Title
Efficient Numerical Implementation and Vadose Zone Application of the Method of Anchored Distributions

Author
Over, Matthew William

Publication Date
2013

Peer reviewed|Thesis/dissertation

Efficient Numerical Implementation and Vadose Zone Application of the Method of Anchored Distributions

By

Matthew William Over

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yoram Rubin, chair
Assistant Professor Sally Thompson
Associate Professor Alan Hubbard
Associate Professor Daniel P. Ames

Fall 2013

Efficient Numerical Implementation and Vadose Zone Application of the Method of Anchored Distributions

By

Matthew William Over

Abstract

Efficient Numerical Implementation and Vadose Zone Application of the Method of Anchored Distributions

by

Matthew William Over

Doctor of Philosophy in Civil & Environmental Engineering

University of California, Berkeley

Professor Yoram Rubin, Chair

The method of anchored distributions (MAD) is a Bayesian model inversion technique with a high level of flexibility. MAD can jointly invert multiple types of parameter fields conditional on multiple types and scales of measurement data. Moreover, MAD permits simultaneous analysis of macroscopic characteristics of spatial heterogeneity, e.g. correlation scale, and point characteristics within the parameter field. This dissertation focuses on MAD.

MAD is formulated with a completely generalized, assumption-free likelihood function – a feature that sets it decidedly apart from other inversion and estimation procedures. However, this advantage comes with a considerable increase in computational cost relative to more assumption-laden model inversion techniques. Thus, a section of the following work derives a theoretical approximation to reduce the computational cost of inversion with MAD that has minimal impacts on the accuracy of the results. The approximation utilizes clustering algorithms to combine simulations on a basis of parameter similarities and ultimately limits the computational expense of evaluating the likelihood function, which has significant impact on overall computational cost. The approach is validated in a characterization of a synthetic transmissivity field using concentration data obtained under natural gradient conditions.

MAD is formulated generically and hence is widely applicable to a variety of scientific practice areas. However, previously there has been no software platform for implementing MAD available to the scientific community. Thus, a section of the following work is dedicated to the design and development of the MAD software, which generically evaluates Bayes' rule for different modeling tools, different physical processes, different random physical parameters, and with different statistical tools. The section, focuses on the creation of a graphical user interface (GUI) that helps users define the necessary aspects of the MAD analysis, but is sequenced in a manner that all dependencies are exploited to reduce the possibility of user error in the set up. The flexibility and ease of using the GUI is validated with successful application in a variety of synthetic case studies.

With the existence of a free and publicly available software, the number of studies that could employ MAD has grown substantially. However, MAD, thus far, has never been applied without geostatistics or outside of saturated groundwater studies. Thus, a section of the following work is dedicated to the derivation of MAD with statistical, rather than geostatistical structural models, and application of the framework to a vadose zone soil column characterization. The MAD software is used for the first time on field data (not synthetic) and reasonable conditional results of the Mualem – van Genuchten parameters. The important outcome of the experiment is the first ever validation of a likelihood function in vadose zone parameter inversion.

In total, this dissertation has focused on generic and rapid numerical implementation of MAD. The case studies in saturated and vadose zone flow and transport are used to provide experimental confidence that the generalization and computational gains are successful.

To my wife,

my parents,

& my brother

TABLE OF CONTENTS

Acknowledgements

"Alone we can do so little; together we can do so much."

-Helen Keller

This dissertation has been made possible by a wealth of supportive people, a great university, and good fortune. I have encountered so many wonderful people during my graduate work; I can only hope to recall a fraction of them.

First and foremost, I must thank my parents Susan & Jerry, my brother Michael, and my wife Mallory for being willing to experience all the joys, breakthroughs, and sorrows of research. No one else has had the patience to deal with the highs and the lows of my academic career and for that I am grateful. Your love and confidence in my ability to succeed is remarkable. I love you all very much.

Second, I must thank my advisor Yoram Rubin. You believed in me and recognized my potential in ways that even I could not see. Your wisdom and experience were an invaluable addition to my time at Cal. Nearly four years ago, you asked me a formative question, "What are you doing after this?" while proctoring a final examination. In my innocence, I literally thought you were asking about later that evening, not presenting me with a golden opportunity to continue studying at Cal. To look back at the road to a Ph.D. that began with so little experience on my part, it has been amazing journey and progression. It has been a pleasure to master and champion the method of anchored distributions under your guidance.

Third, I must thank those who I have collaborated with and those who have helped support my career. I would like to thank the Jane Lewis Fellowship, the UFZ, the Roy G. Post Foundation, the NSF, NERSC, and Yoram Rubin's support in providing me with resources, tuition and salary over the past few years. I would like to respectfully acknowledge my co-workers, in no particular order: Ute Wollschläger, Peter Dietrich, Daniel P. Ames, Carlos Andres Osorio-Murillo, Daniel Gunnell, Shelley Okimoto, Heather Frystacky, Bradley Harken, Michelle Newcomer, Yarong Yang, Xingyuan Chen, Haruko Murakami, Felipe de Barros, Wolfgang Nowak, Alan Hubbard, Hassan Astaneh, Tina Chow, Mark Stacey, & Sally Thompson. Each of you have walked alongside me at some point during my career – you have advocated for, collaborated with, defended and supported me, but most importantly, you have been my teachers.

Finally, I must thank my friends. You make me laugh, you trust me, and you make me feel like family, even when I am so far away from my relatives. In no particular order, I would like to thank: Julien Cohen-Waeber, Joseph Weber, Madison Weber, Justin Hollenback, Josh

Zupan, Roozbeh Mikola, Tonguc Deger, Will Trono, Kati Miller, Ken Ogorzalek, Jörg Haussmann, Jeff Jackman, Joseph Tullmann, Liz Bokermann, Gopal Penny, & Mike George.

I could never have accomplished so much without the kindness and generosity of so many others. With my enduring love and absolute respect, thank you all so much for helping me reach this milestone.

1. Introduction

This dissertation focuses on the improvement, generalization, and extension of the method of anchored distributions (MAD) [*Rubin et al.,* 2010; *Murakami, et al.*, 2011; *Chen et al.,* 2012; *Yang et al.,* 2012; *Over et al.,* 2013]. The improvement to MAD is an approximation technique utilizing clustering for faster implementation. The generalization is the design and development of an open-source graphical user interface (GUI) for generic MAD implementation. The extension is the first application of MAD to a vadose zone experiment. In the remainder of the introduction, fundamental background material and important literature are presented to supplement understanding of the subsequent research findings.

As this dissertation is highly focused on MAD, it is essential at the outset to present and define this Bayesian framework for inversion and uncertainty quantification of heterogeneous parameter fields. But first, it will be useful to review a few other relevant techniques for optimization, calibration, and parameter field estimation that preceded the development of MAD. In this regard, it will be possible to highlight the differences of MAD from these other theories as well as possible advantages of MAD. Specific limitations of the MAD approach are discussed as they are encountered in the subsequent chapters, but the primary drawback is the very large computational cost of MAD.

1.1 Maximum Likelihood, Maximum *a Posteriori*, Generalized Likelihood Uncertainty Analysis, Shuffled Complex Evolution, and the Pilot Point Method

In this section, some of the historically important techniques related to MAD are reviewed. The goal of all of these techniques is the estimation of numerical model parameters that most reliably simulate the measured data for a site.

Amongst the first applications of stochastic methods in hydrogeological characterization problems was the work of *Kitanidis & Vomvoris* [1983]. Their insightful work introduced maximum likelihood (ML) estimation of hydrogeological field structural parameters, such as the integral scale of the transmissivity field, constrained by measurements of the permeability or pressure head. ML was computationally improved by *Kitanidis & Lane* [1984] using the Gauss-Newton method and extended to transient flow problems by *Carrera & Neuman* [1986]. ML has been in myriad hydrological and hydrogeological

applications since this introductory work [*c.f. Hollenbeck & Jensen*, 1998; *Pardo-Iguzquiza*, 1998; *Neuman*, 2003; *Ye et al.,* 2004].

The basic formulation of ML is to define the likelihood function, a joint probability distribution function (PDF) of the measurement data conditional on the model parameters $f(\boldsymbol{z}|\boldsymbol{\rho})$, where $\boldsymbol{z}$ is a vector of measurements and $\boldsymbol{\rho}$ is a vector of generic parameters, and determine the estimate of $\boldsymbol{\rho}$ – denoted by $\boldsymbol{\rho}^*$ - that maximizes the likelihood function.

ML is an important pre-cursor to MAD because it sets the stage for utilizing measurements of indirect state variables (i.e. pressure head), which are often cheaper to attain in the subsurface, to condition the intrinsic variables (i.e. hydraulic conductivity), which are often much more expensive or difficult to attain in the subsurface. Moreover, ML yields unbiased and minimum variance estimates of the parameters.

However, ML has two distinct disadvantages. The first is that the form of the joint PDF of the state variable measurements must be known [*Kitanidis & Vomvoris,* 1983]. The common workaround to this requirement is to make an assumption of a multivariate Gaussian (MG) distribution, but without replicate experiments to measure the same state variables under the same conditions, little can be known about this distributional form and the working assumption of MG is not possible to validate [*Ginn & Cushman,* 1990; *Hollenbeck & Jensen*, 1998]. The second is that because the estimation approach is a maximization problem, without restriction, the parameters can take on non-physical values [*Sorooshian et al.,* 1983].

Another technique common in stochastic inverse approaches is maximum *a posteriori* (MAP) estimation. MAP provides a straightforward, Bayesian way of incorporating prior information into the parameter estimation procedure [*Carrera & Neuman,* 1986; *Bates & Townley* 1988; *McLaughlin & Townley*, 1996]. The approach presented above is now modified slightly to maximize

$$f(\boldsymbol{\rho}|\boldsymbol{z}) = cf(\boldsymbol{z}|\boldsymbol{\rho})f(\boldsymbol{\rho}), \tag{1.1}$$

where $f(\boldsymbol{\rho}|\boldsymbol{z})$ is the posterior joint PDF of the parameters conditional on the measurement data, $f(\boldsymbol{\rho})$ is the prior joint PDF of the parameters, and c is a normalization constant.

One advantage of including a prior in the maximization (right hand side of Equation 1.1) is the ability to confine the parameter within physically reasonable boundaries [*Martin & Stedinger*, 2000], which is an enhancement relative to ML. A key disadvantage of MAP is the introduction of bias in the parameter estimates [*Kitanidis*, 1996], which is a worsening relative to ML.

An extension of ML and MAP is generalized likelihood uncertainty estimation (GLUE) [*Beven & Binley,* 1992; *Beven & Freer,* 2001]. GLUE utilizes a likelihood measure to quantify parametric uncertainty, which is a user defined function that increases as the agreement between model simulations and observations improves. In GLUE, model parameter vectors that have a likelihood measure above a certain threshold (user specified) are classified as "behavioral" and those that do not are classified as "non-behavioral". The behavioral ensemble is then weighted by the likelihood measure and used to attain a cumulative

distribution of model simulations from which quantiles or other statistical measures can be calculated. GLUE has been the topic of recent reviews as well as hydrological and hydrogeological applications [*Mertens et al*., 2004; *Li et al.,* 2010; *Rogiers et al.*, 2012].

A major advantage of GLUE is the method addresses the problem of equifinality, in which multiple model parameter vectors result in best, but identical model performance as quantified by the likelihood measure. ML and MAP on the other hand cannot address equifinality, as there is no way to "break a tie" between the best performing parameter vectors in maximization approaches. However, GLUE is typically much more computationally demanding than ML or MAP [*Mertens et al.,* 2004].

There are three significant shortcomings of GLUE. First, the user-specified threshold is a rejection-acceptance criterion, which has clear implications for the bias of the resulting behavioral simulation ensemble [*Li et al.,* 2010]. Relaxing the criterion increases the behavioral simulation ensemble size and includes simulations that do not agree as stringently with the observations, which typically widens the confidence intervals (CIs) in the cumulative distribution, and vice versa. Thus CIs reported with GLUE are irreversibly a function of the threshold and there is no good rule-of-thumb for setting the threshold value [*Mertens et al.,* 2004; *Rogiers et al.,* 2012]. Second, as argued analogously by *Rubin et al*. [2010] about the pilot point method (summarized shortly hereafter), the exclusion of simulations from the cumulative distribution in the behavioral ensemble of GLUE on the basis of the threshold results in an under-sampling of the parameter space and the subjectivity of the likelihood measure [*Beven & Freer,* 2001] injects bias into the resulting quantiles. Third, there is no way in GLUE to establish that the parameter space has been adequately sampled, i.e. that the results are convergent [*Mertens et al.,* 2004; *Li et al.,* 2010].

At this point, the section shifts focus from stochastic approaches and the remaining two approaches presented can be described as optimization or fitting approaches.

In a pure sense of optimization, an important method to summarize is shuffled complex evolution (SCE). SCE is an extension of local optimization approaches to ensuring that maximization of the objective function does in fact attain a global maximum instead of incorrectly identifying local maximums [*Duan et al.*, 1992]. Users define an objective function and use SCE to determine the model parameter vector which maximizes this function. SCE has been extended and applied heavily in recent years: *Yapo et al.* [1998] extended SCE to multi-objective solutions of the Pareto surface; *Vrugt et al.* [2003] improved the robustness of SCE in multi-objective problems with the inclusion of a Metropolis-Hastings step to prevent the algorithm from collapsing into small regions of the parameter space; *Mertens et al.* [2004] showed in application that a 2-D Pareto surface can balance goodness of model fit with prior information attained about vadose zone van Genuchten parameters at a site; *Vrugt et al.,* [2008] developed a framework for simultaneous execution of the multiple Markov Chain Monte Carlo (MCMC) schemes that share information to actively adapt the chain evolution to further enhance efficiency of SCE.

The main advantages of SCE are the efficiency with which the optimization is executed relative to other parameter space search algorithms [*Duan et al.,* 1992]. There are however a few weakness of SCE and the extensions of it. First, SCE itself does not address the

problem of equifinality as it is still a single optimization problem, so it does not extend beyond GLUE; however, the recent work of *Vrugt et al.* [2008a] introduced an MCMC sampling step in SCE which yields a posterior PDF and circumvents this issue, because a single estimate is no longer the goal of the analysis. Second, the optimal parameter set determined by SCE is always dependent on the objective function selected or is a compromise between various solutions of the separate objective functions that define the Pareto surface [*Boyle et al.,* 2000] – there is always a component of SCE that requires users to exercise judgment or expertise to define these functions. Finally, the significant improvement of the speed of implementation not withstanding [*Vrugt et al.,* 2009], there are still large concerns about the convergence diagnostics not identifying the failures they are intended and designed to identify [*Cowles & Carlin,* 2012], which are embedded in the MCMC SCE approaches for chain termination.

The introduction of multiple objective functions extends the single function maximization approach of ML and MAP and traditional SCE, but also extends the computational cost of multi-objective SCE relative to these methods [*Vrugt et al.,* 2003]. The inclusion of convergence criteria in SCE methods, if they can be validated, would represent an improvement over GLUE.

The final technique worth mentioning in this brief overview is the pilot-point method (PPM) [*Certes & de Marsily*, 1991; *Doherty*, 2003; *Kowalsky et al.,* 2004; *Alcolea et al.,* 2006]. PPM is based on the maximization of an objective function, which means there is some overlap with the ML and MAP principles, but there are two key differences. First, PPM provides a method of localizing information in the parameter field; second, PPM is not a statistical procedure, the objective function is not a PDF. While ML and MAP are usually applied in hydrogeological studies for the estimation of statistical moments of the physical parameter field, e.g. mean and covariance of transmissivity, PPM is geared towards the localized estimation of the physical parameters, e.g. transmissivity at specific coordinates.

The ability of PPM to characterize physical parameters locally is a significant conceptual advance from the ML and MAP approaches, but there are drawbacks to the PPM approach. First, there are problems of regularization and instability that can plague PPM analysis when the number of pilot-points grows large [*Alcolea et al.,* 2006]. Second, there is significant concern that the implementation under-samples the probability space, by virtue of setting a threshold for the objective function and introduces bias in the results (similar to GLUE) [*Mertens et al.,* 2004; *Rubin et al.,* 2010]. Third, pilot points can introduce artifacts in the characterization, because of over-fitting [*Cooley,* 2000].

In closing, this section has presented five methods for parameter estimation: ML, MAP, GLUE, SCE, and PPM. As these methods are subsequently contrasted with MAD, the most significant aspect of ML, MAP, and SCE will be that these approaches derive *one* best estimate of the parameters. The key aspect of GLUE is that convergence is not definable. The important characteristics of PPM are that it is neither a Bayesian nor statistical approach. Finally, with all five methodologies, there exists a user-definition of a likelihood function, objective function, or likelihood measure. In the next section, MAD will be presented theoretically and then contrasted with these five methodologies.

1.2 MAD Theory

MAD is a Bayesian technique for characterizing spatially-distributed physical parameter fields conditional upon state variable measurements, which can be of different type and scale. The remainder of this section is devoted to three topics that introduce the elements of MAD and summarizes the work of *Rubin et al.* [2010]. It is appropriate to begin with the tools for characterizing heterogeneous parameter fields, which includes the structural model and the anchored distribution parameters. Next, the categorization scheme for the measurements is presented along with the framework for conditioning. Finally, MAD is briefly contrasted with the five methodologies presented in Section 1.1.

The MAD approach synthesizes global characterization and localized characterization of the physical parameter fields. Global (or regional) characterization is done by the structural model, which is a tool that mathematically represents spatial heterogeneity in the domain. Local characterization is done by anchored distribution parameters, hereafter called "anchors" for short, which are further developed later in the section. For sake of simplicity, the remainder of this section is devoted to the introduction of MAD for a single spatially heterogeneous physical parameter field $\boldsymbol{Y}$, but in general MAD can extended for joint characterization of multiple physical parameter fields $\boldsymbol{Y}_1, \boldsymbol{Y}_2$, and so on.

The field $\boldsymbol{Y}$ must therefore be defined by parameters from both the structural model and the anchors in order to support the global and local characterization objectives. The parameters of the structural model are denoted $\boldsymbol{\theta}$ and the anchors are denoted $\boldsymbol{\vartheta}$.

The structural model $S(\boldsymbol{Y};\ \boldsymbol{\theta})$ is utilized to generate random realizations of the physical parameter field $\widetilde{\boldsymbol{Y}}$, where the accent indicates a realization. In hydrogeological applications, the structural model is often defined as a geostatistical model and the vector $\boldsymbol{\theta}$ can include parameters such as the integral scale or nugget [*Murakami, et al.,* 2010; *Chen et al.,* 2012; *Over et al.,* 2013]. In other application areas, where geostatistics may not be appropriate, a classical statistical approach may be adopted. In these cases, the structural model may be defined as a statistical model, e.g. Gaussian, and the vector $\boldsymbol{\theta}$ can include parameters such as the covariance matrix components. The important role of the structural model is to generate random realizations of the physical parameter field.

The anchors are statistical devices that define a PDF of the physical parameter at localities in the domain. These devices are non-physical – meaning they do not have to be located where measurements were taken. Anchors are intended to capture non-local information available from the state variable measurements and convert it into direct, point information of the physical parameters. Proper usage of anchors is extensively detailed in the work of *Rubin et al.* [2010] and placement of anchors is optimized in the work of *Yang et al.* [2012].

Finally, before describing the measurement classification system, it is important to note that the synthesis of global and local characterization, via the parameter vector $[\boldsymbol{\theta}, \boldsymbol{\vartheta}]^T$, is only possible when the structural model can incorporate point information. For instance,

geostatistical approaches can easily accommodate point information, but classical statistical models, e.g. a Poisson or a Gaussian model, cannot accommodate spatial information of any type. Therefore, MAD can be formulated with or without anchors, but always with some structural model, but this will be carefully explained when MAD is compared with ML and MAP at the end of this section.

Measurements are categorized as Type-A or Type-B data in MAD, on the basis of the measurement type and the support volume of the measuring device. Type-A data $\boldsymbol{z}_a$ is a direct measurement of the physical parameter field. Type-A data can also measure another variable that relates to the physical parameter field on the same support volume as the measurement device, i.e. a permeability measurement can be related to the hydraulic conductivity on the same volume. Type-B data $\boldsymbol{z}_b$ are functions of the physical parameter field on a volume larger than the support volume of the measurement device, i.e. a pressure head measurement at a piezometer can only be related to the hydraulic conductivity using the entire field, initial conditions, and boundary conditions via the flow equation. Type-A data is often called 'direct' data and has the same requirement as anchors of a spatially compatible structural model. Type-B data is often called 'indirect' data.

With the characterization parameters and data types defined, the Bayesian framework can now be specified. The objective of MAD is to characterize spatially-distributed physical parameter fields conditional on the various measurement data, which can be accomplished using the proportionality

$$f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{z}_a) f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a) \tag{1.2}$$

where $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b)$ is the posterior joint PDF of the structural parameters and anchors conditional on Type-A and Type-B data, $f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{z}_a)$ is the likelihood function joint PDF – the probability of observing the Type-B data given the structural parameters, anchors, and Type-A data, and $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$ is the prior joint PDF of the structural parameters and anchors conditional on the Type-A data only. Note that referring to $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$ as prior is appropriate even though it is itself a conditional PDF, because it is antecedent to the inclusion of Type-B data.

Before describing the implementation of Equation 1.2, it is worthwhile to briefly highlight the important aspects of MAD that set it apart from the five methodologies described in Section 1.1. First, MAD is clearly different than ML, MAP, and SCE, because there is no maximization step. The goal of MAD is to obtain a posterior joint PDF of the structural parameters and anchors, not a single best estimate. Thus MAD provides a more complete depiction of uncertainty in the parameters than an estimate can. Second, MAD does not require user definition of the objective function, likelihood function, or the likelihood measure, which sets it apart from GLUE and PPM as well as ML, MAP, and SCE. Unlike all the other approaches, MAD requires no assumptions about the likelihood function by introducing inferential, hierarchical methods, but unfortunately this comes at cost [*Rubin et al.,* 2010]. MAD can be used to completely eliminate the possibility of user bias in the likelihood function or be used to validate assumptions made about the likelihood function. For these two reasons, there is a clear theoretical motivation for the use of MAD for parameter field characterization.

The importance of a research dissertation that improves the numerical efficiency of MAD, generalizes the numerical implementation of MAD, and extends MAD into a new field of application is justified by the important advancements offered by MAD relative to other statistical parameter estimation or optimization approaches,

1.3 MAD Implementation

In this section, two important elements of MAD are developed. First, a generic overview of five variations of Equation 1.2 that are viable formulations of MAD are presented. Second, the numerical implementation of the most comprehensive formulation of MAD is outlined.

1.3.1 Alternative Formulations of MAD

As briefly mentioned in Section 1.2, MAD can be derived with or without random structural parameters, in the absence of Type-A data, and even without anchors (an argument justified in Chapter 4). The use of indirect data is essential to any formulation of MAD.

Equation 1.2 can be considered the most comprehensive formulation of MAD and yields the joint posterior PDF of the random structural parameters and anchors conditional on both Type-A and Type-B data. An example of an appropriate usage of Equation 1.2 would be the global $\boldsymbol{\theta}$ and local characterization $\boldsymbol{\vartheta}$ of a heterogeneous hydraulic conductivity field conditioned on permeability measurements $\boldsymbol{z}_a$ and pressure head measurements $\boldsymbol{z}_b$. Equation 1.2 is not compatible with classical statistical structural models, because they are not compatible with the spatial dependence of Type-A measurements or anchors.

The next obvious variant is the Bayesian proportionality in the absence of direct measurements of the Type-A variables

$$f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}) f(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \tag{1.3}$$

An example of an appropriate usage of Equation 1.3 would be a repetition of the previous hydraulic conductivity field characterization (described above) without having the permeability measurements available. Like Equation 1.2, the usage of anchors in Equation 1.3 requires a geostatistical structural model.

The next variants are the exclusion of random structural parameters or anchors

$$f(\boldsymbol{\theta}|\boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{1.4}$$

or

$$f(\boldsymbol{\vartheta}|\boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta}) \tag{1.5}$$

which eliminates the global or local characterization from the framework respectively. An important note is that in Equation 1.5, even though the structural parameters are not treated as random variables, it is essential to still have a deterministic, geostatistical structural model to infer the likelihood function $f(\boldsymbol{z}_b|\boldsymbol{\vartheta})$; otherwise the form of the likelihood function must be assumed or user-defined, which is not customary to MAD [*Rubin et al.,* 2010]. Equation 1.4 is appropriate for use with either geostatistical or statistical structural models.

Finally, the last two variants incorporate the Type-A data back into Equations 1.4 and 1.5

$$f(\boldsymbol{\theta}|\boldsymbol{z}_a, \boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{z}_a)f(\boldsymbol{\theta}|\boldsymbol{z}_a) \tag{1.6}$$

and

$$f(\boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\vartheta}, \boldsymbol{z}_a)f(\boldsymbol{\vartheta}|\boldsymbol{z}_a). \tag{1.7}$$

Again, like Equation 1.5, Equation 1.7 also requires a definition of a deterministic, geostatistical structural model for likelihood function inference. However, unlike Equation 1.4, Equation 1.6 is not appropriate with classical statistical structural models because they are incompatible with the spatial dependence of Type-A data.

In closing, the purpose of explicitly writing out the Equations 1.3-7 is to highlight the possible formulations of the MAD Bayesian framework. Generic software for MAD, by necessity, must support all the formulations presented in this section and this topic is reiterated in Chapter 3. A case study utilizing each of these equations is developed in the subsequent three chapters.

1.3.2 Numerical Implementation of MAD

In practice, Equation 1.2 is worked from right to left. Equations 1.3-7 can be evaluated in an analogous manner to that described in this section. As a precursor to evaluating Equation 1.2, it is first necessary to define the available Type-A data $\boldsymbol{z}_a$ and Type-B data $\boldsymbol{z}_b$, the structural model and its parameters $\boldsymbol{\theta}$, and the anchors $\boldsymbol{\vartheta}$. Figure 1.1 presents a flow chart intended to accompany the discussion.

The first step of implementation is the definition of the Type-A conditional joint prior PDF $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$ of the structural parameters and anchors. This distribution can be determined using the Bayesian proportionality

$$f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a) \propto f(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \boldsymbol{z}_a)f(\boldsymbol{z}_a|\boldsymbol{\theta})f(\boldsymbol{\theta}). \tag{1.8}$$

The right side of Equation 1.8 is the triple product of: the conditional joint PDF of anchors given structural parameters and Type-A data $f(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \boldsymbol{z}_a)$, the joint conditional PDF of the

Type-A data given the structural parameters $f(\boldsymbol{z}_a|\boldsymbol{\theta})$, and the joint prior PDF of the structural parameters $f(\boldsymbol{\theta})$.

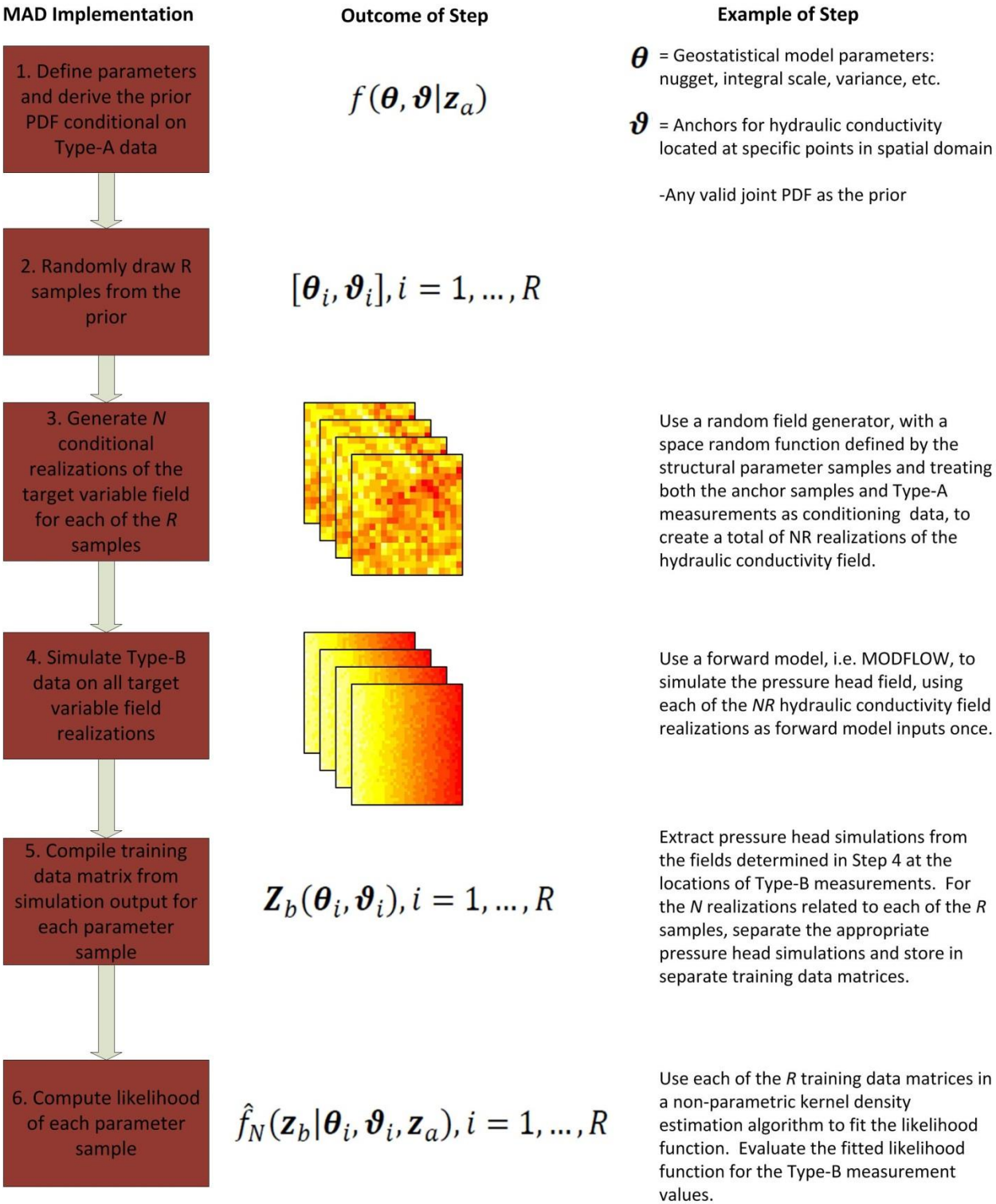

Figure 1.1: Flowchart and example of the numerical evaluation of Equation 1.2. Originally published in *Over et al.* [2013].

The joint prior PDF $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$ can be derived a number of ways, but good statistical practice recommends application of a least subjective principle, to avoid the introduction of bias. The principle of maximum entropy [*Woodbury & Ulrych,* 1993, 1998] or minimum relative entropy [*Woodbury & Ulrych,* 1993, 1998; *Hou & Rubin,* 2005] are two common and well-documented approaches in hydrogeology that can be utilized for definition of $f(\boldsymbol{\theta})$. Additionally, more simplistic approaches, such as non-informative, improper, or a Jeffrey's prior could be invoked for $f(\boldsymbol{\theta})$ [*Wasserman,* 2010]. Analytical expressions of $f(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \boldsymbol{z}_a)$ and $f(\boldsymbol{z}_a|\boldsymbol{\theta})$ can be determined via Gaussian conditioning [*Rubin et al.,* 2010; *Murakami et al.,* 2011; *Chen et al.,* 2012], but of course other approaches may be adopted.

The second step of the MAD implementation is to draw R times from $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$. This sampling can be done in a variety of ways, which are discussed in greater detail in Section 2.1.2. These draws are stored in the matrix $\boldsymbol{D}$, where each row $\boldsymbol{D}_i = [\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i], i = 1, \dots, R$ is a possible structural parameter and anchor configuration – hereafter referred to as 'samples'. The matrix $\boldsymbol{D}$ is of dimension $R \times P$, where P is the total number of anchors plus random structural parameters.

The third step of the MAD implementation is to utilize the structural model and a random field generator to generate N equiprobable conditional realizations – hereafter just 'realizations' - of the parameter field (or transformation thereof) $\widetilde{\boldsymbol{Y}}^{(1)}(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i), \dots, \widetilde{\boldsymbol{Y}}^{(N)}(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ for each of the R samples. The result is R ensembles of size N and a total number of realizations of NR. Note that the realizations are conditional on the values of $\boldsymbol{\theta}, \boldsymbol{\vartheta}$, and $\boldsymbol{z}_a$, but the dependence on $\boldsymbol{z}_a$ is not explicitly defined, because it is the same for all R ensembles, whereas the $(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ are different for each ensemble.

The fourth step of the MAD implementation is to utilize the forward model to simulate the Type-B response in space and time – hereafter simply a 'simulation' - on each of the realizations for all the samples. Obviously, the numerical forward model must simulate physics that are comparable to the conditions under which the Type-B measurements were attained.

The fifth step of the MAD implementation is to extract the output that corresponds with the M Type-B measurement locations and times to be conditioned upon from the temporal and spatial Type-B simulations. Store the extracted simulation data in the matrix $\boldsymbol{Z}_b(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$, where the k^{th} row corresponds with the simulated Type-B response on the k^{th} realization $\widetilde{\boldsymbol{Y}}^{(k)}(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ of the parameter field and the j^{th} column corresponds with the j^{th} component of $\boldsymbol{z}_b$. Store the simulation ensemble matrix $\boldsymbol{Z}_b(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ for each of the R samples. This step requires running the forward model with each of the NR realizations and results in R simulation ensemble matrices of size $N \times M$, where M is the length of $\boldsymbol{z}_b$.

The sixth step of the MAD implementation is to non-parametrically or parametrically estimate the likelihood function using the simulation ensemble matrices for each of the samples. In a non-parametric approach, the simulation ensemble matrices are used to fit the M-dimensional likelihood function $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a), i = 1, \dots, R$. In a parametric approach, the simulation ensemble is used to estimate the statistical parameters of an M-dimensional parametric PDF. For instance, the Gaussian distribution is a possible parametric distribution to employ, in which case the simulation ensemble matrices would

be used to estimate the mean vector and covariance matrix. After employing either approach, simply evaluate the likelihood function for the measurements of the Type-B data obtained from the experimental site.

The last step (not shown in Figure 1.1) is to take the product of the prior joint PDF and the likelihood function values for each of the R samples to obtain the posterior joint PDF values $f(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i | \boldsymbol{z}_a, \boldsymbol{z}_b), i = 1, \dots, R$. From these samples, the complete posterior joint PDF $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta} | \boldsymbol{z}_a, \boldsymbol{z}_b)$ can be regressed using non-parametric methods.

Summarizing, the general formula for a MAD implementation can be condensed to a likelihood function calculation Monte Carlo (MC) procedure nested within an outer MC procedure for regressing the posterior joint PDF [*Maxwell et al.,* 1999].

Before closing this section, it is worthwhile to mention two aspects of implementing MAD that are essential to consider in any application. First, the likelihood function calculation must converge; otherwise, the values obtained are not reliable. This calculation is dimensionally dependent - N should grow as M grows. Second, the posterior joint PDF regression must also converge; otherwise the probability values regressed are not reliable. This calculation is also dimensionally dependent – R should grow as P grows. Since the largest component of computational cost component of implementing MAD is the NR forward simulations, N and R are both significant to the overall cost.

The following three chapters expand upon the material presented in Sections 1.1-1.3. The second chapter motivates and derives an approximation method – called bundling - for the inference of the likelihood function, which reduces the computational cost of the numerical implementation of MAD. The third chapter introduces the MAD software GUI; summarizes how the software supports generalized application of the theory; presents the modular architecture of the software; and a series of brief case studies. The fourth chapter presents the first application of MAD to vadose zone parameter characterization. Each chapter features at least one unique case study: some investigations are synthetic (Chapters 2 & 3) and one utilizes field data from an experimental site (Chapter 4).

2. Bundling

The computational cost of model inversion is often very high, and this cost is a key factor preventing such analysis from becoming the practicing standard, or even becoming more common [*Carrera et al.*, 2005; *Alcolea et al.*, 2006; *Marzouk et al.*, 2007; *Castagna & Bellin,* 2009; *Rubin et al.*, 2010]. The majority of prohibitive computational costs in MAD are affiliated with the large number of forward model (FM) simulations required for likelihood function inference or the number of adjoint model simulations needed to evaluate the sensitivity of the measurements to the model parameters [*Cirpka & Kitanidis,* 2000b]. For example, large FM simulation expense can result from the need to numerically solve the governing equations of saturated or unsaturated flow and reactive transport in three-dimensions, under transient conditions, and in finely-discretized domains. This chapter presents an approximation technique for non-parametric likelihood function inference that can reduce this FM cost without introducing significant error into the results.

There is a body of literature with a focus on reducing the computational cost of model inversion, most of which can be roughly categorized as either model reduction or intelligent sampling. Briefly, model reduction is the replacement of a more expensive numerical FM with a cheaper analog; this category includes techniques like the response surface method (RSM) [*Downing et al.*, 1985], high dimensional model representation (HDMR) [*Rabitz et al.,* 1998], the stochastic response surface method (SRSM) [*Isukapalli* et al., 1998], and the deterministic response surface (DSR) [*Loll & Moldrup,* 1998], or the temporal reduction of transient models to steady state via temporal moments (TM) [*Harvey & Gorelick*, 1995; *Cirpka & Kitanidis*, 2000b; *Leube et al.,* 2012]. The approach presented below is not a model reduction technique, but could be utilized in a complimentary manner, which is detailed after the method is presented. Briefly, intelligent sampling optimizes – relative to random sampling - the search of the model parameter space during the calibration or inversion process; this category includes techniques like the Fourier amplitude sensitivity test (FAST) [*Cukier et al.*, 1978] and Latin Hypercube sampling (LHS) [*McKay et al.*, 1979]. The approach presented in this chapter is also not an intelligent sampling approach, which will also be justified in greater detail after the method is presented.

The technique is a modification of the likelihood function inference procedure in MAD. The approximation strategy developed here infers the likelihood of *sets* of model parameter vectors reproducing the field measurements, which is a departure from the "traditional" inversion approach described in Section 1.3.2 where each model parameter vector is analyzed independently. In this chapter, the necessary modifications to the Bayesian proportionality are shown explicitly when employing a likelihood function that is "bundled" or shared by multiple parameter configurations.

Note that this strategy is not developed to complement ML, MAP, GLUE, PPM, or SCE discussed in Section 1.1. Common to several of the approaches is the use of an assumed analytical expression for the likelihood (ML or MAP), objective function (PPM), or likelihood measure (GLUE), in which a *single* simulation run for a given parametrization can be used to evaluate the likelihood expression. This is fundamentally different than the approach implemented in MAD, where the likelihood function is instead inferred, requiring *multiple* simulation runs for a given model parametrization to first define the likelihood function before it can be evaluated. Moreover, because MAD is not an optimization approach, so this approximation is not developed for SCE or its variants.

The primary motivation for developing this new approximation is simple computational economy: a practical and cheap implementation strategy can first "scan" all possible model parameter configurations on a coarse basis (on the level of sets or volumes) to identify regions that reproduce field measurements with high likelihood, and then one could re-analyze on a fine basis (point by point) only previously-identified "hot" regions. The bundling approach represents this scanning strategy, and does not prohibit the secondary analysis.

Another attractive feature of the bundling strategy is that it can be combined with other techniques to reduce the cost of likelihood function inference, which suffers from the well-known curse of dimensionality. For example, the dimensionality of the likelihood function can be reduced by replacing inversion data with lower-dimensionality representations of the same data – the first few TM of a concentration breakthrough curve (BTC) at an observation well could be used, for instance, instead of the lengthy concentration time series at the same location [*Harvey & Gorelick,* 1995; *Cirpka & Kitanidis*, 2000a]. Besides the improved likelihood function dimensionality, the use of TM is further supported and encouraged by the recent work of *Leube et al.,* [2012] who showed explicitly the high information content contained in TM of drawdown curves. In this chapter – which considers 3-D flow and transport, and extends the work of *Murakami et al.*, [2010] and *Chen et al.*, [2012] – two dimension reduction techniques are employed: TM and vertical flux-averaging of concentration BTC data.

In the following sections, an overview of the necessary modifications to the MAD approach is first presented and then a detailed implementation procedure for the bundling approximation of the likelihood function, where a more detailed comparison of the bundling method with other model inversion efficiency strategies can be found. After the methodology, a synthetic, numerical case study is introduced with details of the physical process being modeled, pertinent site conditions, the parameters being analyzed, and the data used for inversion. Next, the results of the bundling approximation strategy are presented from the case study, compared side-by-side to the results from MAD without modification. This case study is used to evaluate the quality of the approximation, and whether computational savings were attained. The chapter closes with a discussion and summary.

2.1 Methodology

This section presents the bundling approximation for inferring the likelihood function in model inversion. The objective is to develop a method that is equally accurate as, but computationally cheaper than, the traditional MAD approach. In the following sections, the alterations to the MAD equations are presented, and the implementation required for this approximation technique. Next, an extended derivation shows the order of the approximation. After the alterations and the order of the error are introduced, the approach is compared with other methods for reducing computational cost of model inversion. Finally, qualitative expressions for the approximate computational cost of each approach are given.

2.1.1 Bundling

Recalling Section 1.2, here a technique called "bundling" is presented for approximating the likelihood function by altering Equation 1.2. This modification is both numerical and conceptual relative to Equation 1.2, and is based on substituting a likelihood function inferred non-parametrically using a training data matrix of Type-B data generated for a set of samples rather than from a single sample. First, possible benefits and consequences of bundling are presented; followed by the method itself, the derivation of the approximation error, and an example illustrating many of these concepts; this section closes with a brief discussion comparing the implementations of bundling, MAD, and other computationally efficient approaches to model inversion

2.1.1.1 Motivation and Equations

The potential advantage of bundling is to significantly reduce the total number of FM simulations needed for likelihood function inference, which can be the most prohibitive and dominant computational aspect of MAD. The danger – as with all approximation techniques – is the possibility of introducing significant error. With these considerations in mind, the research question is therefore how bundling can reduce computational costs without degrading the quality of the model inversion. The following study suggests that inferring the likelihood function over sets of samples – chosen by some similarity criterion – can yield cheaper computation with manageable amounts of error relative to methods that evaluate just one sample at a time.

With that conceptual motivation in place, the preliminaries and notation of bundling can be established. Define K bundles from amongst the R samples – $\boldsymbol{\Omega}_j, j = 1, .., K$ and $1 < K < R$

– each of which is a set, whose elements are the samples contained in each row of $\boldsymbol{D}$, by first defining a medoid vector for each bundle – $[\boldsymbol{\theta}, \boldsymbol{\vartheta}]_j^*,\ j = 1, \dots, K$, which are samples that are identifiable features of $\boldsymbol{D}$ in the parameter domain [*Kaufmann & Rousseeuw*, 1990] – and then grouping the remaining samples to their nearest medoid in Euclidean distance, r_{ij}. The subscript i indicates an arbitrary sample, and the subscript j an arbitrary medoid. Define n_j as the number of samples in the j^{th} bundle, including the medoid. Additionally, $\sum_{j=1}^{K} n_j = R$, because samples can only belong to one bundle. Euclidean distance is used as the measure of similarity, because it requires the samples, which are vectors, to be alike in both direction and magnitude [*van der Laan & Pollard*, 2003].

