# Extending GLUE with Multilevel Methods to Accelerate Statistical Inversion of Hydrological Models

Max Gustav Rudolph<sup>1</sup>, Thomas Wöhling<sup>1</sup>, Thorsten Wagener<sup>2</sup>, and Andreas Hartmann<sup>3</sup>

<sup>1</sup>Technische Universität Dresden <sup>2</sup>University of Potsdam <sup>3</sup>Dresden University of Technology

April 19, 2024

### Abstract

Inverse problems are ubiquitous in hydrological modelling for parameter estimation, system understanding, sustainable water resources management, and the operation of digital twins. While statistical inversion is especially popular, its sampling-based nature often inhibits its application to computationally costly models, which has compromised the use of the Generalized Likelihood Uncertainty Estimation (GLUE) methodology, e.g., for spatially distributed (partial) differential equation based models. In this study we introduce multilevel GLUE (MLGLUE), which alleviates the computational burden of statistical inversion by utilizing a hierarchy of model resolutions. Inspired by multilevel Monte Carlo, most parameter samples are evaluated on lower levels with computationally cheap low-resolution models and only samples associated with a likelihood above a certain threshold are subsequently passed to higher levels with costly high-resolution models for evaluation. Inferences are made at the level of the highest-resolution model but substantial computational savings are achieved by discarding samples with low likelihood already on levels with low resolution and low computational cost. Two example inverse problems, using a rainfall-runoff model and groundwater flow model, demonstrate the substantially increased computational efficiency of MLGLUE compared to GLUE as well as the similarity of inversion results. Findings are furthermore compared to inversion results from Markov-chain Monte Carlo (MCMC) and multilevel delayed acceptance MCMC, a corresponding multilevel variant, to compare the effects of the multilevel extension. All examples demonstrate the wide-range suitability of the approach and include guidelines for practical applications.













# Extending GLUE with Multilevel Methods to Accelerate Statistical Inversion of Hydrological Models

## M. G. Rudolph<sup>1</sup>, T. Wöhling<sup>2,3</sup>, T. Wagener<sup>4</sup>, A. Hartmann<sup>1</sup>

4	$^1 \mathrm{Institute}$ of Groundwater Management, Technische Universität Dresden, Dresden, Germany
5	$^2\mathrm{Chair}$ of Hydrology, Institute of Hydrology and Meteorology, Technische Universität Dresden, Dresden,
6	Germany
7	<sup>3</sup> Lincoln Agritech, Lincoln, New Zealand
8	<sup>4</sup> Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany

### 9 Key Points:

1

2

3

# We extend the Generalized Likelihood Uncertainty Estimation methodology to a setting with multiple levels of model resolution (MLGLUE) We demonstrate the acceleration with MLGLUE for different spatial (groundwater flow model) and temporal (rainfall-runoff model) resolutions We find that MLGLUE acceleration is comparable to or more efficient than a multilevel extension of Markov-chain Monte Carlo

Corresponding author: M. G. Rudolph, max\_gustav.rudolph@tu-dresden.de

### 16 Abstract

Inverse problems are ubiquitous in hydrological modelling for parameter estima-17 tion, system understanding, sustainable water resources management, and the operation 18 of digital twins. While statistical inversion is especially popular, its sampling-based na-19 ture often inhibits its application to computationally costly models, which has compro-20 mised the use of the Generalized Likelihood Uncertainty Estimation (GLUE) method-21 ology, e.g., for spatially distributed (partial) differential equation based models. In this 22 study we introduce multilevel GLUE (MLGLUE), which alleviates the computational 23 burden of statistical inversion by utilizing a hierarchy of model resolutions. Inspired by 24 multilevel Monte Carlo, most parameter samples are evaluated on lower levels with com-25 putationally cheap low-resolution models and only samples associated with a likelihood 26 above a certain threshold are subsequently passed to higher levels with costly high-resolution 27 models for evaluation. Inferences are made at the level of the highest-resolution model 28 but substantial computational savings are achieved by discarding samples with low like-29 lihood already on levels with low resolution and low computational cost. Two example 30 inverse problems, using a rainfall-runoff model and groundwater flow model, demonstrate 31 the substantially increased computational efficiency of MLGLUE compared to GLUE 32 as well as the similarity of inversion results. Findings are furthermore compared to in-33 version results from Markov-chain Monte Carlo (MCMC) and multilevel delayed accep-34 tance MCMC, a corresponding multilevel variant, to compare the effects of the multi-35 level extension. All examples demonstrate the wide-range suitability of the approach and 36 include guidelines for practical applications. 37

### <sup>38</sup> 1 Introduction

Inverse problems are ubiquitous in hydrological modelling, emerging in the context 39 of parameter estimation, system understanding, sustainable water resources management, 40 and the operation of digital twins (e.g., Leopoldina, 2022). Computational models are 41 are often highly parameterized and non-linear, posing substantial challenges to param-42 eter inversion approaches. Furthermore, observations of system states are affected by mea-43 surement uncertainty and the knowledge about the underlying system is incomplete, re-44 sulting in uncertainties associated with computational models (Beven, 1993; Wagener 45 & Gupta, 2005; Carrera et al., 2005; Beven, 2006; Vrugt et al., 2009; Laloy & Vrugt, 2012; 46 Zhou et al., 2014; Mai, 2023). We need to quantify these uncertainties if models should 47

be used for scientific inquiry or in support of decision making (Blöschl et al., 2019). While 48 process-based spatially distributed models are increasingly used to guide decision-making 49 and to sustainably manage water resources, such modelling approaches are computation-50 ally costly (Doherty, 2015; Herrera et al., 2022), making uncertainty quantification (UQ) 51 and statistical inversion especially challenging (Erdal & Cirpka, 2020; Kuffour et al., 2020; 52 White, Hunt, et al., 2020). There is a need to develop computationally efficient approaches 53 to UQ and statistical inversion to overcome the pressing challenges associated with cli-54 mate change and their impact on water resources. 55

Various approaches to UQ have been developed and applied in that respect; the 56 Bayesian approach to statistical inversion and UQ, however, is especially popular due 57 to the ability to comprehensively treat uncertainties in state variables, parameters, and 58 model output (Montanari, 2007; Vrugt, 2016; Linde et al., 2017; Page et al., 2023). Gen-59 eralized Likelihood Uncertainty Estimation (GLUE) (Beven & Binley, 1992, 2014) - as 60 an informal Bayesian approach - and Markov-chain Monte Carlo sampling (MCMC) (Gallagher 61 et al., 2009; Vrugt, 2016; Dodwell et al., 2019; Brunetti et al., 2023; Lykkegaard et al., 62 2023; Cui et al., 2024) - as a formal Bayesian approach - are frequently applied in the 63 environmental sciences for statistical inversion. The Bayesian framework considers model 64 parameters to be random variables that are associated with prior distributions, which 65 are conditioned on system state observations using a likelihood function to posterior dis-66 tributions. The likelihood function may either be defined formally or informally, depend-67 ing on the belief and assumptions made about sources of error and the intended prop-68 erties of the likelihood function itself, and many different approaches exist to define such 69 functions (Beven & Binley, 1992; Beven & Freer, 2001; Schoups & Vrugt, 2010; Nott et 70 al., 2012; Sadegh & Vrugt, 2013; Beven, 2016; Vrugt & Beven, 2018). 71

Approaches to statistical inversion generally rely on repeatedly running the com-72 putational model with different parameter values (i.e., repeatedly solving the forward 73 problem) to obtain outputs that can be compared to observations of the same variable, 74 if available. With computationally costly models, this approach quickly becomes intractable 75 and there is a need to develop more efficient sampling approaches for statistical inver-76 sion. Different approaches have been developed to reduce computational cost of inver-77 sion, such as using data-driven surrogate or reduced-order models during inversion, which 78 are the often often run instead of the computationally costly high-fidelity model (Doherty 79 & Christensen, 2011; Asher et al., 2015; Burrows & Doherty, 2015; Linde et al., 2017; 80

-3-

Gosses & Wöhling, 2019, 2021; Allgeier, 2022). Reducing model spatial resolution can reduce model complexity and computational cost in general and the effect of horizontal (Wildemeersch et al., 2014; Savage et al., 2016; Reinecke et al., 2020) as well as vertical (White, Knowling, & Moore, 2020) discretization in model performance has been studied before, also in the context of accelerating inversion (von Gunten et al., 2014).

