimensionality of the null space would be about 45, indicating non-identifiability of about 45 combinations of parameters on the basis of this dataset.

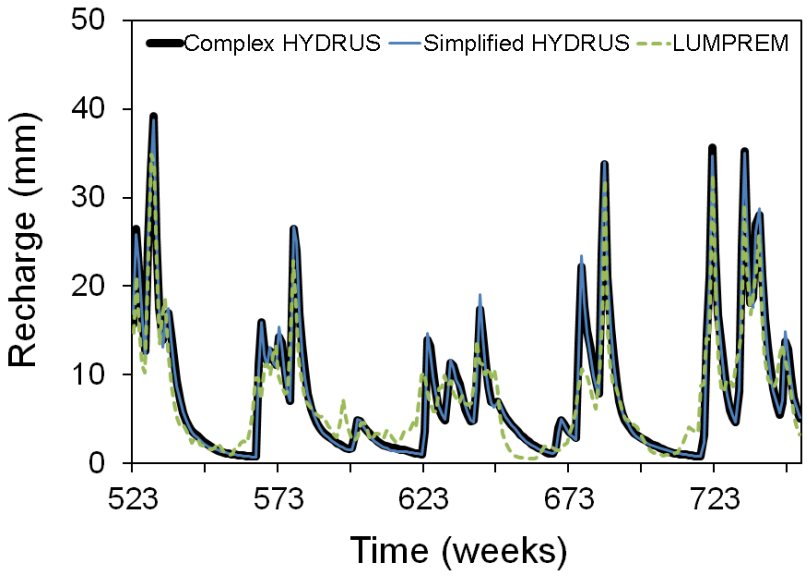

Figure 3.1. Example of the fits attained through calibration of both the simplified HYDRUS and LUMPREM models against complex HYDRUS weekly recharge outputs.

Like the simplified HYDRUS model, the LUMPREM model also employs five parameters. However, while adjustment of these five parameters allow a reasonable fit to be obtained between its outputs and those of the complex HYDRUS model, the fit is far from perfect. This indicates that its parameters do not span the full solution space of the complex HYDRUS model when calibration is undertaken against this particular dataset. Nevertheless, as is evident from Figure 3.1, the salient features of the recharge time series are generally reproduced. Such a fit may be considered acceptable in many real-world modelling contexts where measurement noise would contribute significantly to misfit.

3.4.2 Quality of predictions

As discussed above, predictions of three types were made on the basis of each complex HYDRUS parameter set realization, and then using the two calibrated simplified model counterparts to each such realization. The predictive abilities of the latter were assessed using *s*-versus-$\underline{s}$ scatterplots as described in section 3.3.3. These plots for the three predictions are shown in Figure 3.2. Note that, to enhance the linearity of these scatterplots, the logs of each predicted recharge quantity are employed in lieu of their native values. Regression lines through the scatterplots are calculated as:

$$s = a + b\underline{s} \tag{3.49}$$

where a and b are the regression intercept and slope respectively. 95% prediction intervals are also displayed in Figure 2 (see, for example, Draper and Smith, 1998, eq. 1.4.12 for details of prediction interval calculations). Table 3.2 lists regression parameters and statistics for the s-versus-$\underline{s}$ scatterplots of Figure 3.2.

Table 3.2. Regression coefficients and statistics pertaining to the s-versus-$\underline{s}$ scatterplots depicted in Figure 3.2. a and b are the regression coefficients of equation (49), r^2 is the coefficient of determination and σ is the standard deviation.

	Simplified HYDRUS				LUMPREM			
Prediction	a	b	r^2	σ	a	b	r^2	σ
Total recharge	0.001	1.003	0.995	0.007	-0.008	0.902	0.967	0.018
Max. 4-week recharge	-0.002	0.998	0.989	0.010	0.010	0.998	0.793	0.043
Max. 1-week recharge	-0.009	0.994	0.988	0.016	0.152	1.131	0.724	0.074

Equations describing the relationship between predictions made by a complex model and those made by a simplified model calibrated against a dataset generated by the former are derived in Doherty and Christensen (2011). These equations show that where a simplified model fits the calibration dataset well, vertical scatter of complex model predictions about the line of best fit appearing in plots such as those shown in Figure 3.2 records the null space contribution to predictive uncertainty. Horizontal scatter of simplified model predictions about this same line represents the contribution of measurement noise to predictive uncertainty. A regression line slope of less than unity indicates predictive bias incurred by calibration-induced null-space entrainment. Where a simplified model provides a less-than-perfect fit to complex model outputs, vertical scatter is increased through an artificially expanded null space, while horizontal scatter is increased because of the addition of structural noise to any measurement noise present in the calibration dataset. The extra null space dimensions comprise the solution space components that are missing from the simplified model.

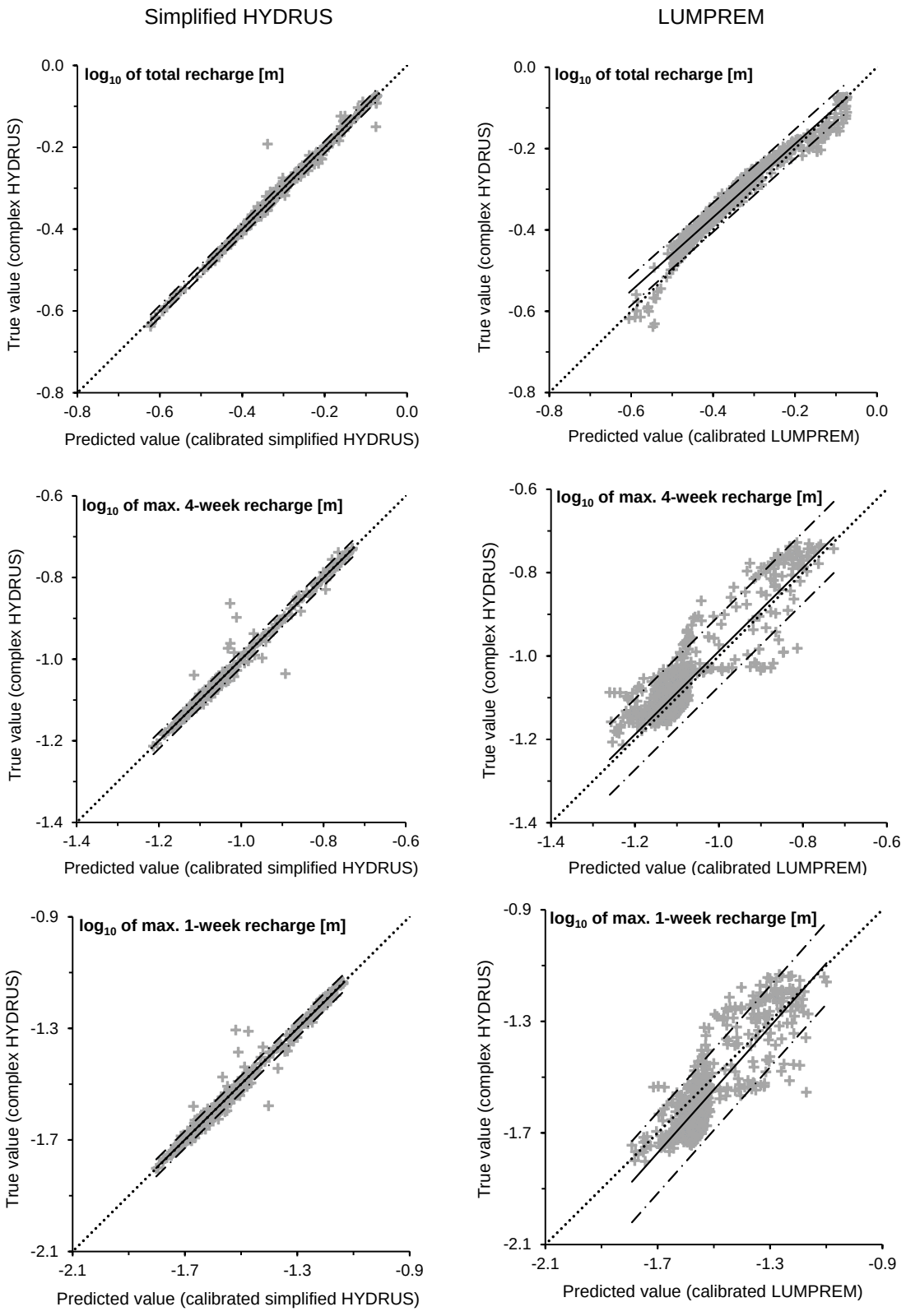

Figure 3.2. *s*-versus-s scatterplots for simplified HYDRUS (left column) and for LUMPREM (right column). The dotted line is the 1:1 line.

The following features of the scatterplots of Figure 3.2 are salient.

1. The limited vertical scatter in the simplified HYDRUS plots indicate limited null space contribution to the predictions which are the subject of the current study.

2. The slopes of the *s*-versus-$\underline{s}$ lines of best fit for all simplified HYDRUS predictions are very close to unity. This is also a reflection of the predominant solution space dependency of the predictions considered in this study. (As will be shown below, it does not indicate the absence of null-space parameter entrainment. However such entrainment is invisible in these plots as the predictions of interest are not sensitive to entrained parameters.)

3. The failure of LUMPREM parameters to span the full solution space of the complex HYDRUS model (highlighted by the less-than-perfect fit of LUMPREM to complex HYDRUS outputs demonstrated in Figure 1) results in an expanded null space and considerable scatter about the *s*-versus-$\underline{s}$ lines of best fit. This scatter is fairly mild for the long term recharge prediction, but much more pronounced for the short-term recharge predictions.

4. The predictive performance of LUMPREM varies with prediction values, this being expressed through variability of scatter about the *s*-versus-$\underline{s}$ line of best fit. This suggests that LUMPREM's performance as a recharge predictor is parameter dependent.

If the complex model is made to include more hydraulic processes, while the complexities of its simplified counterparts are maintained at their present levels, the propensity for bias in simplified model predictions is likely to increase. This results from the need for simplified model parameters to adopt surrogate roles in fitting datasets generated on the basis of more complex processes than they are capable of simulating. This was tested by repeating the above analyses with variability of root depth included in the complex HYDRUS realizations. (As stated above, root depth was fixed at 100 cm in the preceding analyses.) Root depth was varied randomly between realizations using a normal distribution with a mean value of 100 cm and a standard deviation of 25.5 cm. The *s*-versus-$\underline{s}$ regression outcomes for these analyses are provided in Table 3.3. The *s*-versus-$\underline{s}$ scatterplots are not shown for the sake of brevity.

Table 3.3. Regression coefficients and statistics pertaining to *s*-versus-s scatterplots equivalent to Figure 2 but with variable root depth in complex HYDRUS realizations.

	Simplified HYDRUS				LUMPREM			
Prediction	*a*	*b*	r^2	*σ*	*a*	*b*	r^2	*σ*
Total recharge	0.003	0.957	0.978	0.015	0.007	0.861	0.956	0.021
Max. 4-week recharge	-0.024	0.967	0.960	0.019	0.147	1.099	0.742	0.048
Max. 1-week recharge	-0.057	0.953	0.959	0.029	0.406	1.267	0.622	0.086

It was found that the ability of simplified HYDRUS to fit the calibration dataset is only marginally degraded with the addition of root depth variability to complex HYDRUS parameter realizations. At the same time the level of scatter in the corresponding *s*-versus-s plots increases (compare Tables 3.2 and 3.3). Similar considerations apply to LUMPREM.

Comparison of Tables 3.2 and 3.3 demonstrates that the enhanced complexity of the complex model increases the propensity for bias in all predictions made by both simplified models. For simplified HYDRUS, the slopes *b* of the *s*-versus-s regression lines fall below unity. From this it can be inferred that the simplified HYDRUS parameter space is misaligned with the solution space of the complex HYDRUS model. Calibration of simplified HYDRUS against a dataset generated by the process-enhanced complex HYDRUS therefore induces null-space parameter entrainment and the concomitant need for some estimated parameters to adopt compensatory roles to a greater extent. Because the predictions of interest are somewhat sensitive to null-space parameter components (predominantly plant root depth) they thus inherit this bias.

The increase in the *s*-versus-s slopes for the LUMPREM maximum 4-week and 1-week recharge predictions indicates that LUMPREM does not possess the parameter and process sophistication that would enable some of its parameters to adopt surrogate roles when undergoing calibration against the process-enhanced HYDRUS model. Furthermore, complex HYDRUS null-space parameter components that reflect root depth variability are effectively hard-wired at erroneous values in the LUMPREM model. Thus it cannot replicate the increased predictive variability that variability of plant root depth promulgates; an *s*-versus-s slope of greater than unity is the inevitable result.

We now turn to linear analysis. Due to lack of differentiability of HYDRUS-1D model outputs with respect to root depth variability, the latter remains fixed at 100 cm for all subsequent analyses.

3.4.3 Optimal simplification

It was proposed in section 2 that model simplification can be considered optimal when it is undertaken using SVD following transformation of reality model parameters to a parameter space where equation (10a) holds. In this section we demonstrate such simplification as applied to the complex HYDRUS model.

Ranked singular values calculated for the **Y** matrix of equation (3.16) are plotted in Figure 3.3. All 50 singular values are non-zero, this indicating a zero-dimensional null space for a calibration dataset comprising the 230 weekly recharge values discussed above. However singular values beyond approximately the fifth are very small, this implying that adjustment of only 5 parameter combinations (at most) is required for attainment of a very good fit with the calibration dataset. This notion is supported by the fact that the simplified HYDRUS model, with only 5 parameters, is capable of fitting complex HYDRUS calibration outputs very well, as Figure 1 demonstrates.

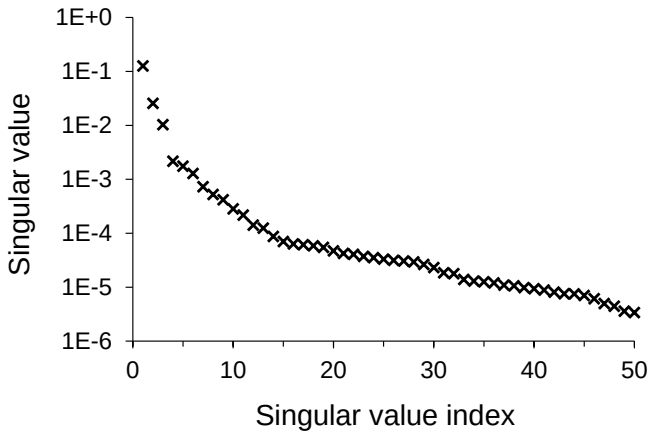

Figure 3.3. Singular values calculated for the complex HYDRUS **Y** matrix.

In practice, a field calibration dataset would be contaminated by measurement noise. As discussed above, this would limit the number of singular values employed in the calibration process. Equation (3.11) can be used to determine the number of singular values at which predictive error variance is minimized in the presence of measurement noise, and to quantify this variance. This is achieved by plotting a curve such as that shown in Figure 3.4a, which pertains to the log maximum 1-week recharge prediction. In order to construct Figure 3.4a a measurement noise standard deviation of 1.5 mm, with no temporal correlation between measurements, was assumed. Similar curves emerge for the other predictions considered in this study. The vertical scale of Figure 3.4a is truncated to clearly show the minimum of the curve and the contributions to

total predictive error variance made individually by the first (null space) and second (solution space) terms of equation (11). Note that the total predictive error variance for zero singular values is 5.21E-3, this corresponding to the prior uncertainty variance of this prediction. Figure 3.4b reproduces the total error variance curve of Figure 3.4a for a further four hypothetical magnitudes of measurement noise standard deviation.

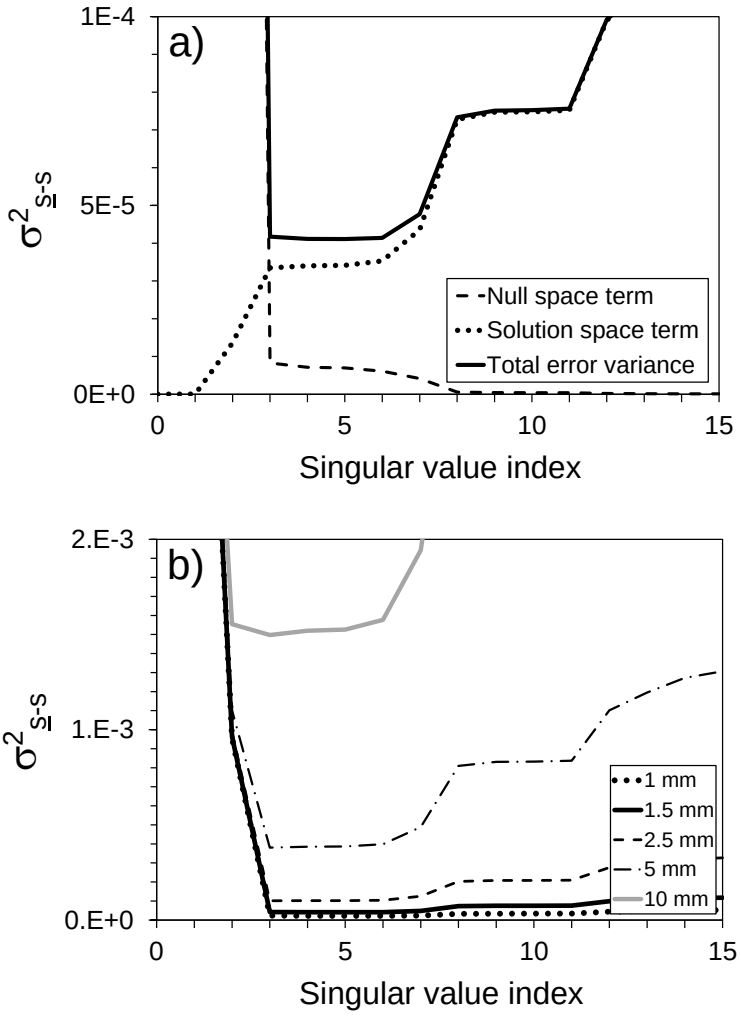

Figure 3.4. (a) Predictive error variance, including contributions from solution space and null space terms, versus number of singular values for the log maximum 1-week recharge prediction; measurement noise standard deviation is 1 mm. (b) The total predictive error variance curve reproduced for four additional measurement noise standard deviations.

It is apparent from Figure 3.4 that the minimum of the predictive error variance curve occurs between 3 and 6 singular values; it is also apparent that this does not vary greatly over a range of realistic values of measurement noise standard deviation. The curve is essentially flat in this area. Hence, with these amounts of measurement noise, a properly designed simplified model need possess as few as three adjustable parameters to attain an acceptable fit with the calibration dataset. The dominant

contribution to the uncertainty of this particular prediction is made by the solution space term. The small contribution to predictive error variance made by null-space parameter components is supported by the small degree of scatter in Figure 3.2 *s*-versus-s plots pertaining to simplified HYDRUS.

The right column of Figure 3.5 depicts estimable combinations of parameters that emerge from optimal simplification (i.e., from SVD of the matrix representing the model after parameters have been appropriately transformed). Denoted as $\mathbf{v}_{1y}$ through $\mathbf{v}_{5y}$, these are the first five columns of the matrix $\mathbf{FE}^{1/2}\mathbf{V}_y$. $\mathbf{V}_y$ is equivalent to $\mathbf{V}$ of equation (3.4) but arises from SVD of $\mathbf{Y}$ instead of $\mathbf{Z}$. $\mathbf{V}_y$ is pre-multiplied by $\mathbf{FE}^{1/2}$ (see equation (3.13b)) so that these estimable combinations of parameters can be presented in **k**-space (and thus in terms of recognizable complex HYDRUS parameters) rather than **m**-space. The left column of Figure 3.5 depicts combinations of observations that are uniquely and directly informative of these parameter combinations. These are the columns of $\mathbf{U}_y$ which are partnered to the columns of $\mathbf{V}_y$ (see equation (3.7) together with text following that equation for further details). The elements of these columns of $\mathbf{U}_y$ (with columns being denoted as $\mathbf{u}_{1y}$ through $\mathbf{u}_{5y}$) are plotted against time as the recharge observations to which they pertain (also shown in this plot) are time dependent.

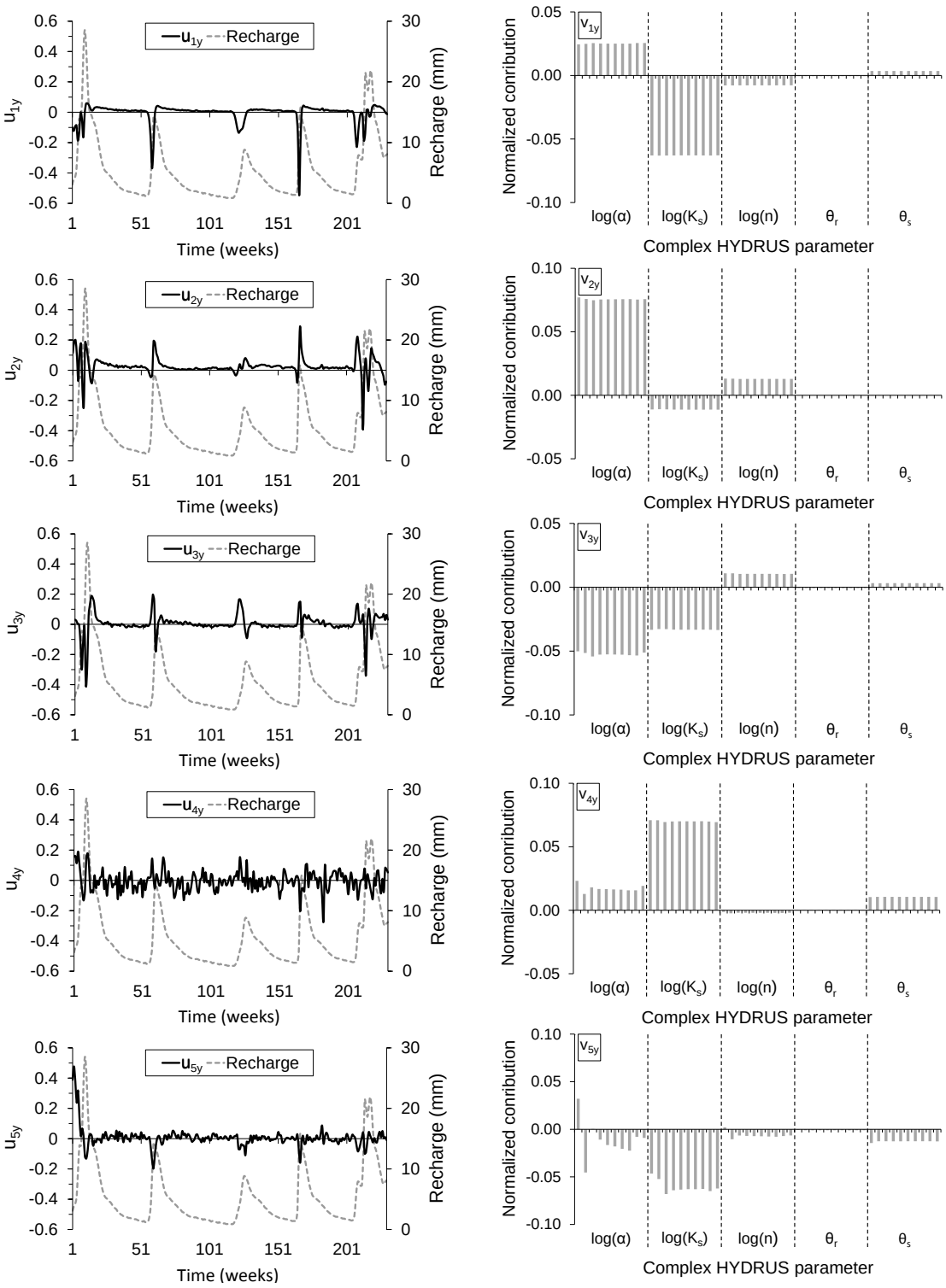

Figure 3.5. The right column shows estimable combinations of parameters emerging from optimal simplification of complex HYDRUS (i.e., columns of the $\mathbf{V}_y$ matrix ($\mathbf{v}_{1y}$ through $\mathbf{v}_{5y}$) calculated through SVD of the $\mathbf{Y}$ matrix of equation (3.16), after transformation to complex HYDRUS model parameter space. The left column shows corresponding combinations of calibration observations comprising the first 5 columns of the $\mathbf{U}_y$ matrix ($\mathbf{u}_{1y}$ through $\mathbf{u}_{5y}$). The ten columns for each complex HYDRUS parameter type represent the different model layers (increasing with depth from left to right).

Salient features of Figure 3.5 are as follows.

1. The most estimable parameter combination (i.e., the parameter combination corresponding to the largest singular value) is dominated by the ratio of log(α) to log(K_s). The information which furnishes this estimate resides in peak recharges and in the recharge decays which follow them.

2. Information within the calibration dataset appears to be much more informative of α, K_s and n than it is of volumetric parameters. The former parameters affect the timing and sharpness of recharge events. Attempts at model simplification in this context should therefore result in a model that exposes these controls to adjustment.

3. No one parameter type dominates any parameter eigencomponent (i.e., column of $\mathbf{V}_y$), with the possible exception of $\mathbf{v}_{2y}$ which features the α parameter prominently. The information on α appears to reside in the steepness of recharge peaks.

4. As the singular value number increases, the amount of detail represented in $\mathbf{u}_{iy}$ increases. Such detail is of high frequency content and may be difficult to distinguish from measurement noise in a real-world situation. The information contained in these combinations of observations is therefore easily lost; the parameter combinations which they inform are therefore likely to be uncertain.

3.4.4 Simplified model parameter composition

3.4.4.1 Complex model parameter contributions

Equation (3.22) can be used to characterize the composition of each simplified model parameter in terms of complex model parameters, thereby elucidating the parameter decomposition implied by model simplification. Figure 3.6 shows the composition of each simplified HYDRUS and LUMPREM parameter in terms of complex HYDRUS parameters as calculated using (3.22). Recall that simplified HYDRUS is generally able to achieve a near perfect fit with calibration data generated by complex HYDRUS (see Figure 1) through adjustment of its five parameters. Because these parameters have the compositions illustrated in Figure 3.6, it follows that it is possible to replicate this particular calibration dataset by mainly adjusting volumetric and drainage

response parameters that pertain to only a small part (namely the shallowest layers) of the subsurface, rather than drainage response parameters pertaining to the whole model domain.