Bundling the likelihood function is based on an "or" statement as follows, presented without loss of generality, for the j^{th} bundle with $n_j = 2, \boldsymbol{\Omega}_j = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$, where for brevity the shorthand $\boldsymbol{\phi}_i = [\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i], i = 1{,}2$ is introduced:

$$f\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) = f(\boldsymbol{z}_b | (\boldsymbol{\phi}_1 \cup \boldsymbol{\phi}_2) \cap \boldsymbol{z}_a)$$

$$f\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) = f(\boldsymbol{z}_b | \boldsymbol{\phi}_1 \cap \boldsymbol{z}_a) + f(\boldsymbol{z}_b | \boldsymbol{\phi}_2 \cap \boldsymbol{z}_a) - f(\boldsymbol{z}_b | \boldsymbol{\phi}_1 \cap \boldsymbol{\phi}_2 \cap \boldsymbol{z}_a). \tag{2.1}$$

If the parameter samples are approximately equal, $\boldsymbol{\phi}_1 \cong \boldsymbol{\phi}_2$, then

$$f\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) \cong f(\boldsymbol{z}_b | \boldsymbol{\phi}_i \cap \boldsymbol{z}_a) + f(\boldsymbol{z}_b | \boldsymbol{\phi}_i \cap \boldsymbol{z}_a) - f(\boldsymbol{z}_b | \boldsymbol{\phi}_i \cap \boldsymbol{\phi}_i \cap \boldsymbol{z}_a)$$

$$f\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) \cong f(\boldsymbol{z}_b | \boldsymbol{\phi}_i \cap \boldsymbol{z}_a) = f(\boldsymbol{z}_b | \boldsymbol{\phi}_i, \boldsymbol{z}_a), \text{ for } i = 1{,}2. \tag{2.2}$$

Generalizing Equation 2.2 for bundles of any size results in an approximation of the likelihood function that is conditional on a *bundle* of - not individual - samples:

$$f(\boldsymbol{z}_b | \boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z_a}) \cong f\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right), \text{ for } \forall\ [\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i] \in \boldsymbol{\Omega}_j\ \&\ j = 1, \dots, K$$

$$\text{where } \boldsymbol{\Omega}_j = \left\{[\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i] : r_{ij} < r_{ik}\right\} \text{ for } i = 1, \dots, R\ \&\ j = 1, \dots, K$$

$$\text{and } r_{ij} = \left\| [\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i] - [\boldsymbol{\theta}, \boldsymbol{\vartheta}]_{\boldsymbol{j}}^* \right\|. \tag{2.3}$$

By Equation 2.3, the *j*th bundle is defined as the set of all samples closer to the *j*th medoid than all other $k \neq j$ medoids. Equation 2.3 holds when $[\boldsymbol{\theta}_m, \boldsymbol{\vartheta}_m] \cong [\boldsymbol{\theta}_n, \boldsymbol{\vartheta}_n]$, for all $m \neq n$, for all samples in a bundle, for all bundles. This final condition is the motivation for using Euclidean distance as the metric for defining bundles. Bundling could easily accommodate any other distance metric by modifying Equation 2.3. When using the Euclidean distance metric, care should be taken to work with structural parameters and anchors that are scaled similarly; otherwise variation in a dominant dimension (i.e. spanning orders of magnitude) of the parameter domain could mask and exclude variations in other parameter dimensions (i.e. within an order of magnitude) when bundles are defined.

Numerical estimation of MAD, after inserting Equation 2.3 into Equation 1.2, is then this:

$$f(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i | \boldsymbol{z}_a, \boldsymbol{z}_b) \cong \frac{\left(\hat{f}_{B_j}\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) + \mathcal{O}(\delta_j)\right) f(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i | \boldsymbol{z}_a)}{f(\boldsymbol{z}_b | \boldsymbol{z}_a)} \propto \left(\hat{f}_{B_j}\left(\boldsymbol{z}_b \middle| \boldsymbol{\Omega}_j, \boldsymbol{z}_a\right) + \mathcal{O}\left(\delta_j\right)\right) f(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i | \boldsymbol{z}_a),$$

$$\text{for } \forall\, [\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i] \in \boldsymbol{\Omega}_j \;\&\; j = 1, \dots, K, \tag{2.4}$$

where δ_j is the largest extent of the bundle, defined as the maximum Euclidean distance across the j^{th} bundle, and N^+ is the number of simulations per sample when bundling is employed. In Equation 2.4, the likelihood function is still M-dimensional as in Equation 1.2; however, it is now fitted using $B_j = n_j N^+$ simulations of the Type-B data in the training data matrix, rather than N simulations. After a discussion comparing Equations 1.2 & 2.4, the order of the error given in Equation 2.4 is derived for the following conditions: FMs that are approximately linear and using a likelihood function joint PDF that is numerically inferred using non-parametric Gaussian kernel density estimation methods with a fixed bandwidth [*Scott & Sain,* 2004].

An important similarity in the implementations of Equations 1.2 and 2.4 is that the posterior and prior joint PDFs are still sampled and evaluated R times – in fact, the majority of the implementation procedures are identical, and the outline given in Section 1.3.2 and Figure 1.1 can quite easily be extended to numerically evaluating Equation 2.4. However, there are two critical differences in implementation between Equations 1.2 and 2.4, which are summarized below and illustrated in Figure 2.1. Figure 2.1 is adapted from Figure 1.1 and shows the additional steps required when applying bundling to a MAD analysis. The red or blue horizontal arrows respectively indicate the output of MAD or bundling at each step.

The first important difference is that the likelihood function is estimated only K times in Equation 2.4, rather than R times in Equation 1.2. It is a reasonable expectation that inferring the likelihood function, whether for a single sample or a single bundle, should require roughly the same number of FM simulations to establish convergence, because the dimensionality of the likelihood function in Equation 1.2 is not changed in Equation 2.4. Therefore, because qualitatively

$$\text{net simulation cost} = (\text{simulation cost per calculation}) * (\#\text{ of calulcations }),$$

bundling should be less expensive than MAD simply because there are fewer calculations to perform – a primary motivation for developing the approximation in the first place.

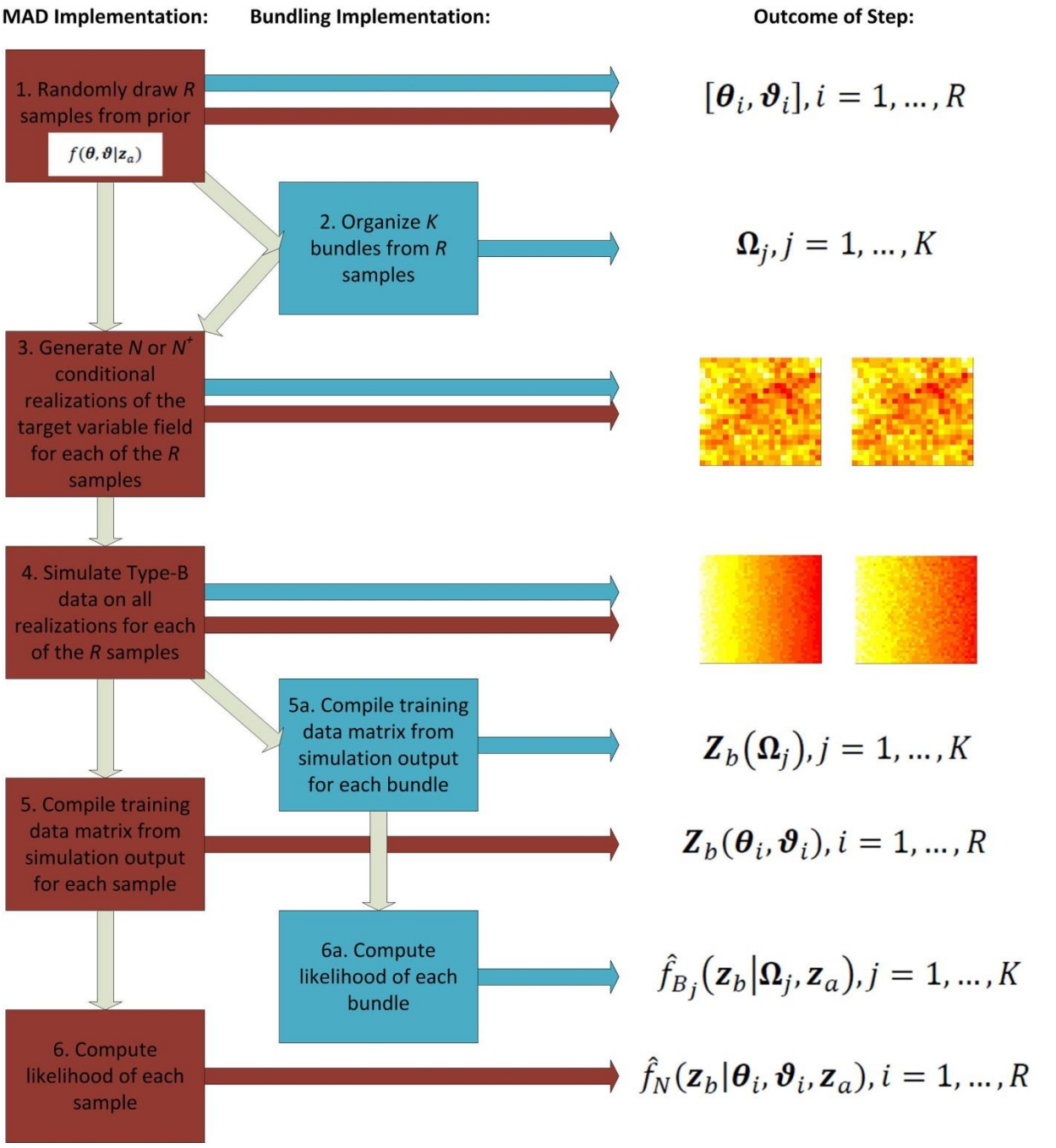

Figure 2.1: Flow chart detailing the similarities and differences between the numerical implementations of MAD and bundling. Originally published in *Over et al.* [2013].

The other major change in the numerical evaluations of Equations 1.2 and 2.4 is their approach to compiling the training data matrices for likelihood function inference, which cannot be observed from the equations directly. When bundling is applied, the B_j Type-B simulations in the training data matrix are uniformly contributed from each of the n_j samples contained in the bundle. Thus, define $\boldsymbol{Z}_b(\boldsymbol{\Omega}_j)$, a matrix of dimension $B_j \times M$, with a dependence not on only one sample, as was the case in Equation 1.2 where

$\boldsymbol{Z}_b(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$, but instead on a bundle of samples. $\boldsymbol{Z}_b(\boldsymbol{\Omega}_j)$ is the Type-B simulation output using N^+ realizations of the target variable field from each of the samples in the bundle, $[\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i] \in \boldsymbol{\Omega}_j$. Put differently, matrix $\boldsymbol{Z}_b(\boldsymbol{\Omega}_j)$ is the aggregation of n_j $\boldsymbol{Z}_b(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ matrices, of size $N^+ \times M$, one for each sample in the bundle. In this study, constant N^+ is used, such that the size of the training data matrix $B_j \times M$ varies linearly with n_j, because this permits direct comparison of N^+ with N. Further computational efficiency may be possible by letting N^+ vary as a function of n_j, but this topic is not further explored in this chapter.

Finally, note that no specific algorithm was declared for defining medoids in Equation 2.3 – medoids can be defined by a user or with a variety of clustering algorithms, c.f. *van der Laan & Pollard* [2003] – nor has a rule or approach yet been specified for establishing how many bundles to use, i.e. the value of K. In Section 2.3.2, a rule for defining the number of bundles is established. In the next subsection, the error term in Equation 2.4 is derived, which is also revisited in Section 2.3.2 with supporting numerical evidence from a case study.

2.1.1.2 Order of the Error

The non-parametrical form of likelihood for M-dimensional $\boldsymbol{z}_b$ conditional on the i^{th} sample - using a Gaussian kernel with fixed bandwidth, h, and a training data matrix $\boldsymbol{Z}_b(\boldsymbol{\theta}_i)$ of dimension $N \times M$ - is

$$\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i) = \frac{1}{Nh^M}\sum_{j=1}^{N}\frac{1}{(2\pi)^{M/2}} exp\left[-\frac{[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_i)_j]^{\mathrm{T}} * [\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_i)_j]}{2h^2}\right], \tag{2.5}$$

where $\boldsymbol{Z}_b(\boldsymbol{\theta}_i)_j$ is the j^{th} row of the training data matrix, $[\quad]^{\mathrm{T}}$ is the transpose of a vector, and standard matrix multiplication rules apply. Note, the derivation in this section is shown for structural parameters only, because it shortens the somewhat lengthy notation, but generally the equations could be expanded to include dependence on anchors as well. All other assumptions are discussed as they are utilized, with a specific attention to assumptions that could be generalized to any application of bundling.

The non-parametrically estimated likelihood for M-dimensional $\boldsymbol{z}_b$ conditional on the k^{th} bundle, without loss of generality for $n_k = 2$ and $\boldsymbol{\Omega}_k = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ is

$$\hat{f}_N(\boldsymbol{z}_b|\,\boldsymbol{\Omega}_k) = \frac{1}{Nh^M}\sum_{i=1}^{n_k}\sum_{j=1}^{N/2}\frac{1}{(2\pi)^{M/2}} exp\left[-\frac{\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_i)_j\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_i)_j\right]}{2h^2}\right]$$

$$= \frac{1}{2}\left\{\hat{f}_{N/2}(\boldsymbol{z}_b|\boldsymbol{\theta}_1) + \hat{f}_{N/2}(\boldsymbol{z}_b|\boldsymbol{\theta}_2)\right\} \tag{2.6}$$

Equation 2.6 also uses a Gaussian kernel with the same fixed bandwidth as Equation 2.5 and a training data matrix $\boldsymbol{Z}_b(\boldsymbol{\Omega}_k)$ which is an aggregation of $\boldsymbol{Z}_b(\boldsymbol{\theta}_1)$ and $\boldsymbol{Z}_b(\boldsymbol{\theta}_2,)$ each of

dimension $N/2 \times M$, a size conveniently selected to allow comparison of the likelihood function with or without the bundling approximation. Further, in Equation 2.6, see that the bundled approximation of the likelihood is equivalent to the arithmetic average of the likelihood of each sample in the bundle. N is taken to be even for simplicity, such that the indexing on the interior summation is valid.

Assuming the simulated Type-B data is multiply differentiable with respect to the parameters, the following Taylor Series approximation is valid, for the j^{th} simulation,

$$\boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j \cong \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j + \sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P} \left(\theta_{1,p} - \theta_{2,p}\right), \tag{2.7}$$

where $\theta_{1,p} - \theta_{2,p}$ is the difference in the p^{th} components of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the dimension of the parameter vector is P, the partial differentiation is carried out for each of the M-dimensions of $\boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j$ such that the summation term is also a vector, and second order or higher terms are neglected.

Expanding Equation 2.6 using Equation 2.7 and defining $\theta_{1,p} - \theta_{2,p} = \delta\theta_p$,

$$\hat{f}_N(\boldsymbol{z}_b|\,\boldsymbol{\Omega}_k) \cong \frac{1}{Nh^M}\left\{\sum_{j=1}^{N/2} \frac{1}{(2\pi)^{M/2}} exp\left[-\frac{\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]}{2h^2}\right] + \right.$$

$$\left. \sum_{j=1}^{N/2} \frac{1}{(2\pi)^{M/2}} exp\left[-\frac{\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j + \sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j + \sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]}{2h^2}\right]\right\}. \tag{2.8}$$

Focusing on the argument of the exponential function that utilized the Taylor Series of Equation 2.7, find that

$$\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j + \sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j + \sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]$$

$$\cong \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right] + 2\left[\sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right] \tag{2.9}$$

where the $\mathcal{O}\left(\delta\theta_p^2\right)$ term has been neglected, because such terms were left out of the expansion in Equation 2.7. Substitution of Equation 2.9 into Equation 2.8 and expanding the exponential function yields

$$\hat{f}_N(\boldsymbol{z}_b|\,\boldsymbol{\Omega}_k) \cong \frac{1}{Nh^M}\left\{\sum_{j=1}^{N/2} \frac{1}{(2\pi)^{M/2}} exp\left[-\frac{\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]}{2h^2}\right] + \right.$$

$$\left. \sum_{j=1}^{N/2} \frac{1}{(2\pi)^{M/2}} \underbrace{\left[1 - \frac{\left[\sum_{p=1}^{P} \frac{\partial \boldsymbol{Z}_b(\boldsymbol{\theta}_2)_j}{\partial \theta_P}\delta\theta_p\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]}{h^2}\right]}_{A} exp\left[-\frac{\left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]^{\mathrm{T}} * \left[\boldsymbol{z}_b - \boldsymbol{Z}_b(\boldsymbol{\theta}_1)_j\right]}{2h^2}\right]\right\}. \tag{2.10}$$

Note the additional factor (term '*A*') on the second summation, relative to the first summation, this is the error introduced by bundling. In general as n_k increases there will be more summations and a greater percentage of them will have affiliated error factors because of bundling. Taking the maximum value of the partial derivatives for all j to be small compared to the differences in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ [Chapter 11, *Rubin*, 2003], find an upper bound on the order of the error and restate Equation 2.10 as

$$\hat{f}_N(\boldsymbol{z}_b|\,\boldsymbol{\Omega}_k) \cong \frac{1}{2}\left\{\hat{f}_{N/2}(\boldsymbol{z}_b|\boldsymbol{\theta}_1) + \left[1 - \mathcal{O}\left(max\{\delta\theta_p\}\right)\right]\hat{f}_{N/2}(\boldsymbol{z}_b|\boldsymbol{\theta}_1)\right\}. \tag{2.11}$$

Finally, introducing $max\{\delta\theta_p\} = \delta_k$, the largest difference across the bundle (called 'bundle diameter' hereafter), and generalizing to any n_k and any parameter vector in $\boldsymbol{\Omega}_k$, gives

$$\hat{f}_N(\boldsymbol{z}_b|\,\boldsymbol{\Omega}_k) \cong \frac{1}{n_k}\left\{\hat{f}_{N/n_k}(\boldsymbol{z}_b|\boldsymbol{\theta}_i) + \sum_{n_k-1}[1 - \mathcal{O}(\delta_k)]\hat{f}_{N/n_k}(\boldsymbol{z}_b|\boldsymbol{\theta}_i)\right\} \cong$$

$$[1 - \mathcal{O}(\delta_k)]\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i), \forall\boldsymbol{\theta}_i \in \boldsymbol{\Omega}_k. \tag{2.12}$$

Thus, the order of the error term in Equation 2.4 is determined as a measure of the bundle diameter using Taylor series expansions. The next section offers a graphical discussion of the bundling approximation.

2.1.1.3 A Brief Example

The simplified two-dimensional parameter domain depicted in Figure 2.2 exemplifies many of the variables defined in the bundling approximation presented in the previous subsection.

The plot shows eight samples partitioned into two bundles (distinguished by color and enclosed in a dotted border), the medoids of both bundles (indicated by solid fill), and an example of the Euclidean distances between a sample and the two medoids (the dashed lines). The borders around the bundles do *not* represent an enclosed volume, but are simply a visual tool for easily identifying the bundles in the graphic. Figure 2.2 also facilitates an explanation of how to compile the training data matrices: $\boldsymbol{Z}_b(\boldsymbol{\Omega}_1)$ would be composed of $5N^+$ Type-B simulations on N^+ realizations generated from each of the five samples in $\boldsymbol{\Omega}_1$, and $\boldsymbol{Z}_b(\boldsymbol{\Omega}_2)$ would be composed of $3N^+$ Type-B simulations on N^+ realizations generated from each of the three samples in $\boldsymbol{\Omega}_2$.

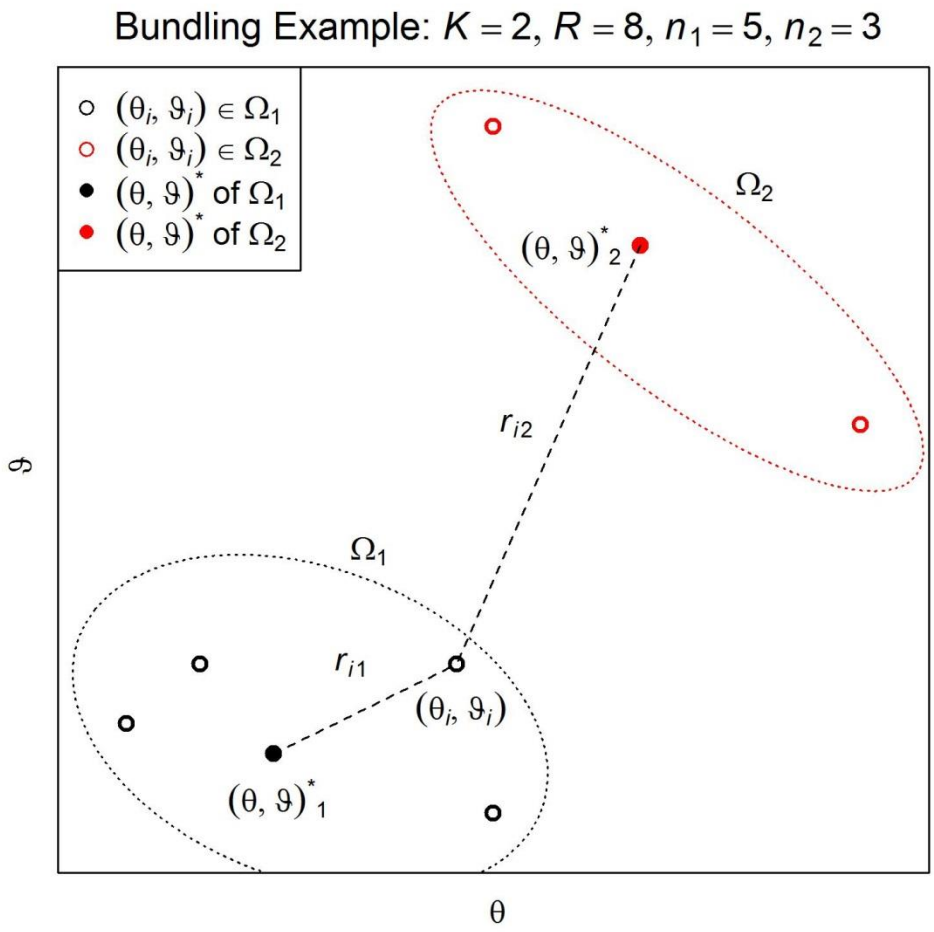

Figure 2.3: Schematic diagram of bundles, medoids, and the Euclidean distance similarity metric. Originally published in *Over et al.* [2013].

2.1.2 Contrasting Bundling with Other Computationally Efficient Inversion Techniques

In this section, justification is presented for the claim made in the opening of the chapter that bundling is not a model reduction or sampling technique. The discussion below relies heavily on Figure 2.3, which schematically shows how the various approaches can be used to infer the likelihood function in MAD. *Rao* [2005] or *Balukrishnan et al.,* [2005] present a much more detailed summary of many of the alternative computational approaches discussed in this section.

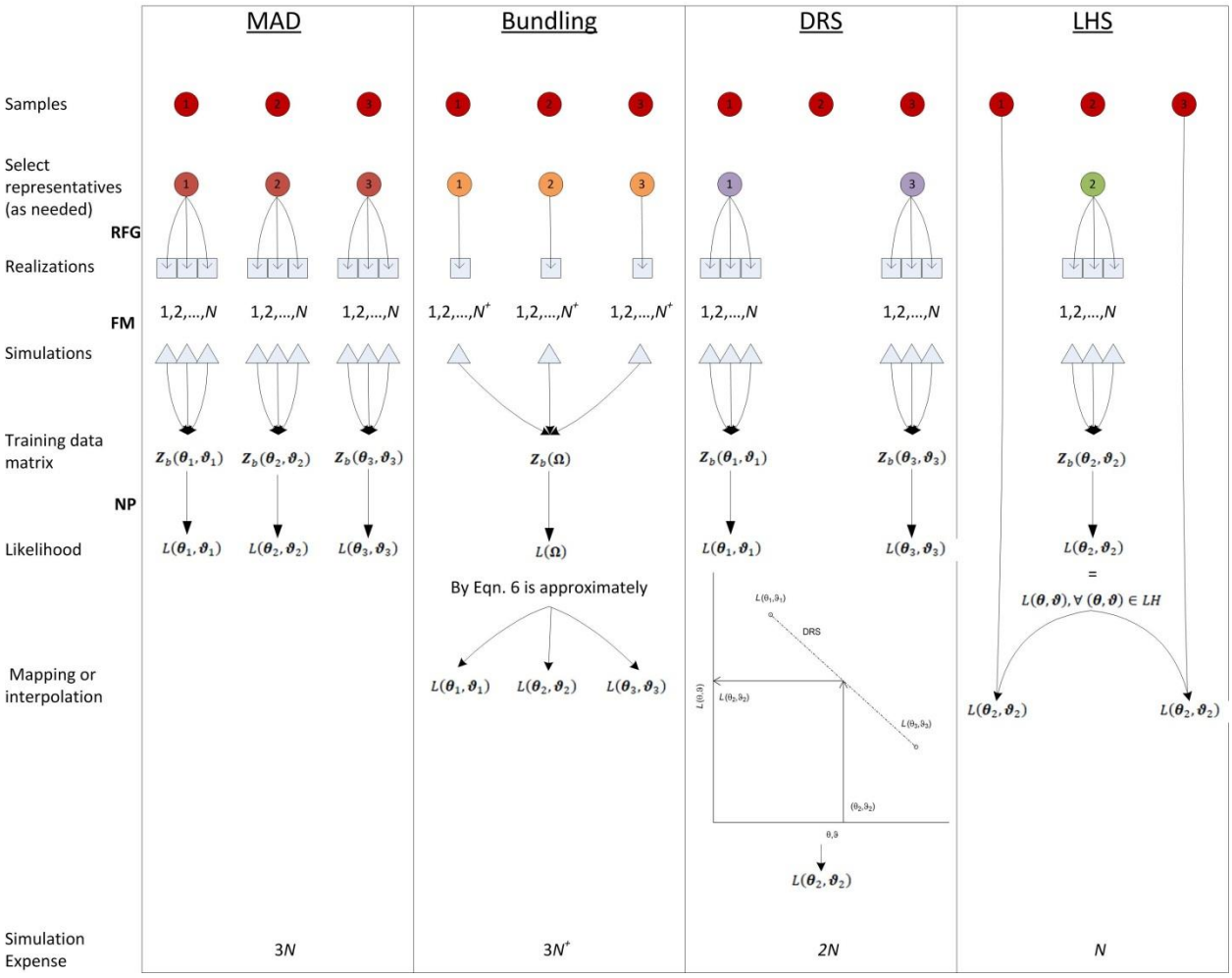

Figure 2.3: Schematic representation of implementations procedure between drawing samples and inference of the likelihood value of a sample using MAD (leftmost), bundling (left-center), DRS (right-center), and LHS (rightmost) approaches. Originally published in *Over et al.* [2013].

In Figure 2.3, four alternative implementations for inferring the likelihood are presented: the "traditional" MAD approach (repeated for reference, leftmost), bundling (left-center), DRS (right-center), and LHS (rightmost). The diagram shows how three samples would be used to generate the training data matrices $\boldsymbol{Z}_b$ for non-parametric inference of the likelihood function and how the likelihood value is assigned to the various samples. Note in this figure, all three samples are assumed to belong to either the same bundle or would fall inside a single Latin Hypercube (LH). The figure introduces the shorthand $L(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i) = f(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$ and $L(\boldsymbol{\Omega}) = f(\boldsymbol{z}_b|\boldsymbol{\Omega}, \boldsymbol{z}_a)$ as well as the acronym "NP" for non-parametric probability density inference. Circles, squares, and triangles respectively indicate samples, realizations, and simulations – the quantities of these symbols depicted in the figure are intended to be generic, i.e. N^+ may or may not actually equal $N/3$, but for bundling to reduce computational cost N^+ must be less than N and is shown as such. An abbreviated reminder of the implementation is provided along the left margin of the Figure (see Figures 1.1 or 2.1 and Sections 1.3.2 or 2.1.1.1 for more detail).

The schematic indicates the conceptual differences between bundling and the other approaches. First, bundling is the only approach that combines simulations from different samples in the training data matrix. Second, compared to LHS or DRS, where a single representative sample's likelihood value is assigned to the entire Latin Hypercube [*McKay et al.*, 1979] or likelihood values are evaluated from the DRS linearly interpolated over the likelihood values of a stratified subset of samples [*Loll and Moldrup*, 1998], bundling utilizes *all* the samples in determining the approximate likelihood value. Finally, with bundling, none of the individual likelihood values of the samples in a bundle $L(\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ are ever directly inferred, but rather it is the likelihood of the bundle which is inferred $L(\boldsymbol{\Omega})$, which is different than "traditional" MAD, LHS, and DRS, where at least one sample's individual likelihood value is always directly inferred. Notice the simulation expense (bottom row), which depending on the relationship between N and N^+ could be minimized for bundling or LHS; experimental results of these two quantities are compared in Section 2.3.1.

Even in the case where bundling is slightly more costly than LHS, bundling may have a conceptual appeal over LHS, because bundling accounts for more of the effects of parameter variation within a region of parameter space (the set defined by a bundle or enclosed in a LH) via the Type-B training data matrix used for inference. Bundling is an arithmetic averaging of the contributions from the various samples within the bundle to the likelihood value (Equation 2.6), whereas the LHS likelihood value is "blind" to all other samples that are within the LH except the representative sample.

Because bundling is a strict average of the likelihood in a bundle under the conditions of Section 2.1.1.2, with no weighting based on location in the parameter space, and because the "volume enclosed" by the set of samples contained in a bundle is at best arbitrarily defined (e.g. hypersphere, hyperellipsoid, etc.) any comparison to the expected value of a linear interpolated DRS is rather ambiguous. Notwithstanding, it is worth defining the condition under which the local (over a bundle) expected value of the linearly interpolated DRS would be equivalent to the arithmetic average of the bundling approximation – shown generically for a bundle $\boldsymbol{\Omega}$ with medoid $[\boldsymbol{\theta}, \boldsymbol{\vartheta}]^*$ containing n samples and spanning a maximum Euclidean distance δ -

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a) = \frac{\int_{\mathcal{V}_\Omega} f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{z}_a)\partial\mathcal{V}_\Omega}{\int_{\mathcal{V}_\Omega} \partial\mathcal{V}_\Omega}, \text{ where } \mathcal{V}_\Omega = \left\{[\boldsymbol{\theta}, \boldsymbol{\vartheta}]: \frac{\|[\boldsymbol{\theta}, \boldsymbol{\vartheta}] - [\boldsymbol{\theta}, \boldsymbol{\vartheta}]^*\|}{\delta/2} \leq 1\right\}. \quad (2.13)$$

In Equation 2.13, the volume enclosed by the bundle $\mathcal{V}_\Omega$ is arbitrarily defined as a hypersphere of radius $\delta/2$ centered at the medoid (since there is no one shape associated with a bundle) and there are no subscripts on the structural parameters or anchors on the integrand of the volume integral to remind that this argument is the linearly interpolated likelihood DRS.

Where bundling may have a conceptual disadvantages relative to LHS and DRS is that the bundles are not required to span the entire parameter space, which means LHS or DRS will provide better coverage of the parameter space in some cases, and the relative complexity

of implementation, in which bundling is arguably the least straight-forward of all the approaches. Note that even though Figure 2.3 may seem to suggest when all bundles are singly populated that LHS and bundling would be equivalent, but this is not the case. Instead, when bundles are all singly populated, bundling is equivalent to random sampling; see Section 1.2.

The model reduction techniques HDMR, RSM, SRSM, and TM are not formally included in Figure 2.3. This is because these approaches involve replacing the FM entirely, which would replace the numerical tool used to generate Type-B data on the realizations, which is not a requirement of bundling. Hypothetically, bundling could be applied in tandem with these other model reduction techniques, provided any modifications to the parameter space (e.g. a partition based on importance rank) are made before bundles are determined.

With respect to the intelligent sampling techniques, it is now also clear from the figure and the preceding discussion that bundling focuses mostly on how to populate the training data matrix for likelihood function inference and this is implemented consecutively *after* samples are drawn whether by FAST, LHS, or just randomly.

2.1.3 Computational Cost

Here, qualitative equations for the approximate net computational cost of both the MAD and bundling approaches to estimating the likelihood function are defined.

Define the cost of a single simulation as $\$_{sim}$, the cost of ensuring adequate convergence of the inferred likelihood PDFs using a sample size of X as $\$_{con}(X)$, and the joint cost of determining the number of bundles and construction of $\boldsymbol{\Omega}_j, j = 1, \dots, K$, as $\$_K$.

Define the cost of evaluating MAD as $\$_{MAD}$, the time logged numerically computing Equation 1.2, by

$$\$_{MAD} \cong R * N * \$_{sim} + \$_{con}(N). \tag{2.14}$$

Define the cost of evaluating bundling as $\$_{BUN}$, the time logged numerically computing Equation 2.4, by

$$\$_{BUN} \cong R * N^{+} * \$_{sim} + \$_{con}(N^{+}) + \$_K. \tag{2.15}$$

In Section 2.3.3, Equations 2.14 and 2.15 are evaluated for the case study in order to compare and assess computational savings (if any) from using bundling relative to MAD for the example inversion.

2.2 Case Study

In this section, the analysis of a synthetic case study developed by *Murakami et al.* [2010] and extended by *Chen et al.* [2012] is presented. In separate subsections, several relevant details are summarized: first the specifics of the numerical implementation, background about the motivating site, and summary of the experiment; then the Type-A and Type-B data used in the inversion; and finally the parameter domain, parameter sampling technique, and assumed prior joint PDF employed.

2.2.1 Synthetic Experiment Summary

The synthetic experiment utilized in this chapter is an extension of the research by *Murakami et al.*[2010] and *Chen et al*. [2012]. In this section, the domain and its numerical discretization, some characteristics of the subsurface at the site, the boundary conditions, and the physical experiment being modeled in the study are summarized.

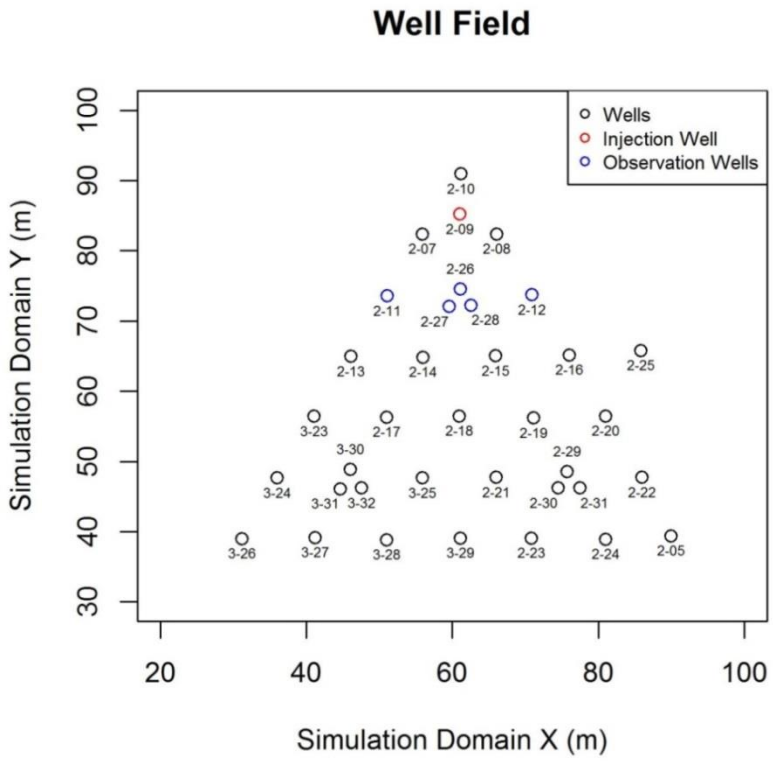

Figure 2.4: The well field of the synthetic case study. The extents of the numerical grid are outside the plot region. The injection well for the transport experiment is highlighted in red, and the observation wells in blue. Originally published in *Over et al.* [2013].

The domain in consideration is an analog of the Hanford 300 Area Integrated Field Research Challenge (IFRC) site located in Richland, Washington, USA, a highly-instrumented site [*Murakami et al.*, 2010*; Chen et al.*, 2012*;* www.ifchanford.pnnl.gov]. Figure 2.4 shows the well field at the site, and the local coordinate system. The identification system for each well (#-##) in Figure 2.4 is used as shorthand to express location of the wells in the discussion that follows. The dimension of the flow domain represented by the numerical grid is 122m by 122m by 10m, and is discretized into 2m by 2m by 0.5m blocks [*Chen et al.*, 2012].

The lithology at the IFRC site consists of poorly-sorted, unconsolidated, and highly-permeable sediments of the Hanford formation [*Bjornstad et al.*, 2009]. The experiment focuses on the saturated portion of the Hanford formation (which is roughly 5-8 meters in depth) between the water table and the underlying lower-permeability Ringold formation [*Murakami et al.*, 2010; *Chen et al.*, 2012].

The water table at the site is highly dynamic, with fluctuations averaging 0.5 meters daily because of the proximity of the Columbia River – which is roughly 250 meters from a boundary of the domain. Therefore, transient head boundary conditions are used in the forward model. The heads along the boundaries have been obtained by interpolating heads measured at observation wells. The bottom and top boundary planes are modeled as no-flux boundaries – because of the sharp decrease in hydraulic conductivity across the

Hanford-Ringold interface for the bottom, and to be similar to recharge conditions during an actual test at the site for the top [*Chen et al.*, 2012].

The physical experiment being modeled in the unconfined aquifer and subject to the boundary conditions is a natural gradient, non-reactive transport test. For the synthetic study, the tracer was injected into well 2-09, and subsequent transport was simulated for 250 hours using a maximum time step of 1 hour. *Chen et al.* [2012] provide additional information about the simulations, including the concentration boundary conditions, all the initial conditions, information about PFloTran [*Hammond and Lichtner*, 2010] - the FM - and so on.

2.2.2 Inversion Data: Type-A and Type-B

Here, a summary of the Type-A and Type-B data used by *Murakami et al.* [2010] the Type-B data used by *Chen et al.* [2012] and the Type-B data in the analysis is presented.

Murakami et al. [2010] utilized 283 electromagnetic borehole flowmeter (EBF) measurements at the Hanford site, and seven pumping tests. The EBF measurements were treated as Type-A data with uncertainty, while the pumping test measurements were treated as Type-B. Specifics of the pumping tests and the measurement locations for each of the seven constant-rate injection tests are detailed in *Murakami et al.* [2010], which determined posterior distributions of structural parameters and anchors describing the hydraulic conductivity in the Hanford 300 Area conditional on EBF and pumping test data.

Chen et al. [2012] explored the assimilation of measurements from the non-reactive tracer transport experiment as Type-B data. The Type-B data inverted are the first-order TM normalized by the zero order TM at multiple wells in the domain. *Chen et al.* [2012] determined the posterior joint PDF of the structural parameters and anchors describing the hydraulic conductivity in the Hanford 300 Area conditional on this tracer test data as well as the same EBF data and pumping test data used by *Murakami et al.*[2010].

In this numerical study, the inverted Type-B data is similar to that of *Chen et al*. [2012], and is collected from the concentration BTCs at the five wells highlighted in blue in Figure 2.4. In order to work with one-dimensional Type-B data at each well, the concentration data is condensed in time via TM [*Cirpka & Kitanidis*, 2000], and along the vertical profile via flux-averaging [*Rubin*, 2003] at each well. This work differs from that of *Chen et al.* [2012] by utilizing vertical flux-averaging and different well locations.

For the case study, a trapezoidal numerical integration scheme was employed. The one-dimensional Type-B data at each of the selected wells is as follows:

$$z_b(\boldsymbol{x}) = \frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})} = \frac{\int_0^{T_f} t\bar{C}(\boldsymbol{x},t)dt}{\int_0^{T_f} \bar{C}(\boldsymbol{x},t)dt}, \tag{2.16}$$

where $\boldsymbol{x} = [x, y]$ is the planar coordinate vector, $m_i(\boldsymbol{x})$ is the i^{th} order TM of the vertically flux-averaged concentration at $\boldsymbol{x}$, $\bar{C}(\boldsymbol{x}, t)$ is the vertically flux-averaged concentration as a function of space and time, t is adjusted time that is set to 0 with the first observation, and T_f is the final observation time. Thus, the full Type-B row vector for the likelihood analysis, with the help of Equation 2.16 and Figure 2.4, can be written as this:

$$\boldsymbol{z}_b = [z_b(\boldsymbol{x}_{2-11}),\ z_b(\boldsymbol{x}_{2-12}),\ z_b(\boldsymbol{x}_{2-26}),\ z_b(\boldsymbol{x}_{2-27}),\ z_b(\boldsymbol{x}_{2-28})], \qquad (2.17)$$

where the subscripts on planar coordinate vectors are well identification numbers.

2.2.3 The Parameter Domain, Parameter Sampling, and Prior PDF

This chapter focuses on inversion of five structural parameters and 325 anchors – that is, $P = 330$. The structural parameter vector is given by $\boldsymbol{\theta} = \{\mu, \sigma^2, \lambda_h, \lambda_v, \nu^2\}$, where μ is the dimensionless mean, σ^2 is the dimensionless variance, λ_h the horizontal integral scale in meters, λ_v the vertical integral scale in meters, and ν^2 the dimensionless nugget of the 3-D geostatistical model for the natural log hydraulic conductivity field – the target variable [*Murakami et al.*, 2010; *Chen et al.*, 2012]. The vector of anchors is the same as *Chen et al.* [2012]. A plan view of the anchor locations in the synthetic domain is shown in Figure 2.4; note that many of the anchors have common planar coordinates, but are at different depths.

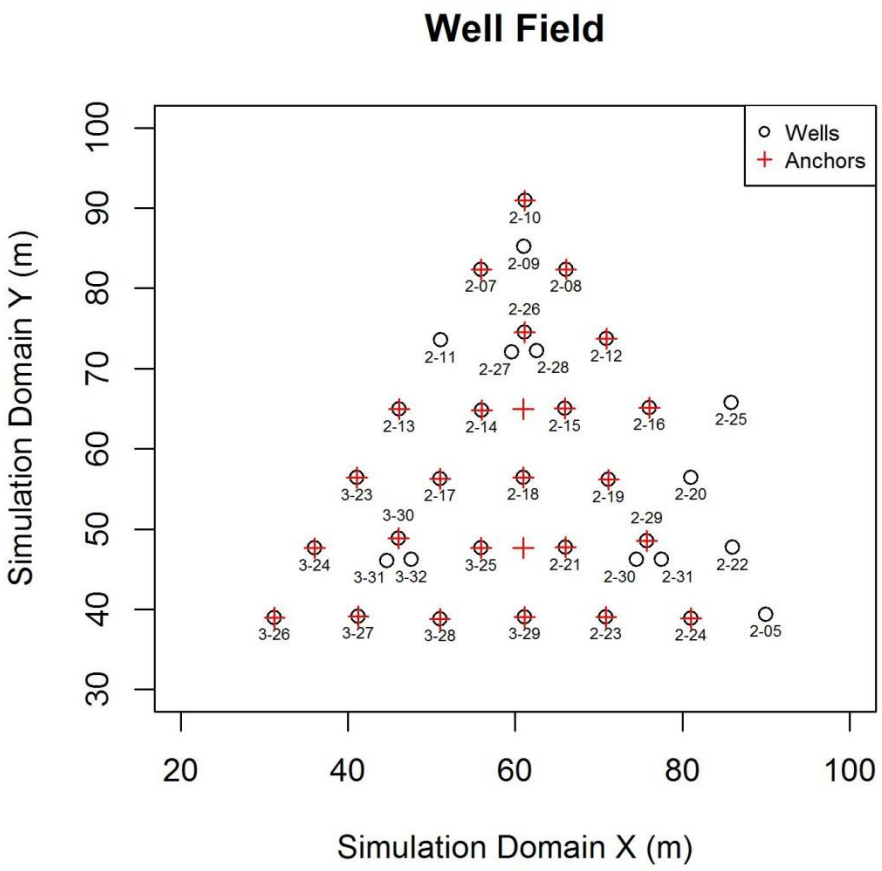

Figure 2.5: The well field of the synthetic case study, with anchor locations depicted. Originally published in *Over et al.* [2013].