Multilevel methods and multilevel Monte Carlo (MLMC) (Heinrich, 2001; Giles, 86 2008; Cliffe et al., 2011; Giles, 2015), with extensions to multilevel MCMC and multi-87 level delayed acceptance MCMC (MLMCMC and MLDA, respectively) (Dodwell et al., 88 2019; Lykkegaard et al., 2023; Cui et al., 2024), were previously introduced with the mo-89 tivation of reducing the computational cost of Monte Carlo estimators. For spatially dis-90 tributed models, multilevel methods utilize multiple levels of spatial domain resolution. 91 Together with the most finely discretized highest level model, several more coarsely dis-92 cretized lower level models are considered. Most solutions to the forward problem are 93 then found using lower level models while the highest level model is executed far less fre-94 quently, harbouring the potential for large savings in overall computation time. Contrary 95 to surrogate- or reduced-order-model-aided approaches to UQ, multilevel methods make 96 no simplifying assumptions about the model and the relevant processes are simulated 97 directly on all resolution levels. Another contrast is that the coarsely discretized mod-98 els are not used instead of the high-fidelity model but they are synergetically used to-99 gether. Linde et al. (2017) summarize first applications of MLMC for the forward prop-100 agation of uncertainties in hydrogeology and hydrogeophysics. We note that multilevel 101 methods can be used with all types of models where a notion of model resolution exists. 102 Typically, multilevel methods are applied to models based on (partial) differential equa-103 tions (PDEs) using different spatial grid resolutions (e.g., in numerical groundwater flow 104 models) or different temporal resolutions (e.g., in rainfall-runoff models). 105

Previous applications of multilevel methods focussed on models with different spa-106 tial resolutions (Cliffe et al., 2011; Linde et al., 2017; Dodwell et al., 2019; Lykkegaard 107 et al., 2023; Cui et al., 2024), entailing challenges when transferring parameter fields from 108 one spatial resolution to another. Geostatistical approaches are often used to (initially) 109 assign parameters for spatially distributed groundwater flow- or other hydrological mod-110 els. This simultaneously reduces overparameterization as the number of geostatistical 111 parameters is much lower than the number of parameters of the computational model. 112 To this end, utilizing point measurements of parameters or the combination with other 113

-4-

predictor variables, Gaussian process regression is frequently used to generate conditioned 114 parameter fields on any desired spatial resolution (Kitanidis & Vomvoris, 1983; Zimmer-115 man et al., 1998; Zhou et al., 2014; Doherty, 2003). Unconditioned random fields are also 116 utilized, where parameter fields are generated on any desired spatial resolution (Y. Liu 117 et al., 2019); using uncorrelated and spatially independent random variables, the Karhunen-118 Loéve expansion is frequently employed to parameterize the random field (Cliffe et al., 119 2011; Dodwell et al., 2019; Lykkegaard et al., 2023; Cui et al., 2024). The definition of 120 hydrological respose units or internally homogeneous zones of parameters represents an-121 other strategy for parameterization (Kumar et al., 2013; Zhou et al., 2014; Anderson et 122 al., 2015; White, 2018). To better constrain the parameter space during inversion and 123 to reduce the aggravating effect of overparameterization, regularization can be employed 124 in combination with different parameterization strategies (Tonkin & Doherty, 2005; Moore 125 & Doherty, 2006; Pokhrel et al., 2008; Moore et al., 2010). Parameter scaling can be used 126 to transfer parameter fields from one spatial resolution to another. While there is no gen-127 erally valid theory for upscaling (i.e., from fine to coarse grids) (Binley et al., 1989; Samaniego 128 et al., 2010), various upscaling operators are used in practice (Binley et al., 1989; Samaniego 129 et al., 2010; Colecchio et al., 2020). 130

While multilevel methods have previously been used to accelerate MCMC algorithms 131 (Dodwell et al., 2019; Lykkegaard & Dodwell, 2022; Lykkegaard et al., 2023; Cui et al., 132 2024) in a formal Bayesian framework, they have not yet been applied in connection with 133 GLUE. In this study, we utilize ideas from multilevel Monte Carlo strategies to accel-134 erate statistical inversion of hydrological models with the GLUE methodology. After in-135 troducing multilevel GLUE (MLGLUE), two example inverse problems are considered. 136 We subsequently apply conventional GLUE and MLGLUE as well as MCMC and MLDA 137 to those problems and compare the results. 138

### 139 2 Methods

140

### 2.1 The Inverse Problem

Consider observations  $\widetilde{\mathbf{Y}} = [\widetilde{y}_1, \dots, \widetilde{y}_k]^T \in \mathcal{Y} \subseteq \mathbb{R}^k$  of a real system and consider a model  $\mathcal{F}$  that simulates the system response  $\mathbf{Y} = [y_1, \dots, y_k]^T \in \mathcal{Y}$  corresponding to  $\widetilde{\mathbf{Y}}$ . The model output also depends on initial and boundary conditions  $\mathcal{C}_i$  and  $\mathcal{C}_b$ , respectively, as well as on model parameters  $\boldsymbol{\theta} \in \mathcal{X} \subseteq \mathbb{R}^n$ 

$$\mathbf{Y} = \mathcal{F}(\boldsymbol{\theta}, \mathcal{C}_i, \mathcal{C}_b) + \boldsymbol{\varepsilon} := \mathcal{F}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$
(1)

 $\mathcal{F}: \mathcal{C}_i, \mathcal{C}_b \to \mathbf{Y} \in \mathcal{Y} \text{ is closed by the parameter vector } \boldsymbol{\theta} \text{ (Kavetski et al., 2006; Vrugt}$ et al., 2009), which is considered a random vector with an associated prior distribution  $p_p(\boldsymbol{\theta}). \ \boldsymbol{\varepsilon} \in \mathbb{R}^k \text{ in this context represents the combined effect of conceptual model er-}$ ror and measurement error (e.g., Kennedy & O'Hagan, 2001; Plumlee, 2017); subsequently we refer to  $\boldsymbol{\varepsilon}$  simply as error and refer to the aforementioned references for more detailed discussions on errors.

Solving the inverse problem in a Bayesian statistical framework means to obtain the posterior distribution of the parameters  $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$  via Bayes' theorem

$$p(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = \frac{p_p(\boldsymbol{\theta})p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta})}{p(\widetilde{\mathbf{Y}})} \propto p_p(\boldsymbol{\theta})p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta})$$
(2)

where  $p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta})$  is the likelihood function and  $p(\widetilde{\mathbf{Y}})$  is the proportionality factor called model evidence, which is the average likelihood of the model to have generated the data.

Assuming that errors  $r_i = y_i - \tilde{y}_i$  are mutually independent, identically distributed (i.i.d.) and follow a Gaussian distribution with constant variance  $\sigma_r^2$ , the log-likelihood takes the form

$$\mathcal{L}(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta}) = -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln(\sigma_r^2) - \frac{1}{2}\sigma_r^{-2} \cdot \sum_{i=1}^k (y_i - \widetilde{y}_i)^2.$$
(3)

The assumptions of i.i.d. errors, however, usually does not hold as these errors of hydrological models often exhibit strong autocorrelation and heteroscedasticity (see, e.g., Beven (2006) for a discussion). Beven and Freer (2001) and Vrugt et al. (2009) give alternative likelihood formulations for non-Gaussian errors that often come at the cost of additional hyperparameters.

163

### 2.2 Multilevel Monte Carlo

We will discuss the notion of multilevel methods from the perspective of multilevel Monte Carlo (MLMC), which is a method to efficiently compute the expectation of a quantity of interest that depends on (model) parameters (Heinrich, 2001; Giles, 2008; Cliffe et al., 2011; Giles, 2015). Consider the situation where we are given a distribution of model parameters,  $p(\theta)$ , and want to compute the expected value of some scalar quantity related to the model output,  $\mathbf{Q} = \mathcal{Q}(\mathcal{F}(\theta))$ , with respect to  $p(\theta)$ . Here,  $\mathcal{Q}$  represents some function of the model output, e.g., it yields the system state at a certain location, or a more abstract quantity. As an example, consider **Q** to represent the groundwater level at some location in the model domain. Propagating the uncertainty contained in the parameter distributions through the model to represent the uncertainty in **Q** is considered a problem of forward propagation of uncertainty, which is the opposite of the inverse problem described in section 2.1. Yet, MLMC builds on a simple intuition that illustrates the idea behind MLGLUE.

For simplicity and without loss of generality consider  $\mathbf{Q} \in \mathbb{R}$  for the remainder 177 of this section. Instead of one single model for the system, assume that there is a hier-178 archy of models (approximations of the real system), which are denoted by  $\{\mathcal{F}_{\ell}\}_{\ell=0}^{\infty}$ , where 179  $\ell$  is the level index. Associated with each model in the hierarchy are values for the quan-180 tity of interest,  $\{\mathbf{Q}_{\ell}\}_{\ell=0}^{\infty}$ , such that  $\widetilde{\mathbf{Q}} = \lim_{\ell \to \infty} \mathbf{Q}_{\ell}$ , where  $\widetilde{\mathbf{Q}}$  represents the true value. 181 In the context of PDE-based models,  $\ell$  may be related to the grid size or time step length 182 of the model, i.e., a larger  $\ell$  corresponds to a higher domain resolution with smaller com-183 putational cells or smaller time steps, for example. We assume that the computational 184 cost for evaluating  $\mathcal{F}_{\ell}$  (or  $\mathbf{Q}_{\ell}$ ) increases while the approximation error decreases as  $\ell \to$ 185 L. Here L is the index of the highest level, which is often associated with the target model 186 and all lower levels have lower resolution. We note that the most common form of the 187 model hierarchy is a geometric series of computational grids, where the factor of refine-188 ment or coarsening between subsequent levels is constant across all levels (Cliffe et al., 189 2011; Giles, 2015). To estimate the expectation of  $\mathbf{Q}$  efficiently, MLMC avoids the di-190 rect estimation of  $\mathbb{E}[\mathbf{Q}_L]$  on the highest level  $\ell = L$ . Instead, the correction of the es-191 timation with respect to the next lower level is computed, based on the linearity of ex-192 pectation: 193

$$\mathbb{E}[\mathbf{Q}_L] = \mathbb{E}[\mathbf{Q}_0] + \sum_{\ell=1}^{L} \mathbb{E}[\mathbf{Q}_\ell - \mathbf{Q}_{\ell-1}]$$
(4)

This approach generally results in substantial computational savings and different multilevel estimators for  $\mathbb{E}[\mathbf{Q}_L]$  exist (Giles, 2008; Cliffe et al., 2011; Giles, 2015; Dodwell et al., 2019; Lykkegaard et al., 2023; Cui et al., 2024). The original MLMC algorithm of Giles (2008) (as well as subsequently applied algorithms) takes a bottom-up approach, i.e., sampling is started on  $\ell = 0$  and  $\ell$  is only incremented if the algorithm has not yet converged on level  $\ell$ . There, efficiency and variance reduction regarding the expectation of **Q** may be optimized by choosing an optimal refinement (e.g., the decrease of cell or time step size when going from  $\ell$  to  $\ell + 1$ ).