The parameter compositions depicted in Figure 3.6 differ markedly from the compositions of optimal parameter components presented in Figure 3.5. Thus both simplified HYDRUS and LUMPREM constitute substantially suboptimal simplifications of the complex HYDRUS model. It follows that unless these models are used to make predictions which are predominantly solution space dependent (which are normally predictions that are very similar in character to those comprising the calibration dataset), predictions made by either of these models may be subject to a high degree of calibration-induced bias. It has been demonstrated that the predictions required of these models in the present case do, in fact, have a high solution space dependency and therefore are not as prone to bias as other predictions made by these models may be. In this sense, despite their suboptimal simplification, the simplified HYDRUS and LUMPREM models are "fit for purpose".

It is pertinent to examine the extent to which the compositions of simplified model parameters are consistent with the complex model parameters that they purport to represent. Ideally, from a parameter estimation perspective, parameter values achieved through calibration of the simplified HYDRUS model should equate to averages over the entire soil column of their complex model counterparts. Equation (3.24) demonstrates that, to the extent to which they can be considered to be estimates of averaged layer properties, these estimates are in error. Figure 3.6 reveals part of the reason for this error. At best, a given simplified HYDRUS parameter has a dominant contribution from the same complex HYDRUS parameter type over only a part of the overall soil column. At worst it almost entirely represents parameters other than that after which it is named. For example, complex HYDRUS parameters α and K_s are essentially absent from the compositions of simplified HYDRUS parameters of the same name.

The situation is a little better for the LUMPREM model. The LUMPREM V_{max} parameter does indeed appear to chiefly reflect complex HYDRUS θ_s - θ_r for the whole soil column. Nonetheless, similar to simplified HYDRUS, shallow complex HYDRUS parameters are generally better represented in LUMPREM parameters than are those associated with deeper parts of the soil column.

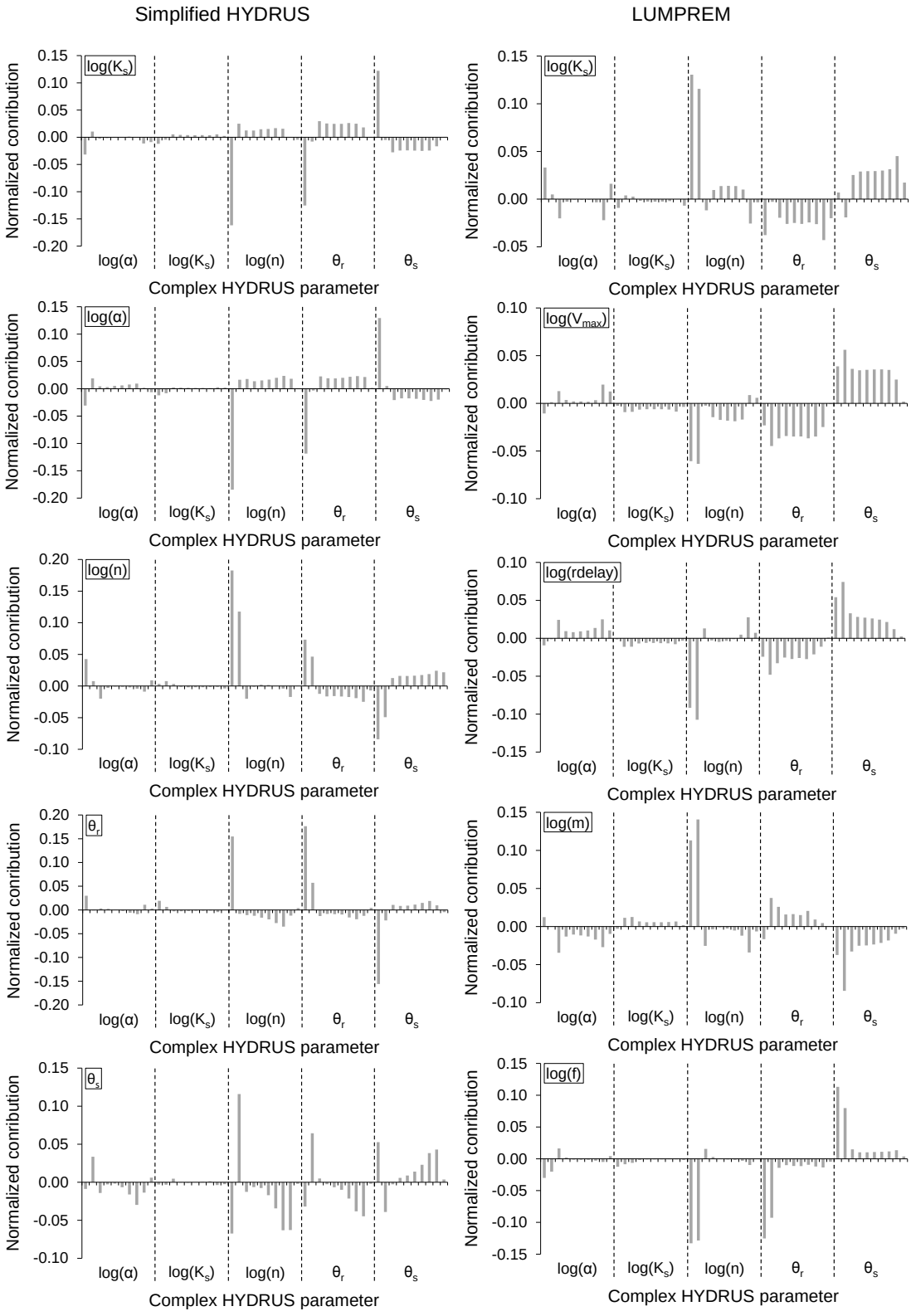

Figure 3.6. Normalized composition of each simplified HYDRUS (left column) and each LUMPREM (right column) parameter in terms of complex HYDRUS parameters (i.e., each row vector, normalized by its total length, comprising the matrix **L** of equation (3.22)). The ten columns for each complex HYDRUS parameter type represent the different model layers (increasing with depth from left to right).

3.4.4.2 Simplified model parameter variability

Equation (3.23) allows calculation of the propensity for variability of simplified model parameters from that of complex model parameters. Presumably the latter are forthcoming from the expert knowledge of the modeller. In practice, expert knowledge may be supplied to the calibration process as a covariance matrix associated with prior information on model parameter values. It can also be used in the setting of parameter bounds in order to limit parameter variability to a level that is considered realistic as parameters are adjusted during the calibration process. In the present synthetic case the covariance matrix of "reality" is known (i.e., the covariance matrix C(**k**) describing the parameter variability that was employed in generating the random complex HYDRUS parameter sets discussed in section 3.1). The C(**p**) matrix computed through equation (3.23), on the other hand, is the expression of "expert knowledge" as it pertains to a simplified model. It is thus the covariance matrix that must be associated with the pre-calibration probability distribution of simplified model parameters. Failure to use this matrix (in particular, if a matrix that expresses a smaller degree of parameter variability is used in its place) may restrict the ability of the calibration process to assign values to simplified model parameters that allow that model to fit any dataset generated by the complex model that is feasible based on the pre-calibration probability distributions of complex model parameters (i.e., true expert knowledge).

Tables 3.4 and 3.5 provide the C(**p**) matrix for simplified HYDRUS and LUMPREM parameters calculated using equation (3.23).

Table 3.4. Prior covariance matrix of simplified HYDRUS parameters calculated using equation (3.23).

	$\log(\alpha)$	$\log(K_s)$	$\log(n)$	θ_r	θ_s
$\log(\alpha)$	0.0326	-0.1062	-0.0026	0.0065	-0.0044
$\log(K_s)$	-0.1062	1.0056	-0.0609	-0.0536	0.0426
$\log(n)$	-0.0026	-0.0609	0.0130	0.0036	-0.0044
θ_r	0.0065	-0.0536	0.0036	0.0055	-0.0034
θ_s	-0.0044	0.0426	-0.0044	-0.0034	0.0030

Table 3.5. Prior covariance matrix of LUMPREM parameters calculated using equation (3.23).

	log(V_{max})	log(*rdelay*)	log(K_s)	log(m)	log(f)
log(V_{max})	0.2913	0.0082	0.0019	-0.0316	0.0158
log(*rdelay*)	0.0082	0.0003	0.0001	-0.0011	0.0001
log(K_s)	0.0019	0.0001	0.0071	-0.0003	-0.0016
log(m)	-0.0316	-0.0011	-0.0003	0.0045	0.0001
log(f)	0.0158	0.0001	-0.0016	0.0001	0.0042

The size of the off-diagonal terms of both of these matrices indicates a high degree of correlation between parameters. Such inter-parameter statistical correlation does not exist for complex HYDRUS parameters. The square root of each diagonal element of the prior covariance matrix is the prior standard deviation of the corresponding parameter (or its log, as indicated). In the case of simplified HYDRUS, prior parameter standard deviations can be directly compared with the true standard deviations of their complex HYDRUS counterparts. Figure 3.7 shows the ratios of simplified HYDRUS parameter standard deviations to corresponding complex HYDRUS parameter standard deviations. It is apparent that the prior standard deviations of some simplified HYDRUS parameters are significantly inflated compared to their complex HYDRUS counterparts. This suggests that the role that true (real-world) expert knowledge can play in parameterization of a simplified model is, at best, unclear. Consider, for example, the question of what bounds a modeller should place on simplified model parameters during the calibration process, and/or what prior probability distribution he/she should award to these parameters in establishing posterior predictive probability distributions through Bayesian analysis.

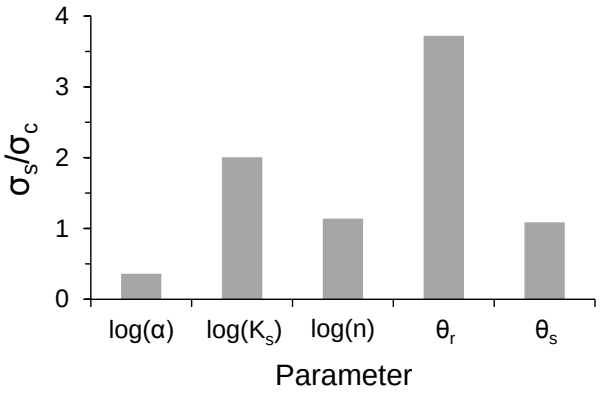

Figure 3.7. Ratio of each simplified HYDRUS model parameter standard deviation (σ_s) to corresponding complex HYDRUS model parameter standard deviation (σ_c).

3.4.5 Back-transformation to complex model parameter space

Equations (3.39) and (3.40) allow us to explore the hypothetical propensity for error in a complex model incurred by its notional calibration through use of a surrogate simplified model. (Note that, because equation (3.10a) does not automatically hold for complex HYDRUS, this analysis required the replacement of **Z** in equations (3.39) and (3.40) with **Y** of equation (3.13a), and the replacement of **S**, **U** and **V** with $\mathbf{S}_y$, $\mathbf{U}_y$ and $\mathbf{V}_y$, respectively.) Ideally, $C(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta})$ of equation (3.39) should be **0**, indicating perfect de facto estimation of complex model solution-space parameter components through simplified model calibration. $C(\underline{\boldsymbol{\beta}}_n)$ of equation (3.40) should also be **0**, indicating the absence of complex model parameter null-space entrainment incurred through simplified model calibration. Figure 3.8 shows the square root of the first five diagonal elements of $C(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for both simplified HYDRUS and LUMPREM. It thus shows the standard error of estimation of each of the elements of **β** that define the values of solution-space parameter projections of the complex HYDRUS model. (Note that both $C(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $C(\underline{\boldsymbol{\beta}}_n)$ are effectively normalized as $C(\underline{\boldsymbol{\beta}})$ is equal to **I**. This follows from equation (3.37) with **m** in place of **k**, and the orthonormality of $\mathbf{V}_1$.)

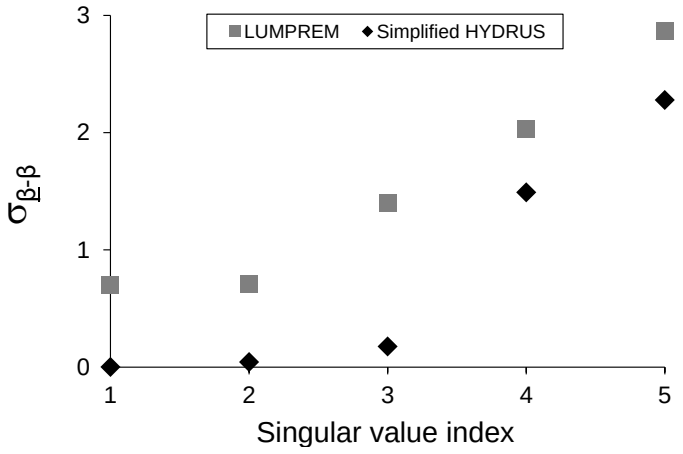

Figure 3.8. Standard deviations of error incurred by simplification in estimation of projections of optimal complex HYDRUS model parameters onto parameter solution space axes.

It is apparent from Figure 3.8 that implicit estimation of complex HYDRUS model parameter solution space components is better achieved through the parameter transformation and decomposition implied by use of simplified HYDRUS than through that implied by use of LUMPREM. This is hardly surprising given the fact that calibration of simplified HYDRUS leads to a better fit with a calibration dataset generated by the complex HYDRUS model than calibration of LUMPREM. What is

surprising, however, is the rapidity with which these errors rise after the third singular value. This suggests (as does Figure 3.4) that a properly constructed model that uses as few as three parameters would be just as suitable for replicating a recharge time series as a model with five parameters, especially where the necessity for a good fit is relaxed through the presence of measurement noise in the calibration dataset.

Figure 3.9 shows the square root of the diagonal elements of $C(\underline{\boldsymbol{\beta}}_n)$ for singular value indices of 6 to 15 (i.e., the first ten components of the complex HYDRUS null space). Ideally, parameter transformation and decomposition implied in calibration of the simplified HYDRUS and LUMPREM models should bestow values of zero on complex model parameter components that correspond to singular values beyond the fifth. Failure to achieve this implies entrainment of complex HYDRUS null-space parameter components through calibration of a simplified replacement model. Figure 3.9 indicates that the risk of calibration-induced parameter bias is relatively high for both the LUMPREM and simplified HYDRUS models. As previously discussed, the degree to which this promulgates predictive bias is prediction specific, for it expresses itself only to the extent that a prediction is sensitive to entrained null-space parameter components.

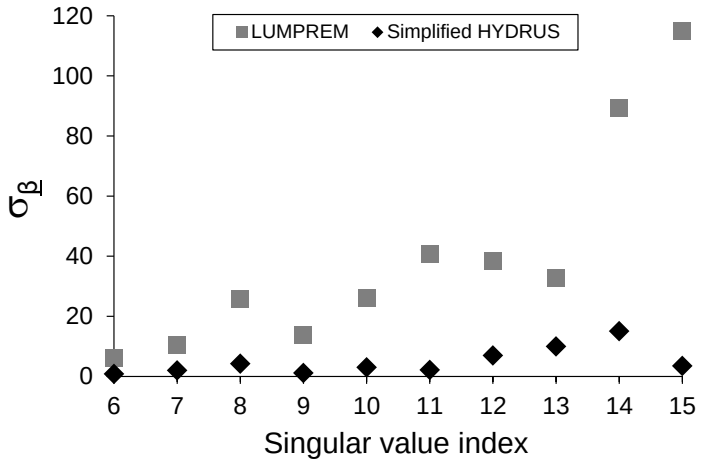

Figure 3.9. Standard deviations of error incurred by simplification in estimation of projections of optimal complex HYDRUS parameters onto parameter null space axes.

3.4.6. The linearity assumption

As is explained earlier in this paper, although an assumption of model linearity underpins the theoretical and experimental work documented herein, conclusions pertaining to optimality of simplification, and to the repercussions of calibrating a suboptimally simplified model, are not affected by this assumption. This is because

the phenomena exposed by our analyses are not related to the linearity (or otherwise) of a model. However, these phenomena are more easily exposed and described where local linearity is assumed. Nevertheless, in order to demonstrate the validity of the outcomes of the preceding analysis, we undertook a number of additional numerical experiments in order to assess the extent to which these outcomes are affected by nonlinearity of the models that we employed. Details are as follows; figures are not presented in the interests of brevity.

- Figure 3.7 was re-plotted based on statistics inferred from actual parameter values estimated through calibration of simplified HYDRUS against complex HYDRUS outputs; see Section 3.3.3. The heights of the bars pertaining to different parameters are comparable. Relativity of these heights is preserved.

- Figures 3.5 and 3.6 were re-plotted following calculations based on **Z** matrices computed using a number of different realizations of complex HYDRUS parameters. The plots differ only in minor details.

- Equation (3.22) was used to calculate a simplified HYDRUS parameter set **p** from each of the 998 stochastic realizations of complex HYDRUS parameter sets **k**. Calibration objective functions computed on the basis of these parameter sets were all acceptably low, this demonstrating their ability to fit corresponding complex HYDRUS-generated calibration datasets.

- Figure 3.8 was re-plotted using sensitivities pertaining to a number of different complex HYDRUS parameter realizations together with those pertaining to corresponding best-fit simplified model parameters. The points corresponding to the fourth and fifth singular value indices showed some variability between realizations, particularly for the simplified HYDRUS model. However, relativity of these values was preserved.

- The above analysis was also undertaken for Figure 3.9, with similar outcomes.

3.5 Discussion

The present study advances the theory and concepts that were introduced in a previous work (namely Doherty and Christensen, 2011), and applies them to a different kind of model from that which was investigated in that work. The theoretical work of Doherty and Christensen (2011) explores both optimality of simplification and the consequences of suboptimal simplification in contexts wherein a model must be calibrated before being employed in a predictive capacity. The present study employs linear analysis to demonstrate and explore these concepts as they apply to a vadose zone model constructed for the purpose of providing time-varying recharge to a groundwater model. The intention of this paper is therefore twofold: (1) to extend the theoretical basis of model simplification from a subspace perspective, thus providing a foundation for further theoretical and numerical research into the confounding but pervasive issue of model simplification; and (2) to contribute to current understanding of the parameter and associated predictive outcomes of typical simplification practice, through the application of this theory, together with other analyses, to some representative synthetic examples.

Model simplification can be considered optimal when calibration of a thus simplified model allows implicit estimation of reality model solution-space parameter components without concomitant assignment of non-zero values to reality model null-space parameter components. Or, to put it another way, simplification is optimal when calibration of a simplified model leads to the same predictive outcomes that would be achieved if the reality model itself were calibrated using truncated SVD in estimation of parameters that have been transformed in accordance with their prior variability. All predictions made by a reality model calibrated in this way would possess minimal bias.

Calibration of a suboptimally simplified model awards non-zero values to at least some reality model null space components; in doing this it implicitly alters at least some reality model parameter values that are not supported by the data. However if calibration of the simplified model achieves a good fit with field data, reality model solution space components are nevertheless implicitly well estimated. Hence predictions made by the calibrated simplified model that are solely dependent on reality model solution-space parameter components will be made with as much accuracy as if the reality model itself were calibrated and then used to make these same

predictions. In fact, if there was no measurement noise associated with the calibration dataset, such predictions would be made without error by the simplified model. If a simplified model has been built and calibrated to make only these kinds of predictions, then suboptimality of simplification matters little, for the model is entirely fit for purpose. (This is not expected to be a common phenomenon, however. Models are usually built to make predictions of system behaviour under conditions that are at least partially different from those that prevailed during its calibration. These predictions are therefore likely to be sensitive to at least some parameters, or combinations of parameters, that are not informed by the calibration dataset, and hence belong to the null space.)

In carefully considering the processes simulated by a model, model simplification strategies often attempt to separate those processes that are either salient to a prediction or are informed by a calibration dataset, from those processes which are not. The former are then represented in a form whereby, through averaging or appropriate definition of lumped process elements, inestimable parameters and/or parameter relationships (these often pertaining to system detail) are eliminated from the model. Such physically based simplification implicitly attempts to follow the precepts of optimal simplification outlined herein, in that it attempts to separate system components which can be informed by the measurement dataset from those which cannot be thus informed. Such separation takes place under an implied assumption that the two components are indeed separable, and hence are orthogonal from a parameter estimation point of view. Rarely will such separation be completely orthogonal however, as is suggested by the examples analysed in the present study. Hence while careful physically based simplification can attain much in terms of the metrics presented herein, it is unlikely that the possibility of calibration-induced bias for at least some predictions will be completely eliminated.

Through linear analysis of a complex vadose zone model, together with two simplifications of this model, we have attempted to illustrate the consequences of suboptimal simplification by demonstrating the implicit and unavoidable transformation of reality model parameters that occurs when a simplified model is calibrated in its place. We have shown that despite this transformation, if a simplified model can fit the calibration dataset well, the legitimacy of estimation of the projection of reality model parameters onto the calibration solution space is maintained. In contrast, where simplification is such as to compromise the ability of a simplified

model to fit the calibration dataset well, implicit estimates of reality model solution-space parameter projections are demonstrably in error.

The unintended adjustment of reality model null-space parameter components from their preferred values of zero through calibration of a substitute simplified model has also been demonstrated. Should a prediction required by a simplified model be sensitive to these aspects of reality, the simplified model would incur significant error in making that prediction as a consequence of this.

The present study has demonstrated that even comparatively mild simplification, in the present case applied through replacing hydraulic property heterogeneity by hydraulic property homogeneity, can lead to significant entrainment of reality model null-space parameter components. At the same time, the homogeneous parameters that are estimated through calibration of the simplified model have a complicated relationship with parameters of the reality model that they replace. As estimates of the average values of equivalent reality model parameters they are thus significantly in error. Though beyond the scope of the present paper, this has repercussions for interpretation of experimental data gathered at field sites (such as lysimeter sites) where experiments are designed explicitly to allow estimation, through calibration, of "field scale" hydraulic properties. It also has repercussions for the role of expert knowledge in calibrating a simplified model. The variability of most simplified HYDRUS parameters required to reproduce complex HYDRUS data is greater than the true variability of the complex HYDRUS reality model parameters that they purport to represent. Thus if expert knowledge is applied in defining parameter bounds employed in calibrating a simplified HYDRUS model, its application would compromise the ability of this model to fit the calibration dataset. This would, in turn, compromise the ability of the calibration process to implicitly estimate reality model solution-space parameter components, and would therefore degrade the accuracy of simplified model predictions which depend solely on these parameter components. On the other hand, it would lessen the degree of null-space entrainment, through reducing the degree of parameter surrogacy that is allowed to occur during calibration; in doing so it would reduce the propensity for error in predictions which are sensitive to such entrained components.

It appears, therefore, that an important outcome of suboptimal simplification is the creation of a tension between expert knowledge on the one hand and information

contained in a calibration dataset on the other hand. This tension extends beyond the often-encountered situation whereby expert knowledge and historical measurements of system state may suggest different values for certain parameters. This tension is more fundamental, in that it reflects an inability on the part of a model to adequately respond to information originating from both of these sources simultaneously, for response to one of these sources compromises its ability to respond to the other. It follows that the goodness of fit sought between simplified model outputs and historical measurements of system state should be prediction-specific. Where a prediction required of a model is entirely solution space dependent (i.e., is entirely informed by past system behaviour), a modeller is entitled to fit historical data to a level that is commensurate with measurement noise. On the other hand, where a prediction depends on null-space parameter components that are subject to entrainment through the simplified model calibration process, a modeller should seek a reduced level of fit between model outputs and field measurements in calibrating the model, thus allowing expert knowledge to hold greater sway in the model parameterization process. Unfortunately however, a modeller cannot know the extent to which he/she should "fail to fit" field data to accommodate this imperative and minimize potential predictive error. If simplification were optimal in the way defined above, this would not be an issue, for application of expert knowledge (soft data) would not compromise assimilation of hard data. The tension (bordering on incompatibility) between the two kinds of information required for reduction of predictive error is an outcome of the fact that less-than-optimal model simplification provides receptacles that cannot hold both types of information simultaneously.

3.6. Conclusions

All models are simplifications of reality. Some are made very simple by design. Others are specifically designed to include as much real-world complexity as possible, but are nevertheless simple when compared with reality. Most models employed for environmental management are calibrated against historical behaviour of the system which they simulate. This serves a number of purposes. One of these is to verify that the model can indeed replicate the behaviour of that system. Another is to extract information from the historical record that informs parameters, and thereby reduces the propensity for error associated with predictions of future system behaviour made by the model.

Simplification comes at a cost. Informed model-based decision making requires that this cost be understood and accommodated. In this paper we have cast model simplification as a form of parameter transformation and decomposition. This has allowed us to define optimal simplification as a standard against which other forms of simplification can be judged. Simplification is considered to be optimal when its outcomes are identical to those that would have been achieved through transformation of real world parameters in accordance with the nature of their variability, followed by orthogonal decomposition of these transformed parameters into solution and null subspaces defined on the basis of the available calibration dataset.

If a simplified model can reproduce a historical calibration dataset well, then a prediction which is entirely solution-space dependent (and is thereby completely informed by this dataset) will be made with little error, regardless of the physical basis (or lack thereof) of the simplified model. However, where a prediction is less than totally informed by the calibration dataset, the history-matching process may detract from the credibility of that prediction by introducing distortions into the components of that prediction which must be informed by expert knowledge. Simplification is suboptimal when it allows this to occur. Mathematically, this happens when the decomposition of real world parameter space implied by simplification is not aligned with that provided by the idealized simplification process described above.

Bayesian analysis informs us that, if a model is a perfect simulator of environmental behaviour, its predictive performance can only be enhanced by imposing constraints on parameter values through the history matching process. The same cannot be said for a simplified model unless simplification is optimal in the sense described above (which will rarely, if ever, be the case). Where a model's simulation of real-world processes is defective, the benefits of calibration become prediction-specific. When a simplified model extracts information from an historical observation dataset, it may place that information into incorrect receptacles. For certain types of predictions (namely those bearing greatest similarity to observations comprising the calibration dataset), this may be of no consequence, for extraction of information from the calibration dataset is all that matters; all information receptacles are therefore "good receptacles". For other types of predictions, namely those that require that information from both the calibration dataset and expert knowledge be combined, the "pushing aside" of expert knowledge that occurs when calibration information is placed into suboptimal parameter receptacles which cannot hold both of these types of information

simultaneously, can engender bias in those predictions. Without a complementary complex model through which the performance of a simplified model can be assessed, this (possibly very substantial) bias cannot be quantified.