The computational budget permitted the numerical likelihood analysis of 2,100 samples, limited primarily by the high cost of three-dimensional flow and transport simulation for multiple realizations of the hydraulic conductivity field per sample. This means that $R = 2{,}100$, and the dimension of $\boldsymbol{D}$ is $2{,}100 \times 330$. The prior joint PDF of the structural parameters and anchors used in the analysis – from which the 2,100 samples were drawn randomly with replacement – is the same as given by *Chen et al.* [2012], which is the mechanism that introduces the EBF and pumping test data of *Murakami et al.* [2010] into the analysis. Table 2.1 summarizes some of the details of the case study that were presented in Section 2.2.

Table 2.1: Summary of case study. Originally published in *Over et al.* [2013].

Numerical Domain	
Domain size [m]	[122, 122, 10]
Grid spacing [m]	[2, 2, 0.5]
Data domain	
# of Type-A measurements	283
# of Type-B measurements, M	5
Parameter Domain	
# of structural parameters	5
# of anchors	325
Implementation	
# of samples, R	2,100
Dimension of D	$2{,}100 \times 330$

In the next section, the model inversion results for the case study are detailed.

2.3 Results

In this section, four posterior marginal PDFs of the structural parameters are presented – defined in Section 2.2.3, excluding the nugget – determined using either MAD or the bundling approaches summarized and developed in Sections 1.2 and 2.1. With either bundling or MAD, the Type-B data used for conditioning is defined in Equation 2.17 for locations given in Figure 2.4. The goal of these computations is to determine whether bundling satisfies two criteria: 1) that the posterior marginal PDFs predicted using Equation 2.4 converge and are comparable to those attained using Equation 1.2 (that is, bundling the likelihood function is stable, and does not meaningfully alter the Bayesian inference), and 2) that the cost of the bundling method, given by Equation 2.15, is less than the cost of evaluating MAD, given by Equation 2.14. The unaltered version of MAD is used to establish a baseline case such that bundling can be compared to this standard. The section opens with plots of the results and a metric for convergence, followed by a discussion the quality of predictions made by bundling relative to the baseline case, and closes with an assessment of the computational costs of the two methods.

The numerically-estimated probability densities in this section are computed with non-parametric Gaussian kernels using fixed bandwidths [*Scott & Sain,* 2004]. A data set of 840,000 simulations (400 simulations for each of the 2,100 samples) or subsets of the 840,000 simulations was used for likelihood function inference [*Chen et al.*, 2012]. Unless otherwise stated, when Equation 2.4 is evaluated, the number of bundles is fixed at $K = 250$, and bundles are defined using the algorithm Partitioning Around Medoids (PAM) [*Kaufmann & Rousseeuw,* 1990]. In this case study, $\boldsymbol{Z}_b(\boldsymbol{\Omega}_j), j = 1, \dots, K$, has variable dimensionality of $B_j \times M$, which results in larger training data matrices for certain bundles; this was chosen so that bundling could be compared to MAD using a convergence criterion based strictly on the number of simulations per sample – that is, N^+ vs. N. Hereafter, for brevity, the constant second dimension of the training data matrices is suppressed ($M = 5$ always) and the first dimension of the training data matrix is referred to as its "length".

2.3.1 Convergence

In this section, convergence of the two approaches is examined. First, the bundling and MAD posterior marginal PDFs are assessed with a qualitative, visual check of stability. If reasonable insensitivity of the posterior marginal PDFs to further increases to the number of simulations per sample used for inference is observed, then the posterior marginal PDFs are re-analyzed with a more rigorous, quantitative criterion based on integrated absolute error (IAE). The quantitative metric is then employed to validate the convergence – defining the number of simulations per sample needed with each method. IAE is also used here to determine a truncation point that defines when to stop generating additional simulations per sample during likelihood analysis, and in Section 2.3.3 to evaluate the computational cost.

Figure 2.6 is an example of the qualitative check for convergence using the two approaches – MAD results are shown on the left, and bundling on the right. In both panes, the plot shows a posterior marginal PDF (dimensionless variance) determined using the complete data set ($N, N^+ = 400$), and the posterior marginal PDFs determined using the next 24 smaller data sets ($N, N^+ = 376, \dots, 399$), each of which has one simulation per sample withdrawn randomly from the next-larger data set.

Figure 2.6 qualitatively suggests that the posterior marginal PDFs determined using either MAD or bundling do converge – the family of curves is very similar. Figure 2.6 is evidence that the posterior marginal PDFs are not highly sensitive to either additional or fewer simulations per sample in the training data matrix, so a more quantitative assessment of convergence is undertaken.

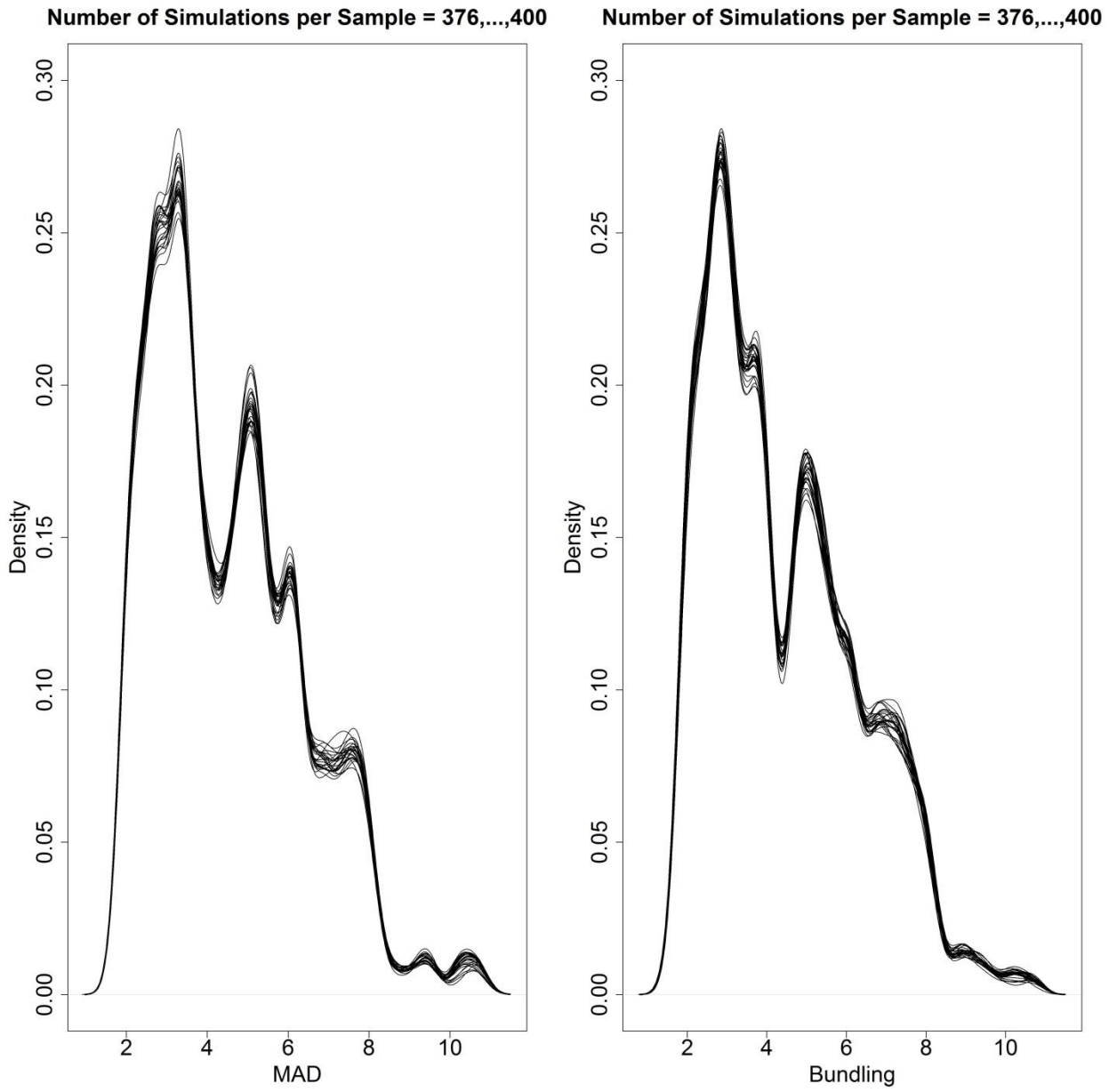

Figure 2.6: Qualitative check of bundling and MAD posterior marginal PDFs of dimensionless variance for possible convergence, computed by sequentially adding simulations to the training data matrices. Originally published in *Over et al.* [2013].

The remainder of this section therefore compares this convergence of MAD and bundling, using metrics established based on IAE. IAE is formally as follows:

$$\text{IAE} = \int \left| f(x) - \hat{f}_n(x) \right| dx, \tag{2.18}$$

which compares the true PDF $f(x)$ and the numerically-estimated PDF $\hat{f}_n(x)$, inferred with a training data matrix of size $n \times 1$ (uni-variate case). The numerically-estimated PDF is said to be a consistent estimator of the true PDF if IAE $\to 0$ as $n \to \infty$ [*Izenman,* 1991].

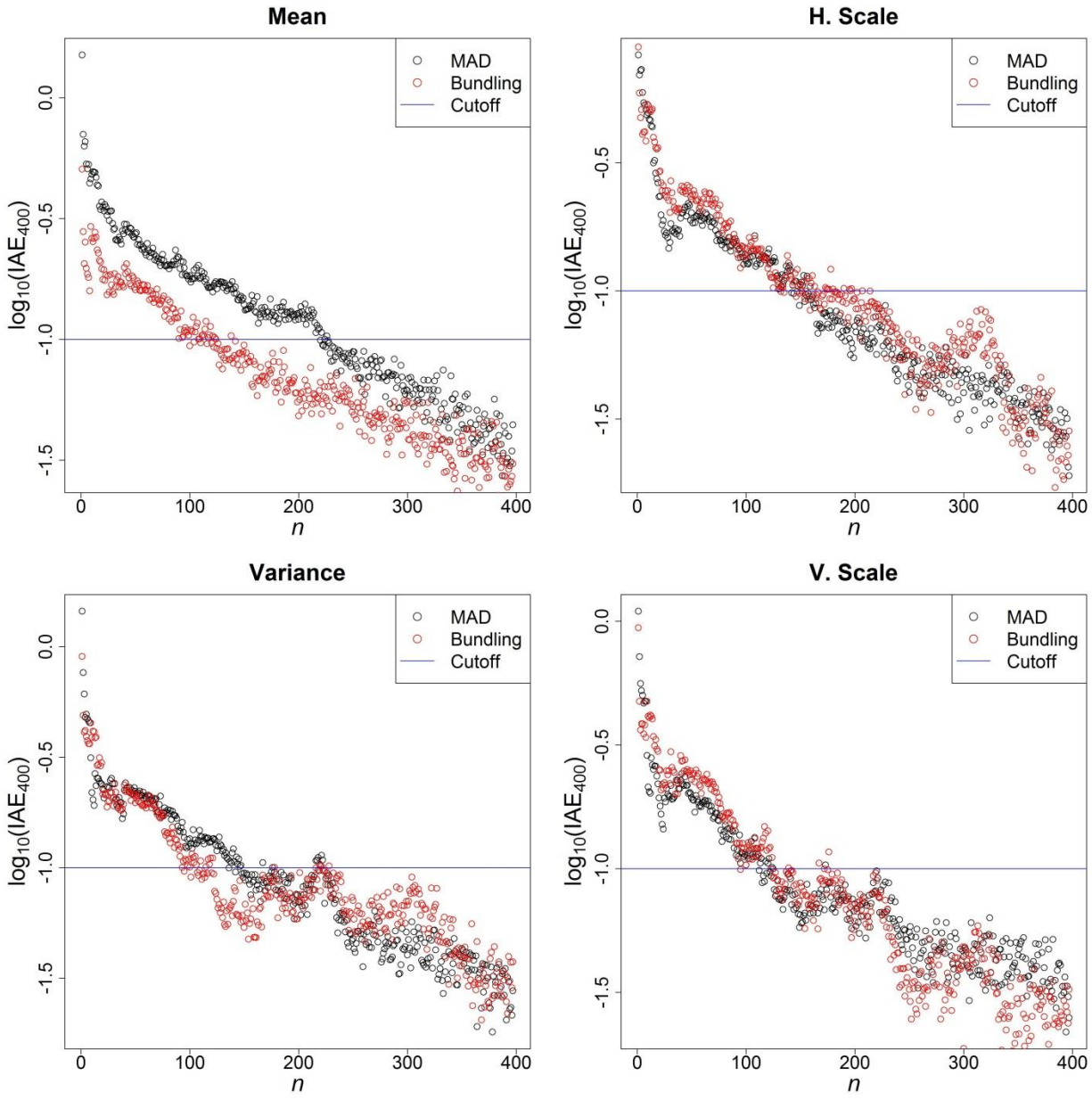

Figure 2.7: Comparison of convergence to the best estimate of the marginal posterior PDFs between the two likelihood function approximation methods. Originally published in *Over et al.* [2013].

To validate MAD and bundling as methods that reasonably converge using this formal IAE definition requires knowing the true posterior marginal PDFs of the structural parameters in question. However, because the true posterior marginal PDF of either method is unknown in this case study, the convergence of the two methods is instead validated using a modification of Equation 2.18. Define the validation metric IAE_{400} – the IAE of the estimated posterior marginal PDFs using training data matrix of length $n = N$ or $n = N^+$ with respect to our best estimates of the posterior marginal PDFs that use $N, N^+ = 400$ – shown explicitly for N, as

$$\mathrm{IAE}_{400}(n) = \int \left| \hat{f}_{N=400}(x) - \hat{f}_{N=n}(x) \right| dx, \text{ for } 1 \leq n < 400. \tag{2.19}$$

Next, the methods are tested for how many simulations per sample are required to satisfy $\text{IAE}_{400} < 0.1$ - an analysis referred to as the "validation approach." The convergence of each estimated marginal PDF to its "true" PDF is shown in Figure 2.7 by evaluating Equation 2.19 as a function of $n = N$ or $n = N^+$.

For each marginal posterior PDF, the smallest number of simulations per sample required to satisfy $\text{IAE}_{400} < 0.1$ was determined and is listed in Table 2.2.

Table 2.2: Convergence in IAE_{400}. Originally published in *Over et al.* [2013].

	$\boldsymbol{\mu}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\lambda_h}$	$\boldsymbol{\lambda_v}$
MAD (N)	223	150	154	114
Bundling (N^+)	112	95	145	116

Comparing the two methods for the limiting values of N and N^+, the maximum number of simulations per sample required with $K = 250$ is 145 for bundling, and 223 for MAD. As such, employing the validation approach results in $(223 - 145)/223 * 100 = 35\%$ relative savings in simulations using Equation 2.4 instead of Equation 1.2 for this case study.

Before moving to a comparison of the shapes of the posterior marginal PDFs from the two methods, however, a convergence rule that does not necessitate knowledge of the true PDF (assumed or known) must be specified. This requires a different metric from the one defined in Equations 2.18 and 2.19, because it is useful to be able to identify a required number of simulations per sample without being restricted to either working with PDFs of known form (Equation 2.18) or generating large data sets before checking convergence (Equation 2.19), the former of which would not require numerical estimation of the PDF and the latter of which would eliminate the possibility of reducing the simulation cost. This alternative metric is important, because it permits an active analysis of convergence that can be performed during the inversion. This alternative metric is referred to as the "truncation approach."

To this end, the IAE expression in Equation 2.18 is modified, so that it compares the posterior marginal PDFs predicted using a small subset of the data, $N, N^+ = 50$, with those predicted using a progressively larger $n = N$ or $n = N^+$. Shown explicitly for N:

$$\text{IAE}_{50}(n) = \int \left| \hat{f}_{N=50}(x) - \hat{f}_{N=n}(x) \right| dx, \text{for } n > 50. \tag{2.20}$$

Figure 2.8 shows IAE_{50} as a function of either N or N^+, evaluated for all four posterior marginal PDFs using both MAD and bundling.

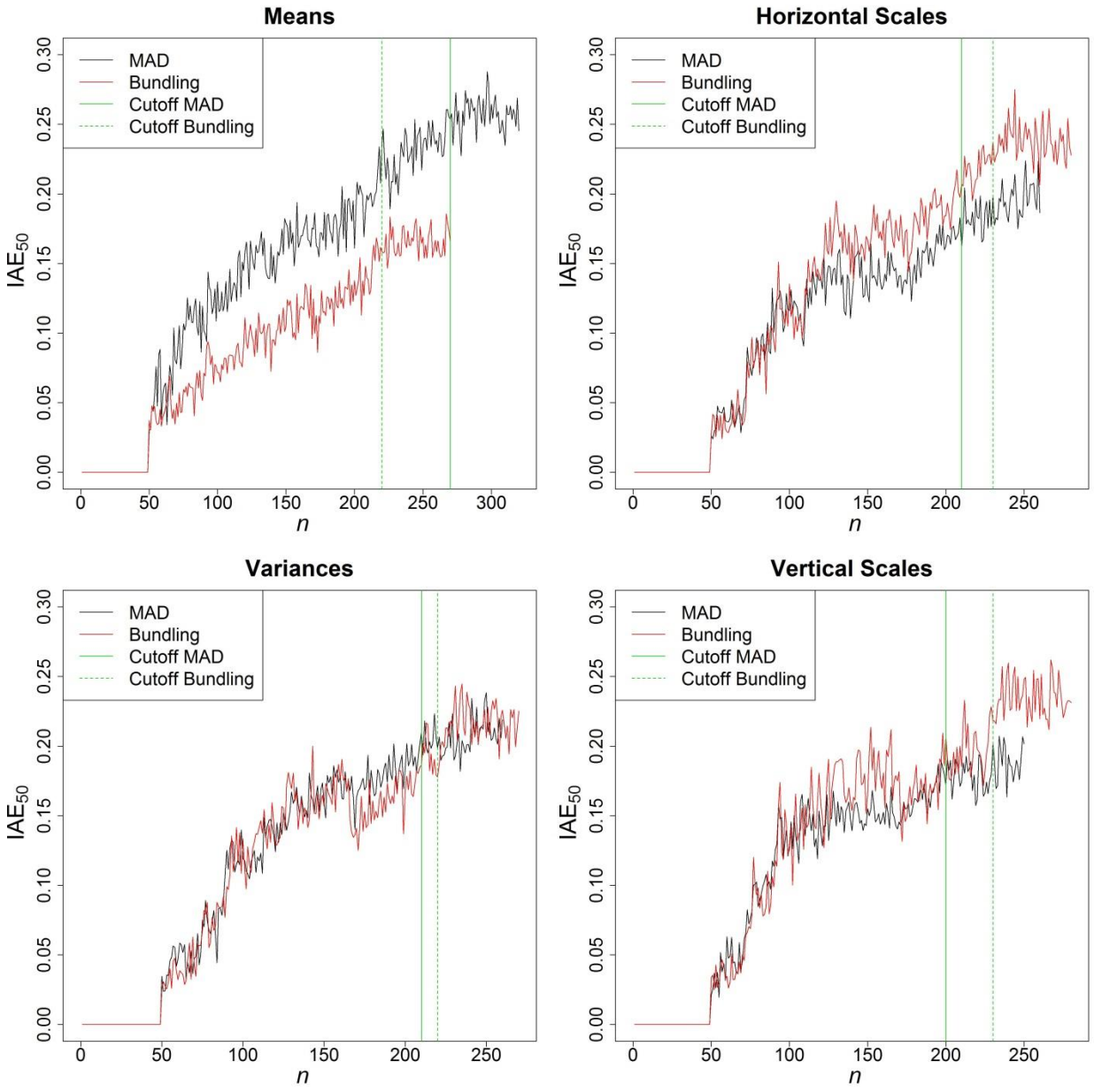

Figure 2.8: Comparison of convergence using a small subset of data data $(N, N^+ = 50)$ to establish a baseline case for the marginal posterior PDFs between the two likelihood function approximation methods. Vertical reference lines indicate the beginning of regions of approximately-constant IAE. Originally published in *Over et al.* [2013].

Figure 2.8 shows for both methods and all four structural parameters that IAE_{50} initially increases with increasing N, N^+, and then levels off to a plateau of approximately-constant IAE_{50}. The increasing IAE_{50} with respect to N, N^+– to the left of the reference lines in each pane – indicates a lack of convergence. The plateau – to the right of the reference lines in each pane – indicates an approach to some stable limit, because the error is constant with respect to N, N^+. The vertical reference lines in Figure 2.8 indicate best estimates of the value of $N, N^+$where the plateau begins, and denote the smallest possible N, N^+ using Equation 2.20. Experience showed that 50 additional simulations per sample ensured that relative plateaus were not misinterpreted as the ultimate limit of the constant behavior of

IAE_{50} – see, for example, the roughly-constant IAE_{50} for $120 < N, N^+ < 160$, followed by increasing IAE_{50} for $160 < N, N^+ < 200$ for horizontal scale. These results of the truncation approach values for N and N^+, determined by identifying plateaus from Figure 2.8 are summarized in Table 2.3.

Table 2.3 shows again that bundling requires fewer simulations per sample than MAD to converge all posterior marginal PDFs – the worst-case values are $N^+ = 230$ and $N = 270$,. However, the truncation approach does still suggest using more simulations per sample than the validation approach to ensure convergence of the estimated PDFs. Accordingly, the percent reduction in simulations per sample is less dramatic in Table 2.3 than in Table 2.2, at only $(270 - 230)/270 * 100 = 15\%$. But, this is still a significant savings for inversions with a large simulation cost. The important outcome of Figure 2.8 and Table 2.3 is that faster convergence of bundling can be confirmed using the truncation approach, at least crudely, without a knowledge of the "true" posterior marginal PDFs.

Table 2.3: Convergence in IAE_{50}. Originally published in *Over et al.* [2013].

	μ	σ^2	λ_h	λ_v
MAD	270	210	210	200
Bundling	220	220	230	230

The significance of the truncation approach is that even during the inversion, convergence can nonetheless be evaluated in a manner that is consistent with the more rigorous – and expensive – validation approach. It is of only minor importance that the savings suggested by the two approaches to setting N and N^+ do not perfectly agree. Actually, the difference in required simulations per samples between Tables 2.2 and 2.3 suggests that requiring a plateau to be roughly constant for 50 additional simulations may be overly strict in the truncation approach, but this requirement was adopted *ad hoc* to be conservative. If the truncation approach came to different conclusions than the validation approach about which method converges faster, it would cast serious doubts on the validity of using visual methods like Figure 2.6 and criteria like Equation 2.18 to establish convergence.

2.3.2 Inferential Quality: Bundling

Here, the numerical results of MAD and bundling in terms of the shape and quality of the posterior marginal PDF that each predicts are compared. Figure 2.9 shows the posterior marginal PDFs of MAD and bundling, the prior marginal PDFs, and the true values of four structural parameters. The posterior results are each determined using $N, N^+ = 400$ simulations, to identify any artifacts in the PDFs imparted by bundling.

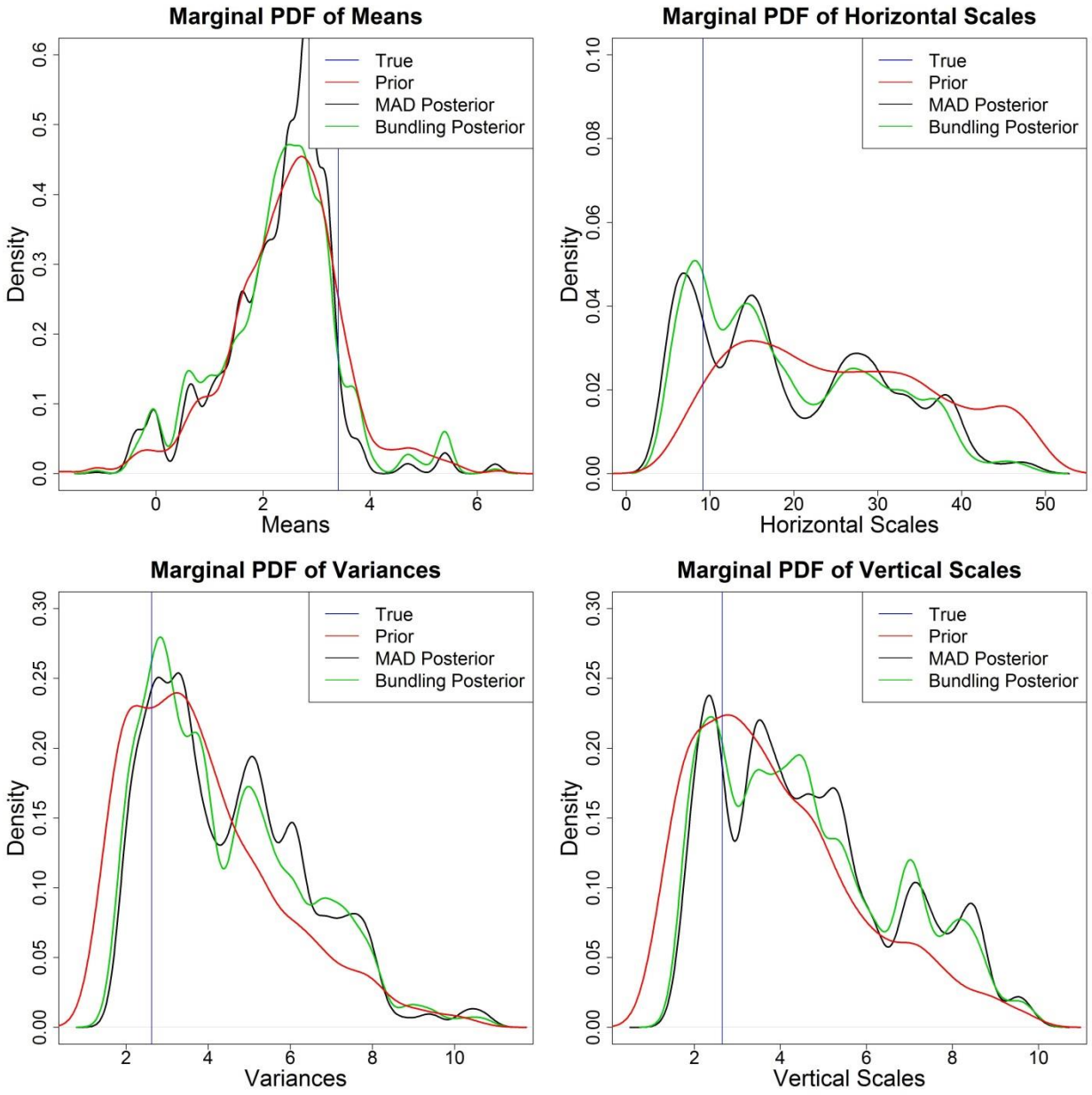

Figure 2.9: Comparison of posterior marginal PDFs calculated using MAD (Equation 2, black) and bundling (Equation 6, green). Prior marginal PDFs are shown in red, and the true values of the parameters for the synthetic study in blue. Originally published in *Over et al.* [2013].

Comparing the results of Equation 1.2 (black lines) with those of Equation 2.4 (green lines) in Figure 2.9, notice that the shapes of the PDFs are similar. Notably, important features of the predictions that use the MAD equation are preserved by the bundling approximation; the single modality in means, the skewed single modality in variances, the tri-modality in horizontal scales, and the peak modality in vertical scales that were found using MAD are also identified using bundling (with some minor variations in the density values). This supports the claim that bundling parameters together before compiling the training data matrices introduces only a very limited amount of error in the likelihood function (at least for properly-sized bundles, discussed next). What remains to be shown, however, is

whether the quality of the prediction is maintained at the earliest convergence limit identified in Section 2.3.1 - a topic revisited at the conclusion of this section after discussing the importance of bundle size.

Next, the question of why the error introduced by bundling is limited only if the number of bundles is properly selected is addressed. Equation 2.3 in Section 2.1.1.1 showed that, in general, bundling is a valid approximation only when the parameter sets included in a bundle are nearly identical. In Section 2.1.1.2 the order of the bundling approximation in the non-parametric Gaussian kernel utilizing a Taylor series expansion of the Type-B data as a function of the parameters was derived. For this case study's numerical implementation, the difference between Equation 1.2 and Equation 2.4 is proportional to the maximum Euclidean distance within a the bundle, which provides the motivation for an *a priori* strategy to determine the number of bundles.

Before generating simulation data, different values of K can be evaluated for their maximum Euclidean distances across bundles, which in combination with the derivation in Section 2.1.1.2, permits a researcher to set K based on an allowable maximum first order error in the likelihood function. Note that each bundle has a maximum Euclidean distance between its two most distant members or is zero for singly-populated bundles.

It is recommended to set K by examining the PDFs of the maximum Euclidean distances across all the bundles, because this permits a clearer understanding of how the number of bundles relates to the error term in Equation 2.4, and also lends itself to easy visualization. Then, set K such that the median maximum Euclidean distance is very small compared to the largest possible maximum distance, but is non-zero for two reasons. First, the smaller the maximum Euclidean distance in the bundles, the smaller the error term introduced by bundling into the likelihood function estimates. See Figure 2.10 for a few posterior marginal PDFs using $K = 2$ relative to the MAD and $K = 250$ results, as well as for a comparison of the maximum Euclidean distance PDFs using $K = 2$ and $K = 250$. Second, any non-zero probability of maximum Euclidean distance equal to 0 implies some $n_j = 1$, and reduces Equation 2.4 to Equation 1.2 at least for some bundles. This, of course, nullifies the bundling approximation and any possible computational savings to result from that approximation – remember from Section 2.1.1.1 that simulation cost accumulates with the number of likelihood-function calculations. Finally, it is recommended to search for medoids and bundles increasing from $K = 2$ - while $K = 2$ may often be an unreasonable choice based on size, it will always provide the largest possible maximum Euclidean distance when using PAM.

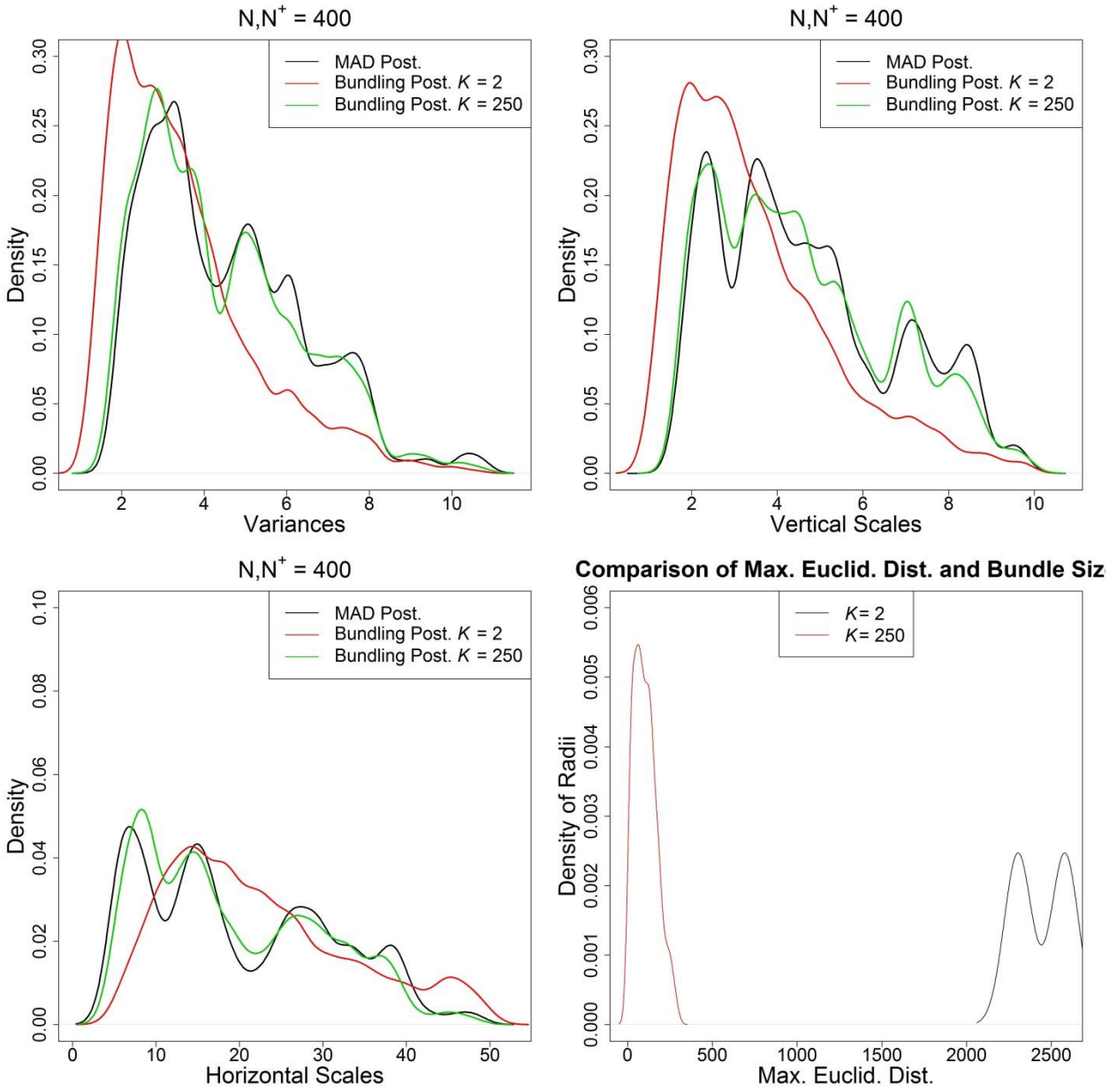

Figure 2.10: Comparison of posterior marginal PDFs of bundling with various values of K to the baseline MAD case and comparison of the densities of maximum Euclidean distance for the selected values of K (bottom right). Originally published in *Over et al.* [2013].

In Figure 2.10, the median maximum Euclidean distances are respectively 4% and 95% percent of the largest possible maximum Euclidean distance. The shapes of the PDFs predicted using bundling with $K = 2$ are quite different from the MAD and the bundling with $K = 250$ results – as indicated by the strong separation of the red ($K = 2$) traces from the green ($K = 250$) and black (MAD) traces in the three panes corresponding to marginal PDFs of structural parameters.. Figure 2.10 shows that the skewness of the bundling with $K = 2$ posterior marginal PDFs is strongly altered, and that there are significantly over-estimated or under-estimated regions in each PDF as well. Both effects are consequences of the poor choice for the number of bundles to use.

This section closes with a comparison of the posterior marginal PDFs using the number of simulations per sample given by the validation approach identified in Table 2.2.

Figure 2.11 echoes Figure 2.9: the pertinent features of the PDFs – for example, the skewness in the variances or vertical scales, and tri-modality in horizontal scale – can still be identified using significantly fewer than N, $N^+ = 400$ with either method. However, the main conclusion to be made from this plot is that bundling makes roughly the same prediction with 35% fewer forward simulations than MAD.

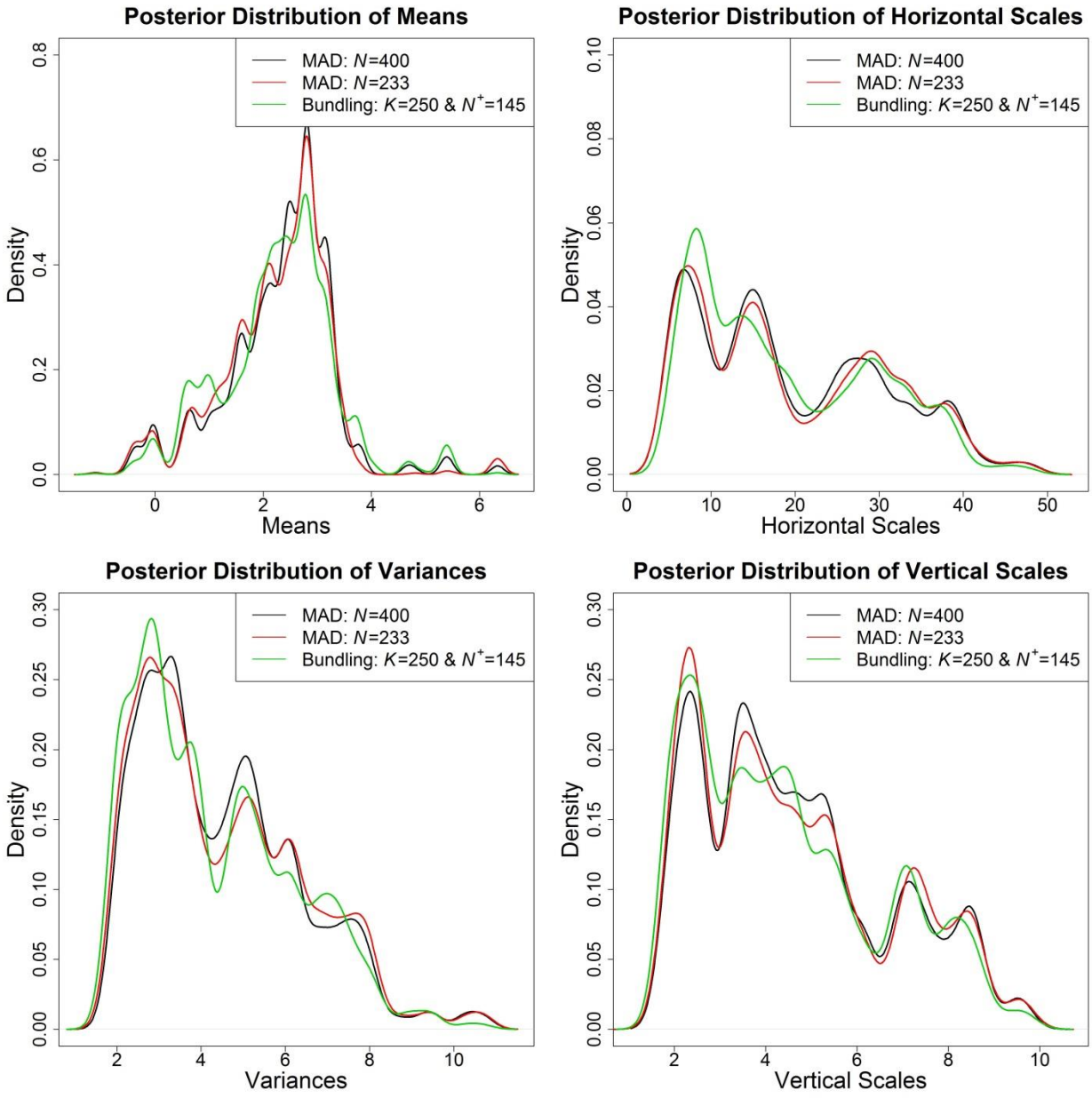

Figure 2.11: Comparison of bundling with $N^+ = 145$ (green) to MAD with $N = 223$ (red) defined using the IAE_{400} criterion of Table 1, relative to the baseline MAD case with $N = 400$ (black). Originally published in *Over et al.* [2013].

An example data set and script to numerically implement bundling (using the R statistical computing language), can be found online at:
http://mad.codeplex.com/wikipage?title=Bundling.

2.3.3 Computational Cost

After establishing a metric for convergence and comparing the quality of the posterior marginal PDFs, it is now possible assess the computational costs of MAD and bundling. Here, Equations 2.14 and 2.15 are evaluated in order to quantify the savings achieved using the bundling approximation instead of MAD.

The cost of a single simulation for this case study using PFloTran is roughly 40 minutes with a single-core AMD "MagnyCours" 2.1 GHz processor on the Franklin supercomputer at the National Energy Research Scientific Computing Center. In Section 2.3.1, the number of simulations required per sample to achieve convergence of bundling and MAD was found to be 145 and 223 respectively, leading to net simulation costs of roughly 203,000 hours and 312,200 hours. Establishing 250 bundles required evaluating PAM for $K = 2, \ldots, 250$ on the parameter sample matrix $\boldsymbol{D}$ and computing the densities of the maximum Euclidean distances in the bundles, which required roughly 37 hours (one-dimensional density estimation has an insignificant cost for this case study – an order of seconds). Finally, verification of convergence of the likelihood function – which was established using two IAE-based metrics (Equations 2.19 and 2.20) – and required comparison of many different values of N and N^+ at a cost of roughly 2 hours per metric for either method. Table 2.4 summarizes the various components of the cost and the total computational expenses using MAD and bundling, all in hours.

Table 2.4: Computational cost . Originally published in *Over et al.* [2013].

	Simulation Cost	**Bundle Determination**	**Convergence Determination**	**Total Cost**
MAD	312,200	-	4	312,204
Bundling	203,000	37	4	203,041

The computational cost of bundling the parameter sets is easily offset by the savings in the simulation expense, and offers a $(312{,}204 - 203{,}041)\ /312{,}204 * 100 = 35\%$ savings over MAD for this case study.

2.4 Conclusions

In this chapter, an approximation method called "bundling" was presented, which evaluates the likelihood of a bundle of samples reproducing the measured data. The approximation strategy was compared with a baseline case of MAD, which individually analyzes the likelihood of each sample, for overall cost and prediction quality. The results suggest that, at least for this numerical case study, bundling converges faster than MAD – over 100,000 hours less simulation time was needed for the bundling approach. Further, it was demonstrated that the predictive quality of bundling is very similar to that of MAD, even when using significantly fewer simulations per sample to infer the likelihood function. The 35% smaller overall inversion cost demonstrated with bundling relative to MAD is significant: this approximation could be a tool that relaxes the typically-prohibitive computational requirement of MAD.

This reduction in cost is the outcome of applying a tool that organizes samples with the primary goal of avoiding redundant likelihood function estimation steps in the parameter space. By setting the requirement that bundles must contain similar samples using Euclidean distance, it was shown in Equation 2.4 and Section 2.1.1.2 that the maximum first order error introduced by bundling is proportional to the maximum Euclidean distance across any of the bundles. Based on this error, a strategy for setting the number of bundles, K, prior to the inversion was identified, based on the density of maximum Euclidean distances of the bundles.

This cost reduction was quantified and validated using two separate approaches, both of which indicated decisively that bundling is cheaper than MAD. The more rigorous validation approach compared predicted PDFs with a PDF using a maximum state of information; however, this cannot be checked during inversion. The more *ad hoc* truncation approach compared predicted PDFs with a PDF using a lower state of information, but can be actively checked during inversion and used to set a stopping point during the simulation process.

A rather stringent convergence criteria was enforced for both measures in order to employ bundling in MAD without subsequent more accurate re-analysis of the samples. However, for studies that use bundling as a "scanning" technique it is recommended to relax the convergence criteria and increase the maximum permissible error in the likelihood function, because both factors should further cheapen the cost of an initial "scan". Such "crude" scans should always be followed by more exact re-analysis of the high likelihood regions before drawing conclusions.

In conclusion, cheaper, but still accurate, numerical implementation costs of MAD can play a huge role in making the theory more accessible to different scientific applications. This chapter identified motivations for such a strategy, developed alterations to the fundamental Bayesian proportionality of MAD, derived order of error in this

approximation, and empirically confirmed that computational savings, with minimal degradation of the quality of the results, could be obtained.

In the next chapter, the central Bayesian proportionality is approached from a different perspective – namely, the numerical implementation of Equation 1.2 (or its variants shown in Section 1.3.1) is fully generalized for implementation with any FM, any RFG, and for any statistically valid PDFs.