In the context of MLMC, the behaviour of the variances  $\mathbb{V}[\mathbf{Q}_{\ell}]$  and  $\mathbb{V}[\mathbf{Q}_{\ell}-\mathbf{Q}_{\ell-1}]$ 202 and expectations  $\mathbb{E}[\mathbf{Q}_{\ell}]$  and  $\mathbb{E}[\mathbf{Q}_{\ell} - \mathbf{Q}_{\ell-1}]$  as  $\ell \to L$  gives an indication of the overall 203 quality and efficiency of the hierarchy  $\{\mathbf{Q}_{\ell}\}_{\ell=0}^{L}$  (Cliffe et al., 2011).  $\mathbb{V}[\mathbf{Q}_{\ell}]$  and  $\mathbb{E}[\mathbf{Q}_{\ell}]$  should 204 be approximately constant as  $\ell \to L$ , ensuring that  $\mathbf{Q}_{\ell}$  is a good enough approxima-205 tion even on the coarsest level  $\ell = 0$ . Furthermore,  $\mathbb{V}[\mathbf{Q}_{\ell} - \mathbf{Q}_{\ell-1}]$  and  $\mathbb{E}[\mathbf{Q}_{\ell} - \mathbf{Q}_{\ell-1}]$ 206 should decay rapidly and be smaller than  $\mathbb{V}[\mathbf{Q}_{\ell}]$  and  $\mathbb{E}[\mathbf{Q}_{\ell}]$ , respectively, as  $\ell \to L$ , en-207 suring that the approximation error decreases with increasing level.  $\mathbb{V}[\mathbf{Q}_{\ell}-\mathbf{Q}_{\ell-1}]$  may 208 be expanded as 209

$$\mathbb{V}[\mathbf{Q}_{\ell} - \mathbf{Q}_{\ell-1}] = \mathbb{V}[\mathbf{Q}_{\ell}] + \mathbb{V}[\mathbf{Q}_{\ell-1}] - 2 \cdot \operatorname{Cov}(\mathbf{Q}_{\ell}, \mathbf{Q}_{\ell-1}),$$
(5)

showing that it should be given that  $2 \cdot \text{Cov}(\mathbf{Q}_{\ell}, \mathbf{Q}_{\ell-1}) > \mathbb{V}[\mathbf{Q}_{\ell-1}]$ , which requires Q<sub>ℓ</sub> and  $\mathbf{Q}_{\ell-1}$  to be sufficiently correlated.

While those relations between levels are not formally required to hold for inversion, they ensure that the multilevel estimator for the expectation of **Q** has reduced variance and is computationally more efficient compared to a single-level estimator (Cliffe et al., 2011; Lykkegaard et al., 2023). While a deviation of the previously described optimal relations between levels does not necessarily indicate a poorly performing model hierarchy, without such a deviation the hierarchy may be said to be well behaved. We discuss the design of the model hierarchy in more detail in section 2.4.2.

219

### 2.3 Multilevel Markov-chain Monte Carlo

The multilevel delayed acceptance (MLDA) MCMC algorithm was developed by 220 Lykkegaard et al. (2023) on the basis of the delayed acceptance algorithm coupled with 221 the randomized-length-subchain surrogate transition (Christen & Fox, 2005; J. S. Liu, 222 2008) and includes many concepts similar to MLMC described in section 2.2. Delayed 223 acceptance MCMC has been employed by Laloy et al. (2013) to accelerate Bayesian in-224 version for groundwater flow models using a generalized polynomial chaos surrogate model. 225 The main functionality of MLDA is shown in Fig. 1 for a case with two levels. We use 226 the Python implementation of MLDA by Lykkegaard (2022) with fixed-length subchains 227

and the option of running a number of  $n_{chains}$  chains in parallel. In the remainder we also assume that the parameter vectors  $\{\theta_\ell\}_{\ell=0}^L$  are comprised of the same model parameters, i.e., we do not consider level-dependent or different coarse and fine (or nested) model parameter vectors.

While other MCMC algorithms sample from a single (posterior) distribution as given 232 in Eq. 2, MLDA considers a hierarchy of distributions  $p_0(\cdot), \ldots, p_\ell(\cdot), \ldots, p_L(\cdot)$  that are 233 computationally cheap approximations of the target density  $p(\cdot)$ , where each  $p_{\ell}(\cdot)$  may 234 be defined according to Eq. 2 corresponding to each model in  $\{\mathcal{F}_{\ell}\}_{\ell=0}^{L}$ . The MLDA al-235 gorithm then gets called on the highest level density  $p_L(\cdot)$ . By recusively calling the MLDA 236 algorithm on level  $\ell - 1$ , subchains with length  $J_{\ell}$  are generated on levels  $1 \leq \ell \leq L$ 237 until level  $\ell = 0$  is reached. We note that different subchain lengths may be used on 238 different levels but the analysis here is restricted to the same  $J_{\ell} = J$  on all levels. On 239 the lowest level  $\ell = 0$ , a conventional MCMC sampler is invoked. The final state of a 240 subchain on level  $\ell - 1$ ,  $\boldsymbol{\theta}_{\ell-1}^{J_{\ell}}$ , is finally passed as a proposal to the higher-level chain 241 on level  $\ell$ . Subsequently, only samples from the highest level are considered for inference. 242 A conventional single-level MCMC sampler may be obtained with using MLDA if only 243 the highest-level model is considered. We note that for MLDA the relation between dif-244 ferent levels is not formally required to show decaying variance and mean as described 245 in section 2.2. Aspects of the design of the model (or posterior) hierarchy are discussed 246 in more detail in section 2.4.2. 247

To assess convergence of the Markov-chains on the highest level, the Gelman-Rubin 248 statistic  $\widehat{R}$  is frequently used for multi-chain samplers (Gelman & Rubin, 1992; Lykkegaard 249 et al., 2023). A value of  $\widehat{R} \leq 1.2$  is often deemed sufficient to ensure convergence (e.g., 250 Vrugt, 2016). MCMC (and MLDA) samples from converged chains are naturally corre-251 lated and may show dependence on initial samples, requiring that an initial number of 252 samples is burned and that that samples are thinned (e.g., every other sample may be 253 omitted to reduce autocorrelation) to obtain approximately independent samples (e.g., 254 Vrugt, 2016; Lykkegaard et al., 2023). The number of approximately intependent sam-255 ples is termed the estimated effective sample size and can be calculated as shown in Geyer 256 (1992, 2011). We obtain effective samples by burning initial samples such that  $\widehat{R} \leq 1.2$ 257 for all chains, followed by thinning such that the resulting number of samples is approx-258 imately equal to the estimated effective sample size. We denote this set of effective sam-259

-9-

ples by matrix **B** with each column representing a single variable and each of the  $N_b$  rows representing a single sample.

262

263

### 2.4 Multilevel Generalized Likelihood Uncertainty Estimation

2.4.1 The MLGLUE Algorithm

The Generalized Likelihood Uncertainty Estimation (GLUE) methodology rejects the formal (Bayesian) statistical basis of inference and instead seeks to identify a set of system representations (combinations of model inputs, model structures, model parameters, errors) that are sufficiently consistent with the observations of that system (Beven & Freer, 2001; Vrugt et al., 2009; Beven & Binley, 2014; Mirzaei et al., 2015).

The likelihood function in GLUE aggregates all aspects of error and consistency as a generalized fuzzy belief. It serves as a decision threshold to separate behavioural (i.e., good agreement between  $\mathbf{Y}$  and  $\tilde{\mathbf{Y}}$ ) and non-behavioural (i.e., poor agreement between  $\mathbf{Y}$  and  $\tilde{\mathbf{Y}}$ ) simulations. Beven and Binley (1992) and (Beven & Freer, 2001) introduced a number of different functions for this purpose. The following likelihood is frequently used (Vrugt et al., 2009):

$$\widetilde{\mathcal{L}}(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) := (\sigma_r^2)^{-W} = \left(\frac{\sum_{i=1}^k (y_i - \widetilde{y}_i)^2}{k-2}\right)^{-W}$$
(6)

where W is a shape parameter of the likelihood function defined by the user. Note that for W = 0, every simulation will have an equal likelihood and for  $W \to \infty$ , the emphasis will be placed on a single best simulation while the other solutions are assigned a negligible likelihood.

Parameter and model output uncertainty is estimated in GLUE by running the model 279 with N parameter samples,  $\{\boldsymbol{\theta}^{(j)}\}_{i=1}^{N}$ , randomly drawn from the prior distribution and 280 evaluating the likelihood function for each sample. The likelihood threshold may either 281 be defined a-priori (as a certain value above which a model realization is considered be-282 havioural) or may be defined as a percentage based on the set of all likelihood correspond-283 ing to the evaluated parameter samples (by setting the threshold to, e.g., the top 10%284 of the likelihood values) (Beven & Binley, 1992; Beven & Freer, 2001; Vrugt et al., 2009). 285 Using only behavioural solutions, (cumulative) probability distributions of model out-286

<sup>287</sup> puts are generated, from which uncertainty estimates are finally computed. Behavioural
 <sup>288</sup> parameter samples are used to estimate the posterior distribution of model parameters.

MLGLUE is generally similar to MLDA (or MLMCMC) as shown in Fig. 1. As with MLDA, a parameter sample  $\theta^{(j)}$  is only finally stored if it is accepted on the highest level. While MLDA makes use of an acceptance probability on all levels (as it is typical in MCMC algorithms), MLGLUE uses a level-dependent likelihood threshold on all levels to distinguish between samples being accepted (i.e., behavioural solutions) and samples being discarded (i.e., non-behavioural solutions).