In the present paper we have studied a complex model of a type that finds common usage in everyday modelling practice, together with two simplified counterparts of that model. Through linear analysis we have attempted to formulate the transformations that are implied in simplification of the complex model, and to then understand the extent to which simplification has, or has not, compromised the performance of the simplified models. To the extent that lessons learned from the analyses documented herein have broader implications, we now state conclusions from this work that can be extended to other modelling contexts.

- Even where the model simplification process is such that a simplified model tries to be faithful to the physics of the processes that it simulates, and employs parameters that attempt to replicate measurable physical properties of the system, what each such parameter actually represents in the calibrated model may be very different from the system property after which it is named.

- As a result, attempts to infer local system properties through calibration of even a highly complex physically based model may lead to highly erroneous estimates of these properties.

- When even a highly complex physically based model is calibrated against a real-world dataset, attainment of a good fit with that dataset may require a greater range of parameter variability than that which would be allowed on the basis of expert knowledge alone.

- Where a model is intended to make predictions that are similar in character to the observations against which it is calibrated, obtaining a good fit with the calibration dataset is more important than adherence to user-informed parameter bounds. In fact a correct physical basis for the model may matter less than whether its parameters collectively span the solution space of the reality model of which it is a simplification.

- Where a model is built to make predictions of many different types (as many models are) the benefits of constraining parameters in order to ensure

replication of past system behaviour become difficult to assess. Even a relatively small amount of non-optimal simplification can force parameters to play roles that compensate for model inadequacies and thereby entrain reality model null-space parameter components as they are adjusted. Any prediction that is sensitive to these null-space parameter components will be biased as a result. It is possible that calibration of the model for the making of those particular types of prediction will do more harm than good.

The conclusions from this study are far from satisfying, and in some ways pose more questions than they answer. These include the following.

- Should the accepted notion (which underpins a great deal of commercial and research-based model usage) that calibration and prediction are entirely separate aspects of model construction and deployment be abandoned (together with the sense of finality that the word "calibration" implies)?

- Should a model be calibrated not once, but many times according to different fitting metrics, in order to thereby optimize its ability to make different types of predictions?

- Where predictions can be demonstrated to be predominantly solution space dependent, should reliance be placed more on lumped parameter models that are easily calibrated than on complex physically based models that are difficult to calibrate?

- Where predictions have a moderate to high degree of null space dependency, should a model that is complex enough to express this null space be calibrated only mildly, or perhaps not at all?

Chapter 4

Outcomes of pilot point-based regularized inversion in a categorically heterogeneous environment

Note: software was developed as part of this work to facilitate stochastic generation of categorical hydraulic property distributions, which forms the basis of the study presented in this chapter. The software allows generation of a user-specified number of random distributions of discrete linear features with flexible control of the variability in length, orientation and number of features including minimum separation threshold specification. It may be used in conjunction with, for example, geostatistical field generation software available in the PEST software suite (Doherty, 2016b) in order to generate ensembles of stochastic hydraulic property fields for Monte Carlo simulation purposes. Examples of the software's outputs are presented as Appendix B.

Abstract

The use of pilot-point-based regularized inversion to calibrate highly parameterized groundwater models is increasingly common practice. Despite the inevitable loss of detail in estimated hydraulic property fields as a cost of attaining a unique solution to the inverse problem, regularized inversion can theoretically provide estimated model parameters and associated predictions that have a minimised potential for wrongness. The subsequent quantification of this potential forms the ultimate goal of modelling in a decision-support context. The current study explores the outcomes of regularized inversion in a subsurface environment comprising discrete preferential flow features ("faults"), wherein theoretically ideal formulation of Tikhonov constraints within a multi-Gaussian framework (which is common groundwater modelling practice) is not possible. Paired model analysis is applied to a synthetic example to quantify the success of the inversion process, at the same time as elucidating the specific contributions to post-calibration potential predictive error, particularly predictive bias.

Several regularization weighting strategies are tested and compared in terms of estimated parameter field characteristics and model performance in making multiple predictions. The presence of faults is shown to induce substantial pre-calibration bias in all predictions. It is shown that for some predictions, ignoring the existence of the faults does not compromise the ability of the inversion process to "calibrate out" the initial bias and markedly reduce (and allow quantification of) potential predictive error. Simultaneously, the calibration process magnifies bias in other predictions, inflating their potential for wrongness far beyond prior uncertainty based on "expert knowledge" alone and thus threatening the integrity of the modelling process. No employed regularization weighting strategy reduces the potential for error in one prediction without simultaneously raising the potential for error in another. Formulation of regularization weights in a heuristic manner demonstrably promotes representation of fault-like features in estimated parameter fields. This is shown to reduce "hardwired" predictive bias, but at the same time inflate bias caused by parameter surrogacy. For the making of some predictions, this renders calibration inevitably fruitless at best and highly detrimental at worst. The present study thus emphasizes the need for prediction-specific tuning of a model calibration process, to the extent that the most pragmatic approach for some predictions may be to forego calibration entirely and quantify uncertainty based solely on the purest possible expression of expert knowledge.

4.1 Introduction

Calibration as a precursor to uncertainty analysis is now considered to be standard practice in environmental modelling for decision support (Hunt et al., 2007; Anderson et al., 2015). The pilot point method (de Marsily, 1978; de Marsily et al. 1984) is of increasing popularity as a spatial hydraulic property parameterization device in groundwater model calibration (Kourakos and Mantoglou, 2012). It allows hydraulic property heterogeneity to arise freely in accordance with the information contained in the calibration dataset, rather than being constrained by predefined zones (e.g., Doherty, 2003; Hunt et al., 2007). Moreover, pilot-point-based parameterization is highly compatible with well-established and relatively computationally efficient methods for undertaking the critical task of estimating the post-calibration propensity for error in model predictions, through for example linear analysis (e.g., Moore and Doherty, 2005; Gallagher and Doherty, 2007a), calibration-constrained Monte Carlo

methodologies (e.g., Tonkin et al., 2007; Tonkin and Doherty, 2009; Herckenrath et al., 2011; Yoon et al., 2013) or Pareto analysis/hypothesis-testing methodologies (e.g., Moore et al. 2010).

Central to the pilot-point approach is the use of a large number of pilot points such that parameter heterogeneity may arise in accordance with the information contained within the calibration dataset, and be explored through uncertainty analysis (e.g., Doherty, 2003). The inevitable ill-posedness of the inverse problem necessitates some form of regularization. Tikhonov regularization is a constrained minimization approach employed extensively in geophysical data interpretation (e.g., Constable et al., 1987; Portniaguine and Zhdanov, 1999; Greenhalgh et al., 2006; Zhdanov, 2010). It allows for inclusion of "expert knowledge" (e.g., expected parameter values or relationships based on site characterization activities) in the inversion process. A number of examples of its use in the groundwater modelling context to regularize pilot-point-based inversion are found in recent literature (e.g., Tonkin and Doherty, 2005; Alcolea et al., 2006, Alcolea et al. 2008; Singh et al., 2008; Hendricks Franssen et al., 2009; Fienen et al., 2009; Herckenrath et al., 2011; Knowling et al., 2015). Supplementing Tikhonov constraints such as preferred parameter values, additional expert knowledge may be included in the inversion process in the form of a covariance-based regularization weighting scheme based on the nature of expected hydraulic property variability (Maurer et al., 1998). As model parameters are adjusted during the calibration process they are thus constrained to respect "geological plausibility" to the greatest extent possible. Expression of expert knowledge in terms of a covariance matrix assumes multi-Gaussian hydraulic property variability. While this assumption is convenient and has been evinced as suitable in many settings (Freeze, 1975; Hoeksema and Kitanidis, 1985), discrete non-Gaussian features are common in natural formations, and may not be identifiable from field data (Wen and Gomez-Hernandez, 1998; Sarma et al., 2008; Zhou et al., 2014).

Pilot-point-based regularized inversion is an implicit simplification device due to the unavoidable cost of attaining a unique solution to the inverse problem; estimated parameter fields are inevitably "smooth" and lack true hydraulic property detail (e.g., McLaughlin and Townley, 1996; Doherty, 2003; Moore and Doherty, 2006). In idealistic circumstances, theoretically optimal regularized inversion (refer to Chapter 2 of the present thesis for details) may nonetheless provide an estimated parameter field that approaches a minimum potential for wrongness, and which thus provides

predictions with a minimum propensity for error (Christensen and Doherty, 2008; Doherty et al., 2010). This provides an optimal foundation for post-calibration predictive uncertainty analysis. Imperfections within the modelling process, however, such as estimation of a continuous pilot-point-based field in the presence of categorically heterogeneous hydraulic properties, may compromise the integrity of model predictions.

Characterization of the predictive ramifications of model simplifications or imperfections (also referred to interchangeably in existing literature as, for example, model structural error, model inadequacies or model defects) has been a subject of extensive study over many years. However, most approaches rely upon a model's imperfections as a simulator of the natural environment being expressed through the calibration process as irreducible model-to-measurement misfit (e.g., Beven and Binley, 1992; Draper, 1995; Gupta et al., 1998; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Vrugt et al., 2005; Ye et al., 2008; Doherty and Welter, 2010; Spaaks and Bouten, 2013; Xu and Valocchi, 2015). Recent literature focuses on the characterization of calibration-induced predictive bias in the case where a model's imperfections do not compromise the ability of the calibration process to achieve a "good fit" (e.g., Doherty and Christensen, 2011; White et al., 2014; Chapter 3 of the present thesis). Calibration-induced predictive bias erodes the gains achieved through history matching and increases the risk of post-calibration uncertainty assessment failing to capture the true prediction value within estimated uncertainty ranges. Underestimation of potential predictive error defines failure of a modelling process according to Doherty and Vogwill (2016). It creates potential for "type II" statistical error (i.e., the false rejection of a true hypothesis), whereby an unacceptable impact of a proposed environmental management strategy occurs despite model-based assurance that it is extremely unlikely (e.g., Downes et al., 2002; Beven, 2010).

Doherty and Christensen (2011) present a methodology designed to allow identification and reduction of predictive bias incurred through calibration of a simplified model, simultaneous with predictive uncertainty quantification. It involves the conjunctive use of a relatively complex model and a simplified model of the system under study. It is proposed as a practical methodology that provides efficiency and numerical stability gains relative to calibration and calibration-constrained uncertainty analysis performed using the former in a standalone fashion, whilst simultaneously accounting for predictive performance degradation that may be incurred through

standalone use of the latter. Chapter 2 of the current thesis presents a comprehensive proof-of-concept study of this methodology, herein referred to as "paired model analysis", validating its efficacy in performing these functions. White et al. (2014) extend the subspace-based theory presented by Doherty and Christensen (2011) to demonstrate a relatively efficient linear approach to quantifying the additional component of predictive error variance attributable to model simplification/defects.

Application of methodologies designed to quantify the additional component of potential predictive error attributable to model imperfections will not always be feasible in practice. Furthermore, methodologies such as those presented by Doherty and Christensen (2011) and White et al. (2014) facilitate quantification of the component of potential predictive error attributable to model imperfections relative to a more complex model. However, even the most complex numerical representation of an environmental system is inevitably imperfect relative to reality. Any predictive bias arising through this discrepancy is beyond the reach of these methodologies. For these reasons, White et al. (2014) advocate synthetic studies emulating real-world modelling contexts, wherein methodologies such as these are employed in order to quantify the predictive consequences of representative model simplifications. The intention of this is to provide best-practice guidance through development of qualitative understanding of the causes of predictive bias, most susceptible prediction types and circumstances, and possible mitigating measures that may be taken by a modeller.

In light of the abovementioned increasing popularity of pilot-point-based regularized inversion, the present study explores the outcomes of the approach in a context wherein the existence of discrete preferential flow features (hereafter referred to as "faults" for the sake of convenience) precludes holistic expression of expert knowledge in multi-Gaussian terms. A simple synthetic example is employed wherein hydraulic conductivity K is characterized both by multi-Gaussian and categorical components of heterogeneity. This is similar to a test case presented by Zimmerman et al. (1998) in which the U.S. Department of Energy's Waste Isolation Pilot Plant (WIPP) site is conceptualised as disconnected high-transmissivity fracture zones or "channels" embedded within a geostatistical "background" field. In their study, Zimmerman et al. (1998) explore the performances of seven different inverse approaches, including the pilot point method, for estimating transmissivity and predicting advective path and travel time in this setting. The focus of their study is on the relative performance of the various inverse methods. They conclude that the pilot point method performs

comparatively well across a broad range of performance evaluation measures, despite results indicating the presence of substantial predictive bias. However, limited conclusions are drawn pertaining to the specific predictive outcomes of employment of the pilot point method.

The present study employs paired model analysis in order to examine in detail the predictive outcomes of pilot-point-based regularized inversion in the presence of categorical heterogeneity. Paired model analysis is chosen due to it being a nonlinear approach and allowing the overall success of the inversion process (i.e., the reduction of potential predictive error relative to prior uncertainty) to be broken down into its various contributions, in particular multiple sources of predictive bias. The performance of the calibrated model in making multiple predictions is assessed, including ungauged hydraulic head, drawdown and advective transport path and travel time. The calibration outcomes based on a number of alternative regularization strategies are tested. These include the application of no stochastic expert knowledge through a regularization weighting scheme, incorporation of "partial" stochastic expert knowledge pertaining only to the background geostatistical field characteristics (this representing a case in which the existence of the faults is ignored, or perhaps unknown), as well as attempted representation of the existence of faults through ad hoc "compromise" regularization weighting. It should be noted that the latter does not reflect the current state of the art of Tikhonov regularization capabilities. It is intended as a rudimentary ad hoc example of the potential outcomes of modified regularization in a context where theoretically optimal regularization is precluded in this case by non-Gaussian hydraulic property features. Doherty (2015) points out the active field of research involving the continuous, differentiable expression of complex non-Gaussian categorical parameter fields based on multiple point geostatistics (e.g., Sarma et al., 2008; Ma and Zabaras, 2011; Vo and Durlofsky, 2014), which are thus compatible with gradient-based optimization algorithms and may facilitate more sophisticated Tikhonov regularization strategies. However, the consideration of methods beyond the standard two-point geostatistical functionality currently offered by widely used inversion software such as PEST (Doherty, 2016a) is beyond the scope of the present study. Nonetheless, as will be discussed, the application of more sophisticated regularization techniques is not expected to affect the general conclusions drawn from the results of the present study.

This chapter is organized as follows. Section 4.2 briefly summarizes the theory and concepts that are central to the analyses conducted in present study, as well as the paired model analysis methodology and predictive error/uncertainty quantification approaches. Section 4.3 describes the synthetic example employed in the present study, including the specifics of the regularized inversion and subsequent predictive analysis. Results from the synthetic example are presented and briefly discussed in section 4.4, while sections 4.5 includes a more thorough discussion as well as insights for modelling practice more broadly. Section 4.6 provides the conclusions of the study.

4.2 Theory, concepts and methods

Subsection 4.2.1 through subsection 4.2.3 contains concepts and equations belonging to widely-applied mathematical inversion theory presented, for example, by Menke (1989), Aster et al. (2005), Moore and Doherty (2005, 2006). A linear relationship between model parameters and model outputs is assumed to apply in the following theory. This provides simplicity and tractability, allowing a model to be represented by a matrix, and model parameters model outputs to be represented by vectors. Further tractability is achieved through formulation of parameter and model output vectors such that they represent perturbations from pre-calibration, expert knowledge-based values.

4.2.1 History matching

Let the vector **k** represent the parameters employed by the model to represent system hydraulic parameters. The Jacobian matrix **X** represents the action of the model on **k** to produce model outputs. Observations of system state comprising the available calibration dataset are contained in **h** such that:

$$\mathbf{h} = \mathbf{Xk} + \boldsymbol{\varepsilon} \tag{4.1}$$

where the vector **ε** contains measurement noise. The data assimilation or "history matching" process seeks to minimise model-to-measurement misfit (to a level commensurate with measurement noise such as to avoid "overfitting" – refer to Chapter 2 of the present thesis for a thorough discussion of this topic). Model-to-measurement misfit is represented by the "measurement objective function" Φ_m, defined as:

$$\Phi_m = (\mathbf{Xk} - \mathbf{h})^t \mathbf{Q}_h (\mathbf{Xk} - \mathbf{h}) \quad (4.2)$$

where $\mathbf{Q}_h$ is the "observation weight matrix". This contains (the squares of) observation weights *q*, and is ideally specified to be proportional to the inverse of the covariance matrix of measurement noise C(**ε**) (e.g., James et al., 2009).

4.2.2 The null space

The increasingly common highly parameterized approach to modelling precludes unique estimation of all model parameters based on the calibration dataset. The inverse

problem is ill-posed when parameters outnumber observations. Even when observations outnumber parameters, ill-posedness will often prevail in a highly parameterised context on the basis of limited information content within the set of observations (e.g., Welter et al., 2015).

The existence of a so-called "null space" follows from ill-posedness of the inverse problem. By definition, a non-zero parameter set $\mathbf{k}_n$ belongs to the null space of **X** if:

$$\mathbf{0} = \mathbf{Xk}_n \quad (4.3)$$

Momentarily ignoring the presence of measurement noise **ε** for convenience, consider a parameter set $\underline{\mathbf{k}}$ that reproduces the calibration dataset **h** perfectly. That is:

$$\mathbf{h} = \mathbf{X}\underline{\mathbf{k}} \quad (4.4)$$

From equation (4.3) and equation (4.4) we can write:

$$\mathbf{X}(\underline{\mathbf{k}} + \mathbf{k}_n) = \mathbf{X}\underline{\mathbf{k}} = \mathbf{h} \quad (4.5)$$

thus demonstrating the nonuniqueness of $\underline{\mathbf{k}}$ due to existence of the null space.

Null-space parameter adjustment is by definition unsupported by the calibration dataset **h** (Doherty and Christensen, 2011). Through an optimal inversion process any parameter components belonging to the null space should thus remain unperturbed relative to their pre-calibration expected values (i.e., $\mathbf{k}_n = \mathbf{0}$). Any such adjustment induces asymmetry in, and thus increases, the potential for error in these parameter components and subsequently in predictions that are sensitive to them.

Regularization provides a means through which uniqueness of the inverse problem is attained though separation of null-space parameter components from those that are estimable on the basis of the calibration dataset (the so-called "solution-space" parameter components). This may conceptually be achieved manually through pre-calibration parameter fixing or lumping, or mathematically through regularization (the latter being central to the highly parameterized modelling context). Two alternative forms of mathematical regularization are constrained minimization regularization (i.e., Tikhonov regularization (Tikhonov 1963a, 1963b; Tikhonov and Arsenin, 1977)) and truncated SVD (e.g., Aster et al., 2005). However, the focus of the present study and thus the theory presented hereafter is limited to Tikhonov regularization.

4.2.3 Tikhonov regularization

The Tikhonov regularization theory presented herein is specific to the way in which it is implemented by PEST. Well-posedness of the inverse problem is achieved by supplementing the calibration dataset **h** with a set of "regularization observations" **r**. These are "preferred" parameter values (or relationships) based on expert geological knowledge (or "soft data") that will prevail unless information contained within the available calibration dataset ("hard data") dictates otherwise. A "regularization objective function" Φ_r is defined as:

$$\Phi_r = (\mathbf{Wk} - \mathbf{r})^t \mathbf{Q}_r (\mathbf{Wk} - \mathbf{r}) \tag{4.6}$$

Here, **r** is a vector containing the abovementioned regularization observations. $\mathbf{Q}_r$ is a "regularization weight matrix". The matrix **W** defines the relationship between **k** and **r**. Where regularization observations consist of preferred parameter values, and these are defined as pre-calibration expected parameter values, $\mathbf{W} = \mathbf{I}$ and $\mathbf{r} = \mathbf{0}$, thus equation (4.6) becomes:

$$\Phi_r = \mathbf{k}^t \mathbf{Q}_r \mathbf{k} \tag{4.7}$$

Tikhonov-regularized inversion constitutes a constrained minimization problem in which Φ_r is minimized subject to the constraint that Φ_m of equation (4.2) is not greater than a user-specified target measurement objective function Φ_m^l (referred to in PEST as the "limiting measurement objective function"). That is:

$$\Phi = \Phi_m + \mu\Phi_r \tag{4.8a}$$

$$\Phi_m \leq \Phi_m^l \quad (4.8b)$$

Here, μ is the "regularization weight factor". This is determined iteratively by PEST as part of the inversion process and is equivalent to a Lagrange multiplier in the solution of the constrained minimization problem (de Groot-Hedlin and Constable, 1990).

The threshold Φ_m^l should be specified to represent "adequate calibration". In an idealised case in which a model is free from structural defects, "adequate calibration" is theoretically defined by a level of model-to-measurement misfit that is commensurate with measurement noise **ε**. If the observation weight matrix of equation (4.2) is specified as the inverse of the covariance matrix of measurement noise C(**ε**), this level of fit is represented by $\Phi_m = N$, where N is the number of observations comprising the calibration dataset.

Where equation (4.7) applies, as it does in the present study, parameters estimated through Tikhonov-regularized inversion are given by:

$$\underline{\mathbf{k}} = (\mathbf{X}^t\mathbf{Q}_h\mathbf{X} + \mu\mathbf{Q}_r)^{-1}\mathbf{X}^t\mathbf{Q}_h\mathbf{h} \quad (4.9)$$

4.2.4 Paired model analysis and predictive bias

The paired model analysis methodology, first presented by Doherty and Christensen (2011), is summarized as follows:

1. A large ensemble of stochastic realizations of "reality" are generated based on expert geological knowledge. For each realization, model-generated "observations" equivalent to the available calibration dataset, as well as outputs pertaining to the prediction(s) of interest, are obtained through forward simulations.

2. A simplified model of the same system is calibrated against the "observations" generated by each of the complex model realizations. Each calibrated simplified model is then used to make the prediction(s) of interest, yielding an ensemble of complex-simple prediction pairs for each prediction.

3. A scatterplot of complex model prediction values (denoted as s) versus calibrated simplified model predictions (denoted as $\underline{s}$) is generated.

Discrepancies between a regression line through the *s*-versus-$\underline{s}$ scatterplot and the unity line represent predictive bias. Regression lines are calculated by Doherty and Christensen (2011) and in the present study as:

$$s = a + b\underline{s} \tag{4.10}$$

where *a* and *b* are the regression intercept and slope, respectively. A measure of scatter about the regression line (e.g., a prediction interval) provides a quantification of post-calibration predictive uncertainty.

4. As the final step in the case of employment of the paired model analysis in practice, the simplified model is calibrated against the available real world dataset and subsequently used to produce the prediction(s) of interest. The *s* verses $\underline{s}$ scatterplot is subsequently used to correct for any bias in the calibrated model prediction and quantify the associated uncertainty. This final step is not undertaken in the present study as no single hypothetical "reality" is considered.

Doherty and Christensen (2011) discuss two general sources of predictive bias as it manifests itself through *s*-versus-$\underline{s}$ scatterplots. Figure 4.1 provides a schematic representation of each of these sources as expressed through a paired model analysis *s*-versus-$\underline{s}$ scatterplot. The two sources are isolated for illustrative purposes in Figure 4.1, but are not mutually exclusive and some combination of the two may arise.

Firstly, an *s*-versus-$\underline{s}$ regression line slope *b* of less than unity indicates parameter surrogacy incurred through the calibration process. The potential for parameters to play surrogate roles during the calibration process in order to compensate for measurement noise and/or model structural defects is widely acknowledged (e.g., Clark and Vrugt, 2006; Beven, 2006; Spaaks and Bouten, 2013; Xu and Valocchi, 2015). Doherty and Christensen (2011) explain that this compensatory parameter behaviour inevitably includes the adjustment of parameter components that belong to the null space. They refer to this process as "null-space entrainment". As explained above, adjustment of null-space parameter components is by definition unsupported by the calibration dataset. It thus adds a component of random error to estimated model parameters, which increases the error variance in model predictions that are sensitive to these parameters. This is expressed as additional horizontal scatter in *s*-versus-$\underline{s}$

scatterplots (as it affects $\underline{s}$ and not s), which serves to decease the regression line slope. This results in discrepancies between the s-versus-$\underline{s}$ regression line and the unity line, and thus predictive bias. This bias varies with $\underline{s}$ and is more likely to affect more extreme predictions (see Figure 4.1).

Secondly, s-versus-$\underline{s}$ scatterplots may exhibit a degree of systematic offset relative to the unity line (which, in terms of equation (4.10), will be represented by a nonzero value of a when $b = 1$). This is due to "hardwired" error in null-space parameter components that are omitted from the simplified model. This component of predictive bias is independent of predictive error variance and is constant for all values of $\underline{s}$ (see Figure 4.1).

For ease of reference the two general sources of bias are herein referred to simply as "surrogacy-induced" bias and "hardwired" bias, respectively.

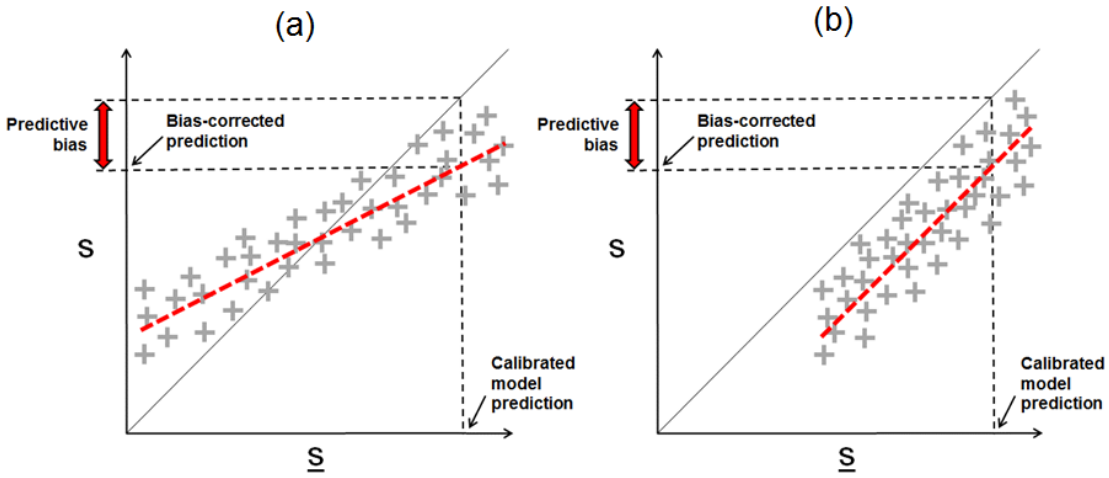

Figure 4.1. Schematic representation of the two general sources of predictive bias as identified through a paired model analysis s-versus-$\underline{s}$ scatterplot: (a) "surrogacy-induced" predictive bias; and (b) "hardwired" predictive bias.