3. The MAD Software

Sixteen years ago, researchers posited that the field of hydrogeology should implement inversion modeling as a standard practice [*Poeter & Hill,* 1997] similar to the push by the USGS in the 1970s that proliferated the use of numerical forward modeling tools, but several functional obstructions have prevented widespread adoption of inversion models. *Carrera et al.* [2005] suggested that the following five issues need to be addressed to encourage widespread use of inversion techniques: 1) incorporation of geological data 2) the flexibility of the code and procedure to handle any and all relevant data types 3) accommodation of uncertainties 4) difficulty of code operation and 5) coupling of techniques with a GIS (Geographic Information System) platform [*Carrera et al*., 2005]. The MAD software addresses these issues.

The statistical capabilities of MAD (outlined in Section 1.2) address the first three issues on *Carrera's* list - these are issues that are fundamentally accounted for by a more robust inversion approach, which is now briefly justified. First, the incorporation of geological data - listed separately from inclusion of other data types to highlight the importance of geological processes that are often unaccounted for by zonation techniques [*Carrera et al.*, 2005] -to this end, MAD is compatible with geostatistics and uses a variety of available models - which *Carrera et al.* [2005] suggested as an alternative to a zonation framework. Second, MAD is a technique capable of assimilating data of any type and on any scale of the inversion problem, thus making it appropriately flexible for any application. Third, MAD is a Bayesian tool that can treat any parameter of interest as a random variable and hence account for uncertainty throughout the inversion process. These three challenges are fundamentally related to the generality of the inversion technique – each of which can be accounted for by MAD - whereas the final two challenges are related to the software platform for the numerical implementation of the inversion technique.

The last two items on the list of *Carrera et al.* [2005] are accounted for by the development of easy-to-use software that lowers the bar for stochastic inversion modeling and is complemented by GIS capabilities. The MAD software has been developed with a GUI that guides users through formulating the inversion problem, manages appropriate data throughout the process, and visualizes spatial data in a GIS mapping environment. The need for scientists to fully understand the statistics of the approach and compile large amounts of simulated data are traditionally requirements of an inversion model application, but, as a default, the MAD software GUI automates the statistics and manages the data; allowing users to focus on specifying a better forward model or the implications of the predictions of the calibrated model, instead of the minutiae of the inversion methodology.

The objective of this chapter is to present the open-source MAD software architecture as an extensible desktop application that further reduces the limitations to the widespread use of

inversion modeling in scientific investigations. The core motivations and objectives of the MAD software development are three-fold: 1) automated, generalized implementations of MAD, 2) simplified setup and configuration processes for end-users via the GUI, 3) a platform that is easily transferrable and adaptable to multiple fields of study.

In the first section of this chapter, the emphasis is on how elements of the graphical user interface (GUI) and the computational core of the software provide functionality to implement MAD. As each function or step of the implementation (in parallel with Figure 1.1 and the discussion in Section 1.3.2) is identified, the utility to automatic implementation and the simplicity of user interactions will be discussed in terms of the three core motivations above. Additionally in the first section, the importance of the sequence, because there are many dependencies, will be explained in greater detail. Each form will be introduced with accompanying screenshots. In the second section of this chapter, the emphasis is on how this software architecture and design are modular as well as generalized for connection with a variety of relevant external tools. After the design is presented, it will be possible to discuss how the MAD software architecture promotes easy transfer between and adaptation to different scientific fields of study. In the final section of this chapter, case studies demonstrating several of the variants of the Bayesian proportionality for MAD (introduced in Section 1.3.1) are presented.

3.1 The GUI and MAD Theory

This section is divided into two subsections: the first provides a coarse overview of the MAD software GUI and the second demonstrates and discusses in detail the generalized implementation of MAD theory that each of the forms in the GUI supports. This section covers both modules of the MAD software GUI: pre-processing and post-processing.

3.1.1 The Forms of the MAD Software GUI

In the beginning of this subsection, the pre-processing module is summarized, which handles all the components of a MAD application up to and including the generation of ensembles of simulated Type-B data (Steps 1-5, Figure 1.1). In the latter portion of the subsection, the post-processing module is overviewed, which handles all of the components of a MAD application affiliated with the likelihood function (Step 6, Figure 1.1.), convergence, and visualization of statistical output. The objective of this section is to briefly give a topical overview of the MAD software modules, before introducing any individual components, connectivity, or workflow in detail.

The pre-processing module guides users through the definition or configuration of all the necessary requisites for ensemble simulation. The main inputs - depending on the nature

of the MAD application, which the GUI can differentiate – can include definition of a spatial domain, $\boldsymbol{z}_a$, $\boldsymbol{z}_b$, $\boldsymbol{\theta}$, $\boldsymbol{\vartheta}$, $S(\boldsymbol{Y};\boldsymbol{\theta})$, and the joint prior PDF. The final, and perhaps most significant, role of the pre-processing module is to establish a connection with the forward model and the random field generator, which are external software components – the MAD software on its own does not possess the capability for simulation or field generation, but can control outside tools to these ends.

Table 3.1 lists the titles of the 9 forms of the pre-processing module and a brief statement of their core functionality.

Table 3.1: Pre-processing module forms and core functionalities.

Form Title	Core Function
Project	Connects MAD software with forward model
Define Domain Area	Configures spatial domain
Define Variables	Defines parameter field(s) for characterization
Measurements	Allows upload of $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$
Structural Model	Defines $S(\boldsymbol{Y};\boldsymbol{\theta})$ and random components of $\boldsymbol{\theta}$
Anchors	Defines $\boldsymbol{\vartheta}$
Prior Distribution	Upload of samples from the joint prior PDF
Likelihood Setup	Allows aggregation of, subset selection from $\boldsymbol{z}_b$
Simulation	Executes ensemble simulation in batch

The post-processing module guides users through analysis of the simulation ensemble, convergence analysis, and visualization. The key elements produced by the post-processing module include the likelihood function and the posterior joint PDF. Another significant role of the post-processing module is diagnostics of the size of the simulation ensemble N. Table 3.2 lists the titles of the 5 forms of the post-processing module and a brief statement of their core functionality.

Table 3.2: Post-processing module forms and core functionalities.

Form Title	Core Function
Project	Connects databases containing simulations
Data Organization	Creates variations of $\boldsymbol{z}_b$, visualize simulations
Compute Likelihood	Computes likelihood function
Convergence	Evaluates the likelihood value as a function of N
Posterior Analysis and Visualization	Visualizes posterior marginal PDFs

In Section 3.1.2, the purpose of each pre-processing and post-processing form will be explained in greater detail with accompanying screenshots.

3.1.2 Detailed Description of Each MAD Software Form

In this subsection, each of the forms listed in Tables 3.1 and 3.2 is presented in more detail. The objective is to identify the essential elements of MAD applications each form is intended to support.

The following 9 forms from pre-processing are necessary for the MAD software to produce the training data matrices $\boldsymbol{Z}_b$. These forms complete steps 1-5 of Figure 1.1, which was described in Section 1.3.2

3.1.2.1 The "Project" Form

Figure 3.1 shows the *Project* form of the MAD software. This form has several critical responsibilities, which are numbered in the graphic: 1) construction of the database in which the MAD software stores all relevant project information, 2) connection of the MAD software to the forward model executable and forward model project file, 3) determination of time dependency for all subsequent forms, and 4) connection of the MAD software to the random field generator executable.

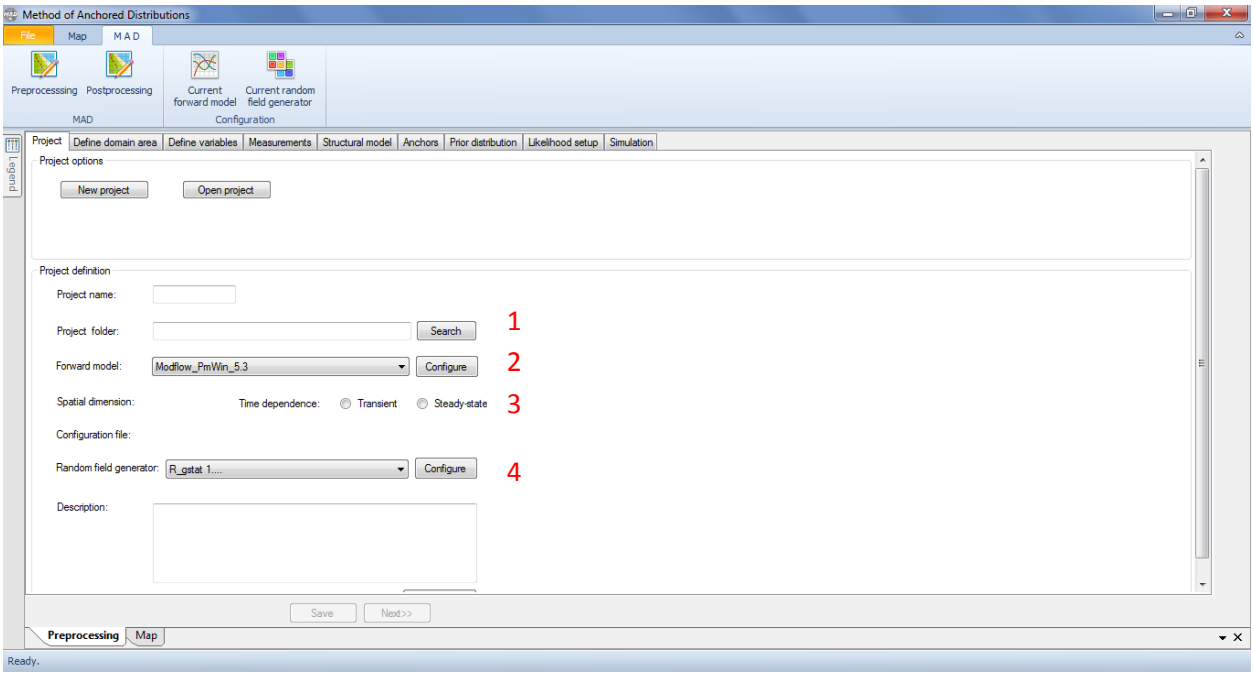

Figure 3.1: The *project* form of the MAD software pre-processing module.

The connections to the random field generator and forward model are important early steps in the MAD software experience, because these constrain the possibilities of a MAD application. The forward model project file establishes constraints for the spatial domain

(dimensionality), the possible kinds of Type-A and Type-B measurements, and units of measurement (when available). The random field generator is used to establish possible structural model types. More obviously, without attaching these executable files, the MAD software would not be able to automate the random field generation or simulation processes affiliated with implementation. Finally, recall Table 3.1 and note that the core functionalities of the next five forms in the pre-processing module all depend on the information (listed above) attained from these connections in the *Project* form, hence the sequence.

There are several particularly user-friendly elements of this form. The GUI adapts to the specific forward model project as well as the random field generator and populates the next five forms with case-specific options only, which is not visible to the user. Second, the connections to the random field generator and forward model are very easy to configure, because they only require identification of directory paths. Finally, because the random field generator and forward model have a common connection to the MAD software, the user does not have to worry about how to store, format, or utilize random fields in the forward model because the process is fully automated by the MAD software.

The next form is used to define the spatial domain of the MAD application.

3.1.2.2 The "Define domain area" Form

Figure 3.2 shows the *Define domain area* form of the MAD software. The form has two critical responsibilities, which are numbered in the graphic: 1) establishment of the coordinate system for all spatial information of the MAD application and 2) definition of the discretization for the random field generator.

There is a fundamental need for coordinates in MAD applications, but there is also a strong motivation for this form to come early in the pre-processing module sequence. Three of the next four forms' core functionalities require a coordinate system: upload of measurement data, definition of structural models with spatial correlation, and definition of anchor locations.

There are several user-friendly features to this form. First, there is a map tab that visualizes the user input spatial discretization (for reasons that are made apparent in the next three sections, a graphic of the map is not presented until Section 3.1.2.6) and accesses multiple types of base maps, which are useful for comparison to site maps or satellite images. Second, the dimensionality of the configured domain is required to match the dimensionality of the forward model, which prevents discrepancies between user input and existing information in the forward model project file. Finally, the data-entry experience is simple and fast. The main drawback to this form is that it does not support variable density discretization or non-rectangular elements, which restricts the types of

forward model projects that can be connected to the MAD software. This is not a limitation of the theory, but of the technical capability of the MAD software.

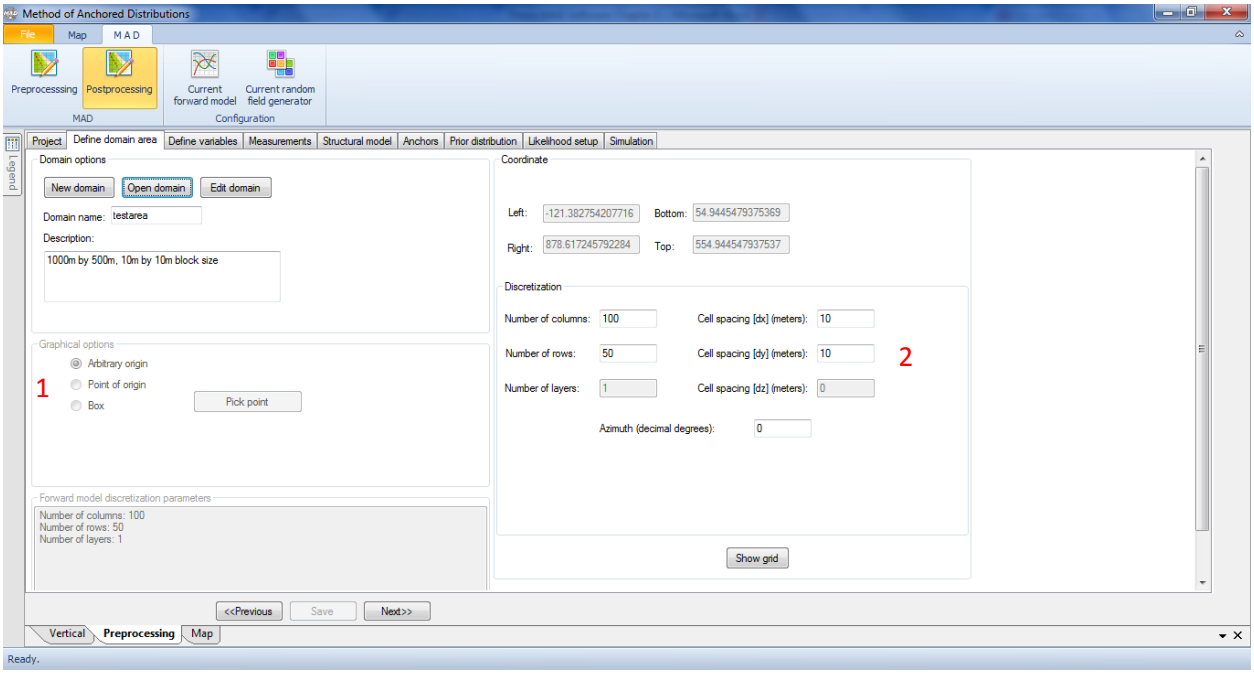

Figure 3.2: The *Define domain area* form of the MAD software pre-processing module. 1. Tool for establishing the coordinate system. 2. Tool for defining the discretization.

The next form is used to select the parameter field or fields that are characterized in the MAD application.

3.1.2.3 The "Define variables" Form

Figure 3.3 shows the *Define variables* form of the MAD software. The form has two critical responsibilities, which are numbered in the graphic: 1) definition of the parameter field or fields that will be represented by some combination of a structural model, anchors, and direct measurements and 2) definition of the indirect measurement data types that will be used to condition the parameter field in the Bayesian analysis. Additionally, this form also determines what inputs of the forward model project file that the MAD software will overwrite with random field realizations as well as the outputs from the forward model that the MAD software will store in databases for analysis in the post-processing module.

This form is also placed early in the sequence, because the measurement data, structural model, and anchors must all be grouped by data type during the random field generation and simulation processes.

The form is indirectly accountable for one half of the most user-friendly aspect of the MAD software: the formatting, reading, writing, and storage of the correct forward model input

files, which can require a tremendous effort in applications done “by hand”. Also appealing is the simple interface that is populated with lists of options pertinent to the specific forward model connected in the *Project* form.

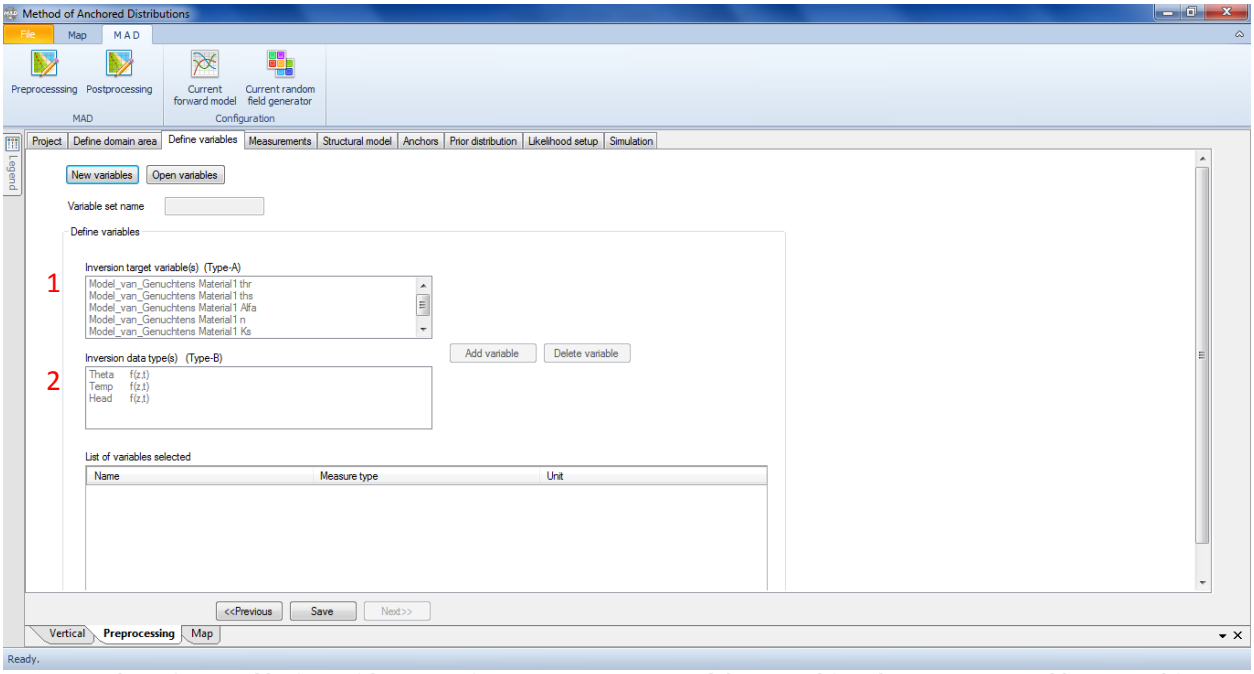

Figure 3.3: The *Define variables* form of the MAD software pre-processing module. 1. Tool for selecting target variables. 2. Tool for selecting conditioning data.

The next form is used to manage and upload measurement data.

3.1.2.4 The “Measurements” Form

Figure 3.4 shows the M*easurements* form of the MAD software. The primary responsibility of this form is the management of all relevant information about a measurement, which includes measurement values, types, coordinates, and times. Also, this form establishes the spatial as well as the temporal coordinates at which the MAD software copies the forward model output from each simulation. As mentioned in the previous two sections, the measurement type and coordinates require the *Define domain area* and *Define variables* forms; however, the information uploaded and defined in this form is utilized in the *Structural model* and *Likelihood Setup* forms, thus the order of the *Measurements* form in the sequence.

This form provides the first concrete link with the equation of MAD presented in the Section 1.1.2, by defining explicitly $\boldsymbol{z}_a$. As mentioned in Section 1.1.2, Type-A data is only compatible with certain structural models (an attribute that is established by the connection of a random field generator in the *Project* form) and the MAD software will not permit Type-A data upload unless this requirement is satisfied.

The form has two key user-friendly attributes. First, text files containing time series of measurement can be directly uploaded. Second, the form automatically visualizes the measurement locations in the map tab (as before, for reasons that are made apparent in the next two sections, a graphic of the map is not presented until Section 3.1.2.6).

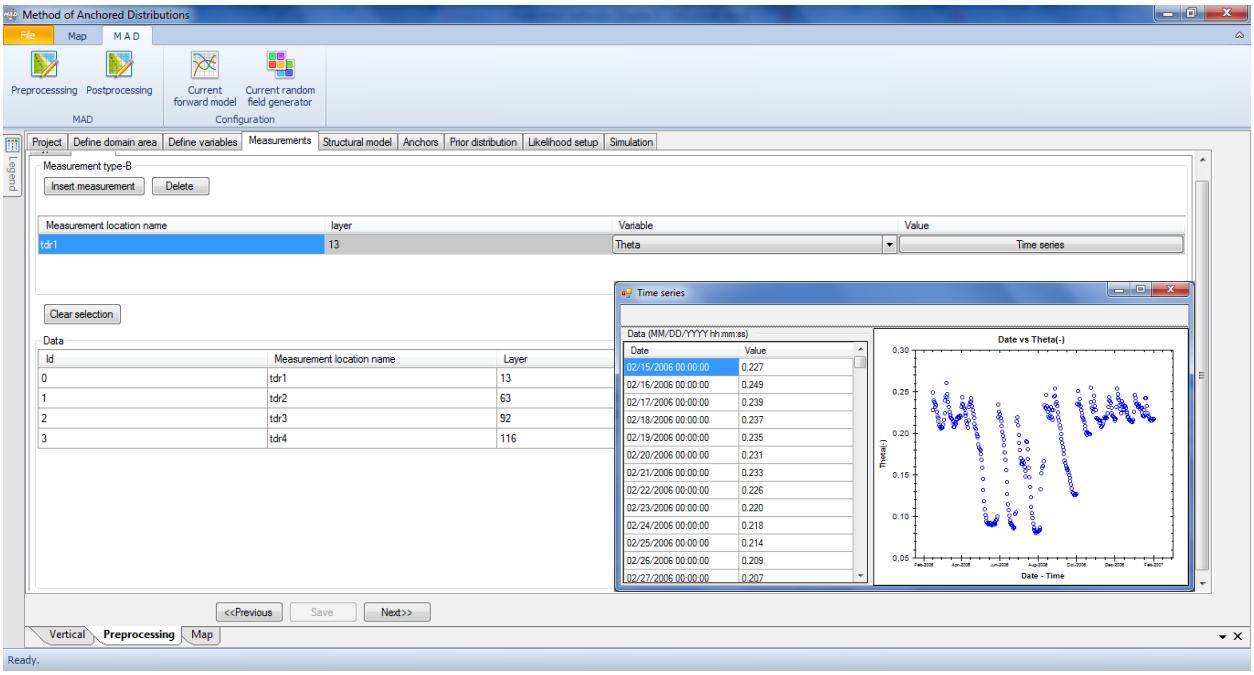

Figure 3.4: The *Measurements* form and the measurement time series editor of the MAD software pre-processing module.

The next form is used to define the structural model and any transformations of the parameter fields.

3.1.2.5 The "Structural model" Form

Figure 3.5 shows the *Structural model* form of the MAD software. There are three critical responsibilities of this form, which are numbered in the graphic: 1) definition of any transformations, 2) configuration of the structural models, and 3) identification of random structural parameters.

Obviously, a single structural model cannot represent all the different possible physical parameter fields. This form allows each parameter field to have a unique structural model definition. This form is sequenced after the *Define variables* form, such that the parameter field or fields to be characterized are already selected before structural model configuration. Additionally, this form is sequenced after the *Measurements* form such that users do not have to transform their Type-A data before upload.

This form defines several key elements of the MAD application: the structural model $S(\boldsymbol{Y};\boldsymbol{\theta})$; the structural parameters that are treated as random variables; and, if required by

the application, the transformations of $\boldsymbol{Y}$, respectively $\boldsymbol{Y}^t$. Note that when a transformation is applied, the structural model should be configured to generate random realizations of the transformed physical parameter. The MAD software will automatically apply the inverse transformation to the generated parameter fields before using them as input to the forward model.

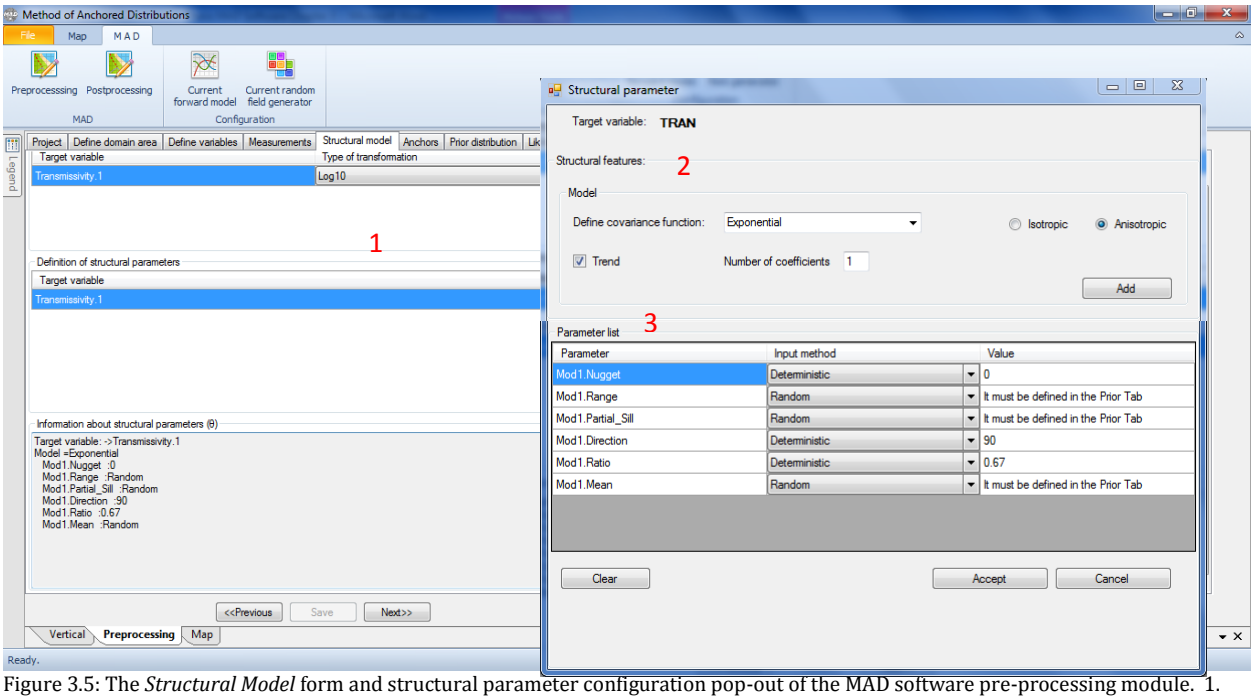

Figure 3.5: The *Structural Model* form and structural parameter configuration pop-out of the MAD software pre-processing module. 1. Tool for transformations. 2. Tool for choosing structural model. 3. Tool for randomizing parameters of structural model.

Because a structural model can be used to represent either heterogeneous physical parameters fields (seen previously in Chapter 2 – e.g. geostatistical model) or block by block physical parameters (coming in Chapter 4 – e.g. classical statistical model); this form adapts to the random field generator connected to the MAD software. The inset window in Figure 3.5 shows a variation of the structural parameter configuration pop-out for a geostatistical random generator. Figure 3.6 shows the structural parameter configuration pop-out for a purely statistical random generator. For classical statistical models, there is additionally a grouping tool that allows multiple physical parameters to jointly be assigned to a multivariate structural model.

There are three key user-friendly features in this form. First, the MAD software manages and applies all transformations and inverse transformations that must be applied before or after random field generation. Second, the form will not permit configuration of any structural models that are not supported by the connected random field generator. Finally, the definition of structural model parameters as random or deterministic is very straightforward.

For reasons described in Section 1.2, this form is only accessible when a spatially compatible random field generator is connected to the MAD software. Additionally, because anchors are PDFs of a given physical parameter, this form must follow the *Define variables* form. This form defines the vector $\boldsymbol{\vartheta}$ for the MAD application.

As mentioned in a few earlier sections, all spatial information is automatically added to the map. After anchors are defined, a complete catalog of the spatial information in a MAD application is available. Figure 3.8 shows the map tab of the MAD software from an example case study application. The map contains the discretization along with the locations of anchors, Type-A and Type-B measurements – the icons for each are listed in the legend on the left hand side.

The user friendly features of this form include the simple interface for defining anchor type and locations as well as the option to overlay a grid of anchors onto the domain.

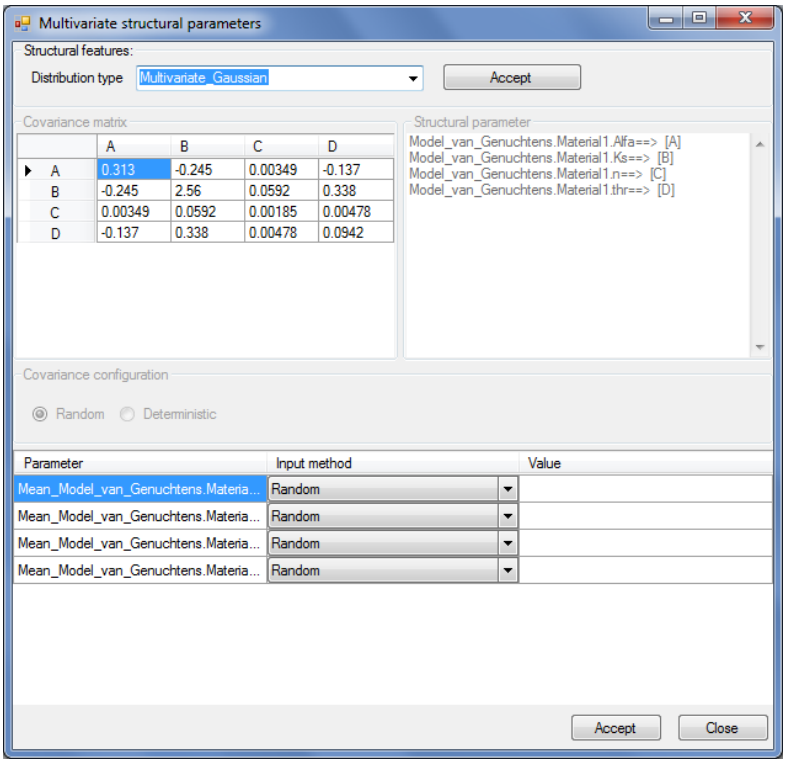

Figure 3.6: The structural parameter configuration pop-out of the MAD software pre-processing module for statistical structural models.

The next form is used to define anchors.

3.1.2.6 The "Anchors" Form

Figure 3.7 shows the *Anchors* form of the MAD software. The key responsibility of this form is the definition of the anchor locations and types.

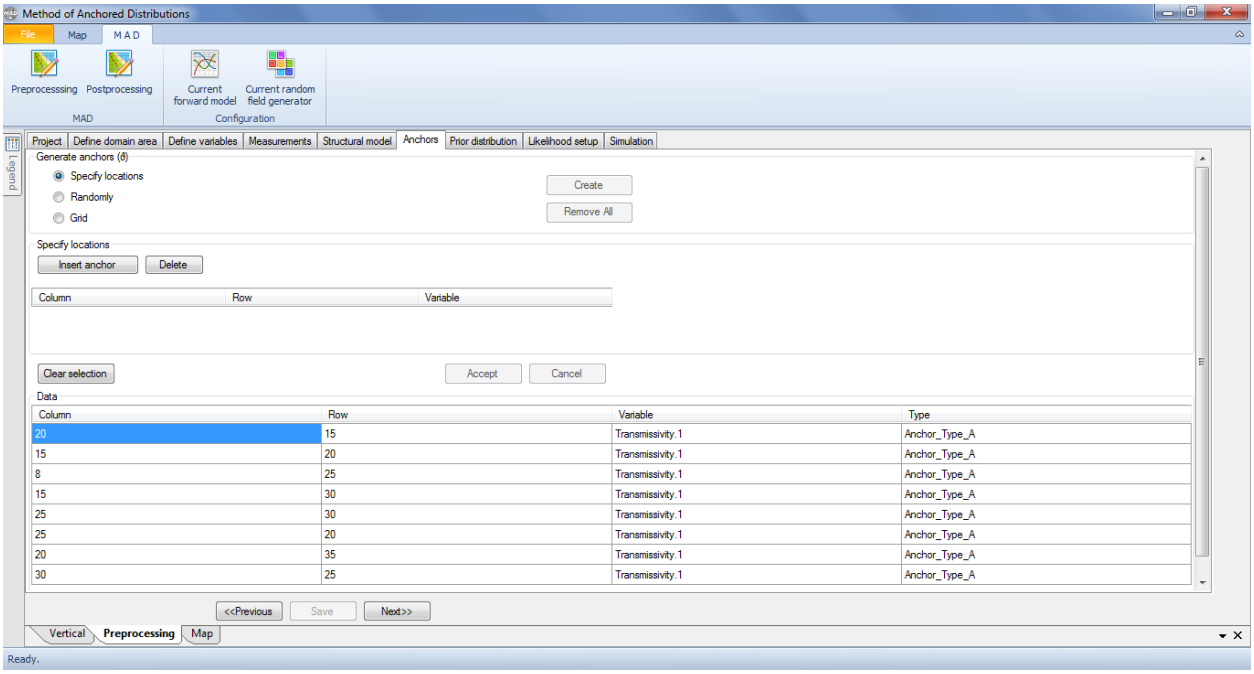

Figure 3.7: The *Anchors* form of the MAD software pre-processing module.

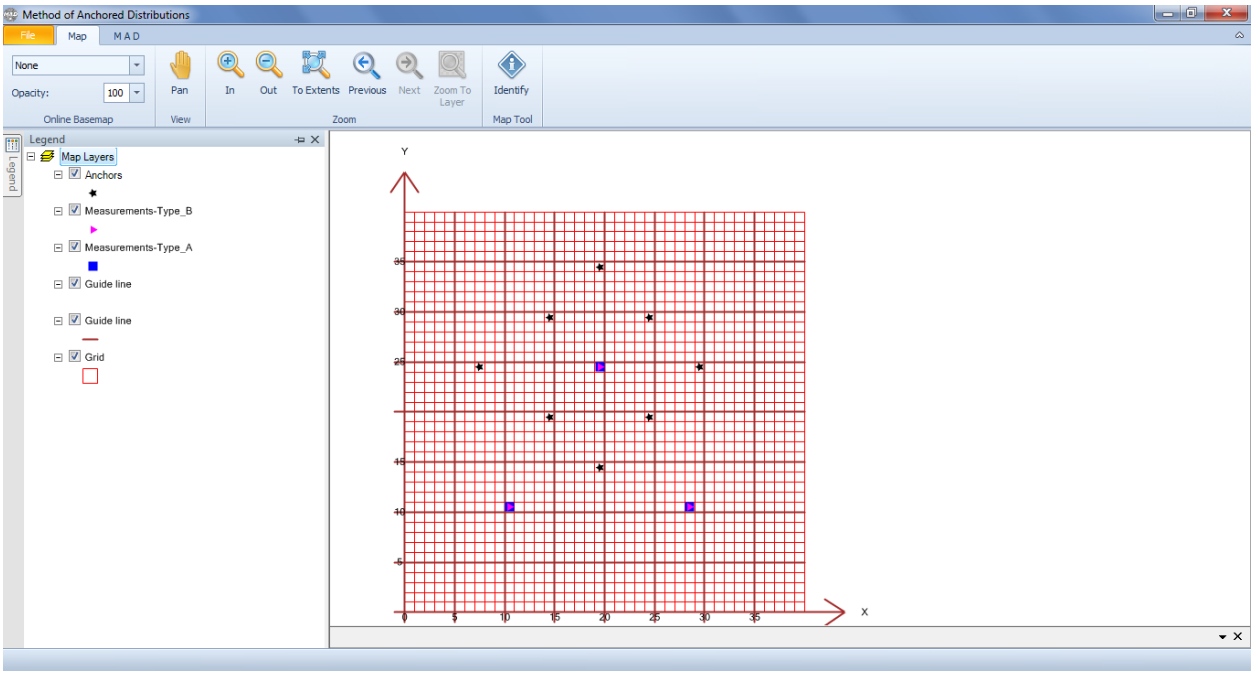

Figure 3.8: The map tab of the MAD software pre-processing module.

The next form is for upload of samples drawn from the joint prior PDF.

3.1.2.7 The "Prior distribution" Form

Figure 3.9 shows the *Prior distribution* form of the MAD software. The key responsibility of this form is to manage samples from the joint prior PDF. Note that this form is not actually used to define a PDF, which is simply a non-negative function that integrates to unity, because of the immense number of functions that can satisfy these requirements. Additionally, there are myriad methods, such as random, stratified, or Latin Hypercube sampling [*McKay et al.,* 1979] for drawing samples from a probability distribution. Therefore, it is left entirely to the user to define a function for the prior joint PDF of the appropriate random anchors and structural parameters and select an optimal sampling approach external to the MAD software.

This form requires knowledge of all the random structural parameters and anchors; therefore, it must follow the *Structural model* and *Anchors* forms. Additionally, this form must precede any simulation or random field generation process, because as shown in Figure 1.1, these procedures require the samples from the prior joint PDF.

This form is used to upload the matrix $\boldsymbol{D}$, which is of dimension $R \times P$, where it is reminded from Section 1.1.3 that R is the number of samples and that P is the number of anchors plus the number of random components in $\boldsymbol{\theta}$. The MAD software then regresses the prior joint PDF non-parametrically using kernel density methods [*Scott & Sain*, 2004], such that an approximate value for $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$, $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, $\hat{f}_R(\boldsymbol{\theta}|\boldsymbol{z}_a)$, $\hat{f}_R(\boldsymbol{\vartheta}|\boldsymbol{z}_a)$, $\hat{f}_R(\boldsymbol{\theta})$, or $\hat{f}_R(\boldsymbol{\vartheta})$ is attained from $\boldsymbol{D}$ depending on the application. The MAD software performs this computation using the package *np* from the R statistical computing project [*Hayfield & Racine,* 2008], using a fixed bandwidth calculated from $\boldsymbol{D}$ using the rule of thumb of *Silverman* [1986] and a second order Gaussian kernel.

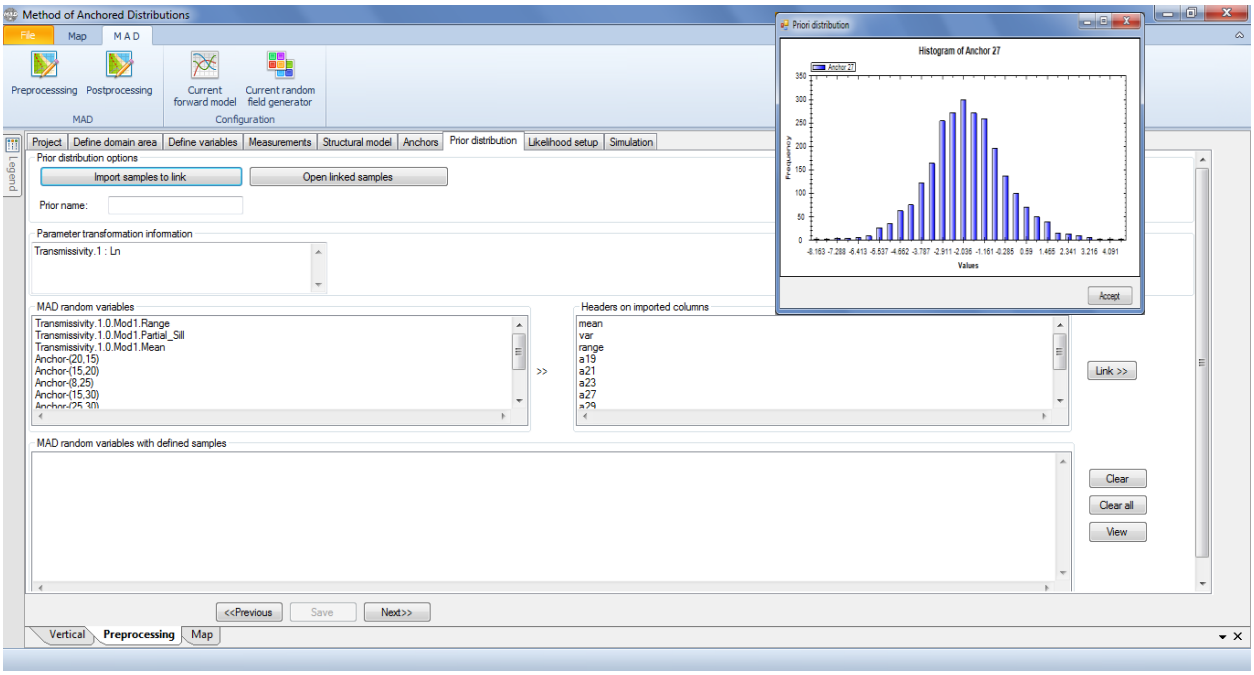

Figure 3.9: The *Prior distribution* form of the MAD software pre-processing module.

It is important to note that, if there is conditioning on Type-A data in the prior joint PDF, the user is expected to incorporate this externally to the MAD software. However, the MAD software does differentiate between cases that utilize a combination of structural parameters and anchors from those where just anchors or structural parameters are employed. Finally, it is also suggested to set R artificially large, as it will improve the quality of the approximation for the values of the joint prior PDF.

There are 3 user friendly features of this form. First is the linking tool for matching columns of an uploaded text file of samples to the random anchors and structural parameters specified in the two previous MAD software forms. Second is the histogram viewer of samples, shown in the inlaid pop-out in Figure 3.9. Third is the list of the parameter transformations specified in the *Structural model* form. The summary of transformations is especially important, because as mentioned in Section 3.1.2.5, the structural model should be configured to generate random realizations of the transformed physical parameter field. This can lead to marked changes in terms of the supports of the random variables, for instance consider the expected value of a hydraulic conductivity field is thought to be between 10^{-3} and 10^{3}, but after a $\log_{10}$ transformation is applied the expected value should be between -3 and 3.

The next form is used for aggregation and subset selection of the Type-B data.

3.1.2.8 The "Likelihood setup" Form

Figure 3.10 shows the *Likelihood setup* form of the MAD software. The key responsibility of this form is the definition of the vector $\boldsymbol{z}_b$ for use in the conditioning of the random structural parameters and anchors. For steady-state applications, this form manages selection of the measurement locations to be utilized for conditioning. For transient applications, this form additionally manages the selection of any subset from or polynomial regression on intervals of the measurement time series.

This form only depends on information uploaded in the *Measurements* form, but must precede any simulation process. It is essential to fully define the Type-B vector before simulation, as this informs the MAD software exactly what data to extract from each run of the forward model.

There are two user-friendly features to highlight in this form. First, the form allows for easy management of time series data at each measurement location. Users can select perhaps to include key events in the time series; to exclude events that are spurious or are the result of high signal noise; or with a single click choose to utilize the entire time series. The dynamic graph highlights the selected measurement times. Second, the form allows for simple configuration of a regressed representation of the time series. The polynomials can be regressed over any combination of intervals, which can overlap. The dynamic graph overlays the time series with the regression fits and the coefficients are added to the "Aggregate Data" list for selection into the Type-B vector, which is shown in Figure 3.11. The purpose of such aggregation is to limit the dimensionality of the likelihood function, which, as discussed in Chapter 2, has computational cost implications.