Figure 1. Schematic representation of multilevel sampling strategies for the case of three levels; (a) MLGLUE approach, green rings indicate a likelihood that is above the level-dependent threshold, red rings indicate the contrary; (b) Multilevel Delayed Acceptance MCMC; circles represent the state or current parameter sample

295 296

297

298

MLGLUE requires that likelihood thresholds are available for every level prior to sampling, although pre-defined likelihood thresholds can optionally be used. MLGLUE considers a simple Monte Carlo estimator to compute likelihood thresholds, where the same set of parameter samples is evaluated on each level using the likelihood function.

The number of those parameter samples,  $N_t$ , should be substantially smaller than the 299 overall number of samples being evaluated with MLGLUE, N. We denote the set of cor-300 responding likelihoods on a single level by  $\{\widetilde{\mathcal{L}}^{(i,\ell)}\}_{i=1}^{N_t}$  and the combined set for all lev-301 els by  $\{\{\widetilde{\mathcal{L}}^{(i,\ell)}\}_{i=1}^{N_t}\}_{\ell=0}^L$ . The likelihood thresholds on the different levels are then obtained 302 by computing a pre-defined percentile estimate from the level-dependent likelihood sam-303 ples (for example, for a threshold corresponding to the top 5% the 95%-percentile is com-304 puted). We denote the set of likelihood thresholds on each level by  $\{\mathcal{L}_{T,\ell}\}_{\ell=0}^{L}$ . We re-305 fer to this step as tuning. For two example problems we discuss the choice of  $N_t$  (see sec-306 tion 4). We also note that the tuning phase can be omitted entirely if level-dependent 307 likelihood thresholds can be pre-defined, e.g., from expert knowledge. 308

From the set of likelihood values on each level,  $\{\{\widetilde{\mathcal{L}}^{(i,\ell)}\}_{i=1}^{N_t}\}_{\ell=0}^L$ , sample estimates of  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$ ,  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell}]$ ,  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}]$ , and  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}]$  for  $\ell = 0, \ldots, L$  are computed to analyze the relation between levels regarding the likelihood. This is equvalent to setting  $\mathbf{Q}_{\ell} = \widetilde{\mathcal{L}}_{\ell}$ , bridging the gap between MLMC and MLGLUE in this context (see section 2.2).

Afterwards, sampling is started and parameter samples  $\boldsymbol{\theta}^{(j)}$  are initially evaluated 313 with the model on the coarsest level,  $\ell = 0$ . If the corresponding likelihood is greater 314 or equal to the level-dependent threshold, the sample is passed to the next higher level 315 and is evaluated again. This process is repeated until the highest level is reached and 316 the sample is finally considered behavioural or non-behavioural. If the likelihood is smaller 317 than the level-dependent threshold on any level, the sample is immediately regarded as 318 non-behavioural and rejected. Therefore, samples with low likelihoods are already dis-319 regarded on lower levels, leading to substantial computational savings. In the support-320 ing information, the reasoning for using level-dependent likelihood thresholds as well as 321 the structure of the algorithm is clarified in more detail. The MLGLUE algorithm is pre-322 sented in algorithm 1 with tuning excluded and schematically shown in Fig. 2. 323

324

### 2.4.2 Designing the Model Hierarchy

During multilevel inversion, no explicit approach exists yet to optimally pre-define the number of levels or the difference in resolution between the levels. In their example applications of multilevel MCMC and MLDA, Dodwell et al. (2019) and Lykkegaard et al. (2023) arbitrarily pre-define the coarsening as well as the number of levels considered but give some analysis of the effect regarding the number of levels. In similar examples

### Algorithm 1: Multilevel Generalized Likelihood Uncertainty Estimation

- 1 Draw a sample  $\Theta_0$  of N points from the (typically uniform) prior distribution  $p_p(\boldsymbol{\theta})$ **2** for j = 0, ..., N do for  $\ell = 0, \ldots, L$  do 3 Compute the likelihood  $\widetilde{\mathcal{L}}^{(j,\ell)} = \widetilde{\mathcal{L}}(\boldsymbol{\theta}^{(j)}|\widetilde{\mathbf{Y}})$  with sample  $\boldsymbol{\theta}^{(j)}$  from  $\boldsymbol{\Theta}_0$  and 4 with the model on level  $\ell$ if  $\ell = L$  and  $\widetilde{\mathcal{L}}^{(j,\ell)} \geq \widetilde{\mathcal{L}}_{T,\ell}$  then 5 Store  $\boldsymbol{\theta}^{(j)}$  in matrix **B**, store the corresponding simulation results **Y** in 6 **S**, increment  $j \leftarrow j + 1$ , and break the loop over the levels  $ext{ if } \widetilde{\mathcal{L}}^{(j,\ell)} \geq \widetilde{\mathcal{L}}_{T,\ell} ext{ then }$ 7 Increment  $\ell \leftarrow \ell + 1$ , continuing the loop over the levels for sample  $\boldsymbol{\theta}^{(j)}$ 8 if  $\widetilde{\mathcal{L}}^{(j,\ell)} < \widetilde{\mathcal{L}}_{T,\ell}$  then 9 Increment  $j \leftarrow j + 1$ , breaking the loop over the levels 10 11 for  $\mathbf{b}^{(i)}, i = 1, ..., N_b$  in **B** do Normalize the corresponding likelihood as  $\widetilde{\mathcal{L}}'(\mathbf{b}^{(i)}|\widetilde{\mathbf{Y}})$  such that 12
- 13 for  $\mathbf{Y}^{(i)}, i = 1, ..., N_b$  in **S** do
- 14 Assign the corresponding weight  $\widetilde{\mathcal{L}}'(\mathbf{b}^{(i)}|\widetilde{\mathbf{Y}})$
- 15 Sort the  $\mathbf{Y}^{(i)}$ ,  $i = 1, ..., N_b$  increasingly according to their weights and create uncertainty estimates from the empirical distribution obtained this way (e.g., as quantiles)



Figure 2. Schematic representation of the multilevel Generalized Likelihood Uncertainty Estimation algorithm; tuning refers to the (optional) Monte Carlo estimation of likelihood thresholds, sampling refers to the repeated evaluation of parameter samples (see the description of algorithm steps)

- to our subsequently considered benchmark example of groundwater flow (see section 3.2),
  Cliffe et al. (2011) consider 5 levels for MLMC, Dodwell et al. (2019) consider up to 5
  levels for multilevel MCMC, Lykkegaard and Dodwell (2022) consider 2 levels with MLDA,
  and Lykkegaard et al. (2023) consider 3 levels with MLDA. In the following we give guidelines on how to design a hierarchy of models and also show directions for further research.
- A geometric series of resolutions for the computational grids (in space or time or both) is often most suitable in the context of MLMC (also see section 2.2), where the factor of grid refinement (when going from  $\ell$  to  $\ell+1$ ) or coarsening (when going from  $\ell$  to  $\ell-1$ ) between subsequent levels is constant (Giles, 2015). We also adopt this method in this study.
- In MLGLUE, a parameter sample that is accepted on the highest level with the highest resolution model is evaluated on all lower levels with lower resolution models be-

fore. Therefore, the number of levels in the model hierarchy should be as low as possi-342

ble and the coarsening factor as large as possible to obtain a high computational efficiency 343 of the multilevel hierarchy. Those aspects are then restricted by the quality of the coarsest-344 level model being sufficiently high, by the required resolution on the highest level, and 345 by the requirement for sufficiently high correlation between subsequent levels. Those cri-346 teria can be analyzed via the relations between levels regarding  $\{\{\widetilde{\mathcal{L}}^{(i,\ell)}\}_{i=1}^{N_t}\}_{\ell=0}^L$  (see also 347 section 2.2). 348

In this study we consider cases where a target resolution is given for the highest 349 level model and lower resolution models are obtained by subsequent coarsening. After-350 wards, in practical applications, the coarsest possible model resolution for the lowest level 351 should be determined approximately. With the highest and lowest resolutions specified, 352 the number of levels is determined through finding an appropriate coarsening factor that 353 results in sufficiently high correlation between the levels (see section 2.2). We investi-354 gate and discuss those aspects in more detail for the results of the test problems in sec-355 tion 4. 356

An alternative strategy for the design of the hierarchy is presented in Vidal-Codina 357 et al. (2015) and Giles (2015) for non-geometric MLMC. It relies on generating a set of 358 test models for a large number of levels,  $\{\mathcal{F}_{\ell}\}_{\ell=0}^{L}$ , and then selecting a subset of levels 359 that satisfy some conditions on the relation between levels, similar to the conditions used 360 in the tuning phase of MLGLUE. In any case, this approach requires additional com-361 putational resources to optimize the hierarchy, being associated with a large number of 362 degrees of freedom in the design. This strategy can potentially be applied for MLGLUE 363 as well but is not the focus of the current study. This approach is left open for further 364 research as it has become apparent in this study that a geometric series generally serves 365 as a robust starting point under various conditions. 366

367

### 2.4.3 Parallelization

368

Like the conventional formulation of GLUE, MLGLUE can be parallelized in a straightforward manner to accelerate computation. We utilize Ray v2.2.0 (Team, 2022) for par-369 allelization with its multiprocessing. Pool API. Parallelization is achieved by using Ray 370 Actors instead of local processes. For MLGLUE and GLUE, the function (or task) be-371 ing parallelized corresponds to the evaluation of a single parameter sample, starting on 372

-15-

 $\ell = 0$  and including all subsequent model runs on higher levels (see the MLGLUE algorithm). MLGLUE considers running the hierarchy of models  $\{\mathcal{F}_0(\theta_i), \ldots, \mathcal{F}_L(\theta_i)\}$  for a single parameter sample  $\theta_i$  as one iteration. As the parallelization is implemented on the level of these iterations, it allows for evaluating multiple parameter samples in parallel. For the case of using MLGLUE with a single level (i.e., conventional GLUE), the iteration reduces to running the target model,  $\{\mathcal{F}_L(\theta_i)\}$ , for multiple parameter samples in parallel.