4.2.6 Predictive uncertainty

4.2.6.1 Nonlinear analysis

The key metric by which the degree of success of the model inversion process is judged in the present study is the reduction in potential predictive error relative to prior predictive uncertainty. Prior uncertainty is calculated herein as the (sample) variance

of the prediction *s* calculated on the basis of an ensemble of expert-knowledge based stochastic realizations of the "reality" model. That is:

$$\sigma_s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[s_i - \frac{1}{n}\sum_{j=1}^{n}(s_j) \right]^2 \qquad (4.11)$$

where *n* is the number of realizations in the ensemble.

As described above, scatterplots of *s*-versus-$\underline{s}$ obtained through paired model analysis facilitate quantification of post-calibration predictive uncertainty (e.g., through 95% prediction intervals characterizing scatter about the *s*-versus-$\underline{s}$ regression line). In accordance with the function of paired model analysis, this uncertainty is bias-corrected and is thus smaller than the actual propensity for error in calibrated model predictions (see Chapter 2 of the present thesis for a thorough discussion and demonstration of this concept). As explained above, the aim of the present study is to characterize model predictive performance in the absence of a bias-correction methodology such as paired model analysis, rather than to utilise the bias-reducing benefits of the analysis itself. For this reason, *s*-versus-$\underline{s}$ scatterplots are not used to quantify potential predictive error, with outright discrepancies between calibrated model predictions and "true" predictions being analysed for this purpose instead (Chapter 2 of the present thesis provides a validation of predictive error variance estimates obtained through paired model analysis in this manner).

Post-calibration predictive error variance is calculated herein based on the ensemble of *n* complex-simple prediction pairs obtained through paired model analysis as:

$$\sigma_{s-\underline{s}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(s_i - \underline{s}_i) - \frac{1}{n}\sum_{j=1}^{n}(s_j - \underline{s}_j) \right]^2 \qquad (4.12)$$

As described above, predictive bias caused by parameter surrogacy-induced adjustment of null-space parameter components during the calibration process increases potential predictive error through increased random scatter. This increases predictive error variance. However, error variance does not account for systematic or average error in predictions. The total penchant for predictive error is the sum of both its variance and its systematic component, this being quantifiable through mean square error (MSE) as (e.g., Parkin et al., 1988):

$$\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2 \tag{4.13a}$$

MSE is calculable simply as:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \underline{s}_i)^2 \tag{4.13b}$$

From equation (4.13b) and equation (4.12), equation (4.13a) can be rewritten as:

$$\mathrm{MSE} = \sigma^2_{s-\underline{s}} + \left[\frac{1}{n}\sum_{i=1}^{n}(s_i - \underline{s}_i)\right]^2 \tag{4.13c}$$

It is clear that the "bias" term of equation (4.13a) is equal to the mean predictive error (the square of which is summed with predictive error variance to give MSE). It should be emphasised that this term accounts exclusively for what is referred to herein in as "hardwired" bias, caused by consistent (i.e., nonzero mean) predictive error. The influence of "surrogacy-induced" bias, as discussed above, is to inflate the "variance" term of equation (4.13a) rather than the "bias" term.

4.2.6.2 Linear analysis

An efficient linear estimate of prediction uncertainty variance of a (scalar) prediction $\underline{s}$ made by a calibrated model is given by (for details the reader is referred to, for example, Christensen and Doherty, 2008; Dausman et al., 2010; Doherty 2015):

$$\sigma^2_{\underline{s}} = \mathbf{y}^t\mathrm{C}(\mathbf{k})\mathbf{y} - \mathbf{y}^t\mathrm{C}(\mathbf{k})\mathbf{X}^t[\mathbf{X}\mathrm{C}(\mathbf{k})\mathbf{X}^t + \mathrm{C}(\boldsymbol{\varepsilon})]^{-1}\mathbf{X}\mathrm{C}(\mathbf{k})\mathbf{y} \tag{4.14}$$

where $\mathbf{y}$ is a vector comprising sensitivities of s to model parameters $\mathbf{k}$, $\mathrm{C}(\mathbf{k})$ is the covariance matrix describing prior parameter variability based on expert knowledge, and $\mathrm{C}(\boldsymbol{\varepsilon})$ is the measurement noise covariance matrix. The first term on the right-hand-side of equation (4.14) represents prior uncertainty (and thus is the linear equivalent to equation (4.11)). The second term on the right-hand-side of equation (4.14) represents the amount by which this uncertainty is reduced through the constraints imposed by the available (noisy) calibration dataset.

4.3 Synthetic case study

4.3.1 Model description

Figure 4.2 shows the synthetic example model domain. It is a single-layer unconfined aquifer with dimensions 800 m by 800 m. Water enters the system as uniform diffuse recharge at 3.8E-5 m/day and fixed inflow through the northern boundary at 9.1E-3 m^3/day per meter length of boundary. Groundwater exits the domain through the southern boundary where heads are fixed at 5 m. The western and eastern edges of the domain are defined by no-flow boundaries. Steady-state groundwater flow within the domain is simulated using MODFLOW 2000 (Harbaugh et al., 2000) with a finite-difference grid comprising 6400 cells with dimensions of 10 m × 10 m.

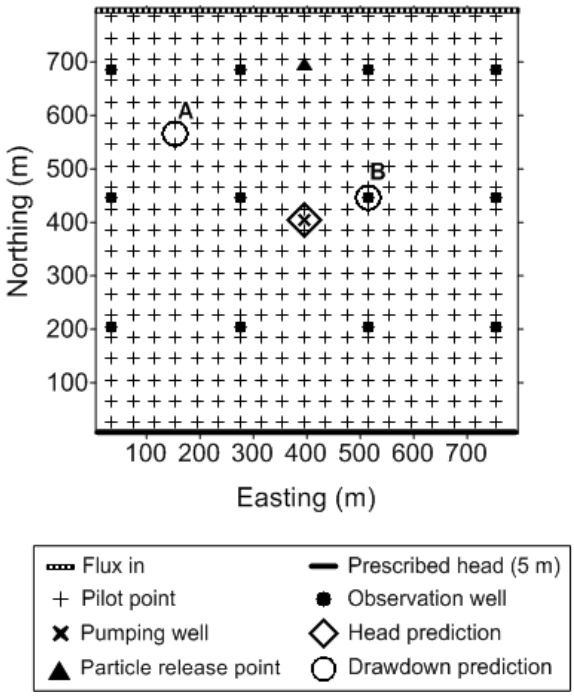

Figure 4.2. Synthetic example model domain and boundary conditions, including locations of pilot points, observation wells, pumping well, particle release point, head prediction point and drawdown prediction points.

4.3.2 Stochastic *K* field generation

One thousand "reality" $\log_{10}K$ field realizations were generated stochastically based on synthetic expert knowledge for use in the paired model analysis described in section 4.2.4. As described above, the nature of subsurface geology in the present synthetic example is similar to one of the test cases considered by Zimmerman et al. (1998). It includes discrete faults (acting as preferential flow features) embedded in a "background" geostatistical field.

Background $\log_{10}K$ variability was generated using the sequential Gaussian simulation method (Deutsch and Journel, 1998), and is defined by a log-exponential variogram with a mean of -1 $\log_{10}$(m/day), a sill of 0.1 $\log_{10}$(m/day) and a range of 300 m.

The number of faults per stochastic $\log_{10}K$ field realization was generated as a random number between 1 and 3 with a central tendency represented by a discrete probability distribution of (0.25, 0.5, 0.25). The length and strike of each fault were also randomized, the former from a uniform distribution between 300 m and 800 m and the latter from a uniform distribution between 35°T and 55°T. The *K* within each fault is a homogeneous value generated from a log-normal distribution with a mean of 1 $\log_{10}$(m/day) (i.e., two orders of magnitude greater than the mean of the geostatistical background field) and a standard deviation of 0.17 $\log_{10}$(m/day). Additional constraints on the stochastic fault generation process included a specified minimum separation of 100 m between fault centres, as well as a minimum distance of faults from lateral no-flow boundaries of 50 m.

Illustrative examples of the stochastically generated $\log_{10}K$ fields are provided in Figure 4.3. This clearly demonstrates the influence of the presence of the faults upon the potentiometric surface (under non-pumping/calibration conditions), advective transport behaviour, and drawdown distribution under pumping conditions.

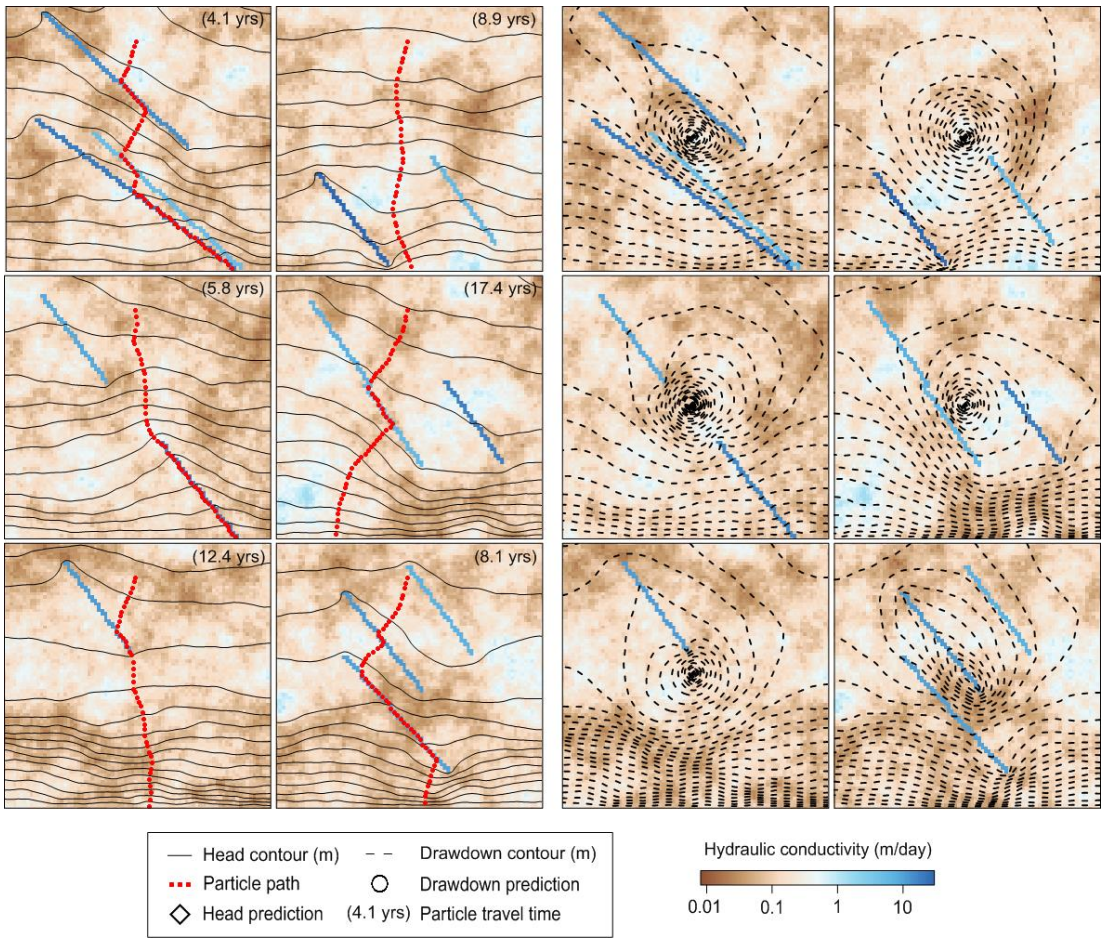

Figure 4.3. Examples of stochastic $\log_{10}K$ field realizations, including, left, steady-state hydraulic head distribution under calibration conditions (1 m increments) with associated particle behaviour and, right, steady-state drawdown contours (0.05 m increments) under pumping conditions.

4.3.3 Calibration

The calibration dataset in the present synthetic example comprises 12 observations of steady-state hydraulic head at the locations indicated in Figure 4.2. Gaussian noise with a standard deviation of 0.1 m was added to each of the 1000 sets of head observations, generated by the 1000 stochastic "reality" $\log_{10}K$ field realizations described above, in order to generate the ensemble of synthetic calibration datasets required for paired model analysis as described in section 4.2.4.

$\text{Log}_{10}K$ values are estimated at 380 evenly distributed pilot points (see Figure 4.2) through regularized inversion using PEST (Doherty, 2016a). Measurement weights for the 12 observations in the calibration dataset were each set at 10.0, this being the inverse of the measurement noise standard deviation of 0.1 m. Φ_m^l was accordingly set

to 12.0, this representing a level of model-to-measurement misfit commensurate with measurement noise as described above.

4.3.4 Regularization weighting strategies

All 380 pilot points were assigned a "preferred value" (represented by the regularization observation vector **r** of equation (4.6)) of -1 $\log_{10}$(m/day), this being equal to the mean of the variogram used to generate the Gaussian "background" $\log_{10}K$ fields.

PEST allows direct user supply of a parameter covariance matrix to supplement Tikhonov regularization observations, the inverse of which is calculated and employed as a regularization weight matrix (i.e., $\mathbf{Q}_r$ of equation (4.6)). A discussed above, the categorical nature of expert geological knowledge in the present synthetic case precludes its holistic expression in terms of a covariance matrix. There thus exists no theoretically optimal $\mathbf{Q}_r$ matrix for regularized inversion in the present synthetic case (the reader is referred to Chapter 2 of the present thesis for a discussion of optimal regularization). The parameter and predictive outcomes arising through application of four alternative (suboptimal) regularization weighting strategies are thus examined.

The first weighting strategy involves uniform regularization weights. That is, no parameter covariance matrix is supplied to PEST for regularization weighting purposes. This is equivalent to defining $\mathbf{Q}_r = \mathbf{I}$. All pilot point parameters therefore possess an equal and independent propensity for variability during the inversion process (i.e., no preferential spatial correlation structure controls deviations from preferred values). This is referred to hereafter as the "uniform" regularization weighting strategy.

For the second regularization weighting strategy, a covariance matrix was constructed based on the log-exponential variogram that was used to generate the stochastic "background" $\log_{10}K$ fields of "reality". This covariance matrix is denoted as $C(\mathbf{k}_b)$. In the absence of the existence of faults in the synthetic system under study, this strategy would represent the theoretically optimal formulation of $\mathbf{Q}_r$. This is hereafter referred to as the "background" regularization weighting strategy.

For the third regularization weighting strategy a covariance matrix was derived empirically from the suite of stochastic fields. All 1000 fields (defined by cell-by-cell variability) were upscaled (via least squares) to pilot point resolution, based on which

a covariance matrix was calculated. This is referred to hereafter as the "empirical" weighting strategy.

The fourth regularization weighting strategy is heuristic in nature. (This strategy was provoked by an observed negligible impact of the above "empirical" weighting strategy relative to the "background" weighting strategy, which is discussed in due course.) The empirical covariance matrix C(**k**) was modified according to:

$$C'(\mathbf{k}) = \mathbf{A}C(\mathbf{k}) \tag{4.15}$$

where C'(**k**) denotes the modified empirical covariance matrix, and **A** acts as selection matrix, formulated to effectively perform a filtering operation on the off-diagonal elements of C(**k**). Only the 0.5% (this being an arbitrarily selected threshold) of off-diagonal elements with the largest magnitude were retained, with the remainder reduced to a negligibly small value. In other words, the covariance matrix C(**k**) was manipulated to encompass only the most prominent inter-pilot-point spatial correlations identified empirically based on the 1000 stochastic realizations. This is hereafter referred to as the "heuristic" weighting strategy.

Figure 4.4 displays graphically the three covariance matrices used for the "background", "empirical" and "heuristic" regularization weighting strategies, respectively.

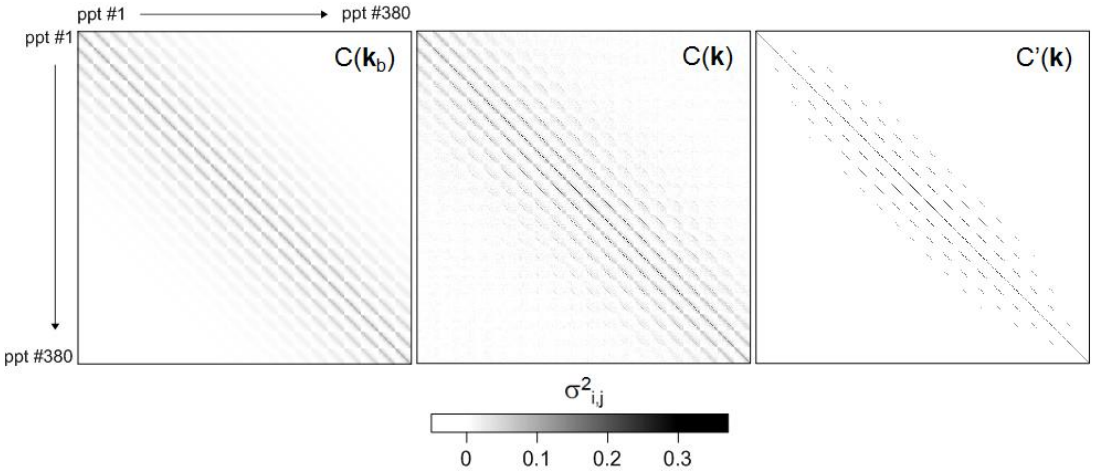

Figure 4.4. (Left) covariance matrix C($\mathbf{k}_b$) pertaining to the log-exponential variogram used to generate true background $\log_{10}K$ field variability; (middle) empirically derived covariance matrix C(**k**); and (right) modified empirical covariance matrix C'(**k**) upon which the "heuristic" regularization weighting strategy is based.

4.3.5. Predictive analysis

Five model predictions are examined in the present study. The movement of a particle released at the point indicated in Figure 4.2 is simulated using the ADV2 package of MODFLOW 2000 (Anderman and Hill, 2001). The time taken for the particle to exit the domain at the southern boundary and its exit location comprise two predictions. Other predictions include the steady-state hydraulic head in the centre of the domain, and steady-state drawdown at two locations as a result of simulated abstraction at a rate of 3.1 m^3/day from the centre of the domain (see Figure 4.2). The simulated abstraction comprises a separate predictive simulation and does not affect the advective transport nor hydraulic head predictions.

Examination of the predictive success or otherwise of the pilot-point-based regularized inversion approach in the presence of categorical heterogeneity is based primarily on the following calculations made for each of the five predictions:

1. Prior prediction uncertainty variance calculated from equation (4.11), based on the ensemble of 1000 stochastic realizations of the "reality" model. This is the benchmark against which the ultimate success or otherwise of the calibration process is judged, as it is the degree of uncertainty that exists based on expert geological knowledge alone.

2. Pre-calibration MSE calculated using equation (4.13b), based on the ensemble of 1000 expert knowledge-based realizations of the "reality" model, where the model prediction $\underline{s}$ is that made by the uncalibrated model. Pre-calibration MSE ideally coincides with prior uncertainty variance (i.e., where the "bias" term in equation (4.13a) is zero). However, as will be demonstrated, the potential for error in model predictions made by the uncalibrated model in the present synthetic example (i.e., based on the initial uniform $\log_{10}K$ field of -1 $\log_{10}$(m/day)) is not symmetrical with respect to prior uncertainty, due to the systematic influence of the faults. Thus the "bias" term of equation (4.13a) is nonzero. As such, pre-calibration MSE elucidates the additional potential for wrongness in model predictions initially inherited through employment of the pilot-point-based regularization method in the presence of categorical heterogeneity. It thus provides additional insight into the gains achieved through the calibration process itself, relative to uncalibrated model

3. A linear estimate of posterior prediction uncertainty variance based on equation (4.14). This is calculated in order to add additional perspective to the outcomes of the calibration process, through demonstration of whether or not the "true" prediction value is successfully captured within estimated uncertainty margins. This, after all, ultimately defines the success or failure of the modelling process as discussed above. This linear uncertainty analysis method was employed instead of a calibration-constrained nonlinear technique such as the null-space Monte Carlo method (see, for example, Tonkin and Doherty, 2009; Herckenrath et al., 2011) due to its efficiency for the sake of the present study, as well as perhaps greater relevance to practical situations due to its applicability as a result of this efficiency. A nonlinear technique through which expression of the faults is made possible is expected to improve the performance of post-calibration uncertainty assessment in terms of the chances of capturing the true prediction. However, this is beyond the scope of the present study.

4. Post-calibration predictive error variance quantified through equation (4.12), based on the 1000 complex-simple prediction pairs obtained through paired model analysis (repeated for each of the five regularization weighting strategies). This is a key component of total post-calibration potential predictive error (see equation (4.13c)). As discussed above, it quantifies the influence of predictive bias caused by calibration-induced parameter surrogacy.

5. Post-calibration prediction MSE calculated through equation (4.13b), based on the 1000 complex-simple prediction pairs obtained through paired model analysis (repeated for each of the five regularization weighting strategies). This is the measure of the total potential for error in predictions made by the calibrated model, accounting for both post-calibration predictive error variance as well as systematic (i.e., nonzero mean) predictive error (see equation (4.13c)).

4.4 Results

4.4.1 Prior uncertainty of predictions

The prior uncertainty of predictions is obtained through unconstrained Monte Carlo analysis based on the suite of 1000 stochastic "reality" $\log_{10}K$ field realizations. Histograms for the ensemble of "true" prediction values are displayed in Figure 4.5. Gaussian probability density functions (overlaid on the Figure 4.5 histograms) based on the calculated prediction variance σ^2_s (obtained through equation 4.11) were adjudged as reasonable approximations of the histograms for all predictions and thus for convenience are used to represent prior uncertainty in the remainder of the study.

Also indicated in Figure 4.5 are the prediction values made by the uncalibrated model (i.e., based on the preferred value-populated homogeneous $\log_{10}K$ field of -1 $\log_{10}$(m/day)). Each prediction made by the uncalibrated model is clearly not central with respect to the prior probability distribution. This represents pre-calibration predictive bias introduced by the presence of the faults.

4.4.2. Estimated parameter field characteristics

Average measurement objective function Φ_m values achieved across the 1000 calibrated model pairs is equal to 12.0 for each of the four regularization strategies, this being equal to the specified target measurement objective function Φ_m^l. Furthermore, a high degree of consistency in Φ_m across the 1000 calibrated model pairs for each regularization strategy is indicated by small Φ_m standard deviations of 0.05, 0.03, 0.03 and 0.2 for the "uniform", "background", "empirical" and "heuristic" regularization weighting strategies, respectively. Thus all models comprising the present analysis can be considered "well-calibrated" from a history-matching point of view.

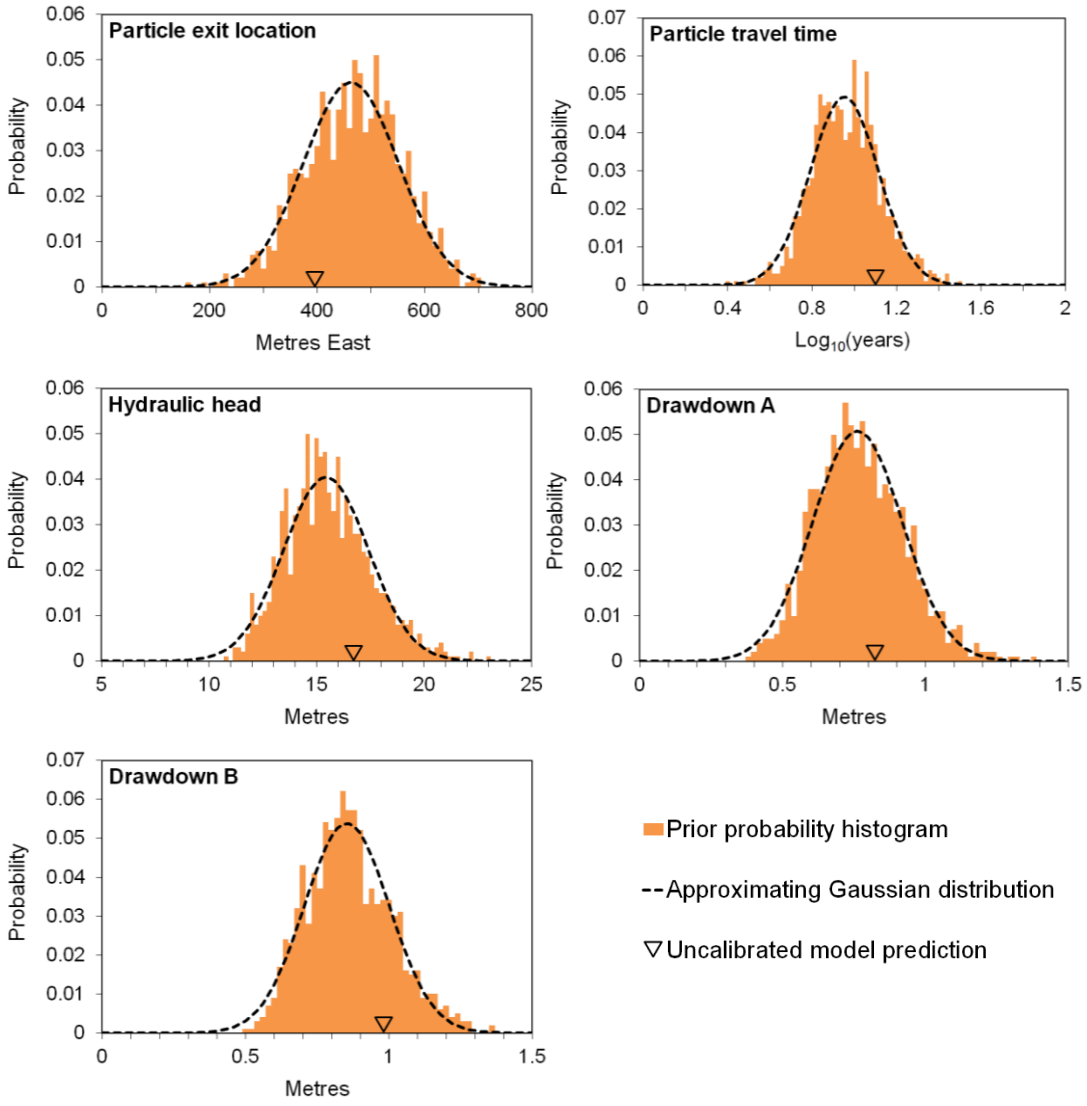

Figure 4.5. Prediction histograms based on unconstrained Monte Carlo analyses using all 1000 "reality" $\log_{10}K$ fields. Overlain are the theoretical Gaussian probability density functions used to approximate each histogram. Also indicated is the pre-calibration value of each prediction (i.e., as made by the uncalibrated model comprising a homogeneous $\log_{10}K$ field of -1 $\log_{10}$(m/day)).