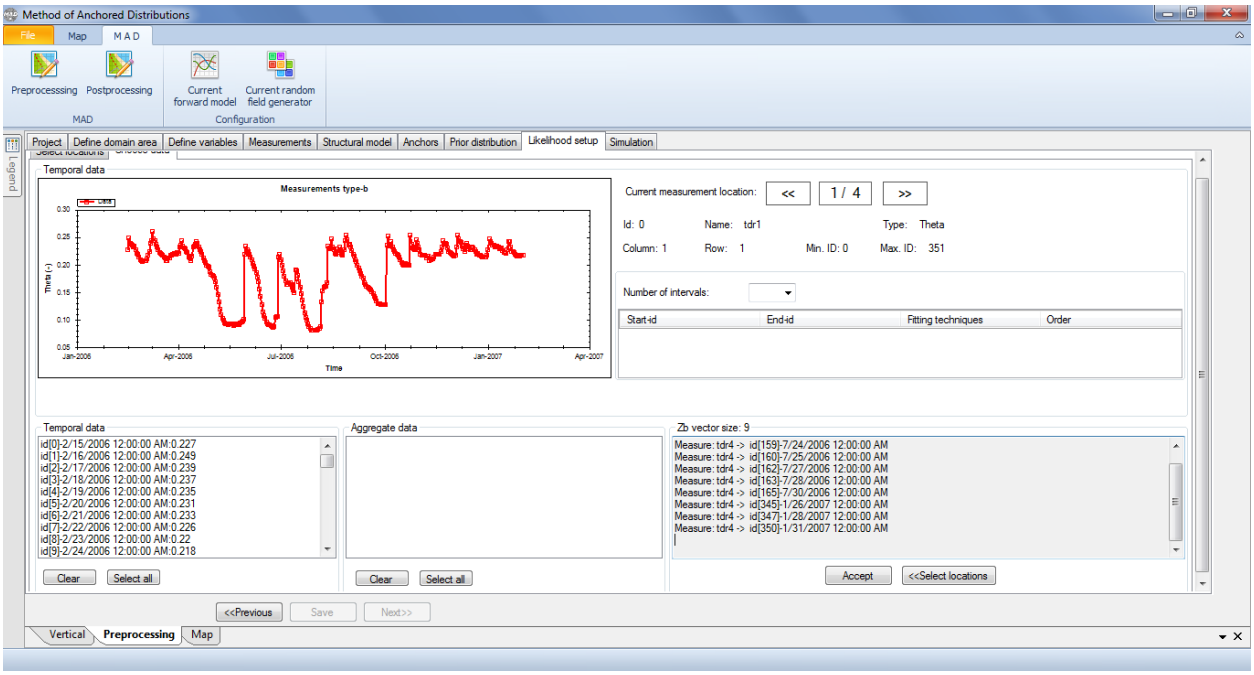

Figure 3.10: The *Likelihood setup* form of the MAD software pre-processing module.

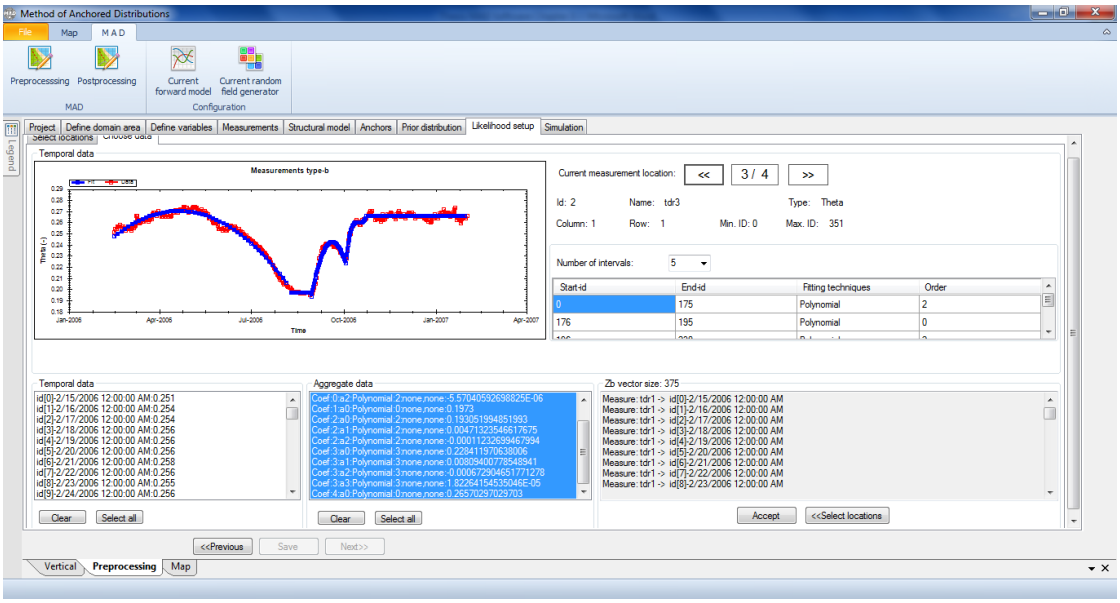

Figure 3.11: Interval polynomial regression of measurement time series.

The next form is used to execute and monitor the simulation process.

3.1.2.9 The "Simulation" Form

Figure 3.12 shows the *Simulation* form of the MAD software. The primary responsibility of this form is to execute simulations. Recalling the flowchart in Figure 1.1, this form internally performs steps 3-5, which are the most demanding computational and data management steps of implementing MAD.

This form must come at the end of the pre-processing module, because it utilizes information from every preceding form to first generate random realizations of the physical parameter fields and second simulate the appropriate Type-B data using the forward model on these realizations.

This form generates the simulation ensembles $\boldsymbol{Z}_b$ for a user specified amount of samples, shown as 1 to 100 in Figure 3.12. Each simulation ensemble is written into its own database, which significantly improves the computational efficiency of the likelihood calculation during post-processing. This form also identifies the calendar date of the first time output of the forward model, i.e. $t = 0$, such that the measurement time series are synchronized with the simulated time series. Finally, this form establishes N the simulation ensemble size.

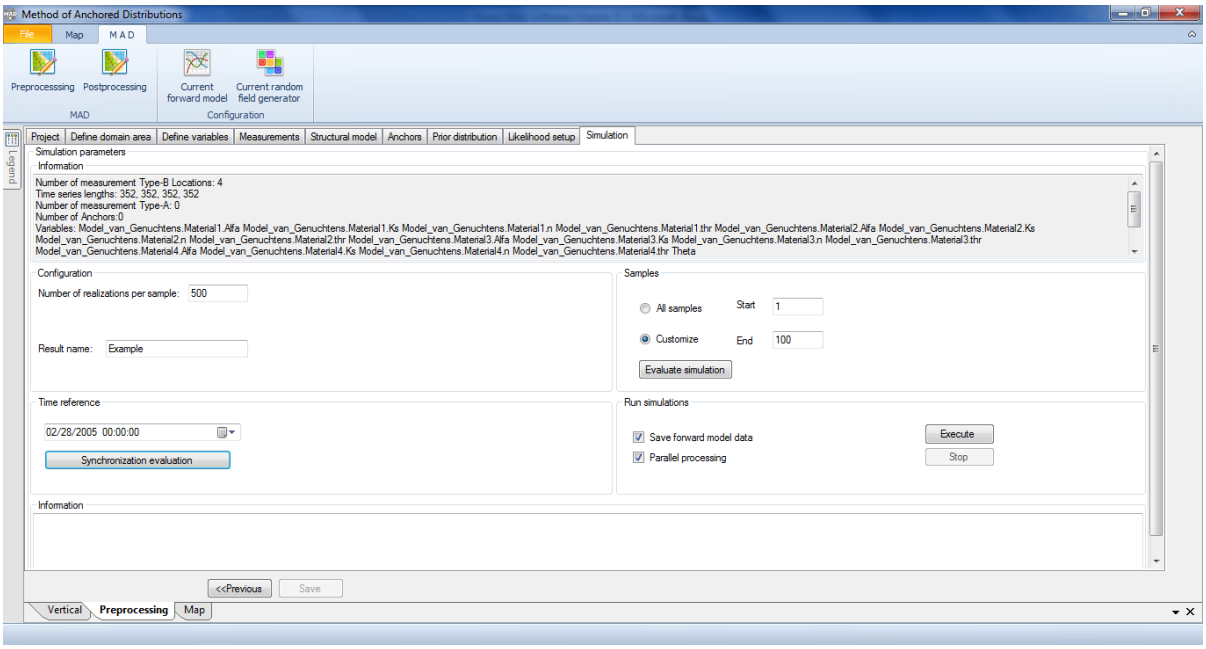

Figure 3.12: The *Simulation* form of the MAD software pre-processing module.

There are several user friendly features in this form. First, the range of samples the user wants to generate simulation ensembles is customizable. Second, there is an option to distribute the simulation and random field generation processes over all the processors of a user's computer, which improves computational speed for MAD projects. Third, the user manually defines N. This value need not be static, but rather it can be increased if necessary following convergence analysis in the post-processing module. There is no loss of data by increasing N, the databases are simply expanded and the additional simulations are appended. Finally, there is an option to extract the complete time series at every measurement location, even if it is not required by $\boldsymbol{z}_b$ for likelihood analysis, which can be convenient if users later decide to expand their Type-B vector, or want to perform different aggregation techniques that require the full time series – e.g. temporal moment calculation - not offered in the GUI.

The *Simulation* form concludes the pre-processing module of the MAD software and the next form begins the post-processing module.

3.1.2.10 The "Project" and "Data organization" Forms

This section contains two forms, because the *Project* form of the post-processing module is almost an exact replica of the same form in the pre-processing module (refer to Figure 3.1), with the exception that projects can only be opened. The main purpose of the *Project* form is to connect the MAD software with the appropriate simulation ensemble databases.

Figure 3.13 shows the *Data organization* form. This form has one key responsibility: the selection of all subsets of the Type-B vector for likelihood calculation. This permits

comparison of many different variations of the likelihood function, which may be used to limit dimensionality-related computational cost, but maintain effective conditioning of the structural parameters and anchors. Permitted options for variation of the Type-B vector at this point do not require re-execution of the simulation processes, because only the previously extracted data populates the list in this form.

The key user-friendly aspect of this form is the output data viewer, which can be used to graphically check the simulation results. This form calculates standard deviations and expected values of the simulated data when the application is steady-state and graphs these quantities in addition to the measured times series in transient applications. The graphic representation is shown in the inlaid pop-out in Figure 3.13.

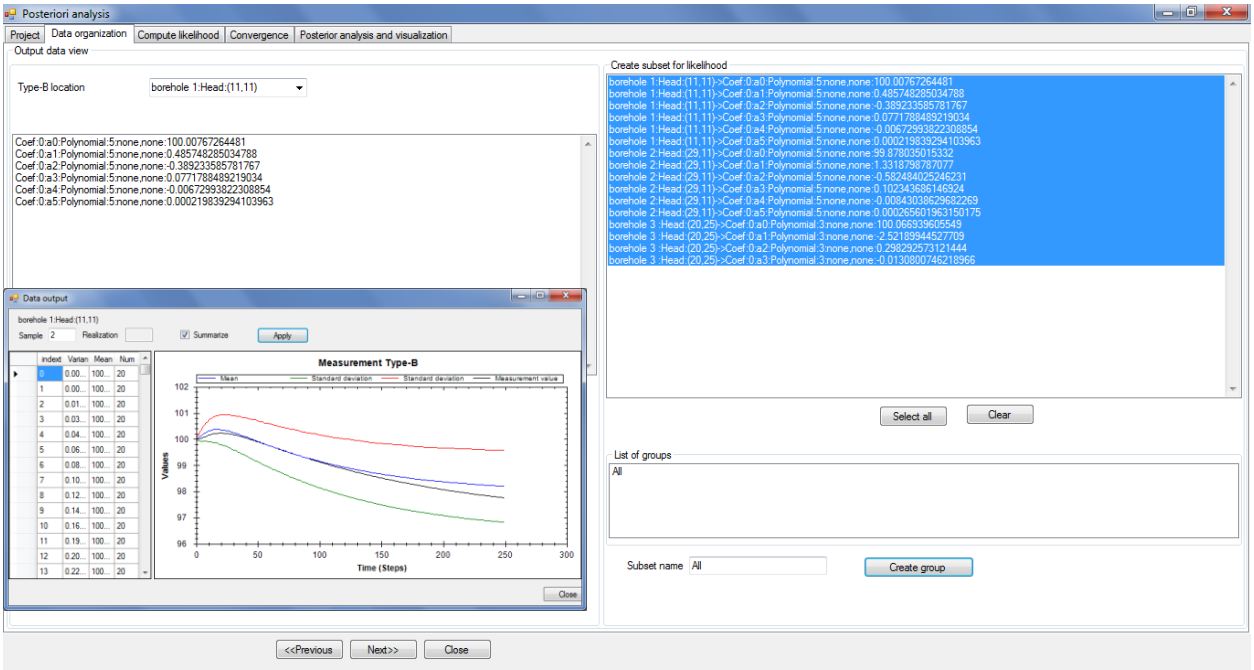

Figure 3.13: The *Data organization* form of the MAD software pre-processing module.

The next form is used to configure the calculation of the likelihood.

3.1.2.11 The "Compute likelihood" Form

Figure 3.14 shows the *Compute likelihood* form of the MAD software. The key responsibility of this form is to compute the likelihood for the samples using the simulation ensembles. Depending on the nature of the application, this function has many possible formulations: $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$, $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$, $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{z}_a)$, $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$, $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i)$, $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\vartheta}_i)$ for $i \subseteq [1, \dots, R]$. An additional important role of this form is the variation of number of samples (convergence of the posterior joint PDF) and number of realizations (convergence of the likelihood function).

The likelihood function calculation in the MAD software is performed non-parametrically using kernel density estimation methods that are similar to those described in Section 3.1.2.7 for the prior joint PDF, with the exception that the bandwidth and regression utilize $\boldsymbol{Z}_b$ not $\boldsymbol{D}$.

The most apparent user-friendly element of this form is the catalog of computed likelihood functions. A less intuitive user-friendly feature is the management of loading the simulation ensembles, non-parametrically regressing and then evaluating the likelihood function joint PDF, and repeating this process rapidly for multiple samples.

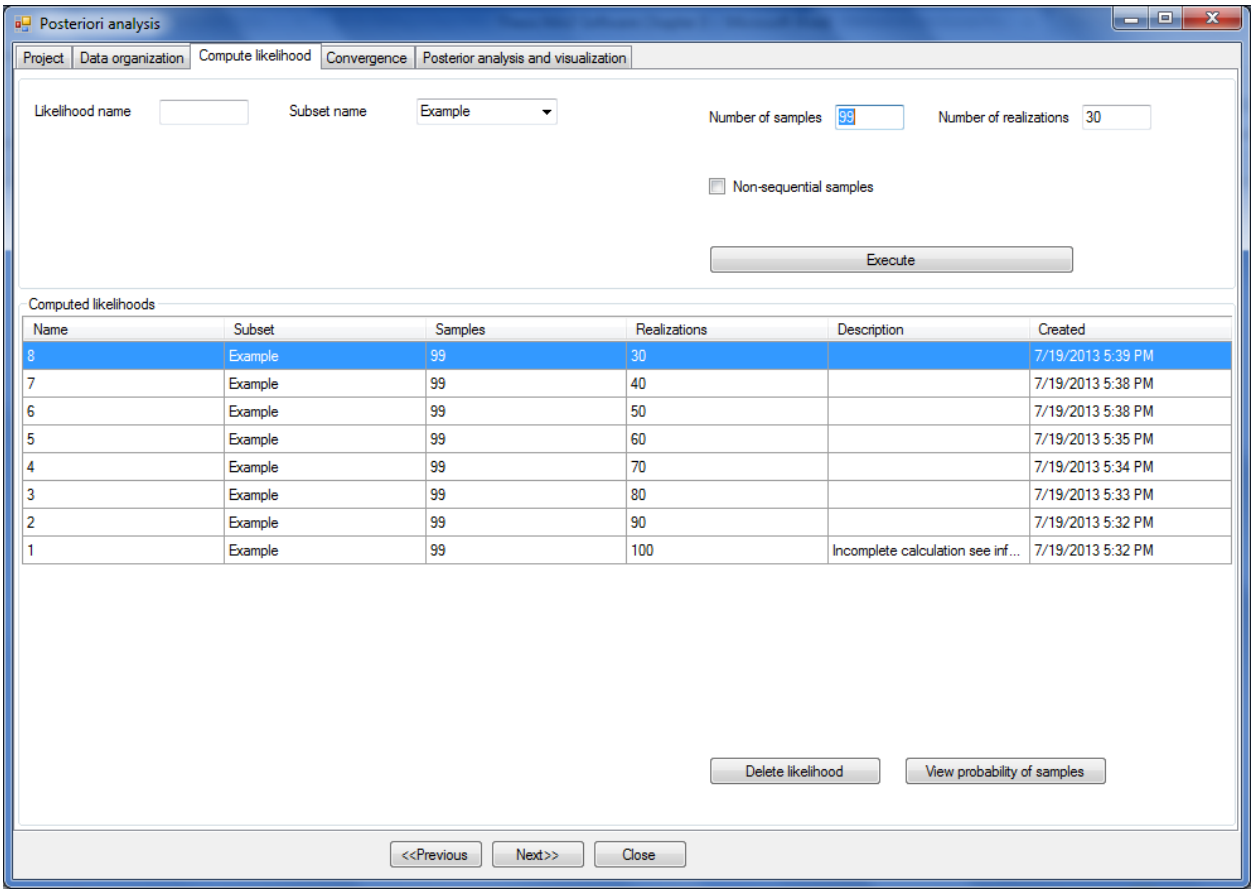

Figure 3.14: The *Compute likelihood* form of the MAD software pre-processing module.

The next form utilizes the variations in N to visually assess the convergence of the likelihood function.

3.1.2.12 The "Convergence" Form

Figure 3.15 shows the *Convergence* form of the MAD software. The key responsibility of this form is providing visual diagnostics for assessing the convergence of the likelihood function. The form offers two different visualization approaches. The first is a comparison of the likelihood function values as a function of N for individual samples, i.e. $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$ vs N where each value of i yields a separate line. The second is a comparison of two likelihood function values using two different ensemble sizes of N_1 and N_2 for all samples, i.e. $\hat{f}_{N_1}(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$ vs $\hat{f}_{N_2}(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$ where each values of i yields a single point and all i are plotted. These features are shown in the upper and lower panes of the form respectively in Figure 3.15.

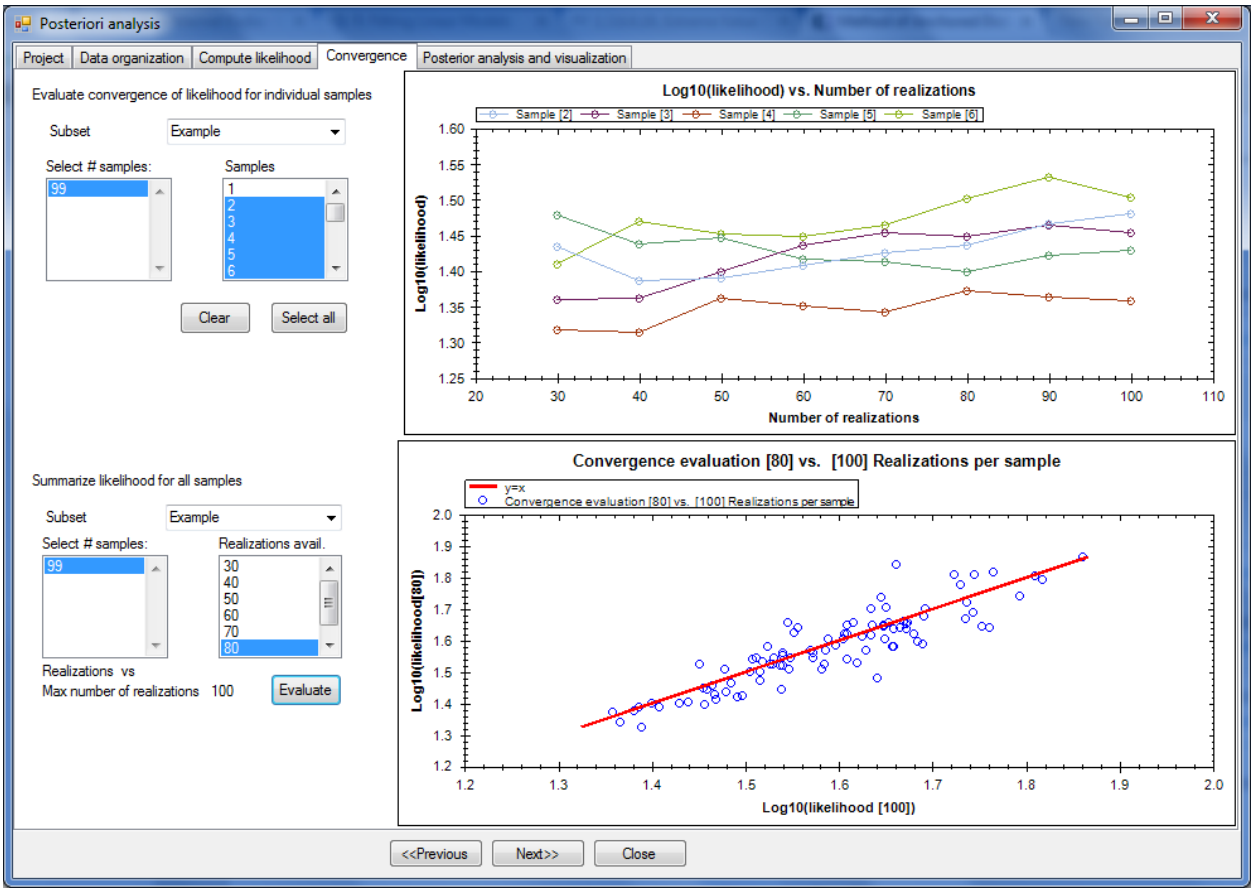

Figure 3.15: The *Convergence* form of the MAD software pre-processing module.

This form is used to evaluate the convergence of the likelihood function, so obviously it must come after the *Compute Likelihood* form where the likelihood function is computed. In Figure 3.15, a reasonably converged likelihood function is displayed by both panels. In cases, where the individual samples (top panel) do not exhibit relatively horizontal behavior as a function of large N, users should return to the pre-processing module and increase the size of the simulation ensemble. Likewise, if the scatterplot does not group around the reference line, $\hat{f}_{N_1}(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a) = \hat{f}_{N_2}(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$, the user should also

increase the simulation ensemble size. Finally, note that the lower pane in this form is intended to be a summary plot of the upper pane, which is most useful when the number of samples grows large. Both forms always graph any value of the base-10 logarithm of any likelihood function values.

This form is extremely important, because the posterior joint PDF cannot be reliably calculated without convergent likelihood function.

The next form is used to calculate the posterior joint PDF.

3.1.2.13 The "Posterior analysis and visualization" Form

Figure 3.16 shows the *Posterior analysis and visualization* form of the MAD software. There are two principle responsibilities of this form: 1) the calculation of the posterior joint PDF values for the samples using a given configuration of the likelihood function and 2) generation of graphics that show marginal distributions from the posterior joint PDF and the prior joint PDF.

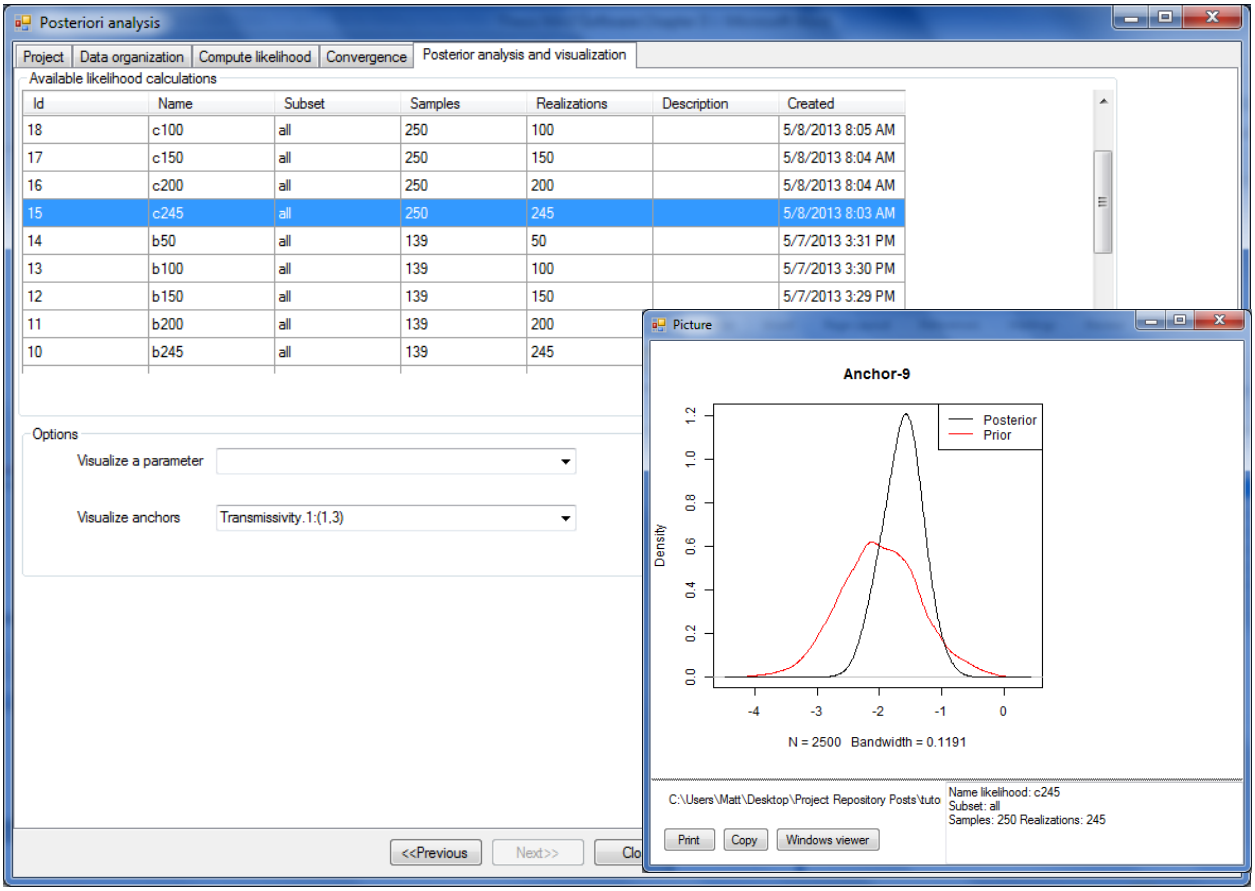

Figure 3.16: The *Posterior analysis and visualization* form of the MAD software pre-processing module.

Based on Equation 1.2, it is clear that this form must follow the *Compute likelihood* and the *Prior distribution* forms. This form calculates the estimates of $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b)$, $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_b)$, $\hat{f}_R(\boldsymbol{\theta}|\boldsymbol{z}_a, \boldsymbol{z}_b)$, $\hat{f}_R(\boldsymbol{\vartheta}|\boldsymbol{z}_a, \boldsymbol{z}_b)$, $\hat{f}_R(\boldsymbol{\theta}|\boldsymbol{z}_b)$, or $\hat{f}_R(\boldsymbol{\vartheta}|\boldsymbol{z}_b)$ depending on the nature of the application.

The marginal distributions are not determined by a numeric integration approach, such as

$$f(x_1) = \int f(x_1, x_2) dx_2 \tag{3.1}$$

but by following the approach of *Wasserman* [2010], which is a three step approach applied as follows: 1) randomly re-sample the parameters with replacement from their posterior joint PDF, 2) collect the marginal components from each of the random draws, 3) apply a univariate kernel density algorithm to regress the marginal distribution from the marginal components.

The user-friendly feature of this form is the automatic creation of the graphics as files in the project directory as well as a pop-out table that shows the computed values of the posterior joint PDF for the samples. This form concludes the MAD software post-processing module.

Over the course of these 13 sections, there have been a few goals: 1) to identify the MAD theory connected to each form of the MAD software, 2) to explicitly state implementation procedures when appropriate, 3) to show the important dependencies between the various forms and the logic of the sequence of each module, 4) to highlight user-friendly features, and 5) to present the flexibility of the MAD software in terms of the possible formulations of MAD it supports.

A very comprehensive user manual that categorizes the inputs required of users for each form is available at https://mad.codeplex.com/.

In the next section, the perspective shifts to the adaptability of the MAD software to interact with other software and transfer to applications in other scientific fields.

3.2 MAD Architecture for General Inversion Applications

As mentioned in Section 3.1, the MAD software is composed of two modules: post-processing and pre-processing. In this section, the software is presented on a very coarse-scale with a focus on how the modules interact with external programs for random field generation and forward modeling.

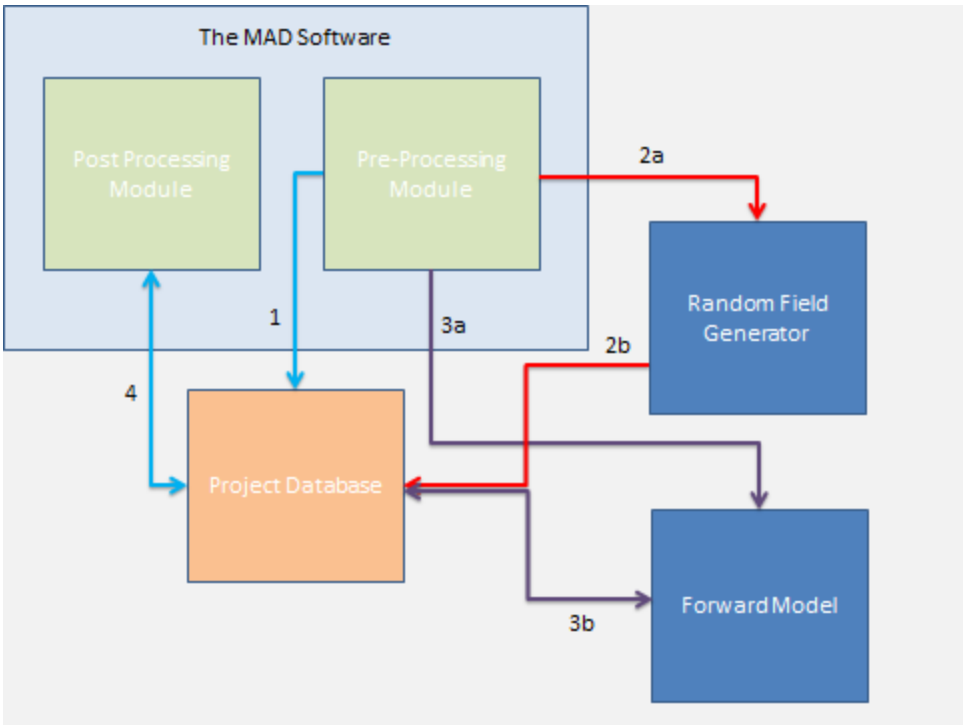

Figure 3.17: Schematic of the MAD software connections to the random field generator, forward model, and project database.

Figure 3.17 is a schematic of how the MAD software is designed to interact with other modeling tools - the colors of each arrow or block are indicators. For the blocks: light green and light blue correspond with the MAD software and its modules; dark blue corresponds with external modeling tools required to implement MAD; orange corresponds with the project database, even though it is created by the MAD software, a separate database is constructed for every new application. For the arrows: red corresponds with a communication pathway involving the random field generator; purple corresponds with a communication pathway involving the forward model; and blue corresponds with a communication pathway involving the project database. Note that some arrows are double-sided and some are one-directional, which graphically indicates the possible direction(s) of information transfer on the pathway, i.e. the pre-processing module sends information to the forward model, but does not receive any. Finally, note the numbering next to the arrows, which indicates the sequence in which each connection is utilized.

The sequence is discussed here briefly. First, fundamental instructions for random field generation process are defined in the pre-processing module and written into the database. Next, the random field generator is configured for a specific sample from the prior joint PDF as well as executed N times (2a) and the realizations of the parameter field are temporarily committed to the database (2b). Next, serially utilizing the N realizations, the forward model is executed (3a) and the Type-B data - at the appropriate times and locations - is written permanently to the database (3b). Finally, the Type-B data is sent to the post-processing for likelihood analysis and the value of the posterior is written to the database. This process is then repeated for all the samples. Note that the project database created by the MAD software is an important, central repository for storing all relevant

project information; the MAD modules, the random field generator, and forward model all communicate with it. Before discussing the flexibility of the modular architecture and sequence, it should also be clearly stated that the MAD software controls the execution of the random field generator as well as the forward model and this requires no user interaction.

The purpose of this architecture and sequence are to maximize the flexibility of the framework. First, the six formulations of MAD presented in Section 1.3.1 all require a random field generator and a forward model to be connected with the MAD software modules, hence this architecture supports all the formulations. Second, the pre-processing module communicates with both external tools, but the post-processing module communicates with neither, which means all the analytical functions in the post-processing module can be easily replaced or modified so long as they revised versions can communicate with the project database, which is written in SQLite. Third, the communication between the pre-processing module, the project database, random field generator, and forward model are facilitated by connections called 'drivers' (red and purple arrows), which means nothing specific to any one random field generator or forward model is coded into the MAD software. This allows for isolation of the core generalized Bayesian framework from application- or forward model-specific details. Moreover the drivers promote the interchangeability of the random field generator and forward model, because any tool can be substituted into the sequence defined above, provided the communication pathway for that specific tool is used.

The drivers are the key to transferability of the MAD software between different scientific fields. Instead of users having to write code to implement MAD from start to finish, they simply write the code to connect appropriate random field generators or forward models for their application. The drivers have simple responsibilities: reading, writing, and interpreting the format of input and output files as well as command line execution of the forward model or random field generator. Thus, creating a driver is significantly less programming effort than creating the entire MAD framework for a given scientific field.

Thus far, drivers exist for two forward models and two random field generators: PMWIN 5.3 [*Chiang & Kinzelbach,* 2001]and HYDRUS 4.14 [*Simunek et al.,* 1998] as well as GSTAT [*Pebesma & Wesseling,* 1998] and R Statistics [*R Core Team,* 2012] respectively. In the next section and next chapter, the MAD software will be used, in conjunction with all four drivers, to conduct multiple case studies.

3.3 Synthetic Case Studies Using the MAD Software

Now, the focus shifts from macroscopic architecture of the MAD software and specific design elements of the GUI to the presentation of a series of applications using the MAD software that highlight its flexibility.

This section features case studies using the MAD software to implement the various formulations of Bayes' rule presented in Equations 1.2, 1.3, and 1.5-7. The MAD software is used in Chapter 4 to implement Equation 1.4 so it is not presented here. Each of the 5 studies presented utilized PMWIN 5.3 as a forward model and GSTAT as a random field generator. The case study is a 1-dimensional problem used to illustrate the conditioning of structural model parameters and anchors in different scenarios. It will be noted in the various subsections what elements of the case study are held fixed: whether structural parameters, anchors, or both are being characterized and also whether Type-A data is included with Type-B data in the conditioning. A comprehensive overview is presented initially and then every subsection will identify the formulation of MAD being studied and discuss the pertinent results.

3.3.1 Case Study Overview

The objective of the experiment is the characterization of the transmissivity T in a strip of confined, saturated aquifer. Available data includes steady-state, error-free pressure head $h(x_b)$ measurements, shown in Figure 3.18 (blue squares), where x_b is the location of the Type-B measurements; error-free measurements of the $T(x_a)$, shown in Figure 3.18 (pink triangles), where x_a is the location of the Type-A measurements; as well as highly reliable estimates of the variogram model parameters and trend of the transmissivity field $\boldsymbol{\theta}$, but these parameters will not always be treated as deterministic in the following Sections. Also shown in Figure 3.18, are the localities at which anchors $\boldsymbol{\vartheta}$ are placed x_ϑ, but these are only relevant to the subsequent studies that involve anchors (black stars). The boundary conditions are known with high confidence and consist of a constant pressure head boundary and a constant flux boundary on the respective ends of the aquifer strip. The transmissivity field is also assumed stationary, ergodic, and its spatial correlation decays exponentially with lag distance.

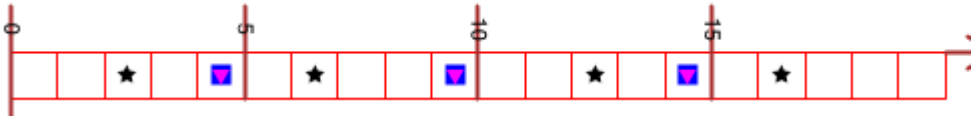

Figure 3.18: Discretized physical domain of the case studies including positions of anchors as well as Type-A and Type-B measurements.

Whenever the objective is local characterization, which requires employing anchors, or the goal is conditioning on Type-A data, it is important that the structural model is, by definition, compatible with spatially dependent data. Therefore, because one or both of these criteria is satisfied in all five studies, the structural model is always geostatistical for the remainder of the chapter. Finally, the mathematical relationship between h and T is the 1-dimensional, steady-state flow equation - a second order ODE, which requires the two boundary conditions listed above.

Table 3.3 lists out the available data from the aquifer strip and details of the case study.

Table 3.3: List of case study details.

Type-B Measurements	
h (m)	x_b (m)
9.59	5
9.40	10
9.25	15
Type-A Measurements	
T (m^2/hr)	x_a (m)
0.0055	5
0.1258	10
0.0137	15
Anchor Parameters (Synthetic Values)	
$\log_{10}(T)$ [dimensionless]	x_ϑ (m)
-1.963	3
-1.316	7
-1.575	13
-2.699	17
Structural Model Parameters for $\log_{10}(T)$ (if deterministic)	
Parameter	*Value*
Nugget (always fixed)	0 [dimensionless]
Sill	0.4 [dimensionless]
Trend	-2 [dimensionless]
Range	3 [m]

As a reminder, the marginal prior PDFs graphed in the following five subsections will often appear to be "rough" compared to their definitions (usually Gaussian or uniformly distributed), but this is because the MAD software reconstructs the prior PDFs from a limited set of samples provided by the user.

3.3.2 Anchor Conditioning Using Type-B Measurements Only

The goal of this case study is local characterization at the anchor locations, using just the available Type-B data and treating the structural model as deterministic. Type-A data is excluded in this study. Therefore, the appropriate formulation of MAD is given in Equation 1.5, which is repeated here for ease of reading

$$f(\boldsymbol{\vartheta}|\boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta}). \qquad \text{(1.5 repeated)}$$

In this case, $P = 4$ and $M = 3$. The joint prior PDF for the anchors $f(\boldsymbol{\vartheta})$ is assumed to be the product of independent Gaussians distributions with the mean (trend in Table 3.3) and variance (sill in Table 3.3) as defined in Table 3.3. The assumption of independent anchors – even within a spatially correlated field – is defensible as the least subjective choice of prior by minimum relative entropy (MRE) [*Woodbury & Ulrych,* 1993, 1998; *Hou & Rubin,* 2005]. Samples were drawn from the joint prior PDF randomly.

Convergence of the non-parametrically calculated 3-D likelihood function $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\vartheta}_i)$ $i = 1, \ldots, R$ is achieved with 150 realizations per sample. The number of samples necessary to stabilize the regression of the 4-D posterior joint PDF $\hat{f}_R(\boldsymbol{\vartheta}|\boldsymbol{z}_b)$ is 300. Plots of the convergence are omitted as they are unnecessarily repetitive: refer to Figure 3.15 or Figures 2.7 & 2.8 for convergence graphics and their accompanying discussion. The $NR = 45{,}000$ total realizations and simulations took an average of 0.86 seconds each on Intel Core 2 Duo CPU E8400 3 GHz Processor for a total run time of just under 11 hours.

Figure 3.19 shows the comparisons of the posterior (blue) and prior (red) marginal PDFs for the four anchor locations characterized as well as the synthetic true values (black vertical lines) of the transmissivity field at the locations. The upper left pane shows the anchor at $x = 3$, upper right pane shows the anchor at $x = 7$, lower left pane shows the anchor at $x = 13$, and lower right pane shows the anchor at $x = 17$. In general the conditioning is very successful: the posterior marginal PDFs are generally narrower (of lower variance) and are more centered on the true values (less biased) than the prior marginal PDFS, with the exception of the anchor at $x = 17$. The results indicate a strong link between the transmissivity field values at the anchor locations and the head measurements at the Type-B locations.

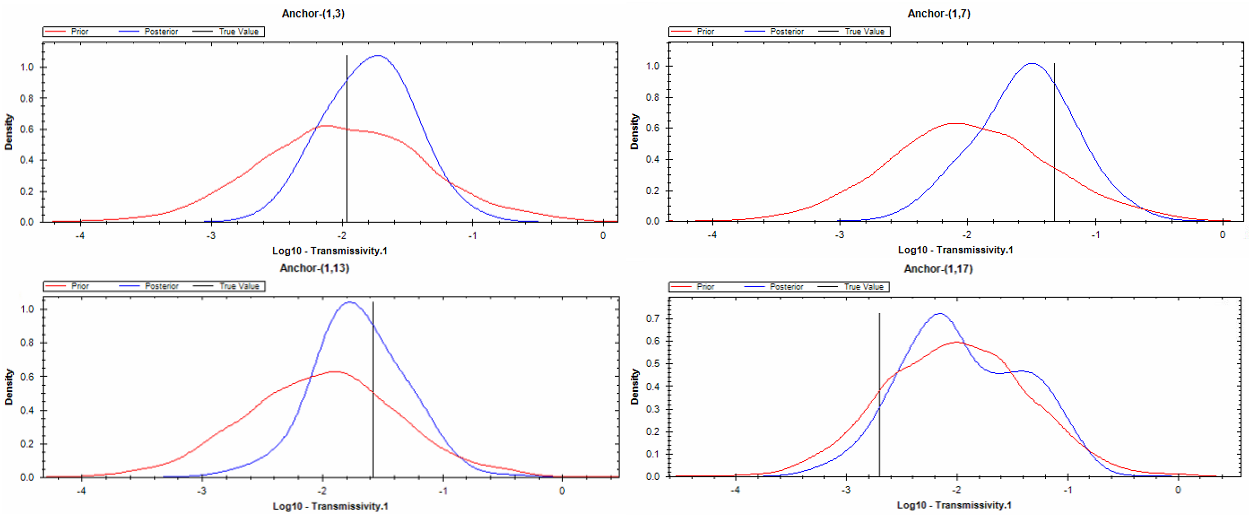

Figure 3.19: Posterior (blue) and prior (red) marginal PDFs for anchors conditional on Type-B data only. Synthetic true values shown as vertical black line.

3.3.3 Anchor Conditioning Using Both Type-A and Type-B Data

In this section, the case study from 3.3.2 is repeated exactly, but this time the available direct measurements of transmissivity are included. This changes the MAD formulation to that of Equation 1.7 repeated here for ease of reading:

$$f(\boldsymbol{\vartheta}|\boldsymbol{z}_b, \boldsymbol{z}_a) \propto f(\boldsymbol{z}_b|\boldsymbol{\vartheta}, \boldsymbol{z}_a) f(\boldsymbol{\vartheta}|\boldsymbol{z}_a). \qquad \text{(1.7 repeated)}$$

Moreover, the prior joint PDF of the anchors is now conditional on the Type-A data $f(\boldsymbol{\vartheta}|\boldsymbol{z}_a)$. The prior used in Section 3.3.2 is still appropriate, but additional constraint can be imposed in this case. Here conditional variance and conditional means for the anchors can be determined using the kriging set of equations to Gaussian condition on the Type-A data, the values of which are listed in Table 3.4. Note the anchors are still assumed independent *a priori*, again because this is a least subjective prior by MRE. Samples are drawn randomly from the joint prior PDF.

Table 3.4: Conditional variances and means for prior PDF of anchors conditional on Type-A data.

Type-A Conditional Mean and Variance of Anchor Locations $\log_{10}(\boldsymbol{T})$

Anchor Location	*Mean*	*Variance*
$x = 3$	-2.13	0.295
$x = 7$	-1.81	0.264
$x = 13$	-1.63	0.264
$x = 17$	-1.93	0.295

The dimensionality of the problem is identical to preceding Section: $P = 4$ and $M = 3$. Thus, similar requirements are expected for R and N, which were also kept the same as the preceding Section respectively at 300 and 150. The overall computational cost for the $NR = 45{,}000$ simulations was again just under 11 hours on an Intel Core 2 Duo CPU E8400 3 GHz Processor.