For MLDA and MCMC, however, the parallelization is implemented on the level 380 of individual chains. While the MLDA implementation (tinyDA v0.9.8, Lykkegaard (2022)) 381 does not use the multiprocessing. Pool API, it still relies on Ray Actors for paralleliza-382 tion, implemented via remote functions. Therefore, the underlying mechanism for par-383 allelization are identical for GLUE, MLGLUE, MCMC, and MLDA. Still, differences re-384 garding the increase in computational efficiency may be observed when comparing se-385 quential and parallelized algorithm run times for GLUE and MLGLUE with those for 386 MCMC and MLDA. This is due to (1) the differences in the implementation of paral-387 lelization and (2) the differences in the algorithms themselves. 388

389

### 2.5 Analysis of Posterior Convergence

In order to compare the different methods of statistical inference in our study, we assess the convergence to a stable posterior distribution and monitor the number of model evaluations and the computational time required for convergence. We introduce a simple way of assessing convergence that works for any method that returns a - possibly ordered - sequence of values in  $\mathbb{R}^n$ , which are assumed here to be samples from a probability distribution. In the context of MCMC, the introduced methodology is not to be mistaken for a way of assessing the convergence of (Markov-) chains.

The central concept of the methodology is to analyze the ratio of mean and variance of the (marginal) posterior distribution, estimated from a subset of the set of all available samples, to mean and variance estimated from the set of all available samples  $(N_b \text{ samples in } \mathbf{B})$ . As the subset gets larger, and eventually becomes equal to  $\mathbf{B}$ , this quantity allows for the analysis of convergence behaviour. The subset is taken to be the first *s* samples from the posterior samples returned by a method of statistical inference. We denote the estimate of the mean or any higher-order moment around the mean by  $\mu_{m}^{s}$ , where *s* represents the size of the subset and *m* represents the moment order. We define the relative deviation  $\mathcal{D}_{m}^{s}$  of moment *m*, computed with a subset of size *s*, from the globally estimated moment as

$$\mathcal{D}_m^s := \frac{\mu_m^s}{\mu_m^{N_b}} - 1 \tag{7}$$

By definition,  $\mathcal{D}_m^s \to 0$  as  $s \to N_b$ ; however, the analysis regarding how and how quickly  $\mathcal{D}_m^s$  tends towards zero as s increases allows for the analysis of convergence behaviour. We assume convergence at  $s = s_c$  if  $-0.05 \leq \mathcal{D}_m^s \leq 0.05$  for all  $s \geq s_c$ . Assuming that the samples are obtained uniformly over time during inference or computation enables the assessment of convergence against computation time instead of sample size.

### 413 **3 Test Problems**

The test problems discussed in sections 3.1 and 3.2 are used to illustrate the dif-414 ferences between the methods of statistical inference (MLGLUE, GLUE, MLDA, MCMC) 415 regarding obtained posterior distributions, uncertainty estimates for model outputs, and 416 computational efficiency. An identical number of prior parameter samples is used for all 417 methods to ensure comparability. For GLUE and MLGLUE, an informal likelihood func-418 tion (Eq. 6) is used for each problem. MCMC and MLDA are used with a formal like-419 lihood function (Eq. 3). We analyze the tuning phase separately for both examples us-420 ing two threshold settings (selecting the top 2 % and 7 %) for different  $N_t$ . 421

For reasons of reproducibility, seeds are used for pseudo-random number generation, which is used in multiple places (e.g., drawing samples form a distribution); for each problem, the same seeds are used for all methods of inference in the example under study.

- All methods of inference are implemented in the Python programming language. The tinyDA v0.9.8 (Lykkegaard, 2022) package is used for MLDA and MCMC sam-
- $_{427}$  pling with a DREAM(Z)-sampler, which is similar to the DREAM(ZS)-sampler (Vrugt,
- <sup>428</sup> 2016; Lykkegaard, 2022), using Ray v2.2.0 (Team, 2022) for parallelization. ArviZ v0.12.1
- (Kumar et al., 2019) is used for the analysis of MLDA and MCMC results regarding chain
- 430 convergence and effective sample size (see section 2.3); in tinyDA, the initial sample is
- returned additionally to the N drawn samples. MLGLUE is implemented as a Python

package and also enabled for parallel computing with Ray v2.2.0 (Team, 2022). We note
that we subsequently refer to the processed posterior samples from MCMC and MLDA
(i.e., after burn-in and thinning, see section 2.3) as effective samples. The same term is
also used for unprocessed GLUE and MLGLUE posterior samples.

436

### 3.1 Rainfall-Runoff Modelling

The first case study considers rainfall-runoff modelling using the conceptual model 437 HYMOD (Boyle, 2001), which is schematically shown in Fig. 3. The model has five pa-438 rameters (explained in Fig. 3), takes time series of precipitation, P(t) [ $LT^{-1}$ ], and po-439 tential evaporation, PET(t) [ $LT^{-1}$ ], as inputs and outputs a time series of discharge, 440 Q(t) [LT<sup>-1</sup>]. This model has been frequently and similarly used in the in the context 441 of statistical inference, uncertainty analysis, and sensitivity analysis (Boyle, 2001; Wa-442 gener et al., 2001; Vrugt et al., 2003, 2005; Blasone et al., 2008; Vrugt et al., 2008, 2009; 443 Herman et al., 2013). We apply the model to data from the Leaf River catchment near 444 Collins, Mississippi, USA, which has been studied with the same model multiple times 445 before (Wagener et al., 2001; Vrugt et al., 2003, 2005; Blasone et al., 2008; Vrugt et al., 446 2008, 2009). We refer the reader to the aforementioned references for detailed descrip-447 tions of the HYMOD model and the study area. Contrary to other studies we consider 448 time series with hourly instead of daily resolution (Gauch et al., 2020, 2021) and use the 449 hydrological year of data from 2009-10-01 to 2010-09-30. The first 25 days are consid-450 ered a warm-up period, being simulated but not used to calculate likelihoods. 451

The model is implemented in the Python programming language following Knoben 452 et al. (2019); Trotter et al. (2022); Trotter and Knoben (2022) and the differential equa-453 tions are solved using the explicit Euler method (e.g., Braun, 1993). The highest-level 454 model uses an hourly time step equal to the data time steps. Two additional lower-level 455 models are considered with time steps of two and four hours, respectively (i.e., time step 456 lengths are doubled when going to the next lower level). On levels  $\ell = 0$  and  $\ell = 1$ , 457 resulting time series of discharge are linearly interpolated to the time steps of the model 458 on level  $\ell = 2$  to allow for the calculation of likelihoods with the original data time steps. 459

The prior distribution  $p_0(\boldsymbol{\theta})$  is chosen to be a uniform distribution over the parameters  $\boldsymbol{\theta} = [C_{max}, \beta, \alpha, \kappa_s, \kappa_q]^T$  with lower bounds  $\boldsymbol{\theta}_l = [1.0, 0.1, 0.0, 0.0, 0.0]$  and upper bounds  $\boldsymbol{\theta}_u = [1000.0, 2.0, 1.0, 0.1, 0.5]$ . Length units of [mm] and time units of [h]



Figure 3. (a) Schematic representation of the HYMOD model (Vrugt et al., 2009);  $C_{max}$  [L] is the maximum catchment storage,  $\beta$  [-] is the spatial variability of soil moisture storage,  $\alpha$  [-] is the distribution factor between reservoirs, and  $\kappa_q$  [ $T^{-1}$ ] and  $\kappa_s$  [ $T^{-1}$ ] are discharge coefficients of the quick-flow and slow-flow reservoirs, respectively; (b) discharge simulated by models on all three levels for two consecutive events, only every fifth time step is marked

are used throughout the model and for all datasets. A total number of  $N_t + N = 5,000 +$ 463 995,000 = 1,000,000 samples are drawn from  $p_p(\theta)$  with each inference method, where 464  $N_t = 5,000$  samples are used to estimate the level-dependent likelihood thresholds (see 465 section 2.4) and to analyze the relations between the levels (see section 2.2) in MLGLUE. 466 The choice of  $N_t$  is discussed in section 4.1. A constant variance equal to the constant 467 additive Gaussian noise variance ( $\sigma^2 = 1.0 \ mm^2 h^{-2}$ ) is used for the Gaussian likeli-468 hood (see Eq. 3); for the likelihood used in MLGLUE and GLUE (see Eq. 6) W = 1469 is used. The likelihood thresholds are estimated to correspond to the best 2% of sim-470 ulations. For MLDA, the sub-sampling rate is set to 5. MLDA and MCMC are run with 471 10 independent chains. All methods are run on 32 dual-core CPUs (64 total threads). 472

473

### 3.2 Groundwater Flow

The second example considers steady-state two-dimensional groundwater flow in an aquifer with inhomogeneous horizontal hydraulic conductivity, Dirichlet-type (fixed potentials), Neumann-type (no-flow conditions, recharge), Robin-type (river), and nodal sink type (wells) boundary conditions:

$$\frac{\partial}{\partial x} \left( K_{xx} \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left( K_{yy} \frac{\partial h}{\partial y} \right) + R = 0 \tag{8}$$

$$h = h_c \quad \forall y \in \partial\Omega, x = 0 \ m \tag{9}$$

$$\frac{\partial h}{\partial y} = 0 \quad \forall x \in \partial\Omega, y \in \{0 \ m, 5, 000 \ m\}$$
(10)

$$\frac{\partial h}{\partial x} = 0 \quad \forall y \in \partial\Omega, x = 10,000 \ m \tag{11}$$

$$f_{riv} = c_{riv}\Delta h \quad \forall 0 \ m \le x \le 10,000 \ m, y = 1,000 \ m$$
 (12)

where K  $[LT^{-1}]$  is the hydraulic conductivity field, h [L] is the hydraulic head field, R  $[LT^{-1}]$  is the recharge flux,  $f_{riv}$   $[LT^{-1}]$  is river inflow, and  $c_{riv}$   $[T^{-1}]$  is riverbed conductance. The model is set up with the finite-differences code MODFLOW-NWT and the reader is referred to Harbaugh (2005) and Niswonger et al. (2011) for a detailed description of the model and boundary condition implementations.