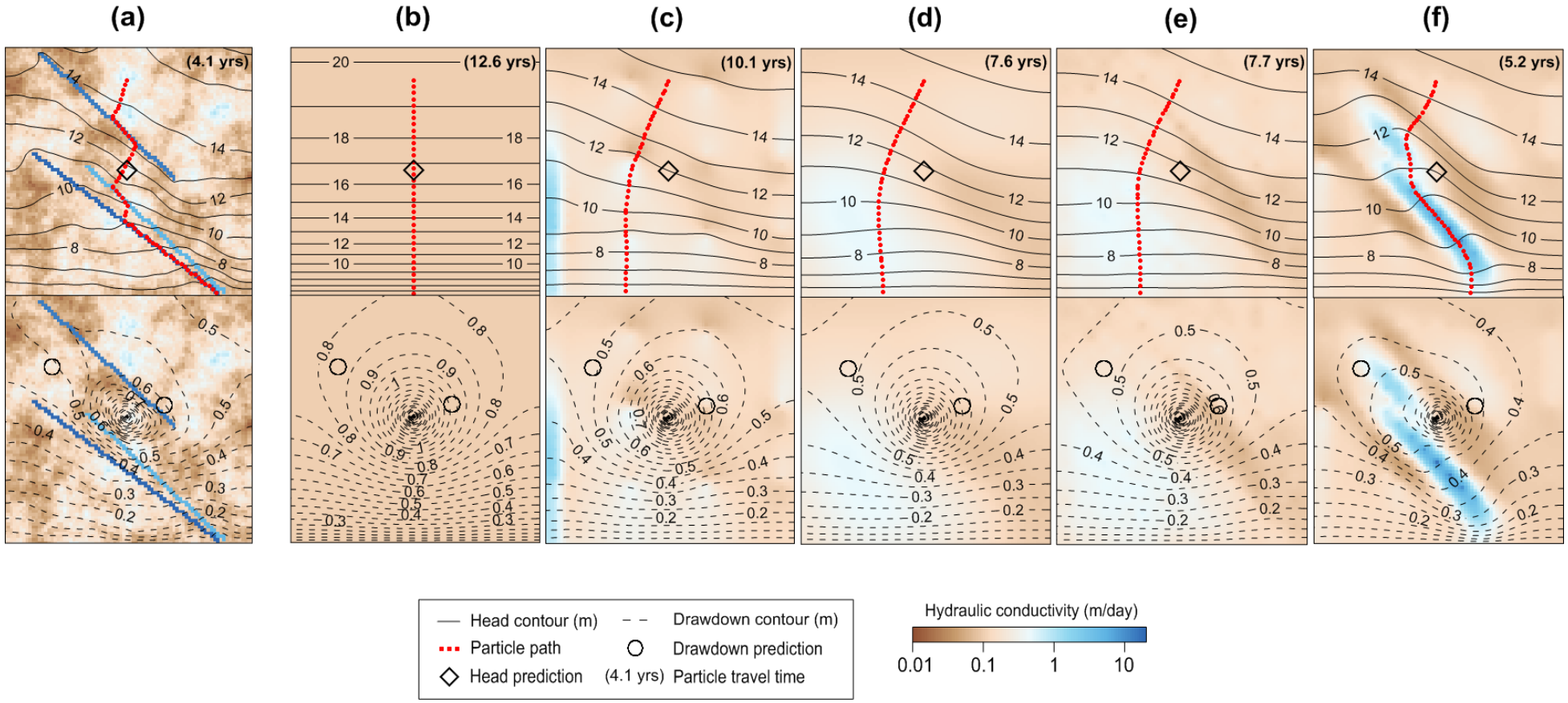

Figure 4.6. (a) arbitrary realization of a "reality" $\log_{10}K$ field, (b) uncalibrated model (based on pre-calibration preferred values), and post-calibration counterparts of the realization shown in (a) based on (c) "uniform", (d) "background", (e) "empirical" and (f) "heuristic" regularization weighting strategies. Information is displayed pertaining to both particle fate prediction under calibration conditions (top) and drawdown prediction under pumping conditions (bottom).

Figure 4.6 depicts an arbitrary stochastic "reality" field realization and its post-calibration counterparts based on each regularization weighting strategy (as well as the uncalibrated model). Prediction-related information for both calibration (non-pumping) and pumping conditions is also displayed.

As discussed above, the inevitable cost of attaining a unique solution to the inverse problem is an estimated parameter field that is smoother than the true field. This is clearly evident for all estimated $\log_{10}K$ fields displayed in Figure 4.6. $\text{Log}_{10}K$ field heterogeneity arising through the "uniform" weighting scheme is clearly a function of observation well locations (the latter are displayed in Figure 4.2). Locally "tweaked" heterogeneity manifesting as "bullseyes" is widely condemned as a data-fitting artefact that is unlikely to constitute a useful representation of the system (e.g., Freyberg 1988; Voss, 2011; Black and Black, 2012; Kourakos and Mantoglou, 2012). These features are free to arise in the absence of a correlation-based regularization weighting scheme and consequential spatial independence of pilot-point adjustment.

The effect of all covariance matrix-based regularization weighting schemes in removing "bullseye"-type heterogeneity is clear from Figure 4.6. The "background" weighting scheme promotes variability in the estimated $\log_{10}K$ field that is in accordance with the large-scale background field heterogeneity defining the true field. The field estimated on the basis of the "empirical" regularization weighting scheme shares very similar large-scale features, including a faint presence of northwest-southeast trending elongation within the $\log_{10}K$ field. In contrast, the influence of the "heuristic" regularization weighting strategy is profound, with it clearly promoting the expression of "fault-like" features. However, these features are obviously of a highly surrogate nature; they are significantly thicker and have not arisen in the same locations as the faults in the "reality" field. Additionally, regions of low $\log_{10}K$ also follow the same "fault-like" correlation structures, which is not a feature of the "reality' field conceptualisation. Despite this, the field may in some respects be considered to be more in harmony with the key geologic features of "reality" and thus more geologically plausible than the estimated fields in which the existence of the faults is neglected.

The characteristics of the estimated $\log_{10}K$ fields are explicated through eigenanalysis. Singular value decomposition (SVD) of the matrix **X** of equation (4.1) yields:

$$\mathbf{X} = \mathbf{USV}^T \tag{4.16}$$

The columns of **U** and **V** are unit vectors that span the output and parameter spaces of **X**, respectively. **S** is a diagonal matrix containing the ranked singular values of **X**. These are displayed in Figure 4.7 on a semi-log plot. The first two singular values are relatively dominant, indicating that adjustment of just two combinations of parameters may be sufficient to achieve the majority of model-to-measurement misfit reduction during the history matching process. (This is particularly the case in the presence of measurement noise – as is artificially included in the present synthetic example – where a less-than-perfect fit is sought in order to avoid overfitting. The reader is referred to Chapter 2 of the present thesis for a thorough discussion of overfitting.)

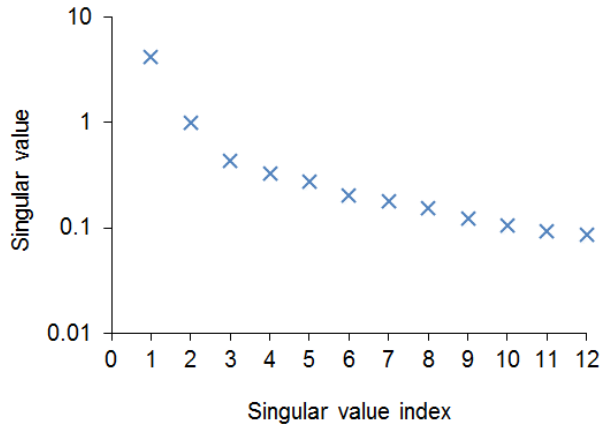

Figure 4.7. Singular values calculated for the Jacobian matrix **X** (based on equation (4.17)).

SVD of any covariance matrix C(**k**) yields:

$$\mathrm{C}(\mathbf{k}) = \mathbf{VSV}^T \tag{4.17}$$

Here, the eigenvectors comprising the columns $\mathbf{v}_i$ of the matrix **V** are the set of orthogonal basis functions spanning the model domain (Doherty et al., 2010).

Figure 4.8 provides a comparison between selected eigenvectors of C($\mathbf{k}_b$), C(**k**) and C'(**k**). The elements of the eigenvectors are arranged in Figure 4.8 in accordance with the spatial distribution of the pilot points within the model domain. (Note that some

values comprising the eigenvectors of C'(**k**) exceed the colour scale presented in Figure 4.8, which is truncated to enhance contrast in the C($\mathbf{k}_b$) and C(**k**) eigenvectors.)

Inspection of Figure 4.8 shows a high degree of similarity between C($\mathbf{k}_b$) and C(**k**) in the eigenvectors pertaining to the largest two singular values in particular (i.e., $\mathbf{v}_1$ and $\mathbf{v}_2$). This indicates that, despite the stochastic inclusion of faults within the empirical covariance matrix calculation process, the eigencomponents corresponding to the largest singular values remain dominated by the larger scale background field correlation characteristics. This explains the similarity of parameter fields estimated through regularized inversion employing either C($\mathbf{k}_b$) or C(**k**), as exemplified by Figure 4.5. More prominent "fault-like" correlation structures are visible in the fourth and fifth C(**k**) eigenvectors in particular. However the features of these less dominant eigenvectors do not find prominence in the estimated parameter fields due to the relative dominance of a very small number of singular values in the inversion problem as discussed above.

Inspection of the eigencomponents of the modified empirical covariance matrix C'(**k**) clearly demonstrates the effect of the ad hoc modification process based on equation (4.15). The eigenvectors of C'(**k**) are wholly comprised of "fault-like" correlation structures occupying different parts of the model domain.

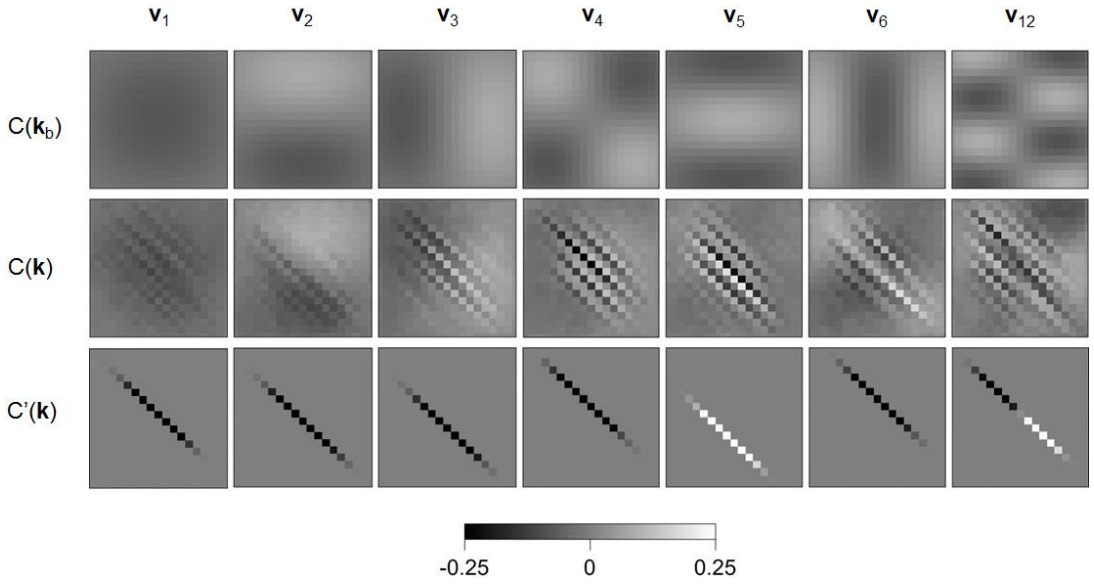

Figure 4.8. Selected eigenvectors (arranged in accordance with the spatial distribution of pilot points within model domain) of the background $\log_{10}K$ covariance matrix C($\mathbf{k}_b$); the empirical $\log_{10}K$ covariance matrix C(**k**); and the modified empirical $\log_{10}K$ covariance matrix C'(**k**).

Tables 4.1 and 4.2 provide a quantitative summary of the parameter outcomes of calibration. For the "reality" $\log_{10}K$ fields and corresponding fields estimated through each regularization strategy, Table 4.1 contains the mean, minimum and maximum standard deviation of $\log_{10}K$ within each cell of the model domain calculated on the basis of all 1000 realizations. Table 4.2 summarizes the standard deviations of error in estimated $\log_{10}K$ field cells relative to the corresponding cell in the "reality" field (denoted as $\log_{10}K_{true}$).

Table 4.1. Summary of standard deviations of $\log_{10}K$ in each model cell within the domain calculated based on all 1000 realizations.

$\sigma_{\log_{10}K}$ [$\log_{10}$(m/day)]	"Reality"	"Uniform"	"Background"	"Empirical"	"Heuristic"
Mean	0.41	0.15	0.19	0.19	0.22
Minimum	0.29	0.01	0.05	0.04	0.00
Maximum	0.63	0.35	0.30	0.29	0.64

Table 4.2. Summary of standard deviations of error in estimated $\log_{10}K$ in each model cell within the domain calculated based on all 1000 realizations.

$\sigma_{\log_{10}K_{true} - \log_{10}K}$ [$\log_{10}$(m/day)]	Pre-cal.	"Uniform"	"Background"	"Empirical	"Heuristic"
Mean	0.41	0.39	0.38	0.38	0.42
Minimum	0.29	0.23	0.22	0.22	0.23
Maximum	0.63	0.62	0.62	0.61	0.75

The mean $\sigma_{\log_{10}K}$ values of Table 1a highlight the relative "smoothness" of all calibrated fields relative to the complex fields, with the values for each estimated field substantially smaller than for the "reality" fields. The influence of the "background" (and "empirical") regularization weighting strategies in removing localised "bullseyes", and inducing broader-scale correlation structures in accordance with expert knowledge, is clear from the increase in mean $\sigma_{\log_{10}K}$ accompanied by a decrease in minimum and maximum $\sigma_{\log_{10}K}$. Despite the "more realistic" correlation structures introduced through the "background" regularization weighting scheme, a high degree of surrogacy observable remains in the estimated field. In their study of the calibration of a defective model, White et al. (2014) discuss the spreading of parameter surrogacy across larger regions of the model domain as a result of pre-calibration (Karhunen-Loève) transformation of parameters in accordance with expert knowledge. In the present study parameter surrogacy is inherent in the chosen calibration approach, with a continuous parameter field replacing discrete features. The spreading of parameter surrogacy across a larger area is visible in Figure 4.6d. In order to reproduce the lowering of the hydraulic head in the southwestern part of the

model domain (caused by the faults in the "reality" model), a relatively large area of higher $\log_{10}K$ is introduced. Table 4.2 indicates that cell-by-cell error in estimated $\log_{10}K$ is barely reduced through the "background" regularization weighting scheme relative to the "uniform" scheme. Thus the error in estimated $\log_{10}K$ at any given point in the domain is effectively unchanged despite the introduction of more realistic broad-scale variability.

The "heuristic" weighting strategy promotes the highest overall variability in estimated $\log_{10}K$, with localised extremes. Some cells within the model domain exhibit a $\sigma_{\log_{10}K}$ as low as 0.00 across the ensemble of 1000 estimated fields, whilst in others the variability is approximately equivalent to the true maximum degree of $\log_{10}K$ variability found in "reality". This is accompanied by a general increase in estimated $\log_{10}K$ error across the domain (resulting in a mean $\log_{10}K$ error that is in fact greater than the mean pre-calibration $\log_{10}K$ error), providing quantitative confirmation of an increased degree of parameter surrogacy incurred through the "heuristic" regularization weighting strategy.

4.4.2 Predictive outcomes

4.4.2.1 s-versus-s scatterplot characteristics

All *s*-versus-*s* scatterplots are displayed in Figure 4.9, accompanied by regression lines as defined by equation (4.10) and 95% prediction intervals (based on Draper and Smith (1998), eq. 1.4.12). Regression statistics pertaining to each *s*-versus-*s* scatterplot are provided in Table 4.3. For the sake of conciseness, the regression statistic *a* is not presented due to limited relevance to discussion herein. (Note: *s*-versus-*s* scatterplots for drawdown predictions are based on 997 realizations, due to "drying" and subsequent deactivation of the cell containing the pump in three stochastic realizations.)

From Figure 4.9 and Table 4.3, the prediction-specific outcomes of the calibration process are clear. Substantial predictive bias exists in some predictions, including both hardwired bias (indicated by a lateral offset in the regression line) and parameter surrogacy-induced bias (indicated by a regression line slope of less than unity). Also clear is the influence of regularization weighting scheme alteration, affecting regression line slope, offset and scatter (as indicated by r^2).

Table 4.3. Regression line slope b and coefficient of determination r^2 pertaining to the s-versus-$\underline{s}$ scatterplots of Figure 4.9.

	Uniform		Background		Empirical		Heuristic	
Prediction	b	r^2	b	r^2	b	r^2	b	r^2
Exit point	0.62	0.21	0.57	0.24	0.59	0.25	0.56	0.14
Log_{10} time	1.31	0.18	0.87	0.20	0.92	0.21	0.59	0.17
Head	1.07	0.96	1.02	0.97	1.02	0.97	0.93	0.93
Drawdown A	1.16	0.81	1.07	0.83	1.08	0.83	1.00	0.84
Drawdown B	1.16	0.82	0.97	0.84	0.97	0.85	0.61	0.60

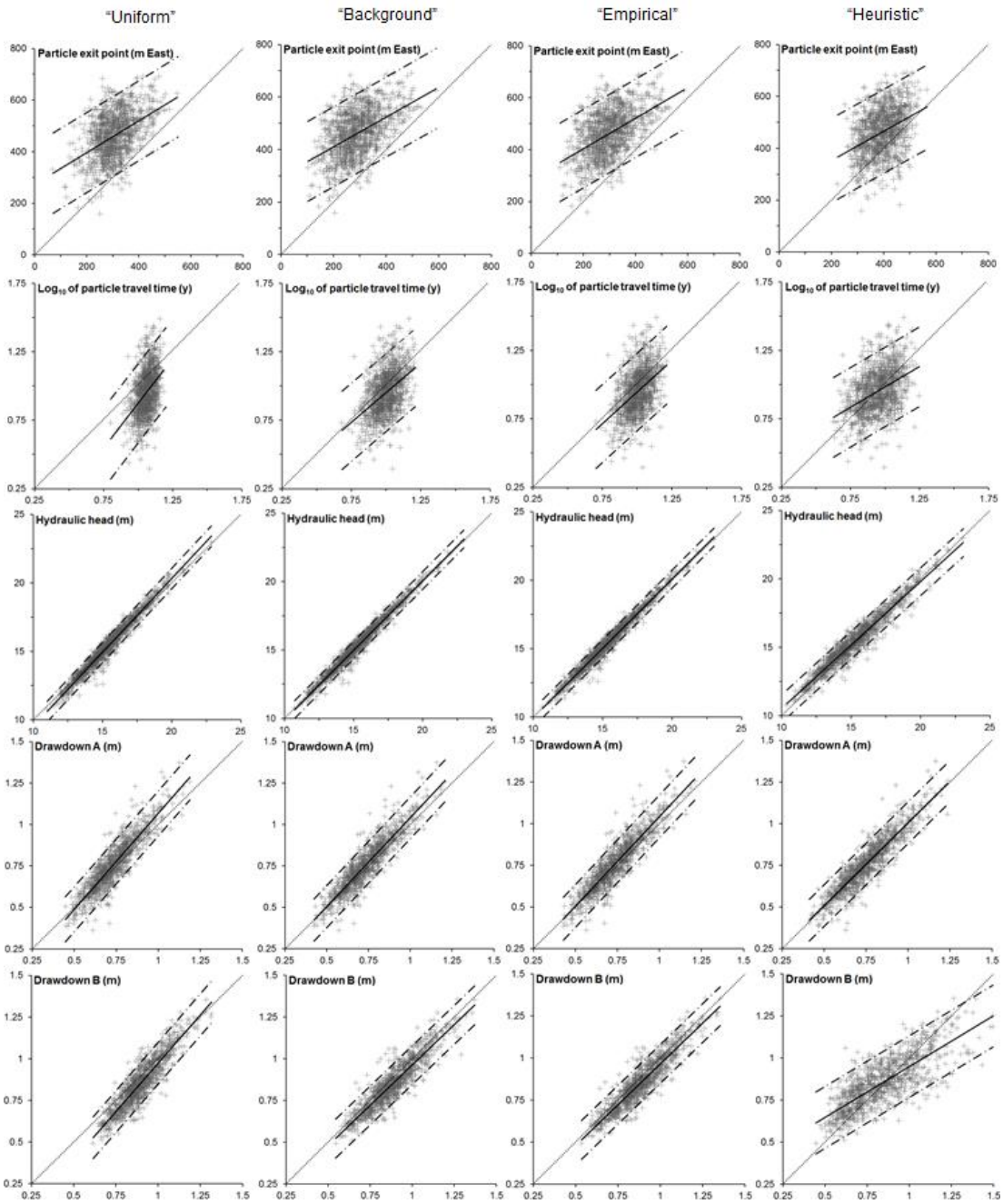

Figure 4.9. s-versus-$\underline{s}$ scatterplots for all five predictions based on each regularization weighting strategy. Axis units are as indicated in the respective plot headings.

Figure 4.9 and Table 4.3 show that all *s*-versus-$\underline{s}$ regression line slopes except that of the particle exit location prediction are greater than unity for calibration based on the "uniform" regularization weighting strategy. This reflects limited predictive range of the ensemble of calibrated models relative to their "reality" model counterparts. This is analogous to the outcome of "underfitting" in a paired model analysis context as discussed in Chapter 2 of the present thesis. In the present case, all models are "well-calibrated" from a goodness of fit point of view (i.e., model-to-measurement misfit is commensurate with measurement noise as explained above). However, the highly localised "bullseye"-type parameter compensation allowed by the lack of a covariance-based regularization weighting scheme serves to achieve the target measurement objective function without the need for broader-scale parameter adjustment. The overall outcome is unrealistically (i.e., compared with "reality") limited domain-wide $\log_{10}K$ variability in the "uniform" case (see Table 4.1), which translates to an unrealistically limited range in predictions. As Figure 4.9 and Table 4.3 demonstrate, *s*-versus-$\underline{s}$ regression line slopes in all predictions are reduced through use of the "background" regularization weighting strategy. This reflects an increased predictive range of the suite of calibrated models attributable to the broad-scale correlation structures of the "background" $\log_{10}K$ variogram. For all predictions other than particle exit location, this results in an *s*-versus-$\underline{s}$ slope that is closer to unity than in the "uniform" case. As will be demonstrated, this corresponds to a reduction in predictive error variance consistent with the theoretical basis of *s*-versus-$\underline{s}$ analysis explored in detail in Chapter 2 of the present thesis. Thus, model performance for the making of most predictions depends upon the representation of a more realistic degree of broad-scale $\log_{10}K$ variability.

The particle exit location prediction is an exception, for which the "background" covariance-based regularization weighting strategy inflates predictive bias relative to the "uniform" strategy. In contrast all other predictions, substantial "surrogacy-induced" bias exists in the exit location prediction following regularized inversion using the "uniform" strategy (i.e., an *s*-versus-$\underline{s}$ slope of 0.62 as shown in Table 4.3). This reflects the extreme sensitivity of this prediction to the presence of the faults, and thus sensitivity to the surrogate nature of the estimated "smooth" parameter fields that do not incorporate faults (resulting in a high degree of random error expressed as "surrogacy-induced" bias). Whilst all other predictions benefit from the more realistic degree of broad-scale $\log_{10}K$ variability introduced through the "background"

regularization strategy through a more realistic predictive range of the calibrated model, the broadened range of exit location predictions serves only to increase the degree of random error (i.e., the *s*-versus-$\underline{s}$ slope is further reduced to 0.57).

Similar to the relative influence of the "background" regularization weighting strategy compared to the "uniform" strategy, the "heuristic" strategy further reduces the *s*-versus-$\underline{s}$ regression line slope for all predictions (see Figure 4.9 and Table 4.3). The translation of these *s*-versus-$\underline{s}$ characteristics in terms of predictive is highly prediction-specific. For some predictions, for example drawdown "A", the *s*-versus-$\underline{s}$ regression line slope remains greater than unity through use of the "background" or "empirical" strategies, despite some reduction relative to the "uniform" strategy. Employment of the "heuristic" weighting strategy further reduces the regression line slope such that it equals unity, this corresponding to further reduction in predictive error variance as will be demonstrated below. Simultaneously, however, the reduction in *s*-versus-$\underline{s}$ regression line slope for other predictions such as drawdown "B" represents "surrogacy-induced" predictive bias and corresponds to inflation of potential predictive error as will be demonstrated below.