Convergence was verified for the regressed 4-D $\hat{f}_R(\boldsymbol{\vartheta}|\boldsymbol{z}_b, \boldsymbol{z}_a)$ and for the non-parametrically calculated 3-D $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\vartheta}_i, \boldsymbol{z}_a)$ $i = 1, \ldots, R$ and again plots are omitted as they are redundant.

Figure 3.20 shows the posterior (blue) and prior (red) marginal PDFs for the four anchor locations along with the true synthetic values (black vertical line). The upper left pane shows the anchor at $x = 3$, upper right pane shows the anchor at $x = 7$, lower left pane shows the anchor at $x = 13$, and lower right pane shows the anchor at $x = 17$.

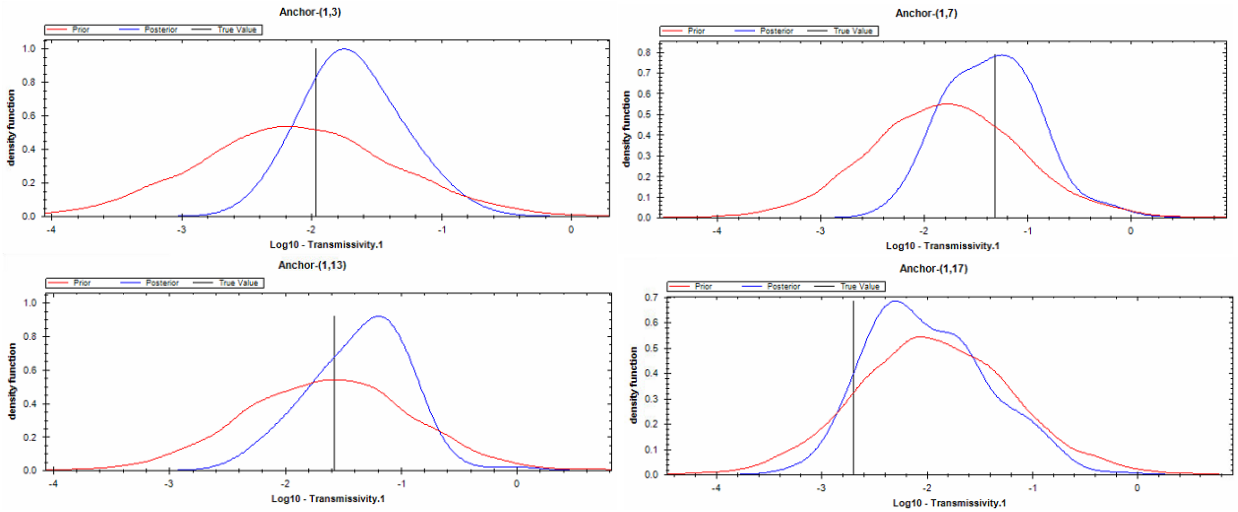

Figure 3.20: Posterior (blue), prior (red) marginal PDFs for anchors conditional on Type-A and Type-B data. Synthetic true values shown as vertical black line.

Given the added Type-A data constraint in the joint prior PDF compared to preceding Section, the expectation is that the posterior PDF results should be even more compelling and indeed the conditioning is more successful than in Section 3.3.2. All four anchor posterior marginal PDFs have less variance and reduced mean squared error (MSE) than their respective prior marginal PDFs.

In comparing the results of this section and the preceding section, note that at $x = 17$ the best results are attained by conditioning on both direct and indirect data sources. This is strong empirical evidence for the MAD framework's ability to work with and the importance of multiple data types: in Section 3.3.2, the Type-B only conditioning of this anchor failed to resolve this $\log_{10}(T)$ value, in Section 3.3.3, Type-A only conditioning - the prior $f(\boldsymbol{\vartheta}|\boldsymbol{z}_a)$ in this formulation – also was not effective at resolving this true value, only the combination of Type-A and Type-B data was successful.

There are subtle changes in the prior with and without Type-A data conditioning, but unfortunately these are very difficult to see in the graphics; comparing the trend and sill values from Tables 3.3 and 3.4 shows that the variance is reduced in all marginal prior PDFs by Type-A data conditioning and that the means are no longer identical for all the anchors when Type-A data is conditioned upon.

3.3.4 Conditioning of Structural Parameters on Type-A and Type-B Data

In this section, the objective switches from local characterization to determination of the uncertainty in the structural parameters. The nugget as mentioned in Table 3.3 is always fixed at zero and the covariance model is known to be exponentially decaying with respect to distance, so only the trend, sill, and range are treated as random variables in this example. The appropriate formulation of MAD for this example is Equation 1.6, repeated here for ease of reading

$$f(\boldsymbol{\theta}|\boldsymbol{z}_b, \boldsymbol{z}_a) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{z}_a) f(\boldsymbol{\theta}|\boldsymbol{z}_a). \qquad \text{(1.6 repeated)}$$

The joint prior PDF of the structural parameters conditional on Type-A data $f(\boldsymbol{\theta}|\boldsymbol{z}_a)$ is constructed as follows. The trend, range, and variance are assumed independent of one another; again the absence of a correlation structure is least biased by MRE. It is assumed that the marginal prior PDFs are uninformative (uniform on a physically plausible support). For the trend, the distribution $U[-5,1]$ is assumed, which is approximately centered on the sample mean of the Type-A measurements $\overline{z_a} = \frac{1}{n_a}\sum_{j=1}^{n_a} z_{a_j} = -1.67$, where n_a is the number of Type-A measurements, and extends roughly ± 4 sample standard deviations $s_{z_a} = \left[\frac{1}{(n_a-1)}\sum_{j=1}^{n_a}\left(z_{a_j} - \overline{z_a}\right)^2\right]^{\frac{1}{2}} = 0.70$. For the sill, the distribution $U[0.1,2]$ is assumed. The sill must be non-negative and this distribution accounts for nearly

a tripling of the sample standard deviation s_{z_a}. For the range, no reliable estimate of the variogram is possible with just 3 measurements of the Type-A data, especially because only two unique lags exist (5 and 10 meters), so the range is assumed $U[1,20]$, which spans from the grid discretization at a minimum to the entire domain length at a maximum. In this manner, the Type-A data was used to conditionally define the prior joint PDF of the structural parameters. Samples are drawn randomly from the joint prior PDF.

In this study, $P = 3$ and $M = 3$. However, even though relative to Sections 3.3.2 and 3.3.3 the parameter dimensionality has been reduced from 4 to 3, the scope of this problem is much more ambitious, because the structural model is now varying for every sample. This provides an excellent empirical example for examining convergence on a case by case basis and not blindly following rules based on dimensionality. Overall convergence of the regression of the 3-D posterior joint PDF $\hat{f}_R(\boldsymbol{\theta}|\boldsymbol{z}_b, \boldsymbol{z}_a)$ required a higher number of samples than before, $R = 400$, but the non-parametric calculation of the 3-D likelihood function $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{z}_a)$ $i = 1, \dots, R$ was comparable to the other variations already investigated and $N = 150$ still. The total cost of the $NR = 60{,}000$ simulations was just over 14 hours on an Intel Core 2 Duo CPU E8400 3 GHz Processor.

Figure 3.21 shows the posterior (blue) and prior (red) marginal PDFs for the three structural parameters inverted along with the true synthetic values (black vertical line). The upper pane shows the trend (labeled 'mean' by the software), the middle pane shows the sill (labeled 'partial sill' by the software), and the lower pane shows the range

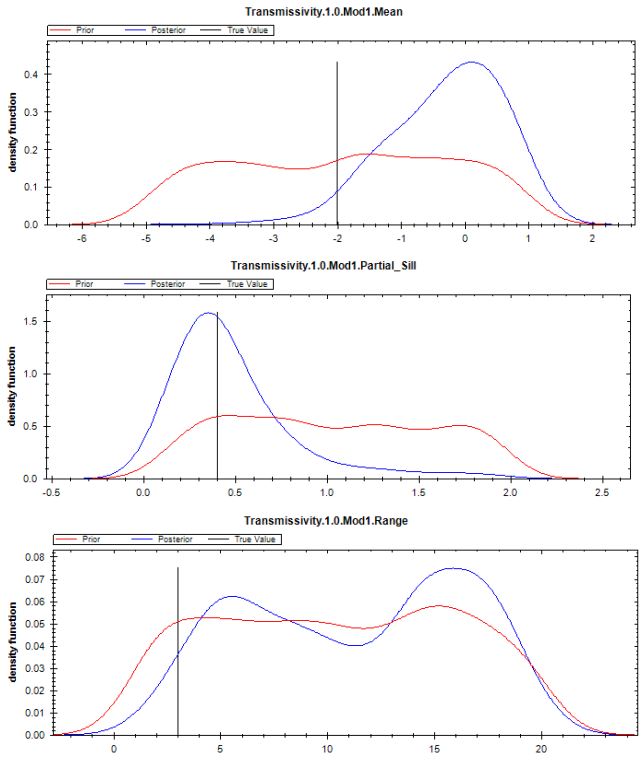

Figure 3.21: Posterior (blue) and prior (red) marginal PDFs for the trend, sill, and range conditional on Type-A and Type-B data. Synthetic true values shown as vertical black line.

The conditioning is only marginally successful. The conditioning of the trend actually favors a more porous formation with higher average transmissivity than the synthetic true value, but the less porous formations (e.g. $-6 \leq \log_{10}(T) \leq -3$) are correctly reduced in probability from the prior values. The posterior marginal PDF of the trend has reduced variance, but increased bias relative to the prior. The posterior marginal PDF of the sill is very accurately conditioned. The posterior and prior marginal PDFs for the range are nearly identical, which suggests the Type-B data cannot improve the identifiability of this structural parameter. In fact, after conditioning, the longer ranges are slightly more favored and the probability of the synthetic true value is reduced.

This example shows that determining the structural model parameters can be a much more challenging task than local characterization (previous two examples) – even with both direct and indirect measurements available.

3.3.5 Global and Local Characterization Using Type-B Data Only

In this section, for the first time the characterization objective is to quantify uncertainty in the structural model parameters and the anchors. The appropriate formulation of MAD for this application is Equation 1.3, repeated here for ease of reading,

$$f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_b) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}) f(\boldsymbol{\theta}, \boldsymbol{\vartheta}). \qquad \text{(1.6 repeated)}$$

A new challenge of defining a joint prior PDF for both the anchors and structural model parameters $f(\boldsymbol{\theta}, \boldsymbol{\vartheta})$must be overcome. Here, an extension of Bayes' raule is incredibly helpful and lends itself to a simple two-stage sampling approach. The joint prior PDF of the structural model parameters and anchors can be expanded as $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\boldsymbol{\vartheta}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$, where $f(\boldsymbol{\theta})$ is the joint PDF of the structural parameters only and $f(\boldsymbol{\vartheta}|\boldsymbol{\theta})$ is the joint PDF of the anchors conditional on the structural parameters. The two-stage sampling immediately becomes clear from the expansion, in which drawing R samples of the vector $[\boldsymbol{\theta}, \boldsymbol{\vartheta}]$ is performed, by first drawing from$f(\boldsymbol{\theta})$ to obtain $\boldsymbol{\theta}_i, i = 1, \dots F$ and then second drawing G times from each $f(\boldsymbol{\vartheta}|\boldsymbol{\theta}_i), i = 1, \dots, F$ where $FG = R$ [*Murakami et al.,* 2011; *Chen et al.,* 2012].

Now, $f(\boldsymbol{\theta})$ can be defined using 'soft' information from other experimental sites with similar geology, by soil texture class minima and maxima, or very broadly as any plausible geostatistical model parametrization. In this example, the case study of Section 3.3.4 is treated as a "comparable" field site and the same uninformative, uniform independent distributions are selected for the three structural model parameters: $U[-5,1]$ for the trend, $U[0.1,2]$ for the sill, and $U[1,20]$ for the range. After sampling these uniform distributions, in the second stage, independent Gaussians defined by the sampled trend values as the mean and the sample sill values as the variance defined $f(\boldsymbol{\vartheta}|\boldsymbol{\theta})$. In this case study, 10 configurations of anchors per structural model sample were selected; in cases where the field is of larger dimensions and/or the spatial arrangement of anchors are not as close in

proximity to each other this number may need to be increased. Therefore, $G = 10$ and F was set by inspecting the 7-D posterior joint PDF of the anchors and structural parameters conditional on the Type-B data $\hat{f}_R(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_b)$ for convergence of the regression and in total $R = 5{,}000$. The assumptions of independence respectively between the structural model parameters (stage 1) and amongst the anchors (stage 2) are again least subjective by MRE. In both stages, samples were drawn randomly from the respective PDFs.

In this example $M = 3$ again and convergence of the non-parametric likelihood function $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i, \boldsymbol{\vartheta}_i)$ $i = 1, \dots, R$ required $N = 150$. The total cost of the $NR = 750{,}000$ simulations was just over a week on an Intel Core 2 Duo CPU E8400 3 GHz Processor.

The results are shown in Figures 3.22 & 3.23. Figure 3.22 shows the posterior (blue) and prior (red) marginal PDFs for the three structural parameters inverted along with the true synthetic values (black vertical line). The upper pane shows the trend (labeled 'mean' by the software), the middle pane shows the sill (labeled 'partial sill' by the software), and the lower pane shows the range. Figure 3.23 shows the posterior and prior marginal PDFs for the four anchor locations along with the true synthetic values, following the same color scheme. The upper left pane shows the anchor at $x = 3$, upper right pane shows the anchor at $x = 7$, lower left pane shows the anchor at $x = 13$, and lower right pane shows the anchor at $x = 17$.

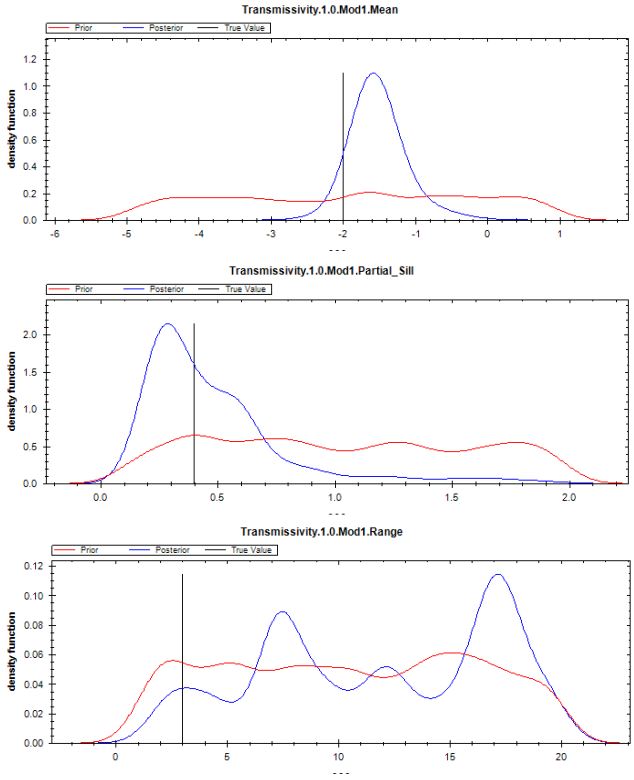

Figure 3.22: Posterior (blue) and prior (red) marginal PDFs for the trend (labeled 'mean'), sill (labeled 'partial_sill'), and range conditional on Type-B data, when jointly inverted with anchors. Synthetic true values shown as vertical black line.

Comparison between Figures 3.21 & 3.22 offers an interesting perspective on the interplay between anchors and structural parameters. Respectively, Type-A data is removed and

anchors are added in this section, but notice the major improvement in the posterior marginal PDF for the trend and the similarity in the range and sill PDFs compared to the last section. Anchors as their name suggests are capable of "anchoring" the transmissivity field in this case and improve the structural model parameter inference. The trend and sill are conditioned quite well, but again there is some difficulty in identifying the true range value, in fact after conditioning longer ranges are again given higher probability.

In general, the conditioning is quite effective. At all four anchor locations variance is reduced and in three out of four locations the bias is also reduced. Similar to Section 3.3.2 where only Type-B data is conditioned upon, the anchor at $x = 17$ is the only location at which the prior marginal PDF has less bias than the posterior marginal PDF. Note that relative to the previous two examples utilizing anchors in Sections 3.3.2 and 3.3.3 how much more diffuse the marginal prior PDFs are, which is a direct outcome of the two stage sampling process with the random structural model parameters.

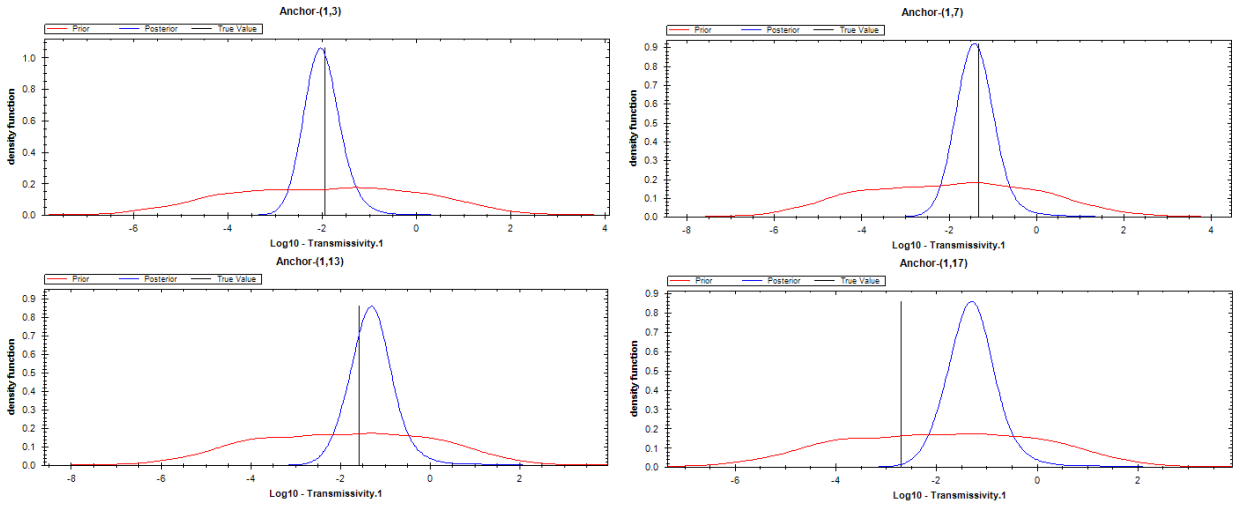

Figure 3.23: Posterior (blue) and prior (red) marginal PDFs for anchors conditional on Type-B data, when jointly inverted with structural model parameters. Synthetic true values shown as vertical black line.

3.3.6 Global and Local Characterization Using Type-A and Type-B Data

In this section, the case study from Section 3.3.5 is repeated, but the available Type-A data is incorporated. This changes the formulation of MAD to that of Equation 1.2, repeated here for ease of reading

$$f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_b, \boldsymbol{z}_a) \propto f(\boldsymbol{z}_b|\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{z}_a) f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a). \qquad \text{(1.2 repeated)}$$

Note that the joint prior PDF is now conditional on Type-A data $f(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z}_a)$, which means that additional constraints can be imposed. Like Section 3.3.5 the prior can be expanded

using Bayes rule $f(\boldsymbol{\theta},\boldsymbol{\vartheta}|\boldsymbol{z}_a) \propto f(\boldsymbol{\vartheta}|\boldsymbol{\theta},\boldsymbol{z}_a)f(\boldsymbol{\theta}|\boldsymbol{z}_a)$, which is well-suited to the previously described two-stage sampling approach. The joint PDF of the structural parameters conditional on the Type-A data $f(\boldsymbol{\theta}|\boldsymbol{z}_a)$ is adopted from Section 3.3.4: respectively, $U[-5,1]$ for the trend, $U[0.1,2]$ for the sill, and $U[1,20]$ for the range. The joint PDF of the anchors conditional on the structural parameters and Type-A data $f(\boldsymbol{\vartheta}|\boldsymbol{\theta},\boldsymbol{z}_a)$ utilizes Gaussian conditioning via the kriging system of equations similar to Section 3.3.3. Similar to the preceding section, 10 anchor configurations per structural model parameter set were employed, so $F = 10$. G again was set by examining convergence of the regression of the 7-D $\hat{f}_R(\boldsymbol{\theta},\boldsymbol{\vartheta}|\boldsymbol{z}_b,\boldsymbol{z}_a)$ and required $R = 5{,}000$. Samples from $f(\boldsymbol{\theta}|\boldsymbol{z}_a)$ and $f(\boldsymbol{\vartheta}|\boldsymbol{\theta},\boldsymbol{z}_a)$ were drawn randomly.

In this example $M = 3$ again, convergence of the non-parametric likelihood function $\hat{f}_N(\boldsymbol{z}_b|\boldsymbol{\theta}_i,\boldsymbol{\vartheta}_i,\boldsymbol{z}_a)$ $i = 1,\dots,R$ required $N = 150$. The total cost of the 750,000 simulations was just over a week on an Intel Core 2 Duo CPU E8400 3 GHz Processor.

The results are shown in Figures 3.24 & 3.25. Figure 3.24 shows the posterior (blue) and prior (red) marginal PDFs for the three structural parameters inverted along with the true synthetic values (black vertical line). The upper pane shows the trend (labeled 'mean' by the software), the middle pane shows the sill (labeled 'partial sill' by the software), and the lower pane shows the range. Figure 3.25 shows the posterior and prior marginal PDFs for the four anchor locations along with the true synthetic values, following the same color scheme. The upper left pane shows the anchor at $x = 3$, upper right pane shows the anchor at $x = 7$, lower left pane shows the anchor at $x = 13$, and lower right pane shows the anchor at $x = 17$.

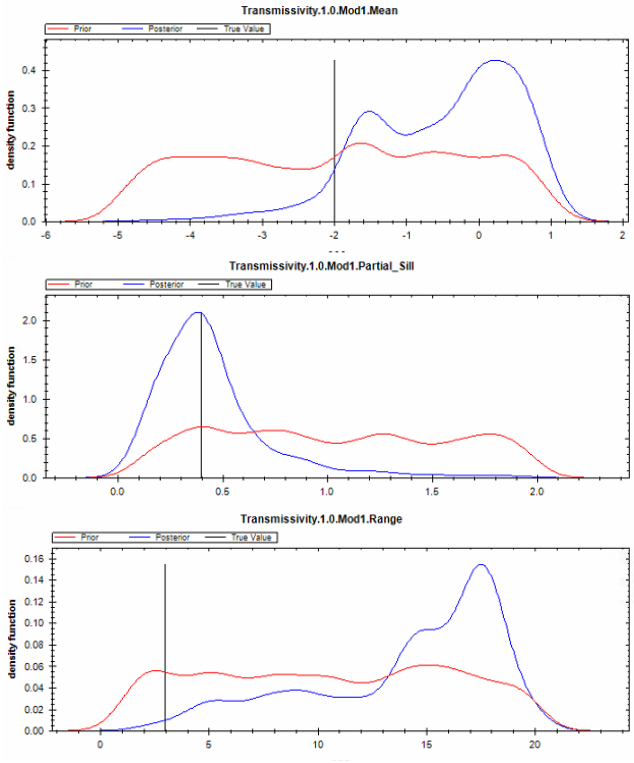

Figure 3.24: Posterior (blue) and prior (red) marginal PDFs for the trend (labeled 'mean'), sill (labeled 'partial_sill'), and range conditional on Type-A and Type-B data, when jointly inverted with anchors. Synthetic true values shown as vertical black line.

The conditioning is most successful for the sill. The variance is reduced in the posterior for both the trend and range, but the bias is increased in the posterior. After conditioning, on the average the medium is likely to be more transmissive (larger trend values) and also more homogeneous (longer range values) than the synthetic medium. A few possible reasons the conditioning of the range is so poor are: 1) the longest lag distance between a pair of measurements (either Type-A or anchors) is 3 m owing to the screening effect [*Rubin,* 2003], which means there are no large lag distances in the raw variogram and 2) the field is only mildly heterogeneous (the variance of the anchors and Type-A measurements is 0.35), which suggests the field could be correlated over large scales; hence the inability of the Type-A data and anchors to effectively condition the range parameter.

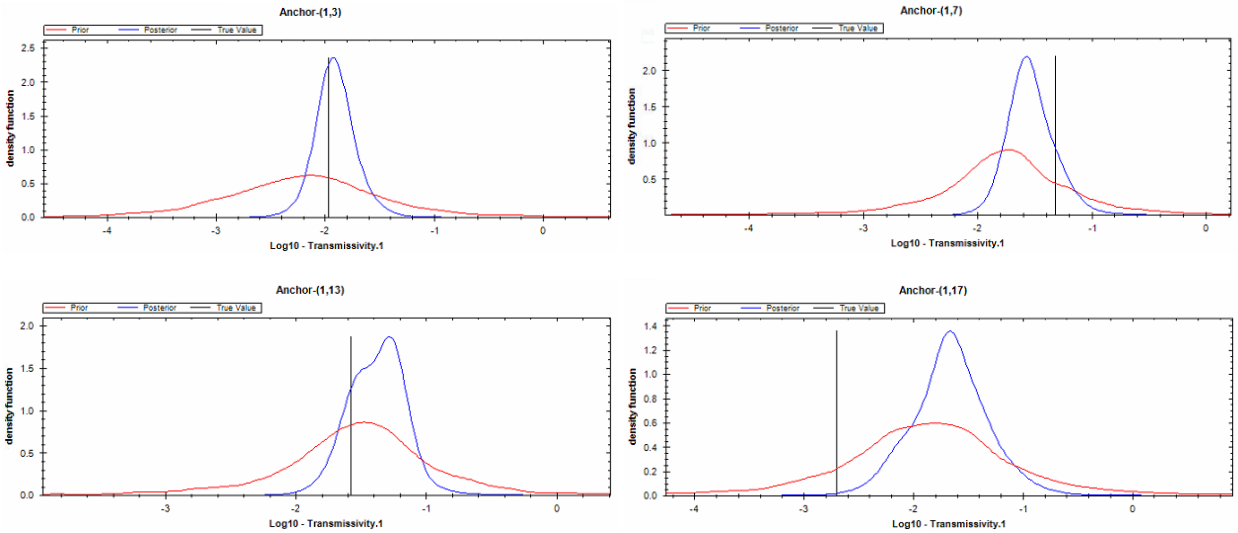

Figure 3.25: Posterior (blue) and prior (red) marginal PDFs for anchors conditional on Type-A and Type-B data, when jointly inverted with structural model parameters. Synthetic true values shown as vertical black line.

Again, the anchors are for the most part accurately conditioned. The same as in Section 3.3.5 the variance is smaller in the posterior marginal PDFs at all four anchor locations However, the bias is increased at $x = 13$ and $x = 17$ in the posteriors. Note that the prior marginal PDFs for the anchors are significantly narrower than their counterparts in Section 3.3.5, which is directly attributable to the Type-A conditioning in the prior joint PDF.

3.4 Conclusion

In Section 3.1, the GUI for the MAD software was introduced. For each form, the supported elements of MAD were identified, such that connections between the theory and implementation were clearly identified. User-friendly features were also highlighted, because a primary objective of the software is reducing the complexity of model-inversion for end-users. Finally, the sequence and dependence of the forms was explained, which

enables the software to independently manage the inversion process. The overall purpose of the section was to show the theoretical and convenient motivation behind the MAD software design.

In Section 3.2, the broader modular architecture of the MAD software is presented. Of note was the interchangeability – via the 'drivers' - of the external components – the random field generator and the forward model - with the modules of the MAD software. The Bayesian inversion functionality of the MAD software is isolated and protected in the modules of the software and is generically developed to support applications in multiple scientific fields. In the next chapter, a case study is conducted utilizing HYDRUS 4.14 and R Statistics that validates the driver design and that the MAD software is adaptable to various scientific disciplines. Finally, the sequence of the automated data generation process and the necessary information exchange between the MAD modules, forward model, and random field generator was outlined.

In Section 3.3, the six subsections presented variations of a case study and were designed to show that the MAD software flexibly supports Equations 1.2-1.7. A simple experiment was selected to allow the focus of the analyses to be the usage of various data types, anchors, and structural parameters. All of the studies employed MRE concepts to define least subjective priors. In general, the anchors were almost always conditioned successfully; even in cases without the direct measurements of transmissivity at nearby locations. The structural model parameters were harder to identify, but with the exception of the range (challenges discussed in Section 3.3.6) the conditioning is fairly successful.

The free open-source software; a manual, which explains how to use each form presented in Section 3.1; and a tutorial, which utilizes the same forward model scenario as the examples in Section 3.3, are all available for download from https://mad.codeplex.com. A more detailed summary of the technical construction of the MAD software can be found in [*Osorio et al.,* 2013].

4. Vadose Zone MAD Theory & Application

In this chapter, a variant of the MAD formulation given in Section 1.2 and the MAD software presented in Chapter 3 are used in a vadose zone hydraulic parameter inversion.

Reliable estimation of unsaturated soil hydraulic properties is essential for accurate, physics-based modeling of vadose zone water dynamics as well as many related processes. It is well known that these properties, which are the soil water retention curve and the unsaturated hydraulic conductivity function, are highly non-linear. In addition, the spatial distribution of these soil hydraulic properties may show significant multi-scale heterogeneity [e.g., *Schaap et al.*, 2001; *Vogel & Roth*, 2003] which makes their appropriate and accurate determination for field-scale applications time consuming and expensive. It has also been shown that unsaturated hydraulic properties measured in the lab are not necessarily transferable to the field scale [e.g., *Dane & Hruska, 1983; Hopmans & Simunek*, 1999; *Ritter et al.*, 2003; *Wöhling et al.,* 2008]. Even if the soil textural class is known, they may vary within a considerable range [e.g., *Carsel & Parrish*, 1988]. Hence, a site-specific determination using appropriate field measurements may be the method of choice in order to obtain a useful structural representation.

As such, there is a body of literature dedicated to the estimation of soil hydraulic properties, which spans from traditional field and laboratory measurements all the way to advanced optimization approaches and stochastic methods [*c.f. Hopmans & Simunek*, 1999; *Vrugt et al.*, 2008; *Cook & Cresswell*, 2008]. These methods incorporate various data types such as soil moisture, pressure head, and geophysical measurements collected in the vadose zone and under natural or forced boundary conditions [e.g., *Kowalsky et al.,* 2005; *Looms et al.,* 2008]. Moreover, these current stochastic and optimization approaches tend to provide very reasonable estimates of soil hydraulic properties and their uncertainty.

Because this chapter has a heavy comparison with an optimization approach [*Wollschläger et al.,* 2009] it is worth expanding a bit the discussion from Section 1.1 about the historical progression of optimization. Over roughly the past decade, SCE-UA has successively been extended. First, it was combined with a Metropolis algorithm and used in a Markov Chain Monte Carlo (MCMC) framework in an approach called shuffled complex evolution Metropolis algorithm (SCEM-UA) [*Vrugt et al.,* 2003]. Subsequently, a genetic algorithm was added for competitive evolution of the Markov chains in an approach called differential evolution adaptive Metropolis (DREAM) [*Vrugt et al.,* 2008a, 2009]. The DREAM approach is a synthesis of the stochastic and optimization approaches and has been applied successfully in many different scenarios [*c.f. Vrugt et al.,* 2008a, 2008b, 2009]. The DREAM-ZS [*Schoups & Vrugt,* 2010] algorithm is an even more computationally efficient version of DREAM owing to a clever recycling of candidate points. As stated before, MAD is not an optimization technique and this chapter will further elucidate this point.

An alternative methodology to optimization that has also shown considerable maturation in the past few decades is a stochastic approach, most applications of which are Bayesian. The work of *Scholer et al.* [2011] contains an excellent overview of the stochastic approaches, so we only briefly present some historical milestones with which we will contrast our following work. Here, some additional historical milestones in this research field that are presented - that were not included in Section 1.1 - with which the following work is contrasted.

Amongst the first stochastic analysis in the unsaturated zone was the application of maximum likelihood estimation (MLE) by *Jury & Sposito* [1985], in which the function to optimize is a probability density function. MLE requires knowledge of the probability distribution of the measurement data used for parameter conditioning [*Kitanidis & Vomvoris,* 1983], which is commonly assumed to be a multivariate Gaussian distribution (MG), but not validated. A critical problems with the maximization process is that the parameters are not required to maintain physical values [*Sorooshian et al.,* 1983].

Another stochastic approach is GLUE [*Beven & Binley*, 1992; *Beven & Freer,* 2001; *Beven & Binley,* 2013]. *Binley & Beven* [2003] applied GLUE for the characterization of vadose zone flow using geophysical measurements for conditioning of the vadose zone model parameters, which is a similar application to this chapter. However, as mentioned in Section 1.1, there are significant challenges associated with GLUE in terms of adequate sampling of the parameter space and in adding bias to the results.

The incorporation of a prior can restrict the parameter space to physically valid or reasonable values only. In a more formal Bayesian framework, *Hou & Rubin* [2005] incorporated least subjective priors and provided analytical solutions for the posterior distribution of Mualem-van Genuchten model parameters. More recently, the efforts of *Scholer et al.* [2011, 2012] and *Scharnagl et al.* [2011] have worked to quantify the importance of prior information and correlation between model parameters. These efforts shifted focus to treating the vadose zone hydraulic parameters as random variables. However, these methods unfortunately still require an assumption about the form of the likelihood function and none of them can be used to directly infer the probability distribution of the residuals.

Commonly, the user-defined likelihood functions or objective functions are assumed to be a MG of the residual difference between model prediction and the observation data. Often the MG is defined by assuming that the residuals are homoscedastic and uncorrelated between all times and locations, which is difficult to justify [*Ginn & Cushman,* 1990; *Hollenbeck & Jensen*, 1998; *Vrugt et al.,* 2008a; *Schoups & Vrugt,* 2010; *Wöhling & Vrugt*, 2011]. An improvement intended to relax these assumptions was the introduction of an autoregressive likelihood function that can accommodate correlation, heteroscedascity, and asymmetry in the distribution of residuals [*Schoups & Vrugt,* 2010; *Wöhling & Vrugt*, 2011].

However, even with more flexible autoregressive form of the likelihood function, there is an underlying issue with all of the aforementioned techniques: all of them require the assumption of the form of the likelihood function and none of them have a mechanism by

which this assumption can be tested. This means that the appropriateness of any form of the likelihood function can never truly be validated. Bayesian approaches are frequently criticized for subjectivity and assumptions in the prior distribution of model parameters leading to bias in the results, but this very same subjectivity and assumptions are ignored in the likelihood and objective functions.

MAD addresses this shortcoming by modifying an approach originally employed for aquifer characterization and developing a framework for validating the appropriateness of the likelihood function.

In the remainder of this chapter, the modifications to the theory in Section 1.2 for inverting soil properties as random variables are presented as well as an application to a real soil column study. The necessary nested simulation procedure and statistical apparatus to implement the data-driven likelihood function are outlined in Section 4.1. The Bayesian framework that analyzes the output of the nested simulation procedure is outlined in Section 4.2. In Section 4.3, the results of the analysis on a natural boundary condition, multi-layered soil profile along with comparison to a previous optimization performed by *Wollschläger et al.* [2009] and a discussion of the findings are presented.

4.1 Physical Parameters of Interest and Statistical Representation of Conceptual Soil Model

Here, the governing equations and the conceptual model used to simulate the soil water dynamics at the field site are defined in Section 4.1.1. Once the hydraulic parameters of interest from the governing equations are defined, Section 4.1.2 presents a method of statistical representation of these physical parameters within the material layers that poses them as distributions of random variables.

4.1.1 Soil Column Representation, Hydraulic Model and Parameterization

In the next three short subsections, the conceptual model of the soil column as delineated by soil horizons, the mathematical model of the one dimensional unsaturated flow physics, and the hydraulic parameters from specific layers that are of interest are presented.

4.1.1.1 Soil Column Conceptual Model

In this section, the pertinent elements of the Gernzhof experimental site (49°25' N, 8°37' E), which is located a few kilometers west of the city of Heidelberg, SW-Germany, the conceptual representation of the soil column, and details of the implementation of this column in HYDRUS-1D (version 4.14) [*Simunek* et al. 2008] are summarized. A much more comprehensive description of the experimental site, the measurement devices and their calibration, and boundary conditions during the monitoring are available in *Wollschläger et al.* [2009] - a more brief, but independent summary is provided here for ease of reading.

Figure 4.1 shows the soil profile as characterized during installation of the TDR probes along with the layering as defined in the HYDRUS-1D implementation of the Richards equation presented in section 4.1.1.2. Also shown in Figure 4.1 are the depths of the 4 installed TDR probes, respectively at 13, 63, 92, and 116 cm below the surface.

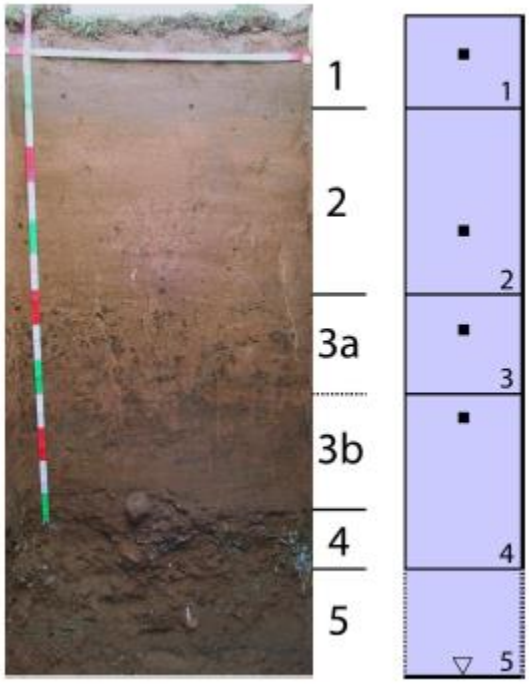

Figure 4.1: Soil profile at the Grenzhof test site and conceptual representation of the layering as applied in the HYDRUS-1D simulations. Black squares indicate TDR installation locations. Layer 5 is not shown to scale. [*Wollschläger et al.,* 2009; reproduced with permission of the authors.]

Figure 4.1 shows the model layer boundaries respectively at depths of 28, 82, 110, and 154 cm, which were determined by transitions observed in the soil texture. The texture of the uppermost four layers of the soil profile is classified as sandy loam according to the USDA-Soil Taxonomy [*Soil Survey Division Staff*, 1993].

The implementation of this conceptual model in HYDRUS-1D is identical to previous work [*Wollschläger et al.*, 2009], with the one exception that we lengthened the spin up time from 45 to 397 days to further decrease sensitivity of our numerical simulations to

uncertainty in the initial conditions (interpolated profile using the 4 available probes). This was done by running the model with the whole cycle of observed surface atmospheric boundary conditions once in advance before the "true" simulation period started.

The water table is artificially set at a depth of 4 m below surface. The Penman-Monteith model evapotranspiration flux is scaled a factor of 0.61 as determined by *Wollschläger et al.* [2009]. The root uptake is modeled following the work of *Feddes et al.* [1978] and *Taylor & Ashcroft* [1972] for grass, with a maximum rooting depth of 0.14 cm.

In the subsequent analysis, the boundary conditions, initial condition, evapotranspiration, and root uptake are models are deterministic for the following reasons: the boundary conditions are well known [*Wollschläger et al.,* 2009]; the model is insensitive to the initial condition because of increased spin-up; the evapotranspiration and root uptake models are used with the optimal parameter values determined by *Wollschläger et al.* [2009] because our objective is to obtain the hydraulic parameters only. The last decision is prompted by computational parsimony and obtaining a characterization of the uncertainty that is directly comparable to previous work using the numerical model and available data.

Many of the modeling choices are directly adopted from the work of *Wollschläger et al.* [2009], because the goal of this study is not to improve on the conceptual model or the physics being modeled, but to develop a comparison analysis and present a more validated statistical framework than earlier works. As such, it is recognized that this forward model has several potential sources of error including, but not limited to: preferential flow, shrinking, swelling, intra-layer heterogeneity, or hysteresis [*Wollschläger et al.,* 2009], but these are issues that are outside the scope of this chapter.

4.1.1.2 One Dimensional Unsaturated Water Dynamics

This chapter focuses on the flow of water through a variably saturated soil column - the appropriate equation for describing this physical scenario is the Richards equation in one dimension:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K\left(\frac{\partial h}{\partial z} - 1\right)\right] - R \tag{4.1}$$

where θ is the volumetric water content $[-]$, h is the matric potential [m], t is time [d], z is depth [m], $K$is the hydraulic conductivity function $[\mathrm{md}^{-1}]$, and R is a sink term $[\mathrm{m}^3\mathrm{m}^{-3}\mathrm{d}^{-1}]$. The Richards equation has three unknowns h, θ, and K; to close the system of equations, the Mualem-van Genuchten model [*Mualem*, 1976; *van Genuchten*, 1980] is introduced

$$\theta(h) = \begin{cases} \theta_r + \frac{\theta_s - \theta_r}{[1+|\alpha h|^n]^{1-\frac{1}{n}}} & h < 0 \\ \theta_s & h \geq 0 \end{cases} \tag{4.2}$$

$$K = K(\Theta) = K_s \Theta^{\frac{1}{2}} \left[1 - \left(1 - \Theta^{n/[n-1]}\right)^{1-1/n}\right]^2 \tag{4.3}$$

with

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} \tag{4.4}$$

where θ_r is the residual water content $[-]$, θ_s is the saturated water content $[-]$, α $[1/cm]$ and n $[-]$ are shape parameters, K_s is the saturated hydraulic conductivity [cm/hr], and Θ is the effective saturation $[-]$.

4.1.1.3 Characterization Target Parameters

In this section, the physical parameters from Equations 4.1 to 4.4 of interest in the characterization of the conceptual model are selected.

In the subsequent application of Bayes' rule, the statistical distributions are quantified for four physical parameters affiliated with Equations 4.1 to 4.4: $\boldsymbol{y} = [K_s, \theta_r, n, \alpha]^T$. Moreover, these physical parameters are characterized in each of the upper four layers of the soil column, such that the composite target parameter vector (denoted by the subscript 'c') can be defined as

$$\boldsymbol{y}_c = \left[\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \boldsymbol{y}^{(3)}, \boldsymbol{y}^{(4)}\right]^T = \left[K_s^{(1)}, \theta_r^{(1)}, n^{(1)}, \alpha^{(1)}, \dots, K_s^{(4)}, \theta_r^{(4)}, n^{(4)}, \alpha^{(4)}\right]^T \tag{4.5}$$

where the superscripts are references to the layer numbers given in Figure 4.1. Note, in this study, the hydraulic properties of the lowest layer are not estimated and are held fixed. The values of the fifth layer parameters are given in Table 4.1, which are defined as the expected values for sand from the database of *Carsel & Parrish* [1988].

Table 4.1: Summary of layer 5 hydraulic parameters for sand as taken from *Carsel & Parrish* [1988]

Hydraulic Parameter	Value in layer 5
K_s [cm/hr]	29.7
α [1/cm]	0.145
n $[-]$	2.68
θ_r $[-]$	0.045
θ_s $[-]$	0.043

Hereafter, the variables given in the vector in Equation 4.5 are always referred to as the 'hydraulic parameters' to distinguish them from the 'structural parameters' affiliated with the implementation of MAD.