The reference model is discretized as a regular structured grid with a cell-size of 483  $25 \ m \times 25 \ m$ , having 200 rows and 400 columns. The aquifer bottom is horizontal at 484 10.0 m above the reference datum; the aquifer top represents a tilted plane falling lin-485 early from  $55.0 \ m$  on the left side of the domain to  $45.0 \ m$  above the reference datum 486 on the right side of the domain. A river crosses the domain along a single row, having 487 a constant water level at 6.0 m below the aquifer top and a river bottom at 9.0 m be-488 low the aquifer top. 5 wells are placed in the model domain with a total extraction rate 489 of 700  $md^{-1}$ . Spatially uniform recharge is applied with a rate of  $2 \cdot 10^{-5} md^{-1}$ . A 490 constant head of  $45.0 \ m$  above the reference datum is assigned to the leftmost column 491 of cells. 12 observation points as well as 1 prediction point are placed in the domain. 492

The hydraulic conductivity in every cell is obtained in the reference model using 493 a regular grid of pilot points (e.g., Doherty, 2003), linearly spaced (5 along columns, 10 494 along rows) starting on the domain boundaries. Reference values of pilot point  $\log_{10}$ -hydraulic 495 conductivities are obtained by sampling from a log-normal distribution with  $\mu = 0.3$ 496 and  $\sigma = 0.7$ . Gaussian process regression (GPR), as implemented in scikit-learn v1.2.0 497 (Pedregosa et al., 2011), is used to interpolate  $\log_{10}$ -hydraulic conductivities at cell cen-498 ters of the reference model with a radial basis function kernel with a fixed length scale 499 of 600 m. The model domain and its main characteristics are shown in Fig. 4 for the 500 models on levels  $\ell = 0$  and  $\ell = 3$ . 501

The reference model is also the highest-level model. Besides this model, three lowerlevel models are considered, resulting in  $\ell = 0, 1, 2, 3$ . Lower-level models are obtained via grid coarsening, where cell sizes are doubled going from  $\ell$  to  $\ell-1$ . Lower-level hydraulic conductivity values at each cell are obtained by using the geometric mean of corresponding higher-level cells.



Figure 4. Groundwater flow model domain; head contours obtained with true parameters on level  $\ell = 0$  (a) and on level  $\ell = 3$  (b); horizontal hydraulic conductivity field on level  $\ell = 0$  (c) and on level  $\ell = 3$  (d); specific characteristics are: constant head cells (blue), river cells (purple), wells (red), observation points (circles), prediction point (diamond)

Besides the 50 pilot point parameters, the GPR length scale is considered a model parameter as well;  $\boldsymbol{\theta} = [\theta_{1,PP}, \dots, \theta_{50,PP}, \theta_{51,GPR}]^T$ . We denote the parameter-to-observable map (i.e., Eqs. 8 to 12) by  $\mathcal{F}(\boldsymbol{\theta})$ . Adding Gaussian random noise to the observations then leads to  $\tilde{\mathbf{Y}} = \mathcal{F}(\boldsymbol{\theta}) + \varepsilon$ ,  $\varepsilon \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$ .

As a prior distribution  $p_p(\theta)$ , a uniform distribution is chosen with lower bounds  $\theta_l = [1 \cdot 10^{-2}, \dots, 1 \cdot 10^{-2}, 5 \cdot 10^2]$  and upper bounds  $\theta_u = [1 \cdot 10^1, \dots, 1 \cdot 10^1, 1 \cdot 10^3]$ . A total number of  $N_t + N = 2,000 + 98,000 = 100,000$  samples are drawn from  $p_p(\theta)$ with each inference method, where  $N_t = 2,000$  samples are used to estimate the leveldependent likelihood thresholds (see section 2.4) and to analyze the relations between the levels (see section 2.2) in MLGLUE. The choice of  $N_t$  is discussed in section 4.2. A constant variance equal to the constant additive Gaussian noise variance ( $\sigma^2 = 1.0 m^2$ ) is used for the Gaussian likelihood (see Eq. 3); for informal likelihoods (see Eq. 6) W =1 is used. The likelihood thresholds are estimated to correspond to the best 7% of all simulations. For MLDA, the sub-sampling rate is set to 5. All methods are run on 32 dual-core CPUs (64 total threads).

### 522 4 Results

For the two examples considered, we now present results of inversion with the methodologies of MLGLUE, GLUE, MLDA, and MCMC. We analyze how models on different levels are related and how the results obtained with a multilevel approach differ from the conventional approach using a single model. Differences between MLGLUE and GLUE on one hand, and between MLDA and MCMC on the other hand, are discussed regarding posterior parameter and model output distributions, as well as computational efficiency.

MCMC chains typically exhibit a transition period where the samples approach the 530 posterior distribution. The samples of this transition period are discarded as burn-in (Gallagher 531 et al., 2009; Brunetti et al., 2023). GLUE and MLGLUE both result in independent pos-532 terior samples, while MCMC and MLDA result in correlated posterior samples. To com-533 pare both groups (GLUE & MLGLUE and MCMC & MLDA) on an equal basis, inde-534 pendent samples are obtained from MCMC and MLDA samples via *thinning*; only ev-535 ery  $\mathcal{K}$ -th sample is considered for subsequent analysis. We apply thinning such that the 536 thinned number of samples is approximately equal to the estimated effective sample size 537 of unthinned samples (see section 2.3). 538

539

### 4.1 Rainfall-Runoff Modelling

In this example, likelihood thresholds are not pre-defined but are estimated during the tuning phase of the MLGLUE algorithm. For two threshold settings the estimated likelihood thresholds are shown in Fig. S1 in the supplementary information for different numbers of tuning samples,  $N_t$ . For the smaller threshold setting of 2 % (i.e., higher likelihood threshold values), likelihood thresholds stabilize at  $N_t = 5,000$  after showing initial oscillations. For the larger threshold setting of 7 %, likelihood values tend to

-22-

decrease successively, stabilizing at  $N_t = 2,000$ . The ratio of the likelihood thresholds on the three levels, however, remains approximately equal for both threshold settings, even for smaller  $N_t$ . From this analysis and with the threshold setting being 2 %, we set  $N_t = 5,000$  in this example.

The relations between the three levels are shown in Fig. S2 in the supplementary information.  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$  and  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell}]$  are approximately constant across all levels and  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}]$  and  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}]$  decay across all levels. The correlation coefficients are 0.9102 between levels  $\ell = 0$  and  $\ell = 1$  and 0.9958 between levels  $\ell = 1$  and  $\ell = 2$  and therefore increase with increasing level index. Consequently, the approximation error of the likelihoods decreases as  $\ell \to L$ .

The sampling efficiencies of all methods are shown in Fig. 5; detailed results of MLDA 556 and MCMC chain convergence (Gelman-Rubin statistic) and the recovery of effective sam-557 ples is described in Text S3 in the supplementary information. With MLGLUE the over-558 all computation time is reduced by  $\approx 58$  % and the number of effective samples per minute 559 is  $\approx 74$  % higher compared to GLUE. With MLDA the overall computation time is re-560 duced by  $\approx 18$  % and the number of effective samples per minute is  $\approx 39$  % lower com-561 pared to conventional MCMC. While the number of effective samples per minute is lower 562 for MLDA compared to MCMC, the ratio between the number of effective samples to 563 the total number of posterior samples on the highest level is higher, indicating lower sam-564 ple autocorrelation before thinning. More detailed alayses of MLDA and MCMC results 565 are presented in the supporting information. 566

The results of convergence analysis (see section 2.5) are shown in Fig. 6. Results 567 are obtained by splitting the original sets of effective parameter samples into 200 con-568 secutive subsets, independently of the method of inference. Multilevel approaches (ML-569 GLUE and MLDA) generally converge after a shorter computation time compared to 570 their conventional counterparts (GLUE and MCMC), respectively. The deviation of mean 571 and variance, however, is larger for small sample sizes with MLGLUE compared to GLUE 572 with the set of prior samples being equal for MLGLUE and GLUE. Compared to MLDA, 573 MCMC results show a larger deviation of the mean even for larger sample sizes. 574

Estimated cumulative distribution functions (CDFs) of the parameter posteriors are shown in Fig. 7 (a) - (d). Posteriors obtained with multilevel methods (MLGLUE and MLDA) are virtually identical to their conventional counterparts (GLUE and MCMC).