It is clear from Figure 4.9 that changes in the degree of consistent offset in the *s*-versus-$\underline{s}$ scatterplots of some predictions accompanies the changes in regression line slope discussed above. The translation of these characteristics in terms of overall predictive performance is explored in the following subsection.

4.4.2.2 Potential predictive error

Post-calibration predictive error probability density functions are displayed in Figure 4.10. For the sake of clarity these are represented by Gaussian distributions (which were adjudged to be reasonable representations of the histograms in the same manner were prior probability histograms) based on $\sigma^2_{s\text{-}\underline{s}}$ calculated using equation (4.12). Also included in Figure 4.10 are the prior probability density functions, as well as the mean predictive error of the uncalibrated model. The latter represents pre-calibration bias as discussed above. (For the sake of clarity, only the mean is represented in Figure 4.10 instead of the entire probability density function, given that the width of the latter is identical to the prior uncertainty probability density function and it can thus easily be visualised as such, with its peak coinciding with the indicated mean.) Finally, Figure 4.10 also includes probability density functions representing linear post-calibration

uncertainty variance quantified through equation (4.14). Note that the latter is based on use of the background covariance matrix C($\mathbf{k}_b$) in equation (4.14). Use of the either the empirical covariance matrix C(**k**) or modified empirical covariance matrix C'(**k**) returns comparable or narrower probability density functions. It is emphasized that these linear estimates are provided solely as an example of the potential performance of a computationally manageable post-calibration uncertainty assessment in the present context. The performance of alternative (e.g., nonlinear) methods is likely to differ, but would not be expected to invalidate the conclusions drawn herein.

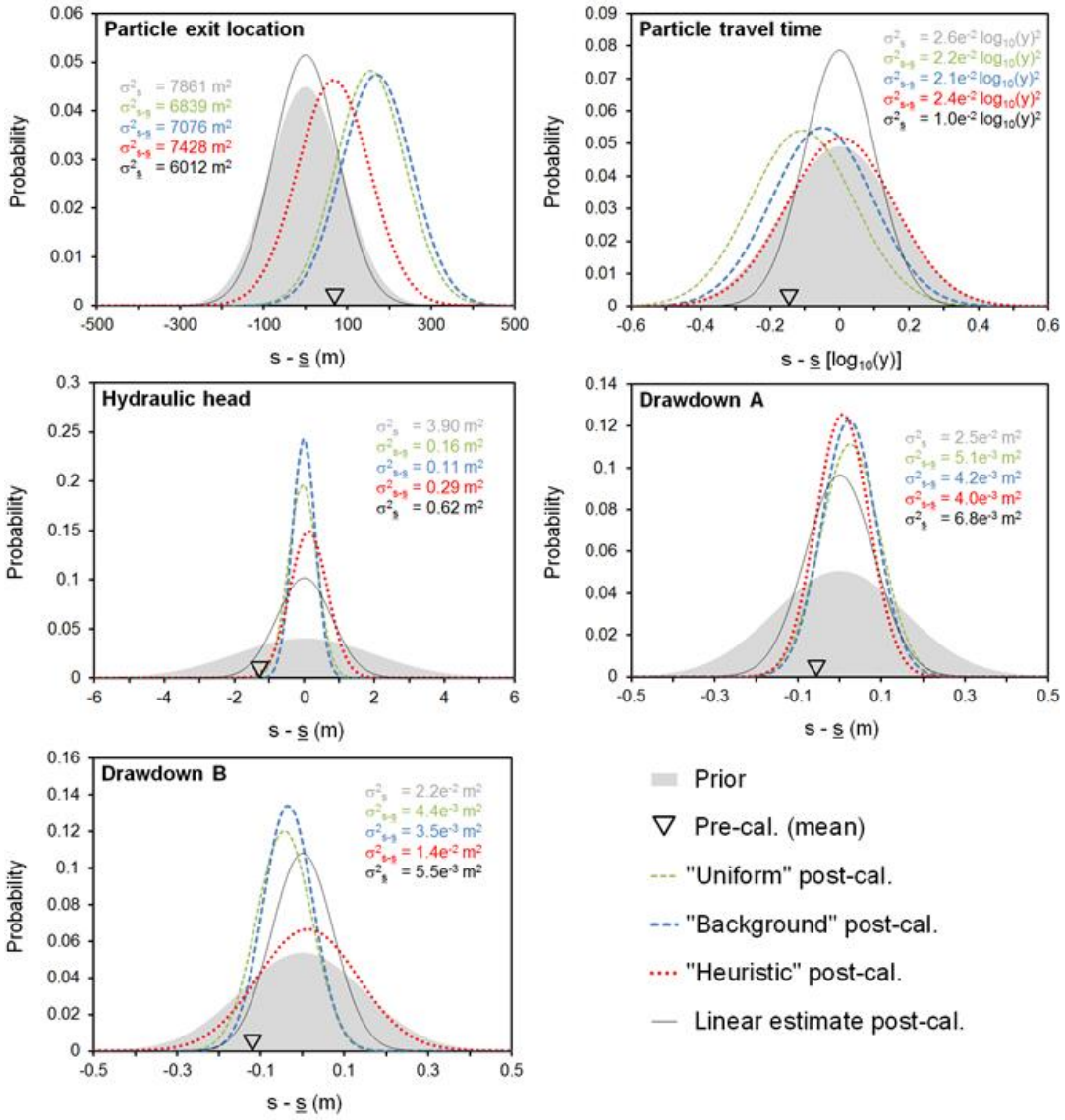

Figure 4.10. Probability density functions representing post-calibration potential predictive error based on each regularization weighting strategy. For the sake of clarity, "empirical" distributions are not displayed due to high degree of similarity with "background" distributions. Also displayed is prior uncertainty variance σ^2_s (shaded), pre-calibration mean predictive error and the linear estimate of post-calibration uncertainty variance $\sigma^2_{\underline{s}}$.

From Figure 4.10 it is clear that, in most cases, post-calibration mean predictive error (represented by the peaks of the respective probability density functions) is of significantly smaller magnitude than pre-calibration mean predictive error. This indicates that pre-calibration predictive bias has largely been "calibrated out" in most cases. For the particle exit location prediction, however, this bias is in fact magnified by the calibration process for most weighting strategies. That is, based on the "uniform", "background" and "empirical" regularization weighting strategies, particle exit location predicted by the calibrated model is on average more in error than particle exit location predicted by the uncalibrated, homogeneous model parameterized by preferred values alone.

The overall success of calibration is immediately evident from Figure 4.10. For some predictions, such as hydraulic head and drawdown "A", post-calibration potential predictive error is substantially smaller than prior uncertainty. Furthermore this potential error is wholly captured by post-calibration uncertainty analysis, irrespective of the employed regularization weighting strategy. For both advective transport predictions, some portion (in some cases a very large portion) of each predictive error probability density function clearly falls outside the range of potential predictive error quantified through post-calibration uncertainty analysis. This, as discussed above, represents failure of the modelling process. Moreover, the post-calibration predictive error probability density functions for the advective transport predictions in most cases exceed the span of prior uncertainty. Thus not only do these predictions exhibit a potential for error beyond range of the post-calibration uncertainty estimate, the potential for wrongness in the predictions made by the calibrated model is greater than the prior uncertainty range defined by expert geological knowledge alone.

The prediction-specific influence of the regularization weighting strategy upon post-calibration potential predictive error is also clear from Figure 4.10. For example, the "heuristic" strategy reduces both mean predictive error and predictive error variance in drawdown "A" (evinced by the more centralised and slightly narrower probability density function). At the same time, the hydraulic head predictive error probability density function obtained through the "heuristic" strategy exhibits both a greater mean error and greater error variance.

For the particle exit location, particle travel time and drawdown "B" predictions, the "heuristic" regularization strategy observably reduces mean post-calibration predictive

error, whilst simultaneously increasing predictive error variance. For the latter two predictions, mean predictive error is effectively eliminated completely. However, the simultaneous increase in predictive error variance renders the post-calibration predictive error probability density functions very similar in width to the prior uncertainty probability density functions (which represents no overall benefit of the history matching process in these cases).

As discussed above, MSE provides a means of quantifying the overall potential for predictive error, accounting for both predictive error variance and mean predictive error (see equation (4.13c)). Figure 4.11 displays MSE for each prediction calculated using equation (4.13b), normalized with respect to the prior uncertainty variance of each prediction calculated through equation (4.11) such as to provide a measure of the overall success of the calibration process in reducing potential predictive error relative to prior uncertainty based on expert knowledge alone.

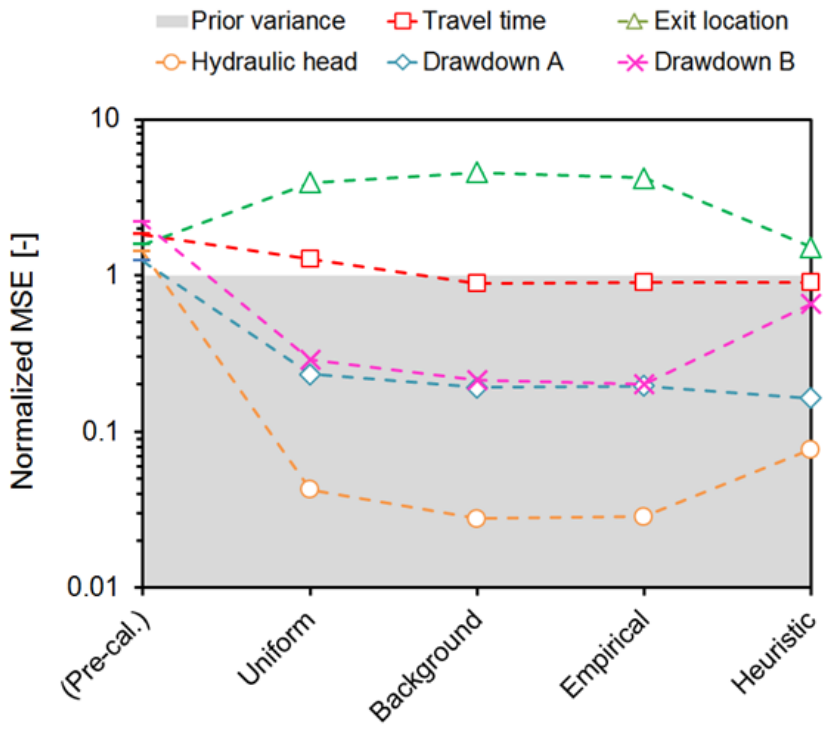

Figure 4.11. Post-calibration prediction mean square error (MSE) for each regularization strategy, normalised with respect to prior prediction variance. The shaded region represents the boundary (i.e., at normalized MSE = 1) between reduction and inflation relative to prior prediction variance. (Note: dashed lines joining the results of each regularization weighting strategy are displayed only to aid relative comparison; there exists no explicit relationship between the separate inversion processes.)

The values at the y-axis in Figure 4.11 are all greater than one, representing the pre-calibration bias in all predictions introduced through the employed regularized inversion approach as discussed above. The magnitude of this pre-calibration bias is substantial in some cases; for example the pre-calibration MSE in drawdown "B" is over 100% greater than prior uncertainty variance.

Despite the pervasive pre-calibration bias, Figure 4.11 indicates that post-calibration prediction MSE is reduced through calibration to a value below prior uncertainty variance for all predictions other than the particle exit location. Reflecting the probability density function characteristics discussed above, post-calibration MSE is generally reduced substantially below prior uncertainty for the hydraulic head and drawdown predictions irrespective of the regularization weighting strategy. An MSE reduction of 70-97% relative to prior uncertainty is achieved in all but one case, which is discussed below.

For the advective transport predictions, Figure 4.11 shows that post-calibration MSE is either comparable to, or greater than, prior uncertainty variance irrespective of the employed regularization weighting strategy. For prediction of travel time, the calibration process has the capacity to eliminate the initial bias introduced by the uncalibrated model. However, MSE is barely reduced relative to prior uncertainty based solely on expert knowledge. In the case of particle exit location, normalized MSE reaches a value of 4.5 at worst. This translates to prediction MSE that is 300% greater than the uncalibrated, homogeneous model, and 350% greater than prior uncertainty variance based on expert knowledge alone. The "heuristic" regularization weighting strategy achieves a slight exit location prediction MSE reduction relative to the MSE of the uncalibrated model, however post-calibration MSE remains greater than prior uncertainty variance. Thus, for these predictions, very little is gained from calibration at best, and at worst the process is highly detrimental to model predictive performance.

Finally, Figure 4.11 highlights the potential complexities in the outcomes of calibration of a simplified model. Firstly, prediction specificity in the success of the calibration process is not limited to predictions of a distinctly different type. For example, employment of the "background" regularization weighting strategy in place of the "uniform" regularization weighting strategy decreases the MSE in the prediction of particle travel time, but simultaneously raises particle exit location prediction MSE.

Similarly, employment of the "heuristic" regularization weighting strategy in place of any other strategy yields a decrease in MSE for the prediction of drawdown "A", whilst at the same time causing a marked inflation in the drawdown "B" prediction MSE. In fact, Figure 4.11 demonstrates that, in the present synthetic example, any reduction in the MSE of a given prediction achieved through a change of the in regularization weighting strategy, is accompanied by an increase in the MSE of at least one other prediction (and vice-versa).

4.5 Discussion

Moore and Doherty (2005) allude to the potentially prediction-specific success of calibration in terms of optimizing model predictive performance. They suggest that multiple specifically tailored calibration processes may be required for the making of multiple predictions. It follows that there exists a need to abandon the notion that a model, once "calibrated", is thereafter fit for making any prediction required of it. A particular focus in recent literature, including Chapter 3 of the present thesis, is the extent to which various model simplifications/defects amplify the degree of prediction-specificity in the success of calibration (e.g., Doherty and Welter, 2010; Doherty and Christensen, 2011; White et al., 2014).

Means of tailoring a calibration process that have received focus in recent literature include observation pre-processing (e.g., Moore and Doherty, 2005; Doherty and Welter, 2010; White et al., 2014), pre-calibration parameter transformation (e.g., Chapter 3 of the present thesis; White et al., 2014), and alteration of the goodness of fit sought between model outputs and the calibration dataset (e.g., White et al., 2014). The present study explores the weighting of Tikhonov regularization constraints as an additional means. The results emphasize the potential importance of undertaking multiple calibration processes for the purpose of making multiple predictions. Through alteration of the regularization weighting strategy in the present synthetic example, no improvement in the ability of the model to make a given prediction is achieved without degrading its ability to make at least one other prediction.

The results of the present study provide transparency to the causes of prediction-specific calibration success through identification of the various components of potential predictive error, in particular the different sources of predictive bias. As explained in section 4.2.4, Doherty and Christensen (2011) describe two forms of

predictive bias in the context of paired model analysis, which are herein distinguished between as "hardwired" bias and "surrogacy-induced" bias. The former is attributable to consistent errors in null-space parameter components that are omitted from the simplified model, and is expressed as a systematic offset in *s*-versus-$\underline{s}$ scatterplots (and a corresponding non-zero mean in the predictive error probability density function). The latter is a consequence of the surrogate roles played by parameters of a simplified/defective model as they compensate for structural inadequacies in order to fit the calibration data. This is inevitably accompanied by inclusion in the parameter estimation process of parameter components that properly belong to the null space. This null-space entrainment induces random error in estimated parameters and thus predictions that are sensitive to those parameters. This manifests as a reduction in the slope of the *s*-versus-$\underline{s}$ scatterplot regression line, which is associated with an increase in predictive error variance (expressed as a broadening of the width of predictive error probability density functions).

Despite the aforementioned high degree of prediction specificity in the outcomes of regularized inversion applied to the present synthetic example, some more general outcomes may be drawn from the results. Firstly, ignoring the presence of the faults does not prevent the calibration process from achieving a substantial reduction in potential predictive error in the hydraulic head-related predictions under both non-pumping and pumping conditions. Moreover, efficient linear uncertainty analysis adequately "captures" the full post-calibration potential for error in each of these predictions.

For all predictions except particle exit location, inclusion of a covariance-based regularization weighting scheme based on the geostatistical background $\log_{10}K$ field of "reality" provides a reduction in post-calibration predictive MSE relative to a lack of a covariance-based regularization weighting scheme (i.e., uniform weights). This is perhaps unsurprising, given that the former represents inclusion of additional, albeit partial, expert knowledge in the calibration process. The ability of the model to make most predictions benefits from an enhanced predictive range provided by the introduction of more realistic broad-scale correlation structures in the estimated parameter fields (despite the fact that the overall accuracy of estimated $\log_{10}K$ values at a cell-by-cell level is barely improved).

The attempted incorporation of "more complete" expert knowledge, though a regularization weighting strategy based on the covariance matrix calculated empirically using the suite of stochastic "reality" fields, has a generally insignificant effect upon the estimated $\log_{10}K$ fields (and thus all predictions) compared with the "background" regularization weighting strategy. This is consistent with the expectation that small-scale $\log_{10}K$ features such as the faults in the present synthetic example lie predominantly within the "true" null space where the calibration dataset comprises relatively sparse hydraulic head observations (e.g., Moore and Doherty, 2005; 2006). This expectation is supported by eigenanalysis in the present study, which suggests that adjustment of a small number (two or three) parameter combinations dominates the history matching process in the present synthetic example. Moreover, SVD of the empirically derived covariance matrix indicates that "fault-like" correlation structures become more prominent only for eigencomponents corresponding to smaller singular values. Thus, introduction of such structures is "not required" during the parameter estimation process in order achieve a satisfactory fit between model outputs and the calibration dataset.

"Heuristic" modification of the empirically derived covariance matrix is demonstrated to effectively force the calibration process to introduce "fault-like" correlation structures through the calibration process. Despite the associated increase in cell-by-cell $\log_{10}K$ error incurred through this approach, minimisation of the potential for error in some predictions depends on the expression of these correlation structures. This occurs mainly through observable reduction in "hardwired" bias in most predictions, for it constitutes surrogate representation of null-space parameter components whose omission from the calibrated simplified models was the original cause of the consistent predictive error. At the same time, the forced inclusion of null-space parameter components in the parameter estimation process is equivalent to null-space entrainment. As a result, the "heuristic" regularization weighting strategy induces an inflation of predictive error variance in most predictions owing to "surrogacy-induced" bias. Thus, the reduction in "hardwired" bias is generally accompanied by an increase in "surrogacy-induced" bias, such that the reduction in the propensity for model predictive error afforded by the former is eroded by the latter.

It is worth noting briefly in further support of the present discussion that an additional paired model analysis process was undertaken (the results of which are not presented for the sake of brevity) wherein the "heuristic" regularization weighting strategy was

accompanied by anisotropic interpolation (with a northwest-southeast principal axis of anisotropy). As per all other paired model analysis processes, 1000 calibrated model pairs ($\bar{\Phi}_m = 11.9$) were evaluated. Post-calibration prediction MSE results are generally similar to those attained through the "heuristic" weighting strategy, but the individual components of MSE exhibit further trade-off between the two sources of predictive bias. For example, "hardwired" bias in the particle exit location prediction is further reduced, whilst "surrogacy-induced" predictive bias is simultaneously inflated, effecting minor net reduction in post-calibration prediction MSE, which remains greater than prior uncertainty variance. (For the purposes of comparison with the results presented herein: mean predictive error is reduced to 37 m, predictive error variance is increased to 8644 m^2 (accompanied by an *s*-versus-$\underline{s}$ regression line slope *b* of 0.41).)

The apparent trade-off between "hardwired" and "surrogacy-induced" predictive bias discussed above prevents reduction of advective transport potential predictive error below prior uncertainty in the present synthetic example. Thus, in this case there is little to be gained and much to be lost through pilot-point-based regularized inversion for the making of these predictions in the presence of categorical heterogeneity. In this particular instance calibration appears to be inevitably unfruitful at best and at worst highly detrimental, and should therefore be abandoned altogether. Nonetheless, predictive uncertainty must still be quantified in some manner, for it is the critical outcome of any modelling process in the decision support context.

Approaches rooted in a classical Bayesian framework allow expert geological knowledge to be expressed in its purest form and eschew the potential for calibration-induced predictive bias (see, for example, Harmon and Challenor, 1997; Kuczera and Parent, 1998; Campbell and Bates, 2001; Qian et al., 2003; Vrugt et al., 2009a; Sadegh and Vrugt, 2014, and references cited therein). The popularity of gradient-based/optimization approaches such as regularized inversion using pilot points is rooted in their computational efficiency relative to a Bayesian framework, with the latter often demanding a prohibitively large number of model runs (Mugunthan and Shoemaker, 2006; Mariethoz et al., 2010a). However, construction of a prediction uncertainty envelope through unconstrained Monte Carlo analysis can be achieved via a far more manageable number of model runs. For example, 1000 expert knowledge-based realizations are used for this purpose in the present study, which is substantially fewer model runs than required for a single calibration process based on the employed

small-scale synthetic example (the number of which rapidly increases for larger models with more parameters). Let us suppose that the present synthetic example represents a case in which a Bayesian approach is computationally infeasible. Expression of advective transport prediction uncertainty is thus most frugally achieved via unconstrained Monte Carlo analysis based solely on stochastic expert knowledge. This provides (in the case of the present synthetic study) an uncertainty range that is at worst comparable, and at best substantially smaller, than potential predictive error obtained through pilot-point-based regularized inversion. Most importantly, it safeguards against underestimation of uncertainty. This is in accordance with a key modelling strategy metric proposed by Doherty and Simmons (2013) in a recent discussion paper on modelling in the decision-making context: the modelling process must be guaranteed to exaggerate the uncertainty associated with a prediction of an unwanted event.

The capacity for stochastic expression of expert geological knowledge for Monte Carlo simulation purposes is ever increasing. Recent literature details a plethora of increasingly sophisticated multiple point geostatistics-based techniques and software that facilitate efficient generation of suites of realistic stochastic fields that conform to complex conceptualizations of subsurface heterogeneity (see, for example, Mariethoz et al. 2010b; Mariethoz and Kelly, 2011; Meerschman et al. 2013; Mahmud et al., 2014; Mariethoz and Lefebvre, 2014; Zahner et al., 2016 and references cited therein).

Finally, as discussed above, Doherty (2015) points out that recent literature also presents ongoing development of state-of-the-art methods for continuous, differentiable representations of complex categorical parameter fields based on multiple point geostatistics (e.g., Sarma et al., 2008; Ma and Zabaras, 2011; Vo and Durlofsky, 2014). These methods provide compatibility with gradient-based optimization algorithms and thus may facilitate more sophisticated Tikhonov regularization schemes that facilitate expression of more complex and realistic geological expert knowledge. Nonetheless, the general insights provided by the present study are not expected to be invalidated by the application of more sophisticated regularization capabilities. The present “heuristic” strategy provides a rudimentary example of the potential outcomes of formulating the inversion process in a manner that emphasizes a certain desired component of parameter variability in an estimated field. More realistic post-calibration parameter fields would be expected to reduce parameter surrogacy (caused though the present “heuristic” strategy by the

aforementioned thickness and "blurriness" of the fault-like features, as well as the unrealistic elongation of low-*K* regions) and thus reduce null-space entrainment to some degree. However a more sophisticated regularization strategy alone cannot reduce the inherent nonuniqueness of the inverse problem, which precludes the faults from being accurately resolved by the available calibration data. Thus, in the same manner as observed in the present study, inclusion of such features in the estimated parameter field may therefore reduce "hardwired" predictive bias but would be expected to increase error variance through "surrogacy-induced" bias caused by null-space entrainment, as discussed above. The influence of more sophisticated regularization approaches is nonetheless of interest and the application of a multiple point geostatistics-based method in the same context is recommended as future work.

4.6 Conclusions

In light of the increasing popularity of pilot-point-based regularized inversion as a means of calibrating highly parameterized groundwater models, the present study explores the outcomes of this approach in a synthetic context that represents an environment where hydraulic property variability cannot be wholly characterised in idealistic multi-Gaussian terms due to the existence of discrete features, which are common in subsurface formations (Wen and Gomez-Hernandez, 1998; Sarma et al., 2008; Zhou et al., 2014).

The presence of preferential flow features (faults) in the present synthetic example is shown to induce substantial pre-calibration predictive bias in all predictions. Nonetheless, failure to account for the presence of the faults does not prevent the calibration process from providing the model with the ability to make hydraulic head and drawdown predictions with little bias and highly reduced potential error (and which is quantifiable through standard linear uncertainty analysis). The pre-calibration bias is "calibrated out", with the history matching process resulting in substantial overall reduction in potential predictive error relative to prior uncertainty (i.e., MSE reductions of 70-97%). For this purpose, a covariance-based regularization weighting scheme limited to broad-scale correlation structures only (i.e., representing "partial" expert knowledge) is most reliable. This yields better predictive performance than uniform weighting of regularization constraints. Whilst further improvement in performance is shown to be possible through heuristic modification of regularization

weights, the risks appear to outweigh the potential benefits with drastic degradation occurring in the ability of the model to make other predictions.

The present study highlights the potential extremes of prediction-specificity in the outcomes of simplified model calibration. For example, calibration employing a regularization weighting scheme based on broad-scale hydraulic conductivity field geostatistics achieves a near-100% reduction (relative to the uncalibrated model) in MSE for the making of an ungauged hydraulic head prediction, whilst simultaneously inducing a 300% inflation in particle exit location prediction MSE. The potential degree of prediction specificity in the outcomes of simplified model calibration is further highlighted by the fact that no alteration of the regularization weighting strategy improves the ability of the model to make one prediction without degrading its ability to make another.

Elucidation of the individual components of total post-calibration potential predictive error across various regularization weighting strategies provides some more general insights into the outcomes of calibration. Through heuristic formulation of regularization weights, the "forced" representation of "fault-like" features in estimated parameter fields is accompanied by a reduction in "hardwired" predictive bias (i.e., consistent predictive error) for most predictions. However, it simultaneously comes at the expense of increased error variance in most predictions due to increased calibration-induced parameter surrogacy and inclusion of null-space parameter components within the parameter estimation process (i.e., "surrogacy-induced" predictive bias). The total potential for error in some predictions is reduced due to the reduction in "hardwired" bias, whilst that in others is drastically inflated due to the increase in predictive error variance caused by "surrogacy-induced" predictive bias. For some predictions, the apparent tendency for a trade-off between these two sources of predictive bias prevents any calibration approach from reducing the potential predictive error below that defined by prior uncertainty. For these predictions the calibration process is thus futile. Extending the notion that a model should be calibrated multiple times for the making of multiple predictions, this suggests that calibration should be abandoned altogether for the making of some predictions. For practical situations wherein constrained uncertainty analysis within a Bayesian framework is computationally prohibitive (which is the reason for employing a more efficient approach such as regularized inversion in the first place), the most pragmatic means of fulfilling the critical requirement of characterizing model prediction

uncertainty may thus be through unconstrained Monte Carlo analysis alone. This can be based upon the purest possible expression of expert knowledge through geologically realistic fields generated using state-of-the-art methods. This will necessarily come at the cost of prediction uncertainty overestimation but, crucially, calibration-induced predictive bias and the associated potential for type II statistical error (and thus ultimate failure of the modelling process) are forestalled.