4.1.2 Structural Model

In this section, a representation of the hydraulic parameters as random variables belonging to a parametric PDF is defined. This statistical representation is a structural model (agreeing with the language of previous MAD applications, even though it is not geostatistical in nature). The parameters of this structural model - the mean vector and covariance matrix of an MG structural model - are the structural parameters. Later, in Section 4.2, the structural model will be updated from initial (prior) estimates of their probabilities, to data-conditional (posterior) estimates of their probabilities via Bayes' rule and demonstrate the important and necessary role of a structural model in this implementation. But first, an argument for the appropriateness of representing the hydraulic parameters as distributed random variables is given.

The spatial distribution of hydraulic properties in natural field soils may be highly heterogeneous [*Carsel & Parrish*, 1988; *Yeh et al.,* 1986; *Schaap et al., 2002*]. Even in the advantageous case where the soil textural class is known, the soil hydraulic parameters may still cover a considerable range *[Carsel & Parrish,* 1988]. Therefore, a reasonable approach to account for this uncertainty in the parameter estimation should be to apply a structural model that represents the respective randomness within the textural class. Justification for an MG structural model is based on the statistical analysis of soil catalog samples.

The work of *Carsel & Parrish* [1988] determined an important set of empirically determined Johnson family transformations [*Johnson & Kotz*, 1970] (grouped by soil texture class and on data collected at hundreds of experimental sites as well as several soil catalogs published by U.S. states) that transform the statistical distributions of hydraulic parameters listed in Equation 4.5 into MG distributions.

Provided the structural model represents the distribution of the hydraulic parameters (or their transforms), there is little justification required for the form of the structural model. For instance, it makes no sense to try to represent a uniformly distributed parameter with any distribution other than a uniform distribution. Therefore, the choice of a structural model that best represents the transformed hydraulic parameters is an MG distribution. However, the transformations that yield such distributions do require some justification.

To broadly accept the transformations derived from samples collected around the world of the same soil texture may at first seem too variable or too generic for any one site. Naturally, the expectation is there is no such thing as a transformation to normality that works universally for all experimental sites. In cases where samples of the hydraulic parameters are obtained at a site, then a 'custom' structural model that most suitably represents the variability at the site could be empirically derived – with or without transformations, which could also be empirically determined for the site. However, when only the soil texture is known at a site, it would be extremely subjective to assume a transformation (or structural model) based on anything except textural classes.

The transformations empirically derived by *Carsel & Parrish* [1988] are specific for each USSI soil category. For the sandy loam texture at the Grenzhof site, the Johnson family transformations for $K_s, \theta_r, n, \& \alpha$ are listed in Table 4.2.

Table 4.2: Johnson family transformations for sandy loam soil texture [*Carsel & Parrish*, 1988]

Hydraulic Parameter	Johnson Family Transformation
K_s [cm/hr]	$y^t = \ln\left(\frac{y}{30 - y}\right)$
α [1/cm]	$y^t = \ln\left(\frac{y}{0.25 - y}\right)$
n [−]	$y^t = \ln(y)$
θ_r [−]	$y^t = \ln\left(\frac{y}{0.11 - y}\right)$

To remind that the structural model is actually being specified for transformations of the hydraulic parameters of interest the superscript y^t is introduced to indicate the transform of a generic hydraulic parameter y. To keep notation brief, different symbols for the transformations, which are unique for each of the hydraulic parameters in consideration, are not introduced (the use of the same superscript is a slight abuse of notation), but this does not limit the subsequent derivations.

Next, the MG structural model is defined for the sandy loam hydraulic parameters at the Grenzhof field site as

$$S(\boldsymbol{y}^t; \boldsymbol{\beta}) = N(\boldsymbol{y}^t; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4.6}$$

which is parameterized by structural parameters $\boldsymbol{\beta}$ - a vector composed of the elements of the mean vector $\boldsymbol{\mu}$ and the unique elements of the covariance matrix $\boldsymbol{\Sigma}$. The mean vector contains four elements, the respective means of the transformed hydraulic parameters. The symmetric covariance matrix is 4×4 and contains the variances of the transformed hydraulic parameters on the diagonal and the covariance of pairs of the transformed hydraulic parameters on the off-diagonal.

The most important role of the structural model is to generate random realizations of the hydraulic parameters. In this case, this requires random realizations of the hydraulic parameters for each of the soil layers, which can subsequently be combined into a composite soil column realization. Therefore, the discussion of structural models is extended to incorporate multiple layers.

For this initial study, as a working assumption the layers are taken to be statistically independent - similar to the work of *Mertens et al.* [2004] - such that the composite structural model for the entire target hydraulic parameter vector of Equation 4.5 can be extended from Equation 4.6 as:

$$S(\boldsymbol{y}_c^t; \boldsymbol{\beta}_c) = N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{4.7}$$

where $\boldsymbol{\mu}_c = \left[\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \boldsymbol{\mu}^{(3)}, \boldsymbol{\mu}^{(4)}\right]^T$ is the composite mean vector of length 16, $\boldsymbol{y}_c = \left[\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \boldsymbol{y}^{(3)}, \boldsymbol{y}^{(4)}\right]^T$ is the composite hydraulic parameter vector of length 16 – in both

cases superscript is an index over layer number, $\boldsymbol{\Sigma}_c$ is the composite covariance matrix of dimension 16×16 and is comprised of four replicates of the 4×4 $\boldsymbol{\Sigma}$ matrices along the diagonal with zeros elsewhere, and $\boldsymbol{\beta}_c$ is formally a vector containing all the components of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$.

Random sampling from the MG model specified in Equation 4.7, after a value for $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are supplied (a topic expanded upon in the next section), is used to generate realizations of the 16-dimensional vector of the transformed hydraulic parameters $\tilde{\boldsymbol{y}}_c^t(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where the accent $\tilde{}$ denotes a random realization and the dependency on the values of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are explicitly expressed. After applying the appropriate inverse Johnson transformations – given in Table 4.3 - for each hydraulic parameter, this comprises a random realization of the hydraulic parameters for the composite of the soil column $\tilde{\boldsymbol{y}}_c(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

Table 4.3: Inverse Johnson family transformations for sandy loam texture class [*Carsel & Parrish,* 1988].

Transformed Hydraulic Parameter	Inverse Johnson family Transformation
K_s^t	$y = \dfrac{30 \exp(y^t)}{1 + \exp(y^t)}$
α^t	$y = \dfrac{0.25 \exp(y^t)}{1 + \exp(y^t)}$
n^t	$y = \exp(y^t)$
θ_r^t	$y = \dfrac{0.11 \exp(y^t)}{1 + \exp(y^t)}$

4.2 Conditioning Data and Bayesian Framework

Before presenting the details of the available measurement data and the Bayesian characterization, it is worthwhile to briefly discuss the conceptual motivation for the structural modeling approach without geostatistics.

In the previous section, the hydraulic parameters of the various soil layers were posed as distributed random variables. Importantly, the exact distribution of these random variables was not fully specified yet; initially, only the MG form is chosen, but with unknown mean vector and covariance matrix (the structural parameters). It was hinted that the objective was to randomize these structural parameters. Now it is explained why.

The purpose of this randomization of the mean vector and the covariance matrix is significant: philosophically, what this states is that some understanding of the distributional form of the target hydraulic parameters $\boldsymbol{y}_c$ is claimed, i.e. what "family" of parametric distributions can represent their distribution in the soil (MG), but that location of their mode ($\boldsymbol{\mu}_c$) and their deviation ($\boldsymbol{\Sigma}_c$**)** are unknown. Via the Bayesian framework, a prior distribution of physically-plausible values for the mode and the deviation (based on soils of similar texture) are specified and then these physically-plausible values are conditioned on their likelihood of reproducing the *in situ* measurements. The output of this

Bayesian implementation is a site-specific distribution of structural model modes and deviations, thereby reducing the physically-plausible "family" from the broader range affiliated with the entire soil textural class to the range suitable just for the Grenzhof field site. Note that the objective is not to obtain an optimal value of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, but rather the conditional PDF.

The uncertainty in $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ can be propagated into $\boldsymbol{y}_c$ and functions of $\boldsymbol{y}_c$. Results of characterization of the uncertainty in the hydraulic parameters, water retention curve, and hydraulic conductivity function are presented in Section 4.3. The remainder of this section introduces the complete Bayesian framework and derivations of specific PDFs necessary to condition the physically-plausible values on the available measurement data at the field site. First, the measured data is introduced.

4.2.1 Available Data

As mentioned in Section 4.1.1.1, within the soil column at the experimental site there are 4 TDR probes that collect water content measurements at different depths: 13, 63, 92, and 116 cm. A full description of the calibration and use of these probes at the site is available in the work of *Wollschläger et al.* [2009]. Conservatively, the measurement error of these devices is taken as 0.015 absolute water content [*Roth et al.,* 1990].

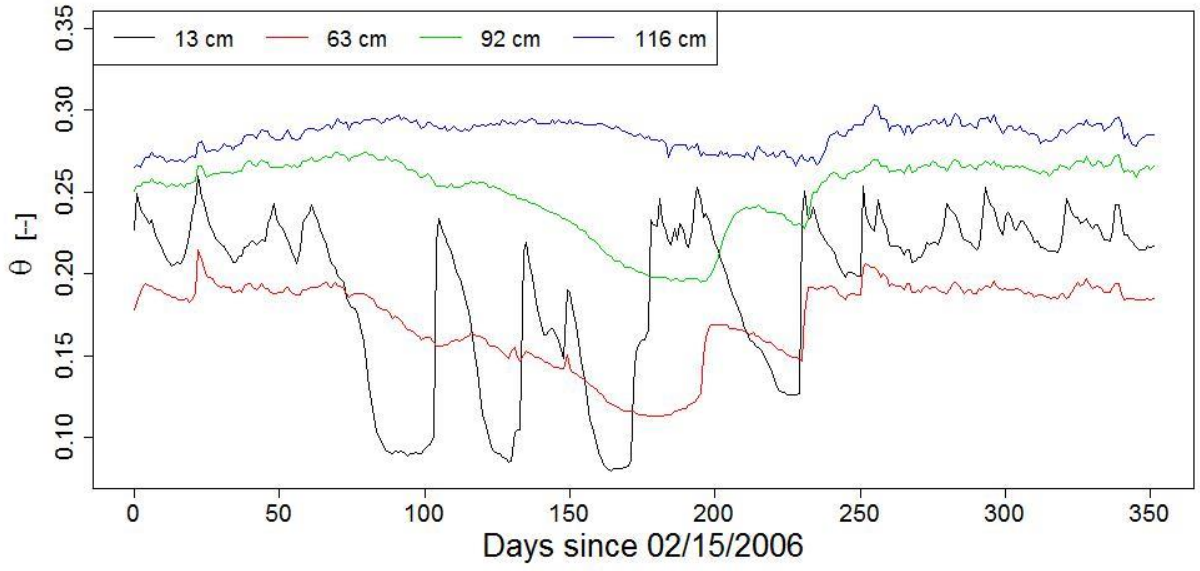

Figure 4.2: Daily averaged water content measurements at 13, 63, 92, and 116 cm depth. Data from *Wollschläger et al.* [2009].

Available to the analysis is nearly one year (352 days from Feb. 15, 2006 until Feb. 1, 2007) of daily-averaged volumetric water content measurements θ. Figure 4.2 shows the time series of the measured values of θ at the four depths mentioned above, which are also

indicated as black squares in Figure 4.1. There is one TDR probe placed in each model layer.

At the Grenzhof site, the most dynamic fluctuations in water content occur in the uppermost layer (the black trace in Figure 4.2) as the soil responds to atmospheric forcing and evapotranspiration. The responses to infiltration and evapotranspiration at the surface are attenuated with depth in the soil column. Listed in Table 4.4 are the ranges of observed water content in each of the 4 layers, which also decrease with depth.

Table 4.4: Minimum, maximum, and range of observed water content by layer.

Layer	$\boldsymbol{min}\{\boldsymbol{\theta}\}$ $[-]$	$\boldsymbol{max}\{\boldsymbol{\theta}\}$ $[-]$	$\boldsymbol{range}\{\boldsymbol{\theta}\}$ $[-]$
1	0.080	0.260	0.18
2	0.113	0.215	0.102
3	0.195	0.274	0.079
4	0.265	0.303	0.038

4.2.2 Stochastic Implementation

In this section, the implementation of Bayes' rule for conditioning $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ is explicitly shown. A statistical property of MG distributions is that the parameters $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are statistically independent variables [*Lukacs,* 1942]. Combining these principles gives

$$f(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|\boldsymbol{z}) \propto f(\boldsymbol{z}|\boldsymbol{\Sigma}_c, \boldsymbol{\mu}_c) f(\boldsymbol{\Sigma}_c, \boldsymbol{\mu}_c) = f(\boldsymbol{z}|\boldsymbol{\Sigma}_c, \boldsymbol{\mu}_c) f(\boldsymbol{\Sigma}_c)\, f(\boldsymbol{\mu}_c) \tag{4.8}$$

where $\boldsymbol{z}$ is the vector of daily-averaged water content measurements in all the layers; $f(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|\boldsymbol{z})$ is the joint posterior PDFs of the structural model modes and variances given the measurement data; $f(\boldsymbol{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is the joint likelihoods of observing the measurement data given the structural model modes and the variances; and $f(\boldsymbol{\mu}_c)$ and $f(\boldsymbol{\Sigma}_c)$ are respectively the joint prior PDFs of the structural model modes and variances.

To determine the joint posterior PDF in Equation 4.8, suitable joint prior PDFs must be selected, based on similar sites and available prior information, and the joint likelihood function must be evaluated. Here, the implementation steps required to work from the joint prior PDF to the joint posterior PDF when employing a statistical structural model as defined in the previous section are expanded.

This requires a nested simulation approach [*Maxwell et al.,* 1999] and it is detailed in the next several paragraphs. Figure 4.3 is a flow chart that augments the discussion.

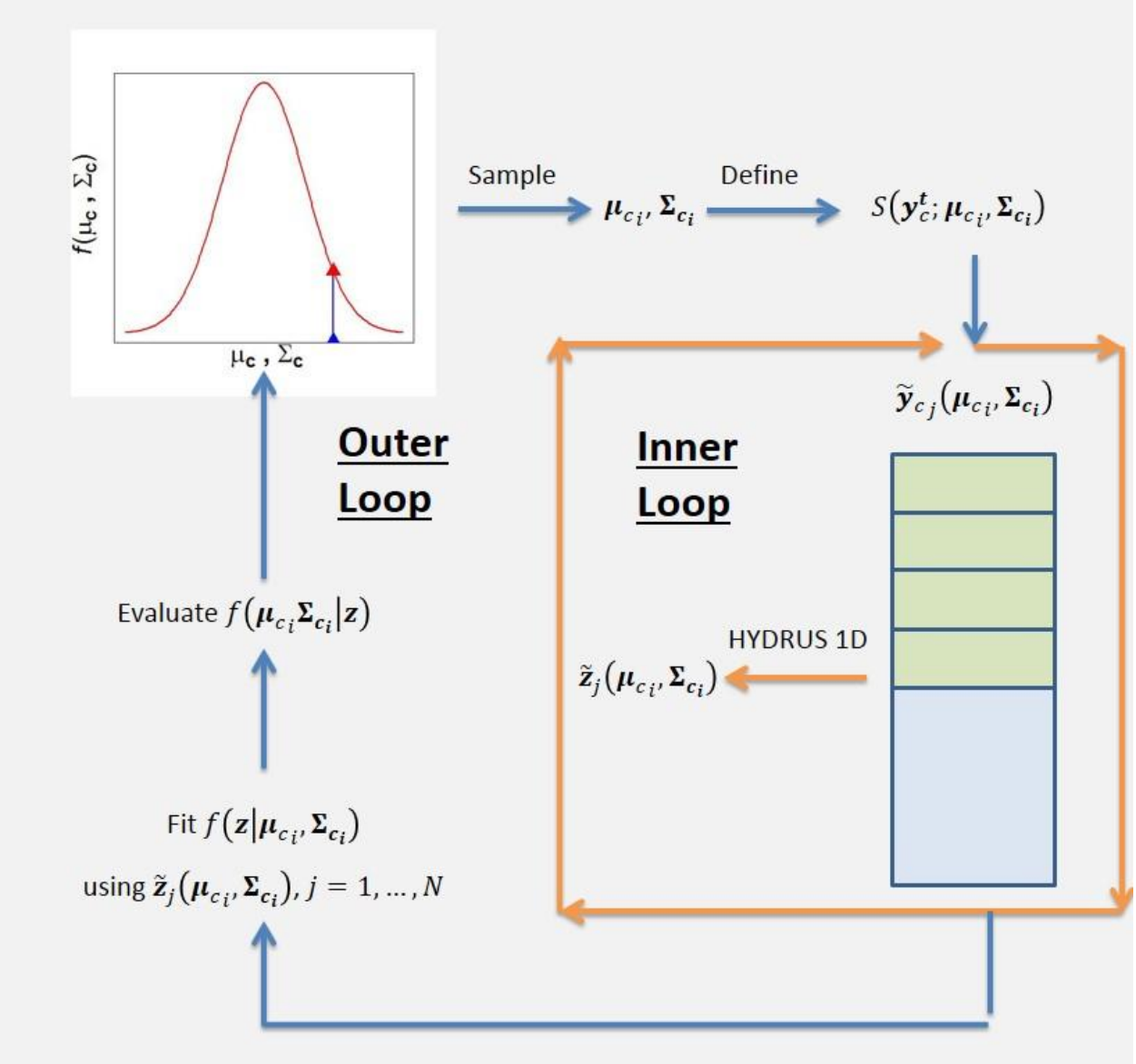

Figure 4.3: A flowchart visualization of the nested simulation procedure implemented for the case study.

Samples of $\boldsymbol{\mu}_{c_i}$ and $\boldsymbol{\Sigma}_{c_i} i = 1, \dots, N$ are obtained from the joint prior PDFs $f(\boldsymbol{\mu}_c)$ and $f(\boldsymbol{\Sigma}_c)$ by sampling from the supports over which $f(\boldsymbol{\mu}_c) > 0$ and $f(\boldsymbol{\Sigma}_c) > 0$. Sampling can be done in a variety of ways: for instance, randomly, in a stratified manner, or with Latin hypercubes, all of which can be compatible with the joint likelihood calculation that follows [*Isukapalli et al.,* 1998, *Over et al.*, 2013]. Each of these N samples of $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ then defines a structural model $S(\boldsymbol{y}_c^t; \boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ for $i = 1, \dots, N$, which are each used to generate their own ensemble of M random composite soil columns following the procedure of Section 4.1..2: $\tilde{\boldsymbol{y}}_{c_j}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ for $i = 1, \dots, N$ and $j = 1, \dots, M$. This results in a total of NM random realizations of the composite soil column.

Numerical simulation of water content on these ensembles of composite soil columns yields ensembles of measurement data $\tilde{\boldsymbol{z}}_j(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$, where the notation $\tilde{}$ is carried to remind that these are simulated on a random realization of the composite soil column and explicitly show the dependence on $\boldsymbol{\mu}_{c_i}$ and $\boldsymbol{\Sigma}_{c_i}$.

Each ensemble of M simulations is used to non-parametrically fit or estimate the necessary statistical parameters after validating a parametric form of the joint likelihood function $f(\boldsymbol{z}|\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$. Note here, the fundamental difference of this approach to assuming a joint likelihood function, because the distributional form can either be validated or directly inferred. Once the joint likelihood function is inferred from the simulation data, it is then evaluated for the measurement data $\boldsymbol{z}$. M must be sufficiently large to obtain converged estimates of either the likelihood function itself or, more cheaply, the statistical parameters that define the validated joint likelihood function. Importantly, simplifications can still be made to this procedure; however, the primary motivation for any simplification would now be computational parsimony rather than because of the need to close the Bayesian expression in Equation 4.8.

Repeating over all N samples and multiplying the joint prior PDFs and likelihood function yields the values proportional to $f(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}|\boldsymbol{z})$ for $i = 1, \ldots, N$. As N becomes large relative to the dimension of $\{\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}\}$, this permits stable regression of the posterior PDF through the points $\{\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}\}$ using the weights $f(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}|\boldsymbol{z})$.

The process described here is analogous to the implementation used in previous applications of MAD [*Rubin et al.*, 2010; *Murakami et al.*, 2011; *Chen et al.*, 2012; *Over et al.*, 2013] discussed in Section 1.2, but without geostatistical structural models. So even though anchors are not applied in this study - because without a geostatistical structural model, point spatial data is incompatible with a statistical structural models – this should still be considered an application of MAD.

Macroscopically, the two loops in Figure 4.3 are both essential and are coupled in a MAD analysis. The outer loop randomizes the possible structural model parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$, which is a step common to most model inversion approaches, but normally would randomize the forward model parameters $\boldsymbol{y}_c$ directly. The inner loop is used to repeat a random experiment that yields random simulations of $\boldsymbol{z}$ conditional on a specific value of $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$ (given by the outer loop), which allows the probability distribution of $f(\boldsymbol{z}|\boldsymbol{\Sigma}_c, \boldsymbol{\mu}_c)$ to be fit directly. This inner loop is not a step in other model inversion approaches.

The inner loop cannot yield random simulations of $\boldsymbol{z}$ without a structural model, because the outer loop would provide a value of the forward model parameters $\boldsymbol{y}_c$ and each simulation of the physics would yield identical $\boldsymbol{z}$ on the soil column identified by the outer loop. Thus in a traditional approach there is no way to randomize $\boldsymbol{z}$, which means there is no way to fit $f(\boldsymbol{z}|\boldsymbol{\Sigma}_c, \boldsymbol{\mu}_c)$ directly – the form must be assumed and there is no way to validate it. Hence, it is not merely the hierarchy that enables a data-driven validation and/or direct fitting of the likelihood function, but also the statistical representation of the forward model parameters – the structural model.

4.2.2.1 The Priors

Even though the complete structural parameter vector $\boldsymbol{\beta}_c$ properly contains all the elements of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$; because of their statistical independence [*Lukacs*, 1942], the prior is split into the two separate components in Equation 4.8. The random components $\boldsymbol{\mu}_c$ - the modes of each of the transformed hydraulic parameters in each layer – and the diagonal terms of $\boldsymbol{\Sigma}_c$ - the spreading of each of the transformed hydraulic parameters in each layer – respectively require the specification of joint prior PDFs $f(\boldsymbol{\mu}_c)$ and $f(\boldsymbol{\Sigma}_c)$.

Here, $f(\boldsymbol{\mu}_c)$ is specified and justified. By the principle of minimum relative entropy (MRE) [*Woodbury & Ulrych,* 1993, 1998; *Hou & Rubin*, 2005], when the first two statistical moments of a group of variables are available (without upper and lower bounds) then the least subjective choice of a prior distribution is MG. Several assumptions are stated and the parameterization of $f(\boldsymbol{\mu}_c)$ is introduced for the study.

Conservatively, the prior is defined such that $\boldsymbol{\mu}_c$ for the soil column can be selected from anywhere in the support of all the reported samples of the transformed hydraulic parameters in the same soil textural class – sandy loam. Moreover, it is assumed that the statistical distribution of the soil catalog samples of sandy loam soil is also appropriate for $\boldsymbol{\mu}_c$, i.e. that the structural model modes should be weighted with high probability for modal values where there is a high probability in the sandy loam soil catalog samples and low probability elsewhere. For the soil catalog samples of sandy loam soils, there are empirical estimates of the first two statistical moments from *Carsel & Parrish* [1988]. Thus the prior for the modes $\boldsymbol{\mu}_c$ can be least subjectively chosen to be MG; combining this conclusion from MRE principles with the assumption about the mode of the samples, define:

$$f(\boldsymbol{\mu}_c) = N(\boldsymbol{m}, \mathbf{E}), \tag{4.9}$$

with mean vector $\boldsymbol{m}$ of length 16 and covariance matrix $\mathbf{E}$ of size 16×16.

Definition of the mean vector $\boldsymbol{m}$ (Table 4.5) and diagonal covariance $\mathbf{E}$ (Table 4.6) – the tables only show the values from one layer, but all the layers are defined identically - completes the specification of $f(\boldsymbol{\mu}_c)$. The $\langle\cdot\rangle$ notation indicates the expected value (equivalent to the mode for MG distributions) of the argument. Note, that $\boldsymbol{m}$ is the mean of the structural model means $\boldsymbol{\mu}_c$ and $\mathbf{E}$ is the covariance matrix of the structural model means $\boldsymbol{\mu}_c$. This repetition could beget the awkward expression "the mean of the means", which is why $\boldsymbol{\mu}_c$ is referred to as the mode (equivalent to the mean for MG distributions) wherever the two may be confused.

Table 4.5: Segment of mean vector $\boldsymbol{m}$ for one layer of the soil column based on estimated means of the transformed hydraulic parameters soil catalog samples collected by *Carsel & Parrish* [1988].

	$\langle K_s^t \rangle$	$\langle \alpha^t \rangle$	$\langle n^t \rangle$	$\langle \theta_r^t \rangle$
$\boldsymbol{m}$	-1.76	-0.937	.634	0.384

Table 4.6: Block of the covariance matrix **E** for one layer of the soil column based on estimated variances of the transformed hydraulic parameters soil catalog samples collected by *Carsel & Parrish* [1988].

E	$\langle K_s^t \rangle$	$\langle \alpha^t \rangle$	$\langle n^t \rangle$	$\langle \theta_r^t \rangle$
$\langle K_s^t \rangle$	2.56	0	0	0
$\langle \alpha^t \rangle$		0.313	0	0
$\langle n^t \rangle$			1.85E-3	0
$\langle \theta_r^t \rangle$				9.42E-2

Briefly, the reason for the choice of $\boldsymbol{m}$ and **E** is discussed. Defining a larger support for $\boldsymbol{\mu}_c$ with a more diffuse $f(\boldsymbol{\mu}_c)$ with diagonal covariance matrix *a priori* [*Woodbury & Ulrych, 1993, 1998; Hou & Rubin,* 2005] is most conservative. If a correlation structure exists within a layer and/or across layers between the components of $\boldsymbol{\mu}_c$ and/or $\boldsymbol{\mu}_c$ should have smaller variance, then these features may be exhibited after conditioning in the posterior PDFs; they do not have to be specified initially. Hence, the final definition for $\boldsymbol{m}$ is simply the sample-based expected values and for **E**, a diagonal matrix, containing the un-scaled sample-based variances from the work of *Carsel & Parrish* [1998] for sandy loam.

Now, the specification of $f(\boldsymbol{\Sigma}_c)$ is introduced. Here, it is assumed that the empirically estimated values of the covariance matrix [*Carsel & Parrish,* 1988] provide limiting maxima on the variance that can be observed within a soil textural class. By definition, the variance must also be non-negative, which is taken as a limiting minimum. When only the upper and lower boundaries for a random variable are specified, the least subjective PDF by the principle of MRE is a uniform distribution. Therefore,

$$f(\boldsymbol{\Sigma}_c) = U[\mathbf{0}, \boldsymbol{E}_{diag}], \tag{4.10}$$

where $\boldsymbol{E}_{diag}$ is the vector of the entries of **E** on the diagonal – repeated four times, once for each layer - that were given in Table 4.6.

All samples are drawn in this study using Latin hypercube sampling [*McKay et al.,* 1979].

4.2.2.2 Likelihood Function

In this section, the joint likelihood function $f(\boldsymbol{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ that must be evaluated to obtain the joint posterior PDF $f(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|\boldsymbol{z})$ is introduced.

The vector $\boldsymbol{z}$ contains all 352 measurement times from the 4 measurement locations available at the Grenzhof site (Figure 4.2), but this is a problem of extremely large dimensionality for non-parametric methods, requiring intractable numbers of simulations for stable inference. Therefore, in this case study, a parametric form of the likelihood is selected and validated - based on the statistical distribution of the simulations of $\boldsymbol{z}$ (inner loop of Figure 4.3.

Investigation of the distribution of simulations of water contents in time and space revealed a very good agreement with an MG structure using the *Shapiro-Wilk* test. The average p-value of these tests was 0.07, which indicates little to no evidence against the hypothesis of normality [*Wasserman,* 2010], so the likelihood function $f(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c|\boldsymbol{z})$ is taken as MG.

To evaluate the likelihood functions, the sample mean vector and sample covariance matrix for the data $\boldsymbol{z}$ using the simulation ensembles generated from the samples of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ need to be calculated. Note that this mean is neither $\boldsymbol{\mu}_c$ or $\boldsymbol{m}$ and this covariance matrix is neither $\boldsymbol{\Sigma}_c$ or $\mathbf{E}$, but rather the sample mean and sample covariance matrix of $\tilde{\boldsymbol{z}}_j(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i})$ over the index j – using the outcome of the inner loop in Figure 4.3. It is important to note that these sample-based calculations are carried out separately over the index i – once for each circuit of the outer loop in Figure 4.3 – and this results in a total of N calculations of this vector and matrix.

The dimension of $\boldsymbol{z}$ helps define the number of simulations M to generate in each ensemble, because estimates of this covariance matrix components are mathematically ill-posed [*Deng & Yuan,* 2009] when the simulation ensemble is smaller than or equal to the length of $\boldsymbol{z}$. Combining the recent work of *Vershynin* [2012] – which importantly showed that logarithmic oversampling is not necessary for sample covariance matrix determination in cases where the data are independent (a condition forced by projection onto an orthogonal basis using singular value decomposition [*Albano et al.,* 1988] before analysis) - with the fact that $\boldsymbol{z}$ is length 1408, $M = 1450$ is set as the ensemble size for the likelihood function analysis.

4.2.2.3 Sequential Partitioning of the Structural Parameter Domain: Computational Parsimony

In this section, a partitioning procedure for the structural parameter randomization (outer loop of Figure 4.3) is introduced to offset the significant computational costs of embedding the likelihood function validation (inner loop of Figure 4.3).

The net computational cost of an analysis can be approximated as NM times the cost of a single simulation – the overall costs contributions of evaluating the PDFs, sampling, etc. are quite insignificant by comparison. M was set at 1450 in the previous section, without a practical strategy for limiting N, the product NM could easily climb into billions or trillions. For reference, even if a simulation cost just one second, a billion simulations would take to over 317 years to compute.

However, N must grow exponentially with the dimensionality of $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$ to ensure convergence of the posterior joint PDF [*Vershynin*, 2012]. Therefore, a strategy common to MCMC algorithms is borrowed that is both computationally affordable and maintains the quality of the statistical results. The strategy is to hold $\boldsymbol{\Sigma}_c$ constant initially and to

randomize just $\boldsymbol{\mu}_c$; then subsequently hold $\boldsymbol{\mu}_c$ fixed and randomize $\boldsymbol{\Sigma}_c$. In MCMC approaches, this process would be repeated multiple times. In the limit of a large number of samples, this sequential, partitioning approach has been shown to determine the complete joint probability distribution of the random variables [*Robert & Casella,* 2004]

Initially, it is assumed that the covariance matrix is fixed using the empirically determined values from *Carsel & Parrish* [1988] given in Table 4.7. Note, the diagonal is identical to Table 4.6 (the limiting maxima) and permits the most conservative search of the structural modal modes - the greatest number of structural models could have non-zero posterior probability, because the hydraulic parameters are the most spread around their respective structural model modes when the covariance matrix diagonal is largest.

Table 4.7: Covariance matrix values for sandy loam texture class (after applying the transformations given in Table 4.2) MG structural models

$\boldsymbol{\Sigma}$	K_s^t	α^t	n^t	θ_r^t
K_s^t	2.56	-0.245	5.92E-2	0.338
α^t		0.313	3.49E-3	-0.137
n^t			1.85E-3	4.78E-3
θ_r^t				9.42E-2

During randomization of $\boldsymbol{\mu}_c$, out of the 19,000 samples searched, one sample had a dominant posterior probability (factor of 10^6 greater than the rest of the samples). Therefore, during the subsequent randomization of $\boldsymbol{\Sigma}_c$, $\boldsymbol{\mu}_c$ was fixed at this dominant value. Owing to the high computational cost (discussed in Section 4.3.3) and the immediate goal of establishing a validated likelihood function for Bayesian inversion in vadose zone characterizations, it is not necessary to proceed with further iterations of partitioning of the random structural parameters. It is acknowledged that this may be a limiting element of the analysis.

4.3 Results and Discussion

In this section, the results of the characterization are presented and discussed. In the first subsection, the complete PDFs of the parameters are presented. Additionally, prior and posterior confidence intervals (CIs) of the water content time series are presented to graphically demonstrate the improvement in the posterior structural models' ability to reproduce the observed water dynamics. The second subsection propagates the uncertainty in the posterior PDFs of the Mualem-van Genuchten parameters into uncertainty in the water retention and hydraulic conductivity functions. In the final short subsection, there is a summary of the computational resources as well as budget. Throughout the sections the strengths and weaknesses of our the approach are discussed.

4.3.1 Posterior Updating of Prior

Figure 4. 4 shows the posterior (red) and prior PDFs (black) of the Mualem-van Genuchten parameters under study. These are obtained by propagating the uncertainty in the prior and posterior Gaussian structural models into uncertainty of the hydraulic parameters. Each row contains one of the four targeted parameters K_s, α, n or θ_r and each row the four layers in the profile (shallowest to deepest, respectively right to left). The expected values of the posterior and prior distributions are graphed as a bold circle following the respective color schemes. Additionally, the plots show the optimized parameter value for the layer attained by *Wollschläger et al.* [2009] as a vertical blue line.

Optimally, conditioning results in a posterior PDF that is narrower (less variance) that is also more centered on the true parameter value (less bias) than its prior PDF. However, be mindful that the true parameters, in this case, are unknown, so bias is not identifiable; a shift of the expected value (the bold circles) towards or away from the optimized value (the vertical line) after conditioning cannot be interpreted as a reduction or increase in bias. The results in Figure 4.4 demonstrate that the water content measurement data used in the likelihood function can condition some, but not all, of the textural class specific (sandy-loam) prior estimates of the structural model (random distribution of the Mualem-van Genuchten parameters) to site-specific (Grenzhof soil profile), posterior estimates. The distributions for $\theta_r, n, \& \alpha$ show significant conditioning effects compared to their respective priors (different modes and skewness in the posterior PDFs), whereas K_s is hardly changed by conditioning.

Additionally, these plots suggest that a prior based on the textural class of the soil is compatible [*Hou & Rubin*, 2005] with the underlying, true parameters for the site, because the posterior PDF is nonzero for all 16 variables, however subsequently it will be shown that this may not be the case. The agreement between the optimization of *Wollschläger et al.* [2009] and the posterior conditioning is also quite successful, although the optimized K_s and θ_r in layer 1 as well as n in layers 3 and 4 are well into the tail regions of their respective PDFs.

It is important to note that the purpose of this comparison with *Wollschläger et al.* [2009] is to show the differences between the MAD approach and an optimization that utilized the same data. The parameters identified by *Wollschläger et al.* [2009] are not considered to be the optimum – although they may be - as this is not a synthetic problem. The most notable improvement from using MAD is the ability to specify the conditional PDFs of the hydraulic parameters, where *Wollschläger et al.* [2009] could only provide optimal values and 95% confidence bounds.

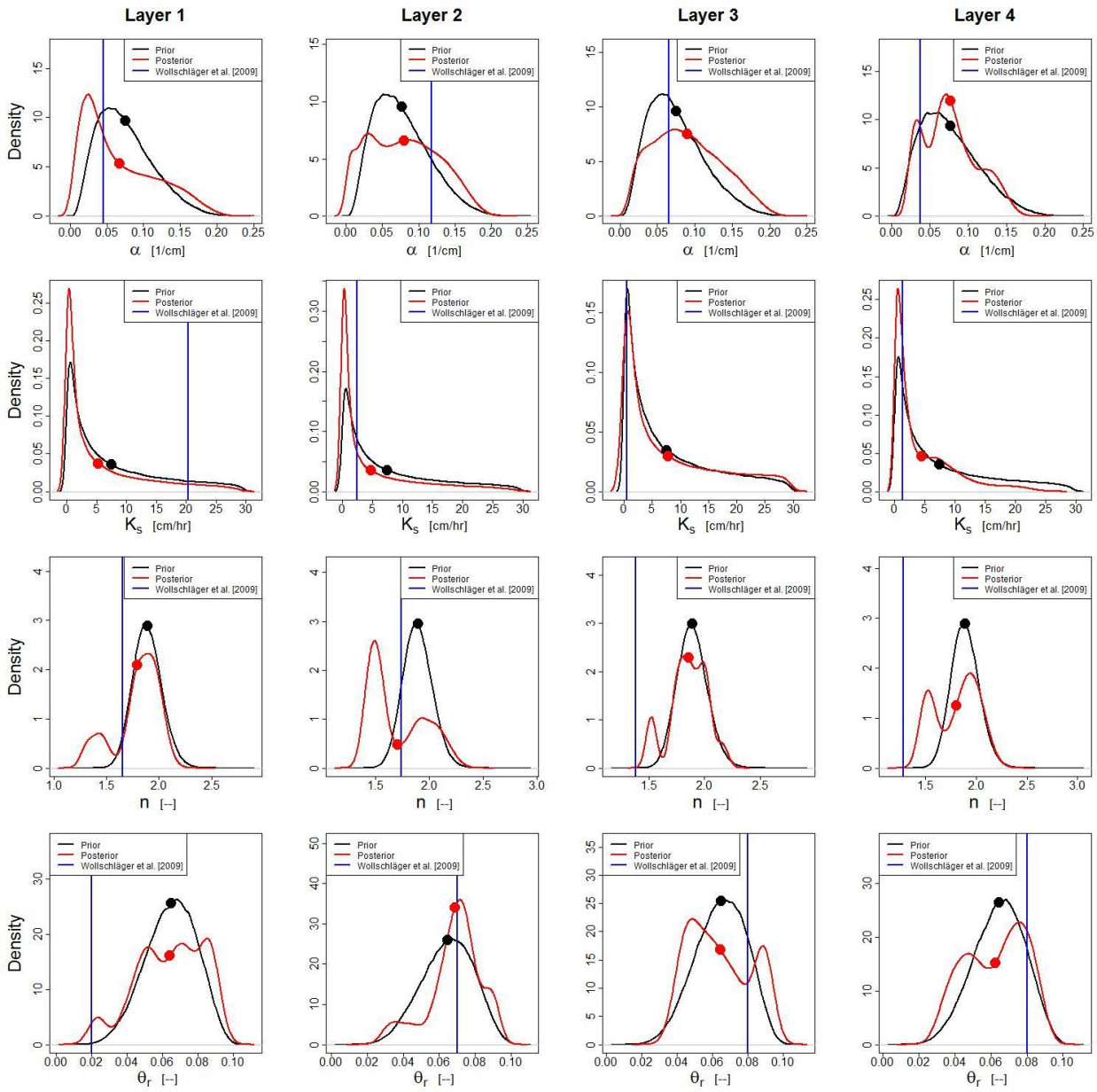

Figure 4.4: Posterior, prior, and optimized values of the Mualem-van Genuchten parameters for the Grenzhof field site [*Wollschläger et al.*, 2009]. The heavy dot represents the expected value of the respectively colored PDF.

The posterior and prior 95% CIs of the water content time series show a similar effect to the variance reduction in the posterior PDFs in Figure 4.4, where for the most part the conditioning improves the identifiability of the observed data.

The results in this section are comparable to the recent Bayesian hydraulic parameter inversions of *Scholer et al.* [2011 & 2012]. Specifically, these studies also investigated layered profiles and showed very similar orders of magnitude in the overall parametric uncertainty as well as similar effects of conditioning, but using water contents estimated from ground penetrating radar measurements instead of moisture content data obtained from TDR. However, the results obtained with MAD utilized a validated likelihood function

rather than an assumption of independent and identically distributed Gaussian residuals [*Scholer et al.,* 2011 & 2012].

Figure 4.5 shows the posterior (blue) and prior PDFs (red) CIs – overlapping portions of the prior and posterior CI are shaded in gray - of the water content time series relative to the observed data (black). The four panels in the figure correspond with the layers of the soil column, from shallowest to deepest moving down the graphic.

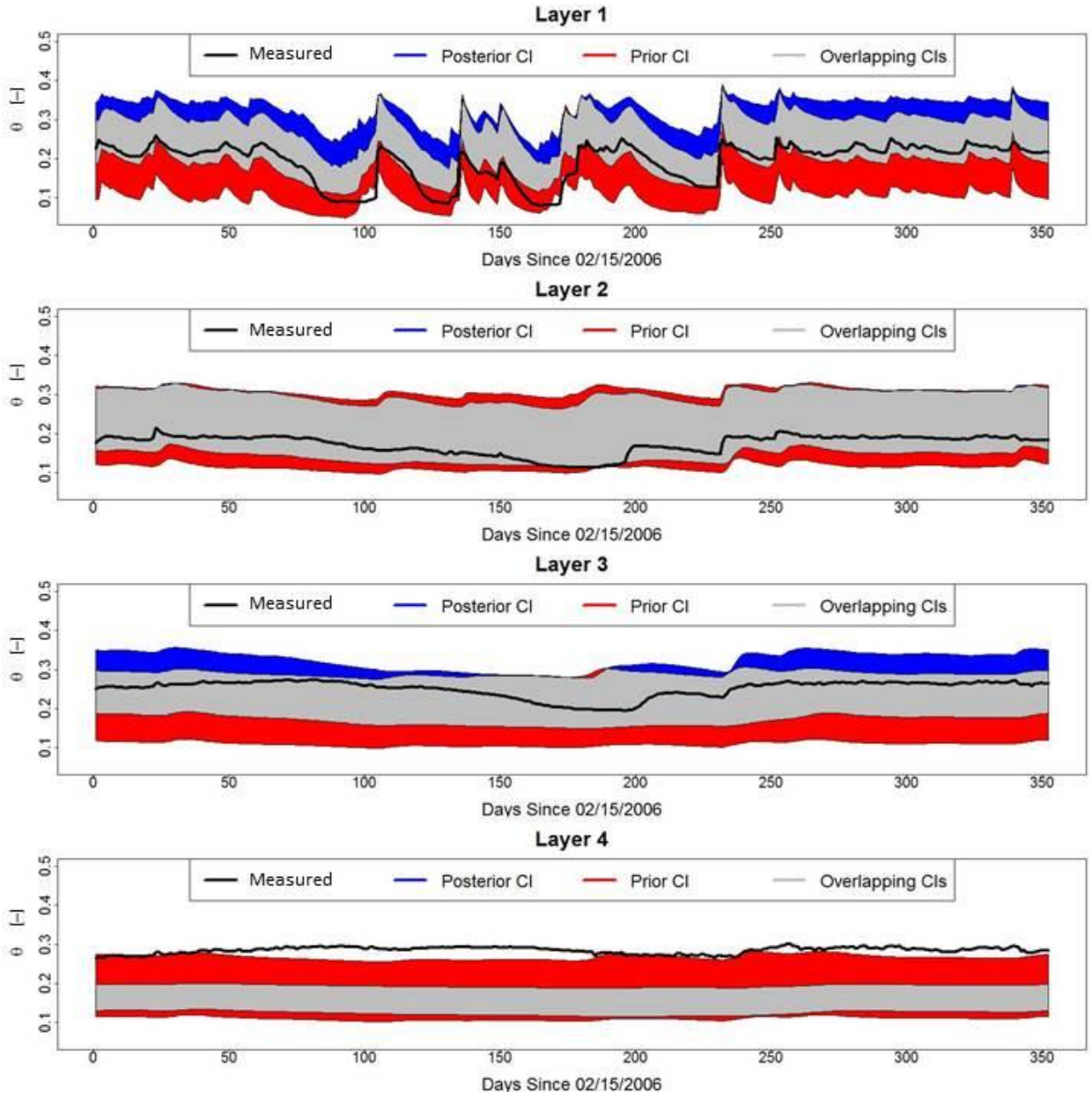

Figure 4.5: Posterior and prior 95% confidence intervals of the water content time series in the four layers (shallowest to deepest reading down).