-23-



Figure 5. Sampling efficiencies for the rainfall-runoff modelling example; (a) computation times with percentual reductions compared to conventional methods; (b) No. of model calls on the highest level (dashed), No. of posterior samples (light colors), No. of effective posterior samples (dark colors); (c) No. of effective posterior samples per minute with percentual increase compared to conventional methods



**Figure 6.** Convergence analysis for the rainfall-runoff modelling example (Eq. 7); for the different methods of inference (a) - (d) shows the deviation of the mean and (e) - (h) shows the deviation of the variance; grey regions represent the region where convergence is achieved; black vertical lines represent the computational time at which convergence is achieved for all parameters

Uncertainty estimates of MLGLUE are different from those of GLUE in that they have smaller range, which is particularly visible at peak flow events (e.g., around 2009-12-17). Uncertainty estimates from MLDA and MCMC are virtually identical, also at peak flow events. The Nash-Sutcliffe model efficiency (Nash & Sutcliffe, 1970), computed with the

- median of the simulations, is virtually identical for MLGLUE and GLUE and slightly
- <sup>583</sup> higher for MLDA compared to MCMC.



Figure 7. CDFs of model parameters for the rainfall-runoff modelling example for MLGLUE and GLUE (a to e), for MLDA and MCMC (f to j) and 99% - 1% uncertainty estimates around the median value for MLGLUE and GLUE (k) and for MLDA and MCMC (l)

### 4.2 Groundwater Flow

584

In this example, likelihood thresholds are not pre-defined but are estimated dur-585 ing the tuning phase of the MLGLUE algorithm. For two threshold settings the estimated 586 likelihood thresholds are shown in Fig. S3 in the supplementary information for differ-587 ent numbers of tuning samples,  $N_t$ . For the smaller threshold setting (2 %, correspond-588 ing to a higher likelihood threshold), the likelihood thresholds on all levels generally in-589 crease as  $N_t$  increases and stabilize at  $N_t = 5,000$ . For the setting with a larger threhold 590 setting (7 %), the likelihood values also increase as  $N_t$  increases but remain at smaller 591 values compared to the smaller threshold setting and stabilize at  $N_t = 2,000$ . The ra-592 tio of the likelihood thresholds on the four levels remains approximately equal only for 593

the larger threshold setting, even for smaller  $N_t$ . See section 4.1 for a more detailed discussion on the tuning phase. With the threshold setting being set to 7 % in this example, we set  $N_t = 2,000$  here to keep  $N_t$  as small as possible to reduce overall computational cost but ensure reasonably stable likelihood threshold estimates.

The relations between the three levels are shown in Fig. S4 in the supplementary information.  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$  and  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell}]$  are approximately constant and  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}-\widetilde{\mathcal{L}}_{\ell-1}]$  and  $\mathbb{E}[\widetilde{\mathcal{L}}_{\ell}-\widetilde{\mathcal{L}}_{\ell-1}]$  decay across all levels. The variance of the sampled likelihoods on level  $\ell = 0$ , however, is smaller than on higher levels. The correlation coefficients are 0.9954 between levels  $\ell = 0$  and  $\ell = 1$ , 0.9989 between levels  $\ell = 1$  and  $\ell = 2$ , and 0.9997 between levels  $\ell = 2$  and  $\ell = 3$  and therefore increase with increasing level index.

The sampling efficiencies of all methods are shown in Fig. 8; detailed results of MLDA 604 and MCMC chain convergence (Gelman-Rubin statistic) and the recovery of effective sam-605 ples is described in Text S4 in the supplementary information. The overall computation 606 time is reduced by  $\approx 63$  % and the number of effective samples per minute is  $\approx 122$  % 607 higher with MLGLUE compared to GLUE. The overall computation time is reduced by 608  $\approx 70$  % and the number of effective samples per minute is  $\approx 206$  % higher with MLDA 609 compared to conventional MCMC. The ratio between the number of effective samples 610 to the total number of posterior samples on the highest level is substantially higher for 611 MLDA compared to MCMC, indicating lower sample autocorrelation before thinning. 612 More detailed analyses of MLDA and MCMC results are presented in the supporting in-613 formation. 614

The results of convergence analysis (see section 2.5) are shown in Fig. 9. Results 615 are obtained by splitting the original sets of effective parameter samples into 200 con-616 secutive subsets, independently of the method of inference. Multilevel approaches (ML-617 GLUE and MLDA) generally converge after a shorter computation time compared to 618 their conventional counterparts (GLUE and MCMC), respectively. The deviation of mean 619 and variance is larger with MLGLUE compared to GLUE, especially for small sample 620 sizes, although the set of prior samples is equal for MLGLUE and GLUE. MLDA and 621 MCMC results show similar convergence behaviour, except for the length scale param-622 eter. MLDA results show larger deviations of the length scale mean and variance for smaller 623 sample sizes. 624

-26-



**Figure 8.** Sampling efficiencies for the groundwater flow example; (a) computation times with percentual reductions compared to conventional methods; (b) No. of model calls on the highest level (dashed), No. of posterior samples (light colors), No. of effective posterior samples (dark colors); (c) No. of effective posterior samples per minute with percentual increase compared to conventional methods



**Figure 9.** Convergence analysis for the groundwater flow example (Eq. 7); for the different methods of inference (a) - (d) shows the deviation of the mean and (e) - (h) shows the deviation of the variance; grey regions represent the region where convergence is achieved; black vertical lines represent the computational time at which convergence is achieved for all parameters

625 626

627

628

629

Estimated CDFs of the parameter posteriors are shown in Fig. 10 (a) - (d). Posteriors obtained with MLGLUE are substantially more conditioned than GLUE posteriors (indicated by the deviations of the cumulative distributions from the straight line representing a uniform distribution). The length scale posterior, however, is similar for MLGLUE and GLUE. MLDA and MCMC posteriors are virtually identical. Uncertainty estimates of MLGLUE are different from those of GLUE as they show slightly larger ranges

and less bias towards higher values, which can be attributed to the differences in the pos-

terior distributions. Uncertainty estimates from MLDA and MCMC are similarly dif-

- ferent in that they have smaller range and less bias towards higher values for MLDA.
- As evaluated with the coefficient of determination  $(R^2)$ , MLGLUE results are slightly
- more accurate compared to GLUE. Similarly, MLDA results are slightly more accurate
- compared to MCMC.



Figure 10. CDFs of model parameters for the groundwater flow example (a, b, c, d) and 99% - 1% uncertainty estimates around the median value for observation points (e) and for the prediction point(f)

### 637 5 Discussion

We applied MLGLUE to two test problems and subsequently compared the results to conventional GLUE as well as to MCMC and MLDA. These applications illustrate the capabilities of the multilevel extension but also identifies aspects that need careful consideration for practical applications. The examples considered here are comparable to other examples used to study multilevel methods found in, e.g., Cliffe et al. (2011),
Dodwell et al. (2019), Lykkegaard et al. (2023), and (Cui et al., 2024). However, although
groundwater flow is a frequently used example case, the system used here (see section
3.2) is far more complex compared to previous applications. Additionally, other previous studies only considered synthetic cases where the underlying truth is known; our rainfall-

<sup>647</sup> runoff modelling example considers a real system.

For both examples it was identified that the number of tuning samples,  $N_t$ , required 648 to obtain stable and accurate estimates of likelihood thresholds increases with decreas-649 ing threshold percentage although the parameter space dimensions were greatly differ-650 ent (n = 5 for rainfall-runoff modelling and n = 51 for groundwater flow). For a thresh-651 old setting of 2 %,  $N_t = 5,000$  tuning samples were needed for accurate estimation in 652 both examples. For a threshold setting of 7 %, however, only  $N_t = 2,000$  tuning sam-653 ples were required for accurate estimation in both examples. This behaviour is in agree-654 ment with the fact that Monte-Carlo estimators generally do not perform well at rare 655 event estimation (e.g., Beck & Zuev, 2015), which can be translated to the present case 656 of estimating values in the tails of the distribution of likelihood values (i.e., estimating 657 large percentiles). We hypothesize that using a Latin hypercube design or quasi-Monte 658 Carlo sampling during the tuning phase increases robustness as well as computational 659 efficiency. 660

The model hierarchies were designed for both examples using a coarsening factor 661 of 2. While for the rainfall-runoff modelling this choice resulted in increased computa-662 tional efficiency of MLGLUE compared to GLUE, a coarsening factor of 3 (results not 663 shown) resulted in a substantially reduced acceptance ratio. This was especially evident 664 from a large difference between highest-level model runs and finally accepted samples. 665 The consideration of a fourth level, being even coarser than the current level  $\ell = 0$ , was 666 not successful as the correlation between the two lowest levels then was found to be very 667 low, again leading to low acceptance rations. Similar behaviour was identified for the ground-668 water flow example, where the likelihood variance on the lowest level with the coarsest 669 resolution was smaller than on subsequently higher levels. As described by Cliffe et al. 670 (2011), further hypothetical grid coarsening beyond the current level  $\ell = 0$  for such a 671 case can result in the graphs of  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$  and  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}-\widetilde{\mathcal{L}}_{\ell-1}]$  to eventually intersect, result-672 ing in  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}] > \mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$  for some  $\ell$ . In the context of MLMC (forward problems), 673 this then leads to an increased computational cost compared to conventional MC. As in-674

-29-

dicated by Eq. 5, if  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell}]$  decreases and  $\mathbb{V}[\widetilde{\mathcal{L}}_{\ell} - \widetilde{\mathcal{L}}_{\ell-1}]$  increases with decreasing  $\ell$ , then 675  $\operatorname{Cov}(\widetilde{\mathcal{L}}_{\ell},\widetilde{\mathcal{L}}_{\ell-1})$  must decrease as well. Insufficient correlation between the likelihood val-676 ues on subsequent levels in MLGLUE would then result in lower acceptance rates on lev-677 els  $\ell > 0$ , affecting the overall computational efficiency of the algorithm. Therefore, the 678 characteristics of the relation between levels as described for MLMC in section 2.2 should 679 also be considered for MLGLUE to ensure computational efficiency. We hypothesize at 680 this point that a non-geometric construction of the hierarchies can potentially further 681 increase computational efficiency (Vidal-Codina et al., 2015; Giles, 2015). The analy-682 sis required for this, however, demands additional computational resources to optimize 683 the design as it is associated with a large number of degrees of freedom. 684