Chapter 5

Conclusions

This thesis addresses the pervasive and important topic of the outcomes of calibrating simplified/imperfect groundwater models. The insights presented herein are founded upon a wealth of experimental modelling data, with the work in total based on over thirty-thousand model calibration processes (comprising over sixty-million individual forward model runs). The accompanying development and application of linear mathematical analysis based on model sensitivities facilitates deeper insight into the interactions between calibration data, model parameters and model predictions through the calibration process.

The key contributions of the thesis include 1) a proof-of-concept study for a recently developed and previously untested bias identification and uncertainty quantification methodology, 2) extension of a recently developed mathematical framework describing the outcomes of calibrating a simplified model, and 3) application of these nonlinear and linear approaches to multiple representative synthetic examples of model simplification and calibration, contributing to the ongoing development of knowledge and best-practice guidance for curtailing calibration-induced predictive bias in everyday modelling. More specific conclusions pertaining to the three studies comprising the present thesis are now summarized.

The first study, presented as Chapter 2, comprises a proof of concept for the "paired model analysis" (PMA) methodology presented by Doherty and Christensen (2011). PMA is designed to identify and allow correction of predictive bias induced through calibration of a simplified model in place of relatively complex model, simultaneous with quantification of post-calibration predictive uncertainty. It has not previously been verified for empirical consistency with theoretical expectation. For this purpose, paired model analysis is applied to a simple synthetic example that is extensively studied in existing literature. Consistency of post-calibration uncertainty quantified

through PMA is demonstrated to be in good agreement with the results of established linear and nonlinear methods. Known sources of predictive bias, namely "overfitting" with respect to measurement noise, and "suboptimal" regularization, are shown to be reliably identified through paired model analysis. The results concomitantly emphasize the potential ramifications of a poorly forged calibration process. Even in the idealised case where a model is structurally flawless, suboptimal regularization (for example, through failure to undertake appropriate pre-calibration parameter transformation) can instil a greater potential for error in a "well-calibrated" model than if the model had not been calibrated at all. Finally, the capacity for reduction of the potential error in model predictions via bias correction achieved through paired model analysis is demonstrated.

The second study, presented as Chapter 3, builds upon the mathematical formulation of simplified model calibration presented by Doherty and Christensen (2011). In particular, a linear subspace-based description of "null-space entrainment" is presented. This concept is introduced by Doherty and Christensen (2011), and refers to the unwitting inclusion of inestimable parameter components within the parameter estimation process as a consequence of the parameter surrogacy that may occur during calibration of a simplified/imperfect model. Sensitivity of model predictions to entrained null-space parameter components is the cause of calibration-induced predictive bias. The developed linear framework is employed together with PMA to thoroughly examine the parameter and predictive outcomes of calibrating two simplified versions of a one-dimensional, Richards equation-based unsaturated zone model used to predict recharge to a groundwater system. The simplification processes are considered typical of modelling practice, these being 1) assumed vertical parameter uniformity, and 2) replacement with a lumped parameter "bucket" model. Substantial calibration-induced parameter surrogacy and consequential null-space entrainment is demonstrated to occur for both levels of simplification. Nonetheless, both simplified models are shown to make largely unbiased predictions of future recharge. This demonstrates that despite potentially poor parameter estimates that compensate for model imperfections, if predictions are of a similar nature to the available field observations, then a model's physical basis becomes less important in the making of future predictions than its ability to achieve a good fit with available calibration data.

The third study, presented as Chapter 4, explores the outcomes of employing the increasingly popular method of pilot-point-based regularized inversion for model

calibration in an environment containing discrete preferential flow features (referred to herein as "faults"). PMA is applied to quantify the success of calibration in terms of the post-calibration potential error in multiple predictions, subject to various regularization weighting strategies. A number of metrics are considered in order to elucidate the various contributions to potential predictive error, particularly different sources of predictive bias. Extending the linear analysis-based insights of Chapter 3, the prediction-specific sensitivity to calibration-induced compensatory parameter behaviour and concomitant null-space entrainment is highlighted. It is shown that for some predictions, ignoring the existence of faults and estimating a surrogate smooth hydraulic conductivity field does not compromise the ability of the inversion process to "calibrate out" pre-calibration bias and provide predictions with greatly reduced, quantifiable potential for error. At worst, predictions made by a "well-calibrated" model may possess a far greater post-calibration potential for error than its prior uncertainty based on expert geological knowledge alone (the present study demonstrates a case in which post-calibration mean square error is 350% greater than prior uncertainty variance). The potential degree of prediction specificity in the outcomes of calibrating a simplified model is highlighted through demonstration that no employed regularization weighting strategy reduces the potential for error in one prediction without simultaneously raising the potential for error in another. Varying degrees of null-space entrainment controlled by the employed regularization weighting strategy is demonstrated. It is shown that certain null-space entrainment may in fact improve model performance, through reduction of "hardwired bias", for the making of predictions that possess a systematic dependence on those null-space parameter components. At the same time, however, the resultant "surrogacy-induced" predictive bias that accompanies null-space entrainment inflates predictive error variance in other predictions. The apparent trade-off between these different sources of predictive bias explicates the inevitable fruitlessness of calibration for predictions that are highly sensitive to null-space parameter components, which by definition cannot be resolved based on information contained within the available calibration data. This work thus emphasizes the need for prediction-specific tuning of a modelling process, to the extent that the most pragmatic approach for some predictions may be to forego calibration entirely and quantify uncertainty based solely on the purest possible expression of expert knowledge using state-of-the-art geological simulation methods.

Appendix A: Derivation of $\Phi_m = N$ for model-to-measurement misfit commensurate with measurement noise

The observation weight matrix $\mathbf{Q}_h$ of equation (2.2) is given (for statistically independent observations contained in **h**) as:

$$\mathbf{Q}_h = \begin{bmatrix} q_1^2 & 0 & \cdots & 0 \\ 0 & q_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & q_N^2 \end{bmatrix} \tag{A1}$$

Where q_i is the weight assigned to the i^{th} observation and N is the number of observations. Where observation weights are specified as the inverse of measurement noise standard deviation, we have:

$$\mathbf{Q}_h = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N^2} \end{bmatrix} \tag{A2}$$

From equation (2.2) defining the measurement objective function Φ_m, model-to-measurement misfit is represented by the term (**Xk – h**). Where model-to-measurement misfit is commensurate with measurement noise, we therefore have:

$$(\mathbf{Xk} - \mathbf{h}) = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_N \end{bmatrix} \tag{A3}$$

Substituting the right-hand sides of equations (A2) and (A3) into equation (2.2):

$$\Phi_m = \begin{bmatrix} \sigma_1 & \sigma_2 & \cdots & \sigma_N \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N^2} \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_N \end{bmatrix} \tag{A4}$$

$$= \sum_{i=1}^{N} \frac{\sigma_i \sigma_i}{\sigma_i^2} \tag{A5}$$

$$= N \tag{A6}$$

Appendix B: Inappropriate parameter transformation and null-space entrainment

This appendix provides a simple demonstration of failure to take account of prior parameter uncertainty when simplification is effected through combining complex model parameters into a smaller number of parameters in order to achieve well-posedness of an inverse problem. It also demonstrates that, like other forms of suboptimal simplification, this can lead to null-space parameter entrainment as the simplified model is calibrated. A linear model is employed in this example; hence the analysis that follows is exact.

Figure B1 depicts a section through an aquifer. Let r_1 and r_2 define the resistances of the two hydrogeological units represented in the figure. We define resistance to groundwater flow through the equation:

$$r = \Delta h/q \tag{B1}$$

where q is the flow through the permeable unit and Δh is the drop in potential across it incurred by this flow. Let us suppose that the only data available for estimation of these resistances is a single upgradient head measurement. The inverse problem of estimating r_1 and r_2 is obviously ill-posed. Let the action of the model on these two parameters be represented by the matrix **Z**; let the vector **r** represent both of these together. Then (ignoring measurement noise for the sake of simplicity of the analysis):

$$\mathbf{h} = \mathbf{Z}\mathbf{r} \tag{B2}$$

where:

$$\mathbf{Z} = q[1 \;\; 1] \tag{B3}$$

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{B4}$$

and:

$$\mathbf{h} = [h] \tag{B5}$$

q in equation (B3) is inflow into the right of the model domain while h in equation (B5) is the head measured in the well at the right of the domain. For convenience we assume that the head at the left boundary of the model domain is fixed at zero.

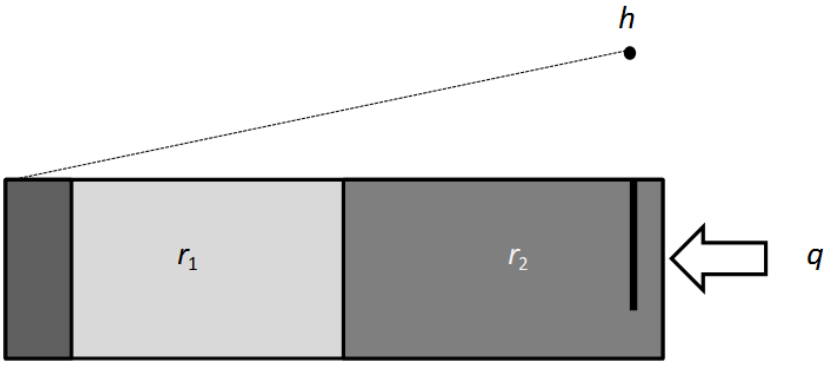

Figure B1. Conceptual model for estimation of two resistances using a single head measurement. The head is fixed at zero at the left of the model domain; inflow into the right is known.

Suppose that geological considerations suggest that r_2 has a greater propensity for variability than r_1. Its prior uncertainty is therefore greater than that of r_1. For illustrative purposes let us assume that it is, in fact, twice as uncertain. If C(**r**) is the covariance matrix of **r**, then:

$$\mathrm{C}(\mathbf{r}) = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \tag{B6}$$

where α is a constant of proportionality. Suppose that, in an attempt to solve this inverse problem, we use SVD without normalizing parameters with respect to their innate variability. This is equivalent to estimating the average value of r_1 and r_2 and assigning it to the whole model domain (a common calibration strategy). The **V** matrix achieved through SVD of **Z** is:

$$\mathbf{V} = \frac{q}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{B7}$$

Because the calibration dataset is comprised of only one observation, the solution subspace of parameter space contains only one dimension. From equation (A7) it is defined by the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$; meanwhile the null space is defined by the vector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

As discussed in the body of this paper, the solution to the inverse problem of model calibration is obtained as the projection of the real (and unknown) parameter vector onto the solution space. We thus seek a value for the sum of r_1 and r_2 while insisting that the difference between r_1 and r_2 be zero. That is, we seek a value for the average r, with "average" defined as $(r_1 + r_2)/2$. Parameter space is depicted in Figure B2.

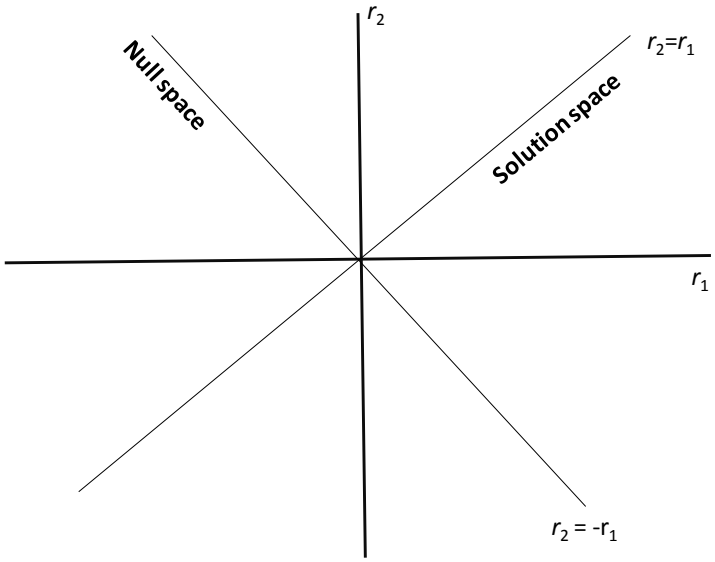

Figure B2. Two-dimensional parameter space showing the one-dimensional solution and null spaces arising from the inverse problem depicted in Figure B1.

Figure B3 depicts a "reality" vector **r**, as well as its projection onto the solution space. A prior probability contour of **r** is also shown; let it be assumed that this is a contour of low probability so that the shaded area defines the area of parameter space in which **r** is most likely to lie. It is apparent that the calibration process endows r_1 with a greater post-calibration potential for error than it had prior to calibration. The same will apply to any model prediction that is heavily dependent on r_1.

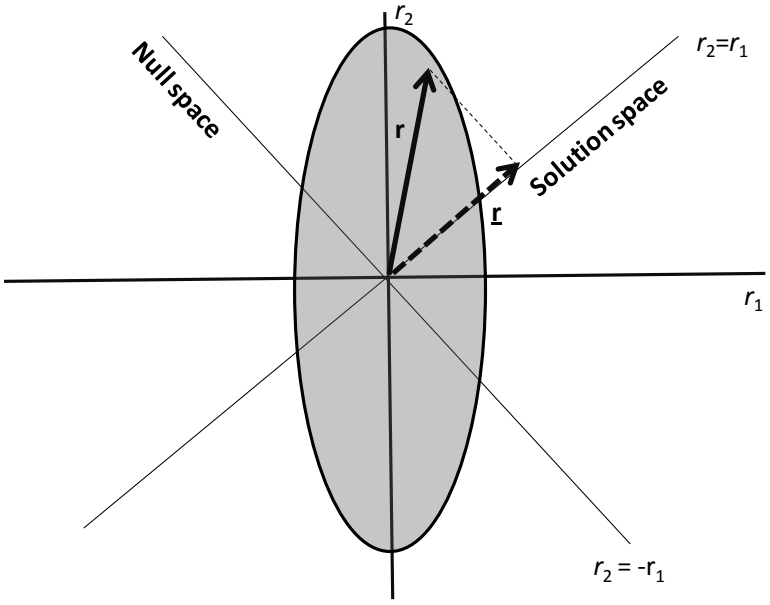

Figure B3. Solution of the inverse problem is obtained as the projection of the true parameter vector onto the solution space. The region of high prior probability of **r** is shown shaded.

The fact that calibration increases, rather than decreases, the potential for error of some model predictions arises because parameters were not normalized with respect to their propensity for variability before being estimated. As a consequence of this, null-space parameter components are entrained as the model is calibrated, as will now be demonstrated. A solution to the inverse problem of model calibration, which conforms to expert knowledge as reflected in the C(**r**) prior parameter covariance matrix, should ensure that r_1 is varied from its pre-calibration expected value less than r_2 is varied from its pre-calibration expected value. In fact, to the extent that such variation is required in order to fit the single head observation *h*, r_2 should be encouraged to vary twice as much as r_1. This can be achieved through estimation of two transformed parameters t_1 and t_2, with the transformation **T** defined as:

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \mathbf{Tr} \tag{B8}$$

The covariance matrix of t is then given by:

$$\mathrm{C}(\mathbf{t}) = \mathbf{T}\mathrm{C}(\mathbf{r})\mathbf{T}^{\mathrm{t}} = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{B9}$$

The model equation then becomes:

$$\mathbf{h} = \mathbf{Zr} = \mathbf{ZT}^{-1}\mathbf{t} = \mathbf{Yt} \tag{B10}$$

where **Y**, the model used for parameter estimation purposes, is defined as:

$$\mathbf{Y} = q[1 \;\; 2] \tag{B11}$$

The null space of **Y** is defined by the unit vector $\frac{q}{\sqrt{5}}\begin{bmatrix}-2\\1\end{bmatrix}$ while its solution space is defined by the vector $\frac{q}{\sqrt{5}}\begin{bmatrix}1\\2\end{bmatrix}$. Obviously, these spaces are orthogonal to each other in **t**-space. However, back-transformation of these to **r**-space using the $\mathbf{T}^{-1}$ transformation leads to the non-orthogonal spaces $\frac{2q}{\sqrt{5}}\begin{bmatrix}-1\\1\end{bmatrix}$ and $\frac{q}{\sqrt{5}}\begin{bmatrix}1\\4\end{bmatrix}$, respectively. Unsurprisingly, the former is aligned with the previous null space. The latter is depicted in Figure B4, together with the projection onto this space implied by calibration of the **Y**-based model. The fact that the solution to the inverse problem is more in harmony with expert knowledge is obvious.

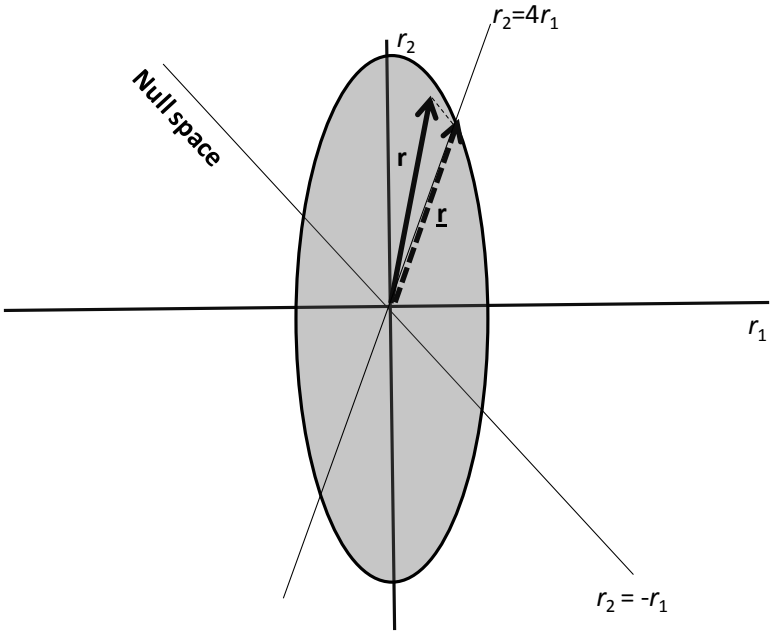

Figure B4. Solution of the inverse problem in **t**-space after back-transformation to **r**-space.

In contrast, the one-dimensional solution space found through SVD undertaken in **r**-space **T**-transforms to the vector **t** shown in Figure B5, which depicts **t**-space. Null-space entrainment of the **r**-space solution in this space is obvious.

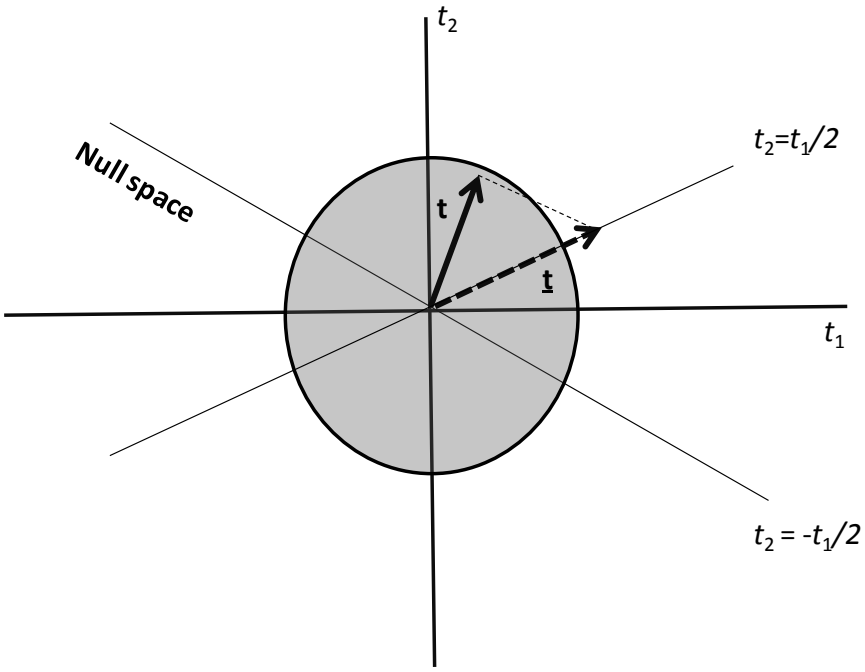

Figure B5. Solution of the inverse problem in **r**-space after transformation to **t**-space. The **r**-space solution to the inverse problem has a non-zero projection onto the **t**-space null space.

The importance of taking expert knowledge into account during simplification is thus obvious. This applies irrespective of the simplification methodology adopted. For example, suppose that a modeller decides to fix one of the resistances in Figure B1 and estimate the other, rather than implicitly or explicitly estimating an average resistance. Obviously he/she should fix r_1 and estimate r_2 as r_1 has less innate variability than r_2; the potential for wrongness in fixing the chosen parameter at its expected value is therefore smaller.

Appendix C: Developed stochastic software

The software developed for the study presented in Chapter 4 facilitates generation of stochastic arrangements of discrete linear features. It allows optional randomization of feature number, length and orientation (within specified dominant angle range(s)), in addition to optional constraints on proximity to domain boundaries and separation of feature centres. Figure C1 provides some arbitrary graphical examples of the software outputs (here used in conjunction with FIELDGEN (Doherty, 2016b) to provide multi-Gaussian parameterization of the "background" fields).

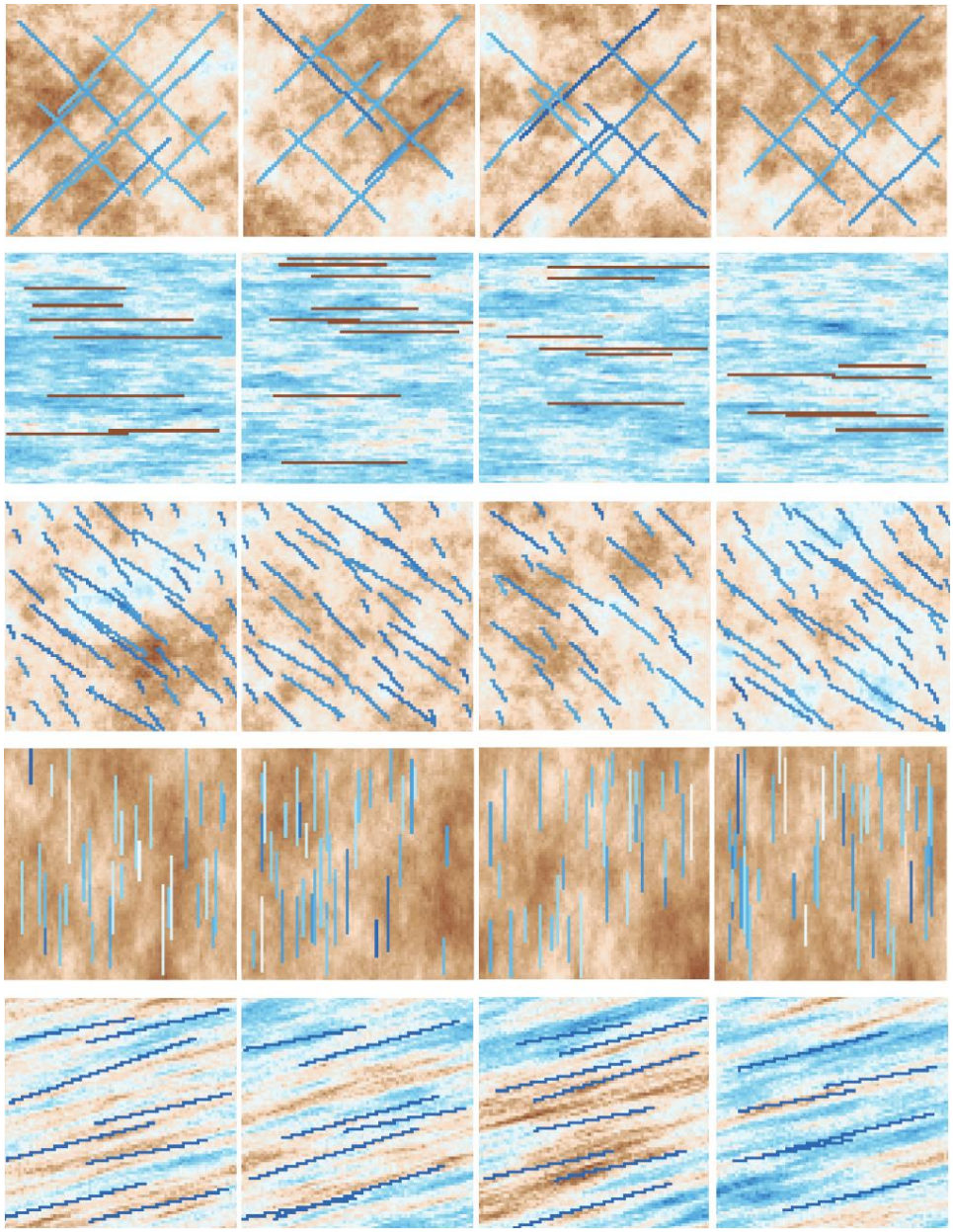

Figure C6. Arbitrary examples of outputs from the developed stochastic software.

References

Aanonsen, S. I. (2008), Efficient history matching using a multiscale technique, SPE Reservoir Eval. Eng., 11(1), 154–164.

Alcolea, A., J. Carrera and A. Medina (2006), Pilot points method incorporating prior information for solving the groundwater flow inverse problem, Adv. Water Resour., 29, 1678–1689.

Alcolea, A., J. Carrera and A. Medina (2008), Regularized pilot points method for reproducing the effect of small scale variability: Application to simulations of contaminant transport, J. Hyrdol., 355, 76–90.