For every one of the 1,408 daily averaged measurements (352 per layer), the posterior CI is narrower than the prior CI, which is a beneficial result of the reduced hydraulic parameter

uncertainty (Figure 4.4). The posterior CIs enclose nearly 90% of the observed water content measurements in the uppermost 3 layers. However, they still have a relatively high level of uncertainty (are quite wide) in terms of the water content of around ~10% , which hypothetically is an outcome of having only one TDR probe per layer and only one type of data to condition upon. Also, in these layers, and most prominently in the uppermost layer, the posterior CI is least effective in capturing the water content dynamics when the soil column is at its driest states – approximately days 80 through 180 in layer 1 and around day 180 in layer 2. This behavior was also demonstrated by *Wollschläger et al.* [2009] and several reasons for the model's inability to reproduce lengthy dry periods were given: most notably, the usage of a steady-state crop factor and an evaporative flux that was not dependent on preceding rainfall or drying events, which could also be responsible for the incorrect behavior in this analysis.

Any observation where the observed water content is not enclosed by either the posterior or prior CIs is indicative of an undesirable model bias. In the lowest layer, both the prior and posterior CI fail to capture the water content time series throughout the year, both underestimate the moisture. This failure could be indicative that the likelihood function, because it evenly accounts for agreement of the simulations and measurements at any depth and time, is also susceptible to "trade-off" between fit in any one layer versus a more favorable fit throughout the column (as is the case with Pareto surface solutions *Wöhling & Vrugt* [2011]). Thus the success in the upper portion of the soil column may offset the poor fit in the fourth layer.

The failure in the fourth layer, where the prior CI is not a good match with the observed time series, is a prime example of a poorly specified prior – here, the texture class based assumptions do not perform well. Determination of the prior CI is only possible after performing a large number of forward model simulations; while it would be correct to refine the prior for this layer and re-analyze the parameters, because of the high computational cost and because the purpose of this chapter is to demonstrate a framework for validating the likelihood function in vadose zone parameter inversion problems, such an effort is beyond the scope of this work.

4.3.2 Soil Hydraulic Properties

In this section, the uncertainty in the hydraulic parameters is visualized as uncertainty in the water retention and hydraulic conductivity functions for the 4 soil layers at the Grenzhof site. Figures 4.6 and 4.7 respectively contain the water retention and hydraulic conductivity functions. The posterior (red) and prior (black) 95% CIs are plotted to show the effect of conditioning. Additionally, the optimized functions from the work of *Wollschläger et al.* [2009] are plotted in blue. The gray background shows the range of water content measurements observed during the experiment for a given layer (listed in Table 4.4).

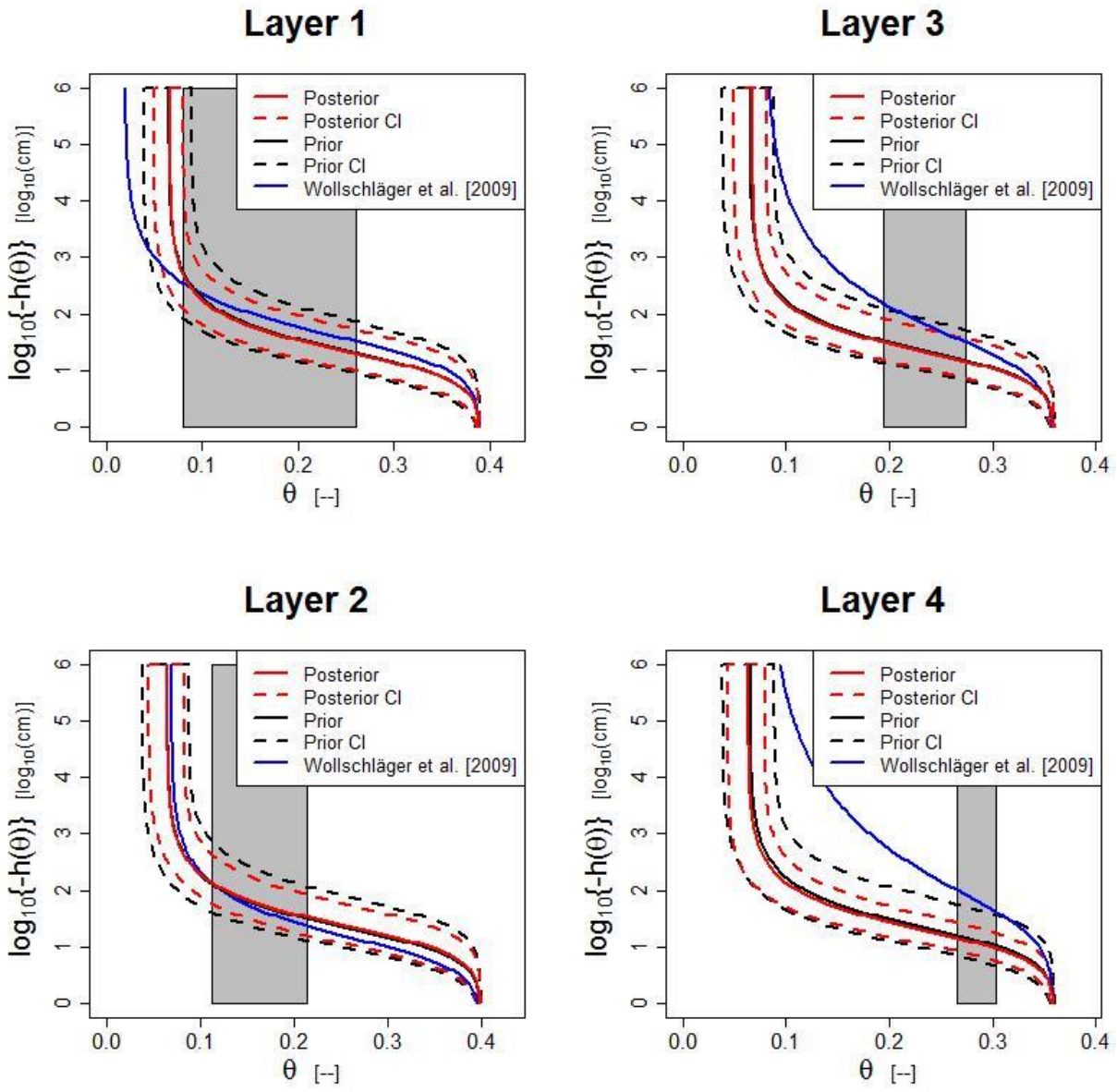

Figure 4.6: Water retention function posterior and prior confidence intervals as well as optimization [*Wollschläger et al.*, 2009] for the four soil layers characterized. The range of observed water content for each layer is shown in gray.

An important note is that the prediction of the water retention and hydraulic conductivity functions are always limited by the range of observed water contents under natural conditions. Thus, since the available conditioning data does not span the complete range of water content states as thoroughly as e.g. a multi-step outflow experiment conducted in the lab does, it should be clearly acknowledged that the optimization results and CIs for states outside the measured ranges are extrapolations.

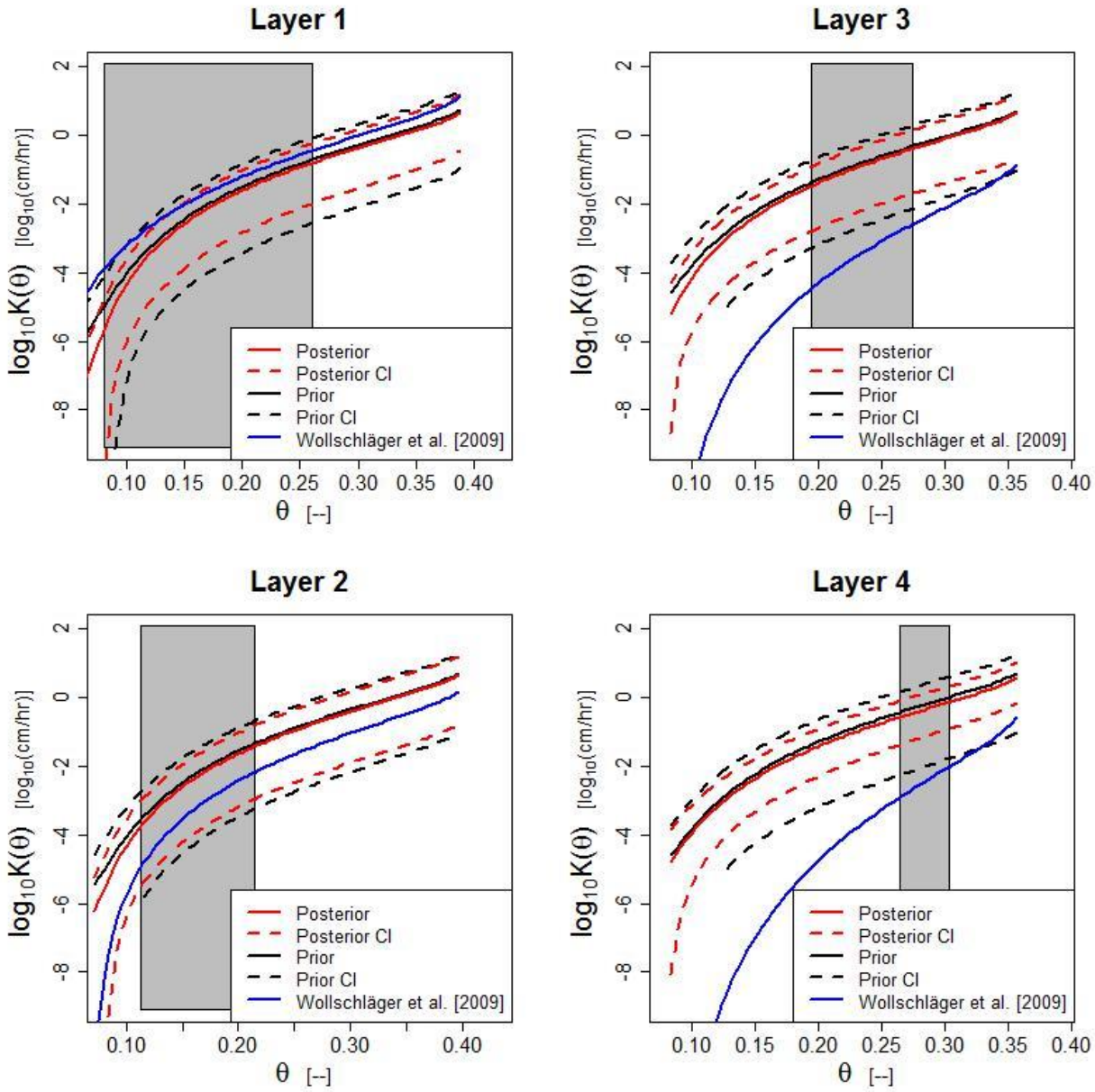

Figure 4.7: Hydraulic conductivity function posterior and prior confidence intervals as well as optimizations [*Wollschläger et al.,* 2009]for the four soil layers characterized. The range of observed water content for each layer is shown in gray.

In Figure 4.6, the positive effect of conditioning is again seen in the reduced variance in the hydraulic parameters of the posterior CIs being narrower than the prior CIs for each layer. The agreement with the optimized water retention functions from *Wollschläger et al.* [2009] is best in the uppermost layers, but deteriorates with greater depth.

Again, Figures 4.6 and 4.7 are also quite comparable to another recent Bayesian inversion of vadose zone properties conducted by *Scholer et al.* [2012]. Specifically, their results obtained with a correlated MG prior show almost identical conditioning effects (albeit using water contents obtained from ground penetrating radar travel times instead of TDR measurements), for the water retention curve. Yet again, it is reminded that the results

presented in this chapter are comparable in terms of results in other recent studies, but that the MAD procedure allowed validation of the likelihood function rather than simply assuming it.

Returning to the failure of the water content to correspond with the observations in the deepest layer (Figure 4.5), the prior and posterior CIs of the water retention function in the deepest layer show a fairly sharp transition from wet to dry conditions (roughly 100 cm of matric potential separates moistures states of $\theta = 0.35$ and $\theta = 0.1$). The recent work of *Wöhling & Vrugt* [2011], which considered a Pareto solution to a multi-layer soil profile vadose zone inversion (in a sandy soil), showed that the forward model uncertainty can lead to CIs for matric potential that are on the order of $\sim$50 cm. Assuming the Grenzhof conceptual model can be considered analogous (albeit with a slightly less porous sandy loam soil and a shallower column), such model uncertainty could easily account for the serious bias ($\sim$15%) in the moisture content time series (given the water retention function CIs) in the lowest layer. This model bias also suggests that an additional measurement of matric potential could help better condition the results (as suggested by *Vereecken et al.,* 2008).

Figure 4.7, also demonstrates the positive outcome of the reduced variance in the hydraulic parameters by the posterior CIs being narrower than the prior CIs for each layer. The agreement with the optimized hydraulic conductivity functions from *Wollschläger et al.* [2009] is best in the two uppermost layers, but deteriorates with greater depth.

In general, there are three critical observations to be drawn from Sections 4.3.1 and 4.3.2. First, the priors that we selected on the basis of a soil textural class and the transformations utilized from soil catalogs [*Carsel & Parrish,* 1988] are in most cases compatible (except the fourth layer) with the observed moisture content data for the Grenzhof field site. Second, the beneficial effects of reducing the variance of the hydraulic parameters by conditioning is better visualized in the CIs of the water content predicted by the model or the hydraulic properties than in the PDFs of the hydraulic parameter directly. Third, the results obtained with MAD compare well with other recent work in vadose zone parameter estimation; but, where using MAD stands apart, is that the statistical framework has allowed for any assumptions about the likelihood function to either be avoided altogether or validated during the analysis.

4.3.3 Computational Considerations

In this section, the computational costs of this application are summarized. During randomization of the structural model means $\boldsymbol{\mu}_c$ 19,000 samples were drawn. During randomization of the structural model variances $\boldsymbol{\Sigma}_c$ 9,000 samples were drawn. Fewer samples were drawn during the analysis of $\boldsymbol{\Sigma}_c$ than $\boldsymbol{\mu}_c$ because $f(\boldsymbol{\Sigma}_c)$ is bounded and $f(\boldsymbol{\mu}_c)$ is not. The convergence in both analyses was verified using a visual jackknife technique [*Wasserman,* 2010]. However, given the single iteration of the partitioning approach for

the parameters, it is hypothesized that the posteriors could be further refined with additional simulations.

With $M = 1450$, the total number of simulations for the analysis was 40,600,000. The overall computation time for the analyses was equivalent to just over one year on a 2 core CPU, based on a 3.0 GHz processor and an average simulation cost of 1.6 seconds. Because the net simulation expense was distributed using Condor High Throughput Computing [*Basney & Livny*, 2000] over 200 CPUs, there is not one set of processor specifications that is exactly representative.

The entire composite profile random generation and simulation process was performed using the MAD# open-source software [*Osorio et al.,* 2013] connected with HYDRUS 1-D version 4.14) [*Simunek et al.* 2008] and the R statistical computing platform version 2.15.1 [*R Core Team*, 2012].

4.4 Summary and conclusions

In this chapter, an analysis of uncertainty in vadose zone hydraulic properties using MAD was developed and demonstrated. The approach is identified PDFs of the hydraulic parameters (Figure 4.4). This uncertainty can be propagated into the hydraulic properties (Figures 4.6 & 4.7), into the forward model's ability to fit the observations (Figure 4.5), and into predictions made by the forward model.

The largest success of the analysis was that the posterior CIs of the moisture content time series, the water retention functions, and the hydraulic conductivity functions were all narrowed relative to their prior CIs. The inversion was more successful in matching the measured data in the shallowest layers, but struggled to fit the lowest layer and prolonged dry conditions nearest to the surface. There were possible statistical (poorly defined prior, no correlation between layers) and forward modeling reasons (constant crop factor and evaporative flux see section 4.1.1.1) for why the fit was poor in the first and fourth layer. Lastly, the CIs (posterior or prior) for moisture content were rather wide, but this is thought to be an outcome of having only a single measurement device and type per layer for conditioning; it is reasonable to expect that additional probes per layer or alternative types, e.g. matric potential measurements, albeit with their well-known difficulties in the dry water content range, could further reduce the parametric uncertainty

Also of importance is that the MAD approach, which utilizes nested simulation (Figure 4.3 allows the likelihood function to be inferred, rather than assumed initially. The form of this PDF can be directly derived from the ensembles of simulation data. If an assumption is made about the shape of the likelihood function, it can be validated with the MAD approach. In this study, an MG form of the likelihood function was validated using the Shapiro-Wilk test statistic to show there was little to no evidence against the hypothesis of

normality. The only reason a parametric form of the likelihood was utilized was to limit the overall computational cost, it was not a requirement of the framework.

There are several areas for improvement in this work. All of which require additional computational resources. First, the incompatible prior in the fourth layer could be revised to better agree with the observation data. Second, the numerical model could be less simplistic and additional elements – such as hysteresis or swelling - could be included that may better represent the *in situ* physics. Third, the sequential analysis of the structural parameters could be iterated or, even better, could be avoided altogether in a joint analysis. Finally, the likelihood function could be non-parametrically inferred and no simplification employed to limit computational cost.

In conclusion, this chapter presented useful modifications to the MAD theory with structural models that are not geostatistical and showed that this framework allows a more extensive analysis of the likelihood function than any other Bayesian inversion, optimization, or calibration technique used previously in the vadose zone.

5. Conclusion

Since its introduction at the beginning of Chapter 1, the method of anchored distributions has been the central theme of this dissertation.

In Chapter 1, MAD was demonstrated to be among the most current and advanced efforts in the field of model inversion – a distinction justified by the comparison with the relative advantages of MAD compared to precursor techniques, such as ML, MAP, GLUE, SCE, SCE-UA, DREAM, and PPM. A common difference which separates MAD from all the other methodologies is a completely generalized, data-driven, assumption-free likelihood function. If it is necessary to invoke assumptions about the likelihood function, for reasons of computational parsimony perhaps, such assumptions can be validated using MAD, which heretofore has never been possible with the other techniques. MAD has this capacity because of the nested simulation framework and the use of structural models.

Also in Chapter 1, the capability of MAD to characterize spatial heterogeneity at multiple resolution scales was introduced, which is a benefit of utilizing anchors and structural models in a complimentary fashion. The technique was also shown to be very flexible in incorporating multiple types and scales of measurement data, which is a benefit of the data categorization in MAD. After introducing the fundamental elements of MAD, the possible variants of the Bayesian proportionality were expanded to show the flexibility of MAD for applications with different characterization goals and available data. Finally, because MAD was presented generically, it was emphasized that MAD is not itself a product of any one scientific field or model, but rather can be applied with great robustness in many areas. However, MAD, because of its flexibility, is not a trivial analysis in terms of computational cost.

The second chapter introduced the 'bundling' approximation technique for MAD that was computationally cheaper, but still quite accurate. Bundling is a modification of the likelihood function that determines the probability of observing the Type-B measurements given a *set* of anchor and structural parameter samples. It was shown that this representation of the likelihood function is approximately representative of the likelihood function for any of the members of the set using a Boolean 'or' statement under the special condition that the anchor and structural parameter samples in the set be very similar. This condition was forced by a pre-screening of the samples using clustering tools, specifically partitioning around medoids. A rule was established for using clustering to set the bundle size, which was shown to have a direct impact on the accuracy of the approximation.

Where bundling gains its efficiency is in the construction of the simulation ensembles, because the overall simulation cost is cheaper than with traditional MAD implementations. Bundling was applied to a synthetic case study of a natural gradient tracer test and shown to be approximately as accurate as the traditional MAD implementation, but 35% cheaper in terms of computational expense. Bundling is suggested as an initial 'screening' step for

applications of MAD that have large parameter spaces, such that high probability regions can be rapidly identified – if greater resolution of these regions of the probability surface are required, it can be achieved with subsequent, traditional applications of MAD, that focus solely on these regions and not the entire parameter space. A drawback of bundling is that it further complicates the implementation of a technique that already required significant skill in data manipulation, programming knowledge, and large amounts of data storage space.

Therefore, in the third chapter, the focus shifted again towards the development of a generic, modularized, open-source GUI that automates and simplifies the implementation of MAD. The MAD software is comprised of a pre-processing and post-processing module, which have the respective responsibilities of configuring a MAD project to build the simulation ensembles and to analyze the ensembles using the correct formulation of the Bayesian proportionality. Since the first two chapters heavily focused on the implementation of MAD, it was appropriate to first present the GUI forms in great detail. For every form, the role, the connection to the theory, and the logic of the form's location in the overall sequence were discussed. After the building blocks of the MAD software GUI were catalogued, it was possible to show how the individual components comprise the two modules of the software and how the software promotes generic attachment to different forward modeling and random field generation tools.

To demonstrate the generic ability of the MAD software to adapt to any of the variants of the MAD proportionality, a synthetic or real field case study using each of Equations 1.2-1.7 was presented either in Chapters 3 or 4. To demonstrate the interoperability of the MAD software to connect with and different forward models, analyses were conducted with both MODFLOW and HYDRUS 1-D. However, even though the software had been constructed to support applications without geostatistics, MAD had never been applied in such a fashion, nor it had been applied outside of aquifer characterization under saturated conditions.

Therefore, in the fourth chapter, the focus shifted towards developing and justifying the theory of MAD without the use of anchors and for characterization of a shallow soil column under variably saturated conditions. Analogously to geostatistical applications, a structural model needed to be specified in the various layers in the profile that represented the targeted hydraulic parameters as random variables. Argumentation was offered on the basis of soil texture class and the extensive empirical evidence of *Carsel & Parrish* [1988] that the randomness in the hydraulic parameters was of multivariate Gaussian form. Prior distributions for the structural parameters – in this study, the mean vector and covariance matrix – were selected using similar site data from the soil catalogs and the principle of minimum relative entropy.

After updating the prior PDFs using the available water-content time series, it was shown that the overall uncertainty in the hydraulic properties and confidence intervals of the forward model were effectively reduced. An important distinction of the conditioning in this application is that, relative to the existing literature in vadose zone model inversion and calibration, it is the first example of an inversion that validates the distributional assumption about the likelihood function. The inversion did have a drawback insofar as the lowest layer moisture content was not well fit after conditioning, but this could be an

artifact of the equal consideration of 1408 measurements in the likelihood function or the relatively sharp water retention function identified in this layer. On the other hand, the conditioning in the more shallow soil layers, the compatibility of the priors that were selected by texture class, and the form of the structural model for vadose zone hydraulic parameters were all successful and novel contributions of this study.

In closing, this dissertation has had at its core the theme of advancing the efficient implementation, usage, theory, and application areas of the method of the anchored distribution.

6. Works Cited

Alcolea, A., J. Carrera, and A. Medina (2006), Pilot points method incorporating prior information for solving the groundwater flow inverse problem, *Adv. Water Res.,* 29, 1678-1689, doi: 10.1016/j.advwatres.2005.12.009.

Albano, A. M., J. Muench, C. Schwartz, A. I. Mees, and P. E. Rapp, (1988), Singular-value decomposition and the Grassberger-Procaccia algorithm, *Phys. Rev. A,* 38(6), 3017-3026, doi: 10.1103/PhysRevA.38.3017.

Balakrishnan S., A. Roy, M. G. Ierapetritou, G. P. Flach, and P. G. Georgopoulos (2005), A comparative assessment of efficient uncertainty analysis techniques for environmental fate and transport models: application to the FACT model, *Journal of Hydrology*, 307(1-4), 204-218, doi: 10.1016/j.jhydrol.2004.10.010.

Basney, J. & M. Livny (2000), Managing network resources in Condor. In *High-Performance Distributed Computing, 2000. Proceedings.,The Ninth International Symposium on,* 298-299, IEEE. doi: 10.1109/HPDC.2000.868666*.*

Bates, B. C. & L. R. Townley (1988), Nonlinear, discrete flood event models, 1. Bayesian estimation of parameters, *J. Hydrol.* 99, 61-76, doi: 10.1016/0022-1694(88)90078-9

Beven, K. J. & A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Process.*, 6, 279-298.

Beven, K. J.& J. Freer (2001), Equifinality, data assimiliation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, *J. Hydrol.,* 249, 11-29, doi: 10.1016/S0022-1694(01)00421-8.

Binley, A, & K. Beven (2003), Vadose zone flow model uncertainty as conditioned on geophysical data, *Groundwater,* 41(2), 119-127, doi: 10.1111/j.1745-6584.2003.tb02576.x.

Beven, K. & A. Binley (2013), GLUE: 20 years on, *Hydrol. Process.*, doi: 10.1002/hyp.10082.

Bjornstad, B. N., J. A. Horner, V. R. Vermuel, D. C. Lanigan, and P. D. Thorne (2009), Borehole completion and conceptual hydrogeologic model for the IFRC Well Field, 300 Area, Hanford Site, PNNL-18340, Pacific Northwest National Laboratory, Richland, WA.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.,* 36(12), 3663-3674, doi: 10.1029/2000WR900207.

Capehart, W. J. & T. N. Carlson (1997), Decoupling of surface and near-surface soil water content: A remote sensing perspective, *Water Resour.Res.* 33(6), 1383-1395, doi: 10.1029/97WR00617.

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

Carrera, J., A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten (2005), Inverse problem in hydrogeology, *Hydrogeology Journal*, 13(1), 206-222, doi: 10.1007/s10040-004-0404-7.

Carsel, R. F. & R. S. Parrish (1988), Developing joint probability distributions of soil water retention characteristics, *Water Resour. Res.,* 24(5), 755-769, doi: 10.1029/WR024i005p00755.

Castagna, M., and A. Bellin (2009), A Bayesian approach for inversion of hydraulic tomographic data, *Water Resour. Res.*, 45, W04410, doi: 10.1029/2008WR007078

Certes, C., & G. de Marsily (1991), Application of the pilot point method to the identification of aquifer transmissivities, *Adv. Water Res.*, 14(5), 284-30, doi: 10.1016/0309-1708(91)90040-U.

Chen, X., H. Murakami, M. S. Hahn, G. E. Hammond, M. L. Rockhold, J. M Zachara, and Y. Rubin (2012), Three-dimensional Bayesian geostatistical aquifer characterization at the Hanford 300 Area using tracer test data, *Water Resour. Res.,*48, W06501, doi: 10.1029/2011WR010675.

Chiang, W. H. & W. Kinzelbach (2001), *3D-groundwater modeling with PMWIN,* Springer, Berlin, Germany.

Cirpka, O. A. and P. K. Kitanidis (2000a), Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments, *Water Resour. Res.,* 36(5), 1221-1236, doi: 10.1029/1999WR900354.

Cirpka O. A. and P. K. Kitanidis (2000b) Sensitivity of temporal moments calculated by the adjoint-state method and joint inversing of head and tracer data, *Adv. Water Res.*, 24(1), 89-103, doi: 10.1016/S0309-1708(00)00007-5.

Cooley, R. L. (2000), An analysis of the pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, *Water Resour. Res.*, 36(4), 1159-1163, doi: 10.1029/2000WR900008.

Cowles, M. K. & B. P. Carlin (2012), Markov chain Monte Carlo convergence diagnostics: A comparative review, *J. Amer. Statist. Assoc.*, 91(434), 883-904, doi: 10.1080/01621459.1996.10476956.

Cukier R. I., H. B. Levine, and K. E. Schuler (1978), Nonlinear sensitivity analysis of multiparameter model systems, *J. Comput. Phys.*, 26 (1), 1-42, doi: 10.1016/0021-

9991(78)90097-9Dagan, G. (1987), Theory of solute transport by groundwater, *Ann. Rev. Fluid Mech.*, 19, 183-215.

Dane, J. H. & G. Clarke Topp (ed.), Soil Science Society of America Book Series, no. 5. Soil Science Society of America, Inc., Madison, WI, 1692 pp.

Dane, J. H. & Hruska, S. (1983), In-situ determination of soil hydraulic properties during drainage, *Soil Sci. Soc. Am. J.*, 47, 619-624.

Deng, X. & M. Yuan (2012), Large Gaussian covariance matrix estimation with Markov structures, *J. Comput. Graph. Stat.*, 18(3), 640-657, doi: 10.1198/jcgs.2009.07170. Doherty, J. (2003), Ground water calibration using pilot points and regularization, *Ground Water,* 41(2), 170-177, doi: 10.111/j.1745-6584.2003.tb02580.x.

Downing D. J., R. H. Gardner, and F. O. Hoffman (1985), An examination of response-surface methodologies for uncertainty analysis in assessment models, *Technometrics*, 27(2), 151-163, doi: 10.1080/00401706.1985.10488032.

Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization methods for conceptual rainfall-runoff models, *Water Resour. Res.*, 28, 1015-1031, doi: 10.1029/91WR02985.

Durner, W., U. Jansen, and S. C. Iden (2008), Effective hydraulic properties of layered soils at the lysimeter scale determined by inverse modeling, *Eur. J. Soil Sci.*, 59(1), 114-124, doi: 10.1111/j.1365-2389.2007.00972.x.

Feddes, R., P. Kowalik, and H. Zaradny (1978), Simulation of field water use and crop yield, *Centre for Agricultural Publishing and Documentation.*

Johnson, N. L. & S. Kotz (1970), *Distributions in statistics: Continuous univariate distributions*, vol. 1, Houghton Mifflin Co., Boston, Massachusetts.

Ginn, T. R. & J. H. Cushman (1990), Inverse methods for subsurface flow: A critical review of stochastic techniques, *Stoch. Hydrol. Hydraul.* 4, 1-26, doi: 10.1007/BF01547729.

Hammond, G. E., and P. C. Lichtner (2010), Field-scale model for the natural attenuation of uranium at the Hanford 300 Area using high-performance computing, *Water Resour. Res.,* 46, W09527, doi: 10.1029/2009WR008819.

Harvey, C. F., and S. M. Gorelick (1995), Temporal moment-generating equations: modeling transport and mass transfer in heterogeneous aquifers, *Water Resour. Res.,* 31(8), 1895-1911, doi: 10.1029/95WR01231.

Hayfield, T. & J. S. Racine (2008), Nonparametric econometrics: The *np* package, *Journal of Statistical Software,* 27(5). http://www.jstatsoft.org/v27/i05.

Hollenbeck, K. J. & K. H. Jensen (1998), Maximum-likelihood estimation of unsaturated hydraulic parameters, *J. Hydrol.*, 210(1-4), 192-205, doi: 10.1016/S0022-1694(98)00185-1.

Hopmans, J. W. & J. Simunek (1999), Review of inverse estimation of soil hydraulic properties. *Characterization and measurement of the hydraulic properties of unsaturated porous media. University of California, Riverside, CA,* 643-659.

Hou, Z. & Y. Rubin (2005), On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling, *Water Resour. Res.,* 41, W12425, doi: 10.1029/2005WR004082.

Isukapalli S. S., A. Roy, P. G. Georgopoulos (2006), Stochastic response surface methods (SRSMs) for uncertainty propagation: Application to environmental and biological systems, *Risk Analysis*, 18(3), 351-363, doi: 10.1111/j.1539-6924.1998.tb01301.x.

Izenman, A. J. (1991), Recent developments in nonparametric density estimation, *Journal of the American Statistical Association,* 86(413), 205-224.

Johnson, N. L. & S. Kotz (1970), *Distributions in statistics: Continuous univariate distributions*, vol. 1, Houghton Mifflin Co., Boston, Massachusetts.

Jury, W. A. & G. Sposito (1985), Field calibration and validation of solute transport models for the unsaturated zone, *Soil Sci. Soc. Am. J.*, 49(6), 1331-1341.

Kaufmann, L., and P. J. Rousseeuw (1990), *Finding Groups in Data – An Introduction to Cluster Analysis*, Wiley Interscience Publication, New York, NY.

Kitanidis, P. K. & E. G. Vomvoris (1983), A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res.*19(3), 677-690, doi: 10.1029/WR019i003p00677.

Kitanidis, P. K. & R. W. Lane (1985), Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method, *J. Hydrol.* (79), 53-71, doi: 10.1016/0022-1694(85),90181-7.

Kitanidis, P. K. (1996), On the geostatistical approach to the inverse problem, *Adv. Water Resour.*, 19(6), 333-342, doi: 10.1016/0309-1708(96)00005-X.

Kowalsky, M. B., S. Finsterle, and Y. Rubin (2004), Estimating flow parameter distributions using ground-penetrating radar and hydrological measurements during transient flow in the vadose zone, *Adv. Water Res.*, 27(6), 583-599, doi: 10.1016/j.advwatres.2004.03.003.

Kowalsky, M. B., S. Finsterle, J. Peterson, S. Hubbard, Y. Rubin, E. Majer, A. Ward, and G. Gee (2005), Estimation of field-scale soil hydraulic and dielectric parameters through joint inversion of GPR and hydrological data, *Water Resour. Res.,* 41, W11425, doi: 10.1029/2005WR004237.

Leube P. C., W. Nowak, and G. Schneider (2012), Temporal moments revisited: Why there is no better way for physically based model reduction in time, *Water Resour. Res.*, 48, W11527, doi: 10.1029/2012WR011973.

Li, L., J. Xia, C.Y. Xu, and V. P. Singh (2010), Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models, *J. Hydrol.,* 390, 210-221, doi: 10.1016/j.jhydrol.2010.06.044.

Loll P. and P. Moldrup (1998), A new two-step stochastic modeling approach: Application to water transport in a spatially variable unsaturated soil, *Water Resour. Res.*, 34(8), 1909-1918.

Looms, M. C., A. Binley, K. H. Jensen, L. Nielsen, and M. Hansen (2008), Identifying unsaturated hydraulic parameters using an integrated data fusion approach on cross-borehole geophysical data, *Vadose Zone J.*, 7(1), 238-248, doi: 10.2136/vzj2007.0087.

Lukacs, E. (1942), A characterization of the normal distribution, *Ann. Math. Stat.,* 13(1), 91-93, http://www.jstor.org/stable/2236166.

Martins, E. S. & J. R. Stedinger (2000), Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resour. Res.*, 36(3), 737-744, doi: 10.1029/1999WR900330.

Marzouk, Y. M., H. N. Najm, and L. A. Rahn (2007), Stochastic spectral methods for efficient Bayesian solution of inverse problems, *J. Comput. Phys.,* 224, 560-586, doi: 10.1016/j.jcp.2006.10.010.

Maxwell, R. M., W. E. Kastenberg, and Y. Rubin (1999), A methodology to integrate site characterization information into groundwater-driven health risk assessment, *Water Resour. Res.,* 35(9), 2841-2855, doi: 10.1029/1999WR900103.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, 21(2), 239-245.

McLaughlin, D., and L. R. Townley (1996), A reassessment of the groundwater inverse problem, *Water Resour. Res.*, 32(5), 1131-1161. Doi: 10.1029/96WR00160.

Mertens, J., H. Madsen, L. Feyen, D. Jacques, and J. Feyen (2004), Including prior information in the estimation of effective soil parameters in unsaturated zone modeling, *J. Hydrol.,* 294, 251-269, doi: 10.1016/j.jhydrol.2004.02.011.

Mualem, Y. (1976), A new model for predicting the hydraulic conductivity of unsaturated porous media, *Water Resour. Res.,* 12, 513-522, doi: 10.1029/WR012i003p00513.

Murakami, H., X. Chen, M. S. Hahn, Y. Liu, M. L. Rockhold, V. R. Vermeul, J. M Zachara, and Y. Rubin (2011), Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 Area, *Hydrol. Earth Syst. Sci.,* 14, 1989-2001, doi: 10.5194/hess-14-1989-2010.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stoch. Env. Res. Risk. A.*, 17(5), 291-305, doi: 10.1007/s00477-003-0151-7.

Osorio-Murillo, C. A., M. W. Over, D. P. Ames, and Y. Rubin (2013), Introducing an extensible open source inversion modeling and uncertainty characterization software framework, *Environ. Modell. Softw.*, submitted.

Over M. W., X. Chen, Y. Yang, and Y. Rubin (2013), A strategy for improved computational efficiency of the method of anchored distributions, *Water Resour. Res.,* 49, 1-19, doi: 10.1002/wrcr.20182.

Pardo-Iguzquiza, E. (1998), Maximum likelihood estimation of spatial covariance parameters, *Math. Geol.* 30(1), 95-108, doi: 10.1023/A:1021765405952.

Pebesma, E. & C. Wesseling (1998), GSTAT: A program for geostatistical modeling, prediction, and simulation, *Comput. Geosci.*, 24(1), 17-31, doi: http://dx.doi.org/10.1016/S0098-3004(97)00082-4.

Poeter, E. P. & M. C. Hill (1997), A necessary next step in ground-water modeling, *Groundwater*, 35(2), 250-260, doi: 10.1111/j.1745-6584.1997.tb00082.x.

R Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* ISBN 3-900051-07-0, URL http://www.R-project.org/

Rabitz H., O. F. Alis, J. Shorter, and K. Shim (1999), Efficient input-output model representations, *Computer Physics Communications*, 117(1-2), 11-20, doi: 10.1016/S00010-4655(98)00152-0.

Rao K. S. (2005), Uncertainty analysis in atmospheric dispersion modeling, *Pure Appl. Geophys.*, 162, 1893-1917, doi: 10.1007/s0024-005-2697-4.

Ritter, A., F. Hupet, R. Muñoz-Carpena, S. Lambot, and M. Van-clooster (2003), Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods, *Agr. Water Manage.*, 59,77-96, doi: 10.1016/S0378-3774(02)00160-9.

Robert, C. P. & G. Casella (2004), *Monte Carlo Statistical Methods vol. 319.* Springer, New York.

Rogiers, B., D. Mallants, O. Batelaan, M. Gedeon, M. Huysmans, and A. Dassargues (2012), Estimation of hydraulic conductivity and its uncertainty from grain-size data using GLUE and artificial neural networks, *Math. Geosci.*, 44(6), 739-763, doi: 10.1007/s11004-012-9409-2.

Rubin Y. (2003), *Applied Stochastic Hydrology*, Oxford University Press, New York, New York.

Rubin Y., X. Chen, H. Murakami, and M.S. Hahn (2010), A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields, *Water Resour. Res.,* 46, W10523, doi: 10.1029/2009WR008799.

Schaap, M. G., F. J. Leij, and M. T. van Genuchten (2001), ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, *J. Hydrol.*, 251(3-4), 163-176, doi: 10.1016/S0022-1694(01)00466-8.

Scharnagl, B., J. A. Vrugt, H. Vereecken, and M. Herbst (2011), Inverse modeling of in situ soil water dynamics: Investigating the effect of different prior distributions of the soil hydraulic parameters, *Hydrol. Earth Syst. Sci.,* 15, 3043-3059, doi: 10.5194/hess-15-3043-2011.

Scholer, M., J. Irving, A. Binley, and K. Hollinger (2011), Estimating vadose zone hydraulic properties using ground penetrating radar: The impact of prior information, *Water Resour. Res.,* 47(10), doi: 10.1029/2011WR010409.

Scholer, M., J. Irving, M. C. Looms, L. Nielsen, and K. Holliger (2012), Bayesian Markov-chain-Monte-Carlo inversion of time-lapse crosshole GPR data to characterize the vadose zone at the Arrenaes site, Denmark, *Vadose Zone J.,* 11(4), doi: 10.2136/vzj2011.0153.

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

Scott, D. W. & S. R. Sain (2004), Multi-dimensional density estimation, in *Handbook of Statistics vol. 23: Data Mining and Computational Statistics,* edited by C. R. Rao and E. J. Wegman, 229-263, Elsevier, Amsterdam, Netherlands.

Shapiro, S. S. & M. B. Wilk (1965), An analysis of variance test for normality (complete samples), *Biometrika,* 52(3-4), 591-611, http://www.jstor.org/stable/2333709.

Silverman, B. W. (1986), *Density Estimation*, Chapman and Hall, London, United Kingdom.

Simunek, J., M. Th. Van Genuchten, and M. Sejna (1998), The HYDRUS-1D software package for simulating the movement of water, heat, and multiple solutes in variably-saturated media, version 2.0, 1, US Salinity Laboratory, Agricultural Research Service, US Department of Agriculture, Riverside, CA, 1998.

Soil Survey Division Staff (1993), *Soil survey manual*, USDA, http://soils.usda.gov/technical/manual.

Sorooshian, S., V. K. Gupta, and J. L. Fulton (1983), Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: influence of calibration data variability and length on model credibility, *Water Resour. Res.,* 19(1), 251-259, doi: 10.1029/WR019i001p00251.

Taylor, S. & G. Ashcroft (1972), *Physical edaphology: The physics of irrigated and non-irrigated soils*, W.H. Freeman & Co., San Francisco, California.

Van der Laan, M. J., and K. S. Pollard (2003), A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap, *Journal of Statistical Planning and Inference*, 117(2) 275-303, doi: 10.1016/S0378-3758(02)00388-9.

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, doi:10.2136/sssaj1980.03615995004400050002x.

Vereecken, H., J. A. Huisman, H. Bogena, J. Vanderborght, J. A. Vrugt, and J. W. Hopmans (2008), On the value of soil moisture measurements in vadose zone hydrology: A review, *Water Resour. Res.*, *44*, W00D06, doi:10.1029/2008WR006829.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing, Theory and Applications,* edited by Y. Eldar and G. Kutyniok, 210-268, Cambridge University Press.

Vogel, H. J. & K. Roth (2003), Moving through scales of flow and transport in soil, *J. Hydrol.*, 272, 95-106, doi: 10.1016/S0022-1694(02)00257-3.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.,* 39(8), doi: 10.1029/2002WR001642.

Vrugt, J. A., C. J. F. ter Braak, H. V. Gupta, B. A. Robinson (2008a), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling, *Stoch. Environ. Res. Risk Assess.*, 23, 1011-1026, doi: 10.1007/s00477-008-0274-y.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008b), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.,* 44, W00B09, doi: 10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlin. Sci. Num.,* 10(3), 273-290, doi: 10.1515/IJNSNS.2009.10.3.273.

Wasserman, L. (2010), *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York, New York.

Wöhling, Th., J. A. Vrugt, and G. F. Barkle (2008), Comparison of three multiobjective optimization algorithms for inverse modeling of vadose zone hydraulic properties, *Soil Sci. Soc. Am. J.*, 72, 305-319, doi: 10.2136/sssaj2007.0176.

Wöhling, Th. & J. A. Vrugt (2011), Multiresponse multilayer vadose zone model calibration using Markov chain Monte Carlo simulation and field water retention data, *Water Resour.Res.,* 47(4), W04510, doi: 10.1029/2010WR009265.

Wollschläger, W., T. Pfaff, and K. Roth (2009), Field-scale apparent hydraulic parameterisation obtained from TDR time series and inverse modeling, *Hydrol. Earth Syst. Sci,* 13(10), 1953-1966, doi: 10.5194/hess-13-1953-2009.

Woodbury, A. D. & T. J. Ulrych (1993), Minimum relative entropy: Forward probabilistic modeling, *Water Resour. Res.,* 29(8), 2847-2860, doi: 10.1029/93WR00923.

Woodbury, A. D. & T. J. Ulrych (1993), Minimum relative entropy and probabilistic inversion in groundwater hydrology, *Stoch. Hydrol. Hydraul.,* 12, 317-358, doi: 10.1007/s004770050024.

Yang, Y., M.W. Over, and Y. Rubin (2012), Strategic placement of localization devices (such as pilot points and anchors) in inverse modeling schemes, *Water Resour. Res.*, 48, W08519, doi: 10.1029/2012WR011864.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *J. Hydrol.,* 30(1-4), 83-97, doi: 10.1016/S0022-1694(97)00107-8.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, 40(5), doi: 10.1029/2003WR002557.

Yeh, T. C. J., L. W. Gelhar, and A. L. Gutjahr (1985), Stochastic analysis of unsaturated flow in heterogeneous soils: 1. Statistically isotropic media, *Water Resour.Res.,* 21(4), 447-456, doi: 10.1029/WR021i004p00447.