Anderman, E. R., and M. C. Hill (2001), MODFLOW-2000, the U.S. Geological Survey modular ground-water model—Documentation of the advective-transport observations (ADV2) package, U.S. Geol. Surv. Open File Rep., 01–54.

Anderson, M. P., W. W. Woessner, and R. J. Hunt (2015), Applied groundwater modeling—Simulation of flow and advective transport (2nd ed.), 610 pp., Elsevier, Amsterdam, Netherlands.

Andreu, J. M., F. J. Alcalá, A. Vallejos, and A. Pulido-Bosch (2011), Recharge to mountainous carbonated aquifers in SE Spain: Different approaches and new challenges, J. Arid Environ., 75, 1262–1270.

Aster, R. C., B. Borchers, and C. H. Thurber (2005), Parameter Estimation and Inverse Problems, 301 pp., Elsevier, Amsterdam, Netherlands.

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. (2006), A manifesto for the equifinality thesis, J. Hydrol., 320(1-2), 18–36, doi: 10.1016/j.hydrol.2005.07.007.

Beven, K. J. (2010), Preferential flows and travel time distributions: Defining adequate hypothesis tests for hydrological process models, Hydrol. Processes, 24, 1537–1547, doi:10.1002/hyp.7718.

Black, G. E., and Black, A. D. (2012), PEST controlled: responsible application of inverse techniques on UK groundwater models. Geological Society, London, Special Publications, 364(1), 353-373.

Borghi, A., P. Renard, and F. Cornaton (2016), Can one identify karst conduit networks geometry and properties from hydraulic and tracer test data?, Adv. Water Resour., 90, 99–115, http://dx.doi.org/10.1016/j.advwatres.2016.02.009.

Brunner, P. J. Doherty, and C. T. Simmons (2012), Uncertainty assessment and implications for data acquisition in support of integrated hydrologic models, Water Resour. Res., 48, W07513, doi:10.1029/2011WR011342.

Burrows, W., and J. Doherty (2015), Efficient calibration/uncertainty analysis using paired complex/surrogate models, Groundwater, 53(4), 531–541, doi:10.1111/gwat.12257.

Burrows, W., and J. Doherty (2016), Gradient-based model calibration with proxy-model assistance, J. Hydrol., 533, 114–127.

Campbell, E. P. and B. C. Bates (2001), Regionalization of rainfall-runoff model parameters using Markov chain Monte Carlo samples, Water Resour. Res., 37(3), 731–739.

Campbell, E. P., D. R. Fox, and B. C. Bates (1999), A Bayesian approach to parameter estimation and pooling in nonlinear flood event models, Water Resour. Res., 35(1), 211–220.

Carrera, J., and S. P. Neuman, (1986), Estimation of aquifer parameters under transient and steady-state conditions. 2. uniqueness, stability, and solution algorithms, Water Resour. Res., 22(2), 211–227, doi: 10.1029/WR022i002p00211.

Castelletti, A., S. Galelli, M. Restelli, and R. Soncini-Sessa (2011), Data-driven dynamic emulation modeling for the optimal management of environmental systems, Env. Mod. and Soft., doi:10.1016/j.envsoft.2011.09.003

Cheng, W.-C., M. Putti, D. R. Kendall, and W. W.-G. Yeh (2011), A real-time groundwater management model using data assimilation, Water Resour. Res., 47, W06528, doi:10.1029/2010WR009770.

Christensen, S., and J. Doherty (2008), Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration, Adv. Water Resour., 31, 674–700.

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.

Constable, S.C., R. L. Parker, and C. G. Constable (1987), Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, Geophysics, 52(3), 289–300.

Cooley, R. L. (2004), A theory for modeling ground-water flow in heterogeneous media, U.S. Geol. Surv. Prof. Pap., 1679, 220 pp.

Cooley, R. L., and S. Christensen (2006), Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media, Adv. Water Resour., 29, 639–656, doi:10.1016/j.advwatres.2005.07.012.

Cui, T., C. Fox, and M. J. O'Sullivan (2011), Bayesian calibration of a large scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm, Water Resour. Res., 47, W10521, doi:10.1029/2010WR010352.

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.

de Groot-Hedlin, C., and S. Constable (1990), Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data, Geophysics, 55(12), 1613–1624.

de Marsily, G., De I'identification des systemes en hydrogeologiques (tome 1), Ph.D. thesis, pp. 58–130, L'Univ. Pierre et Marie Curie-Paris VI, Paris, 1978.

de Marsily, G., C. Lavedan, M. Boucher, and G. Fasanion (1984), Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, in Geostatistics for Natural Resources Characterization. Part 2, edited by G. Verly et al., pp. 831–849, D. Reidel, Dordetcht, Netherlands.

Deutsch, C., and A. Journel (1998), GSLIB Geostatistical Software Library and User's Guide, 2nd ed., 368 pp., Oxford Univ. Press, N. Y.

Doherty, J. (2003), Ground water model calibration using pilot points and regularization, Ground Water., 43(2), 170–177.

Doherty, J. (2015), Calibration and uncertainty analysis for complex environmental models, Watermark Numerical Computing, Brisbane, Australia, 227 pp., ISBN: 978-0-9943786-0-6.

Doherty, J. (2016a), PEST: Model-independent parameter estimation, user manual, Watermark Numerical Computing, available at http://www.pesthomepage.org/Downloads.php.

Doherty, J. (2016b), Groundwater data utilities, part B: program descriptions, Watermark Numerical Computing, available at http://www.pesthomepage.org/Downloads.php.

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., M. N. Fienen, and R. J. Hunt (2010), Approaches to highly parameterized inversion: Pilot-point theory, guidelines, and research directions: U.S. Geol. Surv. Sci. Invest. Rep. 2010-5168, 36 pp., U.S. Geol. Surv., Middleton, Wis.

Doherty, J., and R. J. Hunt (2010), Approaches to highly parameterized inversion—A guide to using PEST for groundwater-model calibration, Sci. Invest. Rep. 2010-5169, U.S. Geol. Surv., Reston, Va.

Doherty, J., and C. T. Simmons (2013), Groundwater modelling in decision support: reflections on a unified conceptual framework. Hydrogeology Journal, 21(7), 1531–1537.

Doherty, J., and R. Vogwill (2016), Models, decision-making and science, In: Solving the Groundwater Challenges of the 21st Century, pp. 95–114.

Doherty, J. and D. Welter (2010), A short exploration of structural noise, Water Resour. Res., 46, W05525, doi:10.1029/2009WR008377.

Downes, B. J., L. A. Barmuta, P. G. Fairweather, D. P. Faith, M. J. Keough, P. S. Lake, B. D. Mapstone, and G. P. Quinn (2002), Monitoring ecological impacts. Concepts and practice in flowing waters, Cambridge Univ. Press, Cambridge, U. K.

Draper, D. (1995), Assessment and propagation of model uncertainty, J. R. Stat. Soc., Ser. B, 57(1), 45–97.

Draper, N. R., and H. Smith (1998), Applied Regression Analysis, 3rd ed., 706 pp., John Wiley, N. Y.

Dripps, W. R., and K. R. Bradbury (2007), A simple daily soilwater balance model for estimating the spatial and temporal distribution of groundwater recharge in temperate humid areas, Hydrogeo. J., 15(3), 433–444. http://dx.doi.org/10.1007/s10040-007-0160-6.

Eggleston, J., and S. Rojstaczer (1998), Identification of large-scale hydraulic conductivity trends and the influence of trends on contaminant transport, Water Resour. Res., 34(9), 2155–2168.

Farmer, C. L. (2002), Upscaling: A review, Int. J. Numer. Meth. Fluids, 40, 63–78, doi:10.1002/fld.267.

Feddes, R. A., P. J. Kowalik, and H. Zaradny (1978), Simulation of Field Water Use and Crop Yield, John Wiley, Hoboken, N. J.

Fienen, M., C. Muffels, and R. Hunt (2009), On constraining pilot point calibration with regularization in PEST, Ground Water, 47(6), 835-844, doi:10.1111/j.1745-6584.2009.00579.x.

Francés, A. P., E. Berhe, and M. Lubczynski (2010), Spatio-temporal groundwater recharge assessment using a lumped-parameter distributed model of the unsaturated zone (pyEARTH-2D), Geophys. Res. Abs., 12, EGU2010-6627-2.

Freeze, R. A., J. Massman, L. Smith, T. Sperling, and B. James (1990), Hydrological decision analysis, 1, A framework, Ground Water, 28(5), 738–766.

Freyberg, D. L. (1988), An exercise in ground-water model calibration and prediction, Ground Water, 26(3), 350–360.

Gallagher, M., and J. Doherty (2007a), Parameter estimation and uncertainty analysis for a watershed model, Environ. Model. Software, 22(7), 1000–1020, doi:10.1016/j.envsoft.2006.06.007.

Gallagher, M. R., and J. Doherty (2007b), Parameter interdependence and uncertainty induced by lumping in a hydrologic model, Water Resour. Res., 43,W05421, doi:10.1029/2006WR005347.

Gerritsen, M., and J. V. Lambers (2008), Integration of local-global upscaling and grid adaptivity for simulation of subsurface flow in heterogeneous formations, Comput. Geosci., 12, 193–208.

Gómez-Hernández, J. J. (2006), Complexity, Ground Water, 44(6), 782–785.

Greenhalgh, S. A., B. Zhou, and A. Green (2006), Solutions, algorithms and inter-relations for local minimization search geophysical inversion, J. Geophys. Eng., 3(2), 101–113, doi:10.1088/1742-2132/3/2/001.

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.

Haitjema, H. (2011), Model complexity: A cost-benefit issue, In: Geological Society of America Annual Meeting, October 9–12, 2011, Minneapolis, Minnesota,

http://gsa.confex.com/gsa/2011AM/finalprogram/abstract_197453.htm (accessed July 4, 2016)

Hansen, N. and A. Ostermeier (2001), Completely derandomized self-adaptation in evolution strategies, Evol. Comput., 9, 159–195.

Hansen, N., S. D. Muller, and P. Koumoutsakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evol. Comput., 9, 159–195.

Harbaugh, A. W., E. R. Banta, M. C. Hill, and M. G. McDonald (2000), The U.S. Geological Survey modular ground-water model: User guide to modularization concepts and the ground-water flow process, U.S. Geol. Surv. Open File Rep., 00-92.

Harmon, R., and P. Challenor (1997), A Markov chain Monte Carlo method for estimation and assimilation into models, Ecol. Model. 101, 41–59.

Hendricks-Franssen, H.-J., A. Alcolea, M. Riva, M. Bakr, N. van der Wiel, F. Stauffer, and A. Guadagnini (2009), A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments, Adv. Water Resour., 32(6), 851–872, doi:10.1016/j.advwatres.2009.02.011.

Herckenrath, D. (2012), Informing groundwater models with near-surface geophysical data, Ph.D thesis, Technical University of Denmark.

Herckenrath, D., C. D. Langevin, and J. Doherty (2011), Predictive uncertainty analysis of a salt water intrusion model using null space Monte Carlo, Water Resour. Res., 47, W05504. doi:10.1029/2010WR009342.

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.

Hunt, R. J., J. Doherty, and M. J. Tonkin (2007), Are models too simple? Arguments for increased parameterisation, Ground Water, 45(3), 254–262, doi:10.1111/j.1745-6584.2007.00316.x.

Hunt, R. J., J. Luchette, W. A. Shreuder, J. Rumbaugh, J. Doherty, M. J. Tonkin, and D. Rumbaugh (2010), Using the cloud to replenish parched groundwater modeling efforts, Ground Water, 48(3), 360–365, doi:10.1111/j.1745-6584.2010.00699.x.

Hunt, R. J., and D. E. Welter (2010), Taking account of "unknown unknowns", Ground Water, 48(4), 477, doi:10.1111/j.1745-6584.2010.00681.x.

Hunt, R. J., and C. Zheng (2012), The current state of modelling, Ground Water, 50(3), 330–333.

James, A. L., and C. M. Oldenburg (1997), Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation, Water Resour. Res., 33(11), 2495–2508.

James, S. C., J. E. Doherty, and A.-A. Eddebbarh (2009), Practical postcalibration uncertainty analysis: Yucca Mountain, Nevada, Ground Water, 47(6), 851–869.

Kanso, A., M. C. Gromaire, E. Gaume, B. Tassin, and G. Ghebbo (2003), Bayesian approach for the calibration of models: application to an urban stormwater pollution model, Water Sci. Technol. 47(4), 77–84.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, Water Resour. Res., 42, W03408, doi:10.1029/2005WR004376.

Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, Water Resour. Res., 46, W10517, doi:10.1029/2009WR008584.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol., 63, 425–464, doi:10.1111/1467-9868.00294.

Knowling M. J., A. D. Werner, and D. Herckenrath (2015), Quantifying climate and pumping contributions to aquifer depletion using a highly parameterised groundwater model: Uley South Basin (South Australia). J. Hydrol. 523, 515-530.

Konikow, L. F., and J. D. Bredehoeft (1992), Ground-water models cannot be validated, Adv. Water Resour., 15(1), 75–83.

Kourakos, G., and A. Mantoglou (2012), Inverse groundwater modeling with emphasis on model parameterization, Water Resour. Res., 48, W05540, doi:10.1029/2011WR011068.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J. Hydrol., 331, 161–177, doi:10.1016/ j.jhydrol.2006.05.010.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm, J. Hydrol., 211, 69–85.

Langevin, C. D., and S. Panday (2012), Future of groundwater modelling, Ground Water, 50, 334–339, http://dx.doi.org/10.1111/j.1745–6584.2012.00937.x.

Langevin, C. D., and M. Zygnerski (2013), Effect of sea-level rise on salt water intrusion near a coastal well field in southeastern Florida, Groundwater, 51(5), 781–803, doi: 10.1111/j.1745-6584.2012.01008.x.

Lewis, M. F., and G. R. Walker (2002), Assessing the potential for episodic recharge in south-western Australia using rainfall data, Hydrogeol. J., 10, 229–237.

Ma, X., and N. Zabaras (2011), Kernel principal component analysis for stochastic input model generation, J. Comput. Phys., 230, 7311–7331, doi:10.1016/j.jcp.2011.05.037.

Mahmud, K., Mariethoz, G., Caers, J., Tahmasebi, P., and A. Baker (2014), Simulation of Earth textures by conditional image quilting, Water Resour. Res., 50 (4), 3088–3107.

Makowski, D., D. Wallach, and M. Tremblay (2002), Using a Bayesian approach to parameter estimation; comparison of the GLUE and MCMC methods, Agronomie, 22, 191–203.

Mariethoz, G., and B. F. J. Kelly (2011), Modeling complex geological structures with elementary training images and transform-invariant distances, Water Resour. Res., 47, W07527, doi:10.1029/2011WR010412.

Mariethoz, G., and S. Lefebvre (2014), Bridges between multiple-point geostatistics and texture synthesis: review and guidelines for future research, Comput. Geosci., 66, 66–80.

Mariethoz, G., P. Renard, and J. Caers (2010a), Bayesian inverse problem and optimization with iterative spatial resampling, Water Resour. Res., 46, W11530, doi:10.1029/2010WR009274.

Mariethoz, G., P. Renard, and J. Straubhaar (2010b), Direct sampling method to perform multiple-point geostatistical simulations, Water Resour. Res., 46(11), W11536. doi:10.1029/2008WR007621.

Martínez-Santos, P., and J. M. Andreu (2010), Lumped and distributed approaches to model natural recharge in semiarid karst aquifers, J. Hydrol., 388, 389-398. doi:10.1016/j.jhydrol.2010.05.018.

Maurer, H., K. Holliger, and D. E. Boerner (1998), Stochastic regularization: Smoothness or similarity?, Geophys. Res. Lett., 25(15), 2889–2892, doi:10.1029/98GL02183.

McLaughlin, D., and L.R. Townley, (1996), A reassessment of the groundwater inverse problem. Water Resources Research, 32(5), 1131–1161.

Meerschman, E., G. Pirot, G. Mariethoz, J. Straubhaar, M. Van Merivenne, and P. Renard (2013), A practical guide to performing multiple-point geostatistical simulations with the direct sampling algorithm, Comput. Geosci., 52, 307–324.

Meisel, W. S., and D. C. Collins (1973), Repromodeling: An approach to efficient model utilization and interpretation, IEEE Trans. Syst., Man,Cybern,, SMC-3, 349-358.

Menke W. (1984), Geophysical Data Analysis: Discrete Inverse Theory, 289 pp., Academic, N. Y.

Mondal, A., Y. Efendiev, B. Mallick, and A. Datta-Gupta (2010), Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump Markov chain Monte Carlo methods, Adv. Water Resour., 33(3), 241-256.

Moore, C., and J. Doherty (2005), The role of the calibration process in reducing model predictive error, Water Resour. Res., 41(5), W05020, doi:10.1029/2004WR003501.

Moore, C., and J. Doherty (2006), The cost of uniqueness in groundwater model calibration, Adv. Water Resour., 29(4), 605–623 doi:10.1016/j.advwatres.2005.07.003.

Moore, C., T. Wöhling, T., and J. Doherty (2010), Efficient regularization and uncertainty analysis using a global optimization methodology, Water Resour. Res., 46, W08527, doi:10.1029/2009WR008627.

Mugunthan, P., and C. A. Shoemaker (2006), Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, Water Resour. Res., 42, W10428, doi:10.1029/2005WR004640.

Oakley, J., and A. O'Hagan (2002), Bayesian inference for the uncertainty distribution of computer model outputs, Biometrika, 89, 769-784.

Pachepsky, Y. A., A. K. Guber, M. Th. van Genuchten, T. J. Nicholson, R. E. Cady, J. Simunek, and M. G. Schaap (2006), Model abstraction techniques for soil-water flow and transport, prepared for Division of Fuels, Engineering and Radiological Research, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission, Rep. NUREG/CR-6884, 135 pp.

Portniaguine, O., and M. S. Zhdanov (1999), Focusing geophysical inversion images, Geophysics, 64(3), 874–887.

Qian, S. S., C.A. Stow, and M. E. Borsuk (2003), On Monte Carlo methods for Bayesian inference, Ecol. Model. 159, 269-277.

Ratto, M., A. Castelletti, and A. Pagano (2011), Emulation techniques for the reduction and sensitivity analysis of complex environmental models, Env. Mod. and Soft., doi:10.1016/j.envsoft.2011.11.003.

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

Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Højberg, and L. Troldborg (2012), Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, Adv. Water Resour., 36, 36–50, doi:10.1016/j.advwatres.2011.04.006.

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.

Sadegh, M., and J. A. Vrugt (2014), Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC), Water Resour. Res., 50, doi:10.1002/2014WR015386.

Sarma, P., L. J. Durlofsky, and K. Azis (2008), Kernel principal component analysis for efficient, differentiable, parameterization of multipoint geostatistics. Math. Geosci., 40, 3-32.

Savenije, H. H. G. (2009), The art of hydrology, Hydrol. Earth Syst. Sci., 13, 157–161.

Scheidt, C., J. Caers, Y. Chen, and L. J. Durlofsky (2011), A multi-resolution workflow to generate high-resolution models constrained to dynamic data, Comput. Geosci., 15(3), 545-563, doi:10.1007/s10596-011-9223-9.

Schoups, G. and J.W. Hopmans. 2006. Vadose Zone Journal Special Issue. Evaluation of model complexity and input uncertainty of field-scale water flow and salt transport. Vadose Zone Journal 5:951-962.

Simmons, C. T., and R. J. Hunt (2012), Updating the debate on model complexity, GSA Today, 22(8), 28–29, doi:10.1130/GSATG150GW.1

Šimůnek, J, M. Šejna, H. Saito, M. Sakai, and M. Th. van Genuchten (2009). The HYDRUS-1D Software Package for Simulating the One-Dimensional Movement of Water, Head, and Multiple Solutes in Variably-Saturated Media. Department of Environmental Sciences, University of California Riverside, Riverside, California.

Singh, A., B. S. Minsker, and A. J. Valocchi (2008), An interactive multi-objective optimization framework for groundwater inverse modeling, Adv. Water Resour., 31, 1269–1283.

Sivakumar, B. (2008), Dominant processes concept, model simplification and classification framework in catchment hydrology, Stoch. Environ. Res. Risk Assess., 22, 737–748.

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.

Stone, N. (2011), Gaussian process emulators for uncertainty analysis in groundwater flow, PhD thesis, Univ. of Nottingham, Nottingham.

Tikhonov, A. N. (1963a), Solution of incorrectly formulated problems and the regularization method, Soviet Mathematics Doklady, 4, 1035-1038.

Tikhonov, A. N. (1963b), Regularization of incorrectly posed problems, Soviet Mathematics Doklady, 4, 1624–637.

Tikhonov A. N., and V. Y. Arsenin (1977), Solution of Ill-Posed Problems, Winston, Washington, D. C.

Tolson, B. A., and C. A. Shoemaker (2008), Efficient prediction uncertainty approximation in the calibration of environmental simulation models, Water Resour. Res., 44, W04411, doi:10.1029/2007WR005869.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, 41, W10412, doi:10.1029/2005WR003995.

Tonkin M., and J. Doherty (2009), Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques, Water Resour. Res., 45, W00B10, doi:10.1029/2007WR006678.

Tonkin, M., J. Doherty, and C. Moore (2007), Efficient nonlinear predictive error variance for highly parameterized models, Water Resour. Res., 43, W07429, doi:10.1029/2006WR005348.

Touhami, I., J. M. Andreu, E. Chirino, J. R. Sánchez, H. Moutahir, A. Pulido-Bosch, P. Martínez-Santos, and J. Bellot (2012), Recharge estimation of a small karstic aquifer in a semiarid Mediterranean region (southeastern Spain) using a hydrological model, Hydrol. Proc., doi: 10.1002/hyp.9200.

Ulugergerli, E.U. (2011), Two dimensional combined inversion of short- and longnormal dc resistivity well log data, J. Appl. Geophys., 73, 130–138.

van Genuchten, M. Th. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Sci. Soc. Am. J., 44, 892–898.

Vermeulen, P. T. M., A. W. Heemink, and C. Stroet (2004), Reduced models for linear groundwater flow models using empirical orthogonal functions, Adv. Water Resour., 27(1), 57–69.

Vermeulen, P. T. M., A. W. Heemink, and J. R. Valstar (2005), Inverse modeling of groundwater flow using model reduction, Water Resour. Res., 41, W06003, doi:10.1029/2004WR003698.

Vermeulen, P. T. M., C. B. M. te Stroet, and A. W. Heemink (2006), Model inversion of transient nonlinear groundwater flow models using model reduction, Water Resour. Res., 42, W09417, doi:10.1029/2005WR004536.

Vo, H., and L. Durlofsky (2014), A new differentiable parameterization based on principal component analysis for the low-dimensional representation of complex geological models, Math. Geosci. 46(7), 775–813, doi:10.1007/s11004-014-9541-2.

Voss, C. I. (2011), Editor's message: Groundwater modeling fantasies—part 1, adrift in the details. Hydrogeol J., doi:10.1007/s10040-011-0789-z.

Vrugt, J. A., C. G. H. 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.

Vrugt, J. A., B. O. Nuallain, B. A. Robinson, W. Bouten, S. C. Decker, and P. M. A. Sloot (2006), Application of parallel computing to stochastic parameter estimation in environmental models, Comput. Geosci., 32, 1139–1155.

Vrugt, J. A., G. H. Schoups, J. W. Hopmans, C. Young, W. W. Wallender, T. H. Harter, and W. Bouten (2004), Inverse modeling of large-scale spatially-distributed vadose zone properties using global optimization, Water Resour. Res., 40, W06503, doi:10.1029/2003WR002706.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Hidgon (2009a), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, Int. J. of Nonlinear Sci., 10(3), 273–290.

Vrugt, J. A., C. J. ter Braak, H. V. Gupta, and B. A. Robinson (2009b), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, Stoch. Environ. Res. Risk Asses., 23, 1011–1026, doi:10.1007/s00477-008-0274-y.

Welter, D. E., J. T. White, R. J. Hunt, and J. E. Doherty (2015), Approaches in highly parameterized inversion—PEST++ Version 3, a Parameter ESTimation and uncertainty analysis software suite optimized for large environmental models: U.S. Geol. Surv. Tech. Methods, Book 7, Chap. C12, 54 pp., http://dx.doi.org/10.3133/tm7C12.

Wen, X.-H., and J. J. Gómez-Hernández (1998), Numerical modeling of macrodispersion in heterogeneous media: A comparison of multi-Gaussian and non-multi-Gaussian models, J. Contam. Hydrol., 30, 129–156.

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, doi:10.1002/2013WR014767.

Xu, T., and A. J. Valocchi (2015), A Bayesian approach to improved calibration and prediction of groundwater models with structural error, Water Resour. Res., 51, 9290– 9311, doi:10.1002/2015WR017912.

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.

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.

Yoon, H., D. B. Hart, and S. A. McKenna (2013), Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods, Water Resour. Res., 49, doi:10.1002/wrcr.20064.

Young, P. C., and M. Ratto (2009), A unified approach to environmental systems modeling, Stoch. Environ. Res. Risk Assess., 23, 1037–1057.

Young, P. C., and M. Ratto (2011), Statistical emulation of large linear dynamic models, Technometrics, 53, 29–43.

Zahner T., T. Lochbühler, G. Mariethoz, and N. Linde (2016), Image synthesis with graph cuts: a fast model proposal mechanism in probabilistic inversion, Geophys. J. Int., 204(2), 1179–90. http://dx.doi.org/10.1093/gji/ggv517.

Zheng, C., M. Bianchi, and S. M. Gorelick (2011), Lessons learned from 25 years of research at the MADE site, Ground Water, 49, 649–662, doi:10.1111/j.1745-6584.2010.00753.x.

Zhou H., J. J. Gómez-Hernández, and L. Li (2014), Inverse Methods in Hydrogeology: Evolution and Recent Trends. Adv. Water Resour. 63, 22–37. doi:10.1016/j.advwatres.2013.10.014.

Zhu, J., and D. Sun (2009), Effective soil hydraulic parameters for transient flows in heterogeneous soils, Vadose Zone J., 8, 301–309, doi:10.2136/vzj2008.0004.

Zimmerman, D. A., et al. (1998), A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, Water Resour. Res., 34(6), 1373–1413, doi:10.1029/98WR00003.