Differences exist in the number of posterior samples between MLGLUE and GLUE. 685 This can be attributed to parameter samples being occasionally discarded on lower lev-686 els with lower resolution models although they would be accepted on higher levels. This 687 is due to the fact that the likelihoods on subsequent levels are not perfectly correlated 688 in both example applications. This effect is reduced as the correlation between subse-689 quent levels increases; it can be controlled through careful design of the model hierar-690 chy (see section 2.4.2). This behaviour is also reflected in the convergence analysis where, 691 using the same set of prior samples, MLGLUE initially shows larger deviations of pos-692 terior mean and variance. Differences in posterior samples also result in small deviations 693 regarding posterior parameter distributions and uncertainty estimates of model outputs. 694

### 695 6 Conclusions

In the hydrological sciences, the popularity of statistical inference and inversion has 696 remained high. However, the applicability of corresponding approaches to more complex 697 models and in the context of digital twins has been limited by the associated computa-698 tional cost of solving inverse problems. The goal of our study was to introduce and test 699 an extension to the GLUE methodology for Bayesian inversion that alleviates the prob-700 lems associated with computationally costly models through considering multiple lev-701 els of model resolution (MLGLUE). Inspired by multilevel Monte Carlo, in MLGLUE 702 most parameter samples are evaluated on lower levels with computationally cheaper low-703 resolution models instead of using a (data-driven) surrogate model that is decoupled from 704 the high-fidelity or target model. Only samples associated with a likelihood above a cer-705 tain threshold, which can optionally be estimated during a tuning phase of the algorithm, 706

-30-

are subsequently passed to higher levels with costly high-resolution models for evaluation. Inferences are made at the level of the highest-resolution model but substantial computational savings are achieved by discarding samples with low likelihood already on levels with low resolution and low computational cost.

MLGLUE is evaluated using example inverse problems involving a rainfall-runoff 711 model and a groundwater flow model. The results of statistical inversion with MLGLUE 712 are compared to the results from GLUE, Markov-chain Monte Carlo (MCMC), as well 713 as multilevel delayed acceptance (MLDA) MCMC. Identical numbers of prior samples 714 are considered for all methods to ensure comparability. We show that the results (pa-715 rameter posteriors, uncertainty estimates, convergence behaviour) obtained with mul-716 tilevel approaches (MLGLUE and MLDA) are highly similar to conventional approaches 717 (GLUE and MCMC), respectively. MLGLUE showed the resulted in the lowest compu-718 tation time and the highest number of posterior samples per minute for both example 719 problems and compared to all other methods of inference. 720

We identified in both example applications that MLGLUE and MLDA generally 721 result in less precise estimates of parameter posteriors for small effective sample sizes com-722 pared to GLUE and MCMC, respectively. This effect, however, vanishes for larger sam-723 ple sizes required in practical applications. For both examples, MLGLUE resulted in the 724 lowest computational time for inversion and the highest number of effective samples per 725 minute compared to all other methods. We expect the computational benefit of using 726 MLGLUE to increase as the computational cost of a single model call increases, which 727 has been previously identified for multilevel Monte Carlo and multilevel inversion (Cliffe 728 et al., 2011; Giles, 2015; Dodwell et al., 2019; Lykkegaard et al., 2023). 729

- 730 Our results demonstrate that:
- By considering a hierarchy of models with decreasing (spatial) resolution, MLGLUE
   can substantially reduce the computational cost of statistical inversion for differ ent kinds of hydrological models.
- MLGLUE is most effective for differential-equation-based models, such as they are
   often encountered in the hydrological sciences; notions of grid or time-step refine ment and coarsening are well understood in such cases and MLGLUE may be di rectly applied.

-31-
- Although rigorous criteria on the choice of the number of levels and the coarsen-738 ing factor do not exist, for MLGLUE there should be as few levels as possible with 739 differences in resolution being as large as possible. Those aspects are restricted 740 by the quality of the coarsest-level model being sufficiently high, the required res-741 olution on the highest level, and the requirement for sufficiently high correlation 742 between subsequent levels. A non-geometric construction of the hierarchy promises 743 to be an alternative, however being associated with elevated computational cost 744 to optimize the hierarchy (see section 2.4.2). 745
- Statistical analysis of model outputs on all levels can potentially reveal various
  aspects such as the impact of model resolution on quantities of interest or the pos sibility for model simplification. This offers an interesting direction for future re search with multilevel methods.
- 750

## Open Research Section

Relevant resources needed to reproduce the results as well as figures are openly avail able and can be found under the DOI 10.5281/zenodo.10963983 (Rudolph et al., 2024).
 The MLGLUE algorithm is available as a Python package under https://github.com/
 iGW-TU-Dresden/MLGLUE.

## 755 Acknowledgments

Funding for Thorsten Wagener has been provided by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. The authors thank Robert Scheichl, Aretha Teckentrup, and Anastasia Istratuca for fruitful discussions and inspiration. We gratefully acknowledge the support by the Open Access Publishing Fund of TU Dresden. Open Access funding is enabled and organized by the DEAL project. The authors have no conflict of interest with respect to the results of this study.

## 763 **References**

 Allgeier, J. T. (2022). Analytical and Stochastic Numerical Methods for the Simulation of Subsurface Flow in Floodplains (Doctoral dissertation, Eberhard Karls Universität Tübingen, Tübingen). Retrieved from http://dx.doi.org/ 10.15496/publikation-76913

768	Anderson, M. P., Woessner, W. W., & Hunt, R. J. (2015). Applied groundwater
769	modeling: simulation of flow and advective transport (Second edition ed.). Lon-
770	don ; San Diego, CA: Academic Press. (OCLC: ocn921253555)
771	Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015, August).
772	A review of surrogate models and their application to groundwater modeling:
773	SURROGATES OF GROUNDWATER MODELS. Water Resources Re-
774	search, 51(8), 5957-5973. Retrieved 2022-07-13, from http://doi.wiley.com/
775	10.1002/2015WR016967 doi: 10.1002/2015WR016967
776	Beck, J. L., & Zuev, K. M. (2015). Rare-Event Simulation. In R. Ghanem, D. Hig-
777	don, & H. Owhadi (Eds.), Handbook of Uncertainty Quantification (pp. 1–
778	26). Cham: Springer International Publishing. Retrieved 2024-02-21, from
779	https://link.springer.com/10.1007/978-3-319-11259-6_24-1 doi:
780	10.1007/978-3-319-11259-6_24-1
781	Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological mod-
782	elling. Advances in Water Resources, $16(1)$ , $41-51$ . Retrieved 2023-05-23, from
783	https://linkinghub.elsevier.com/retrieve/pii/030917089390028E doi:
784	10.1016/0309-1708(93)90028-E
785	Beven, K. (2006, March). A manifesto for the equifinality thesis. <i>Journal</i>
786	of Hydrology, 320(1-2), 18-36. Retrieved 2022-02-25, from https://
787	linkinghub.elsevier.com/retrieve/pii/S002216940500332X doi:
788	10.1016/j.jhydrol.2005.07.007
789	Beven, K. (2016, July). Facets of uncertainty: epistemic uncertainty, non-
790	stationarity, likelihood, hypothesis testing, and communication. $Hydro-$
791	logical Sciences Journal, 61(9), 1652–1665. Retrieved 2023-10-04, from
792	http://www.tandfonline.com/doi/full/10.1080/02626667.2015.1031761
793	doi: $10.1080/02626667.2015.1031761$
794	Beven, K., & Binley, A. (1992, July). The future of distributed models: Model
795	calibration and uncertainty prediction. $Hydrological Processes, 6(3), 279-$
796	298. Retrieved 2022-05-25, from https://onlinelibrary.wiley.com/doi/
797	10.1002/hyp.3360060305 doi: 10.1002/hyp.3360060305
798	Beven, K., & Binley, A. (2014, November). GLUE: 20 years on. <i>Hy</i> -
799	$drological \ Processes, \ 28(24), \ 5897-5918. \qquad \qquad \text{Retrieved } 2023-01-09, \ \text{from}$
800	https://onlinelibrary.wiley.com/doi/10.1002/hyp.10082 doi:

10.1002/hyp.10082

801

802	Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty esti-
803	mation in mechanistic modelling of complex environmental systems using the
804	GLUE methodology. Journal of Hydrology, 19.
805	Binley, A., Beven, K., & Elgy, J. (1989, June). A physically based model of hetero-
806	geneous hillslopes: 2. Effective hydraulic conductivities. Water Resources Re-
807	search, 25(6), 1227-1233. Retrieved 2023-05-26, from http://doi.wiley.com/
808	10.1029/WR025i006p01227 doi: 10.1029/WR025i006p01227
809	Blasone, RS., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., &
810	Zyvoloski, G. A. (2008, April). Generalized likelihood uncertainty esti-
811	mation (GLUE) using adaptive Markov Chain Monte Carlo sampling. $Ad$ -
812	vances in Water Resources, 31(4), 630–648. Retrieved 2022-06-16, from
813	https://linkinghub.elsevier.com/retrieve/pii/S0309170807001856
814	doi: 10.1016/j.advwatres.2007.12.003
815	Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A.,
816	Zhang, Y. (2019, July). Twenty-three unsolved problems in hydrology (UPH)
817	- a community perspective. Hydrological Sciences Journal, 64(10), 1141-
818	1158. Retrieved 2023-10-04, from https://www.tandfonline.com/doi/full/
819	10.1080/02626667.2019.1620507 doi: $10.1080/02626667.2019.1620507$
820	Boyle, D. P. (2001). Multicriteria calibration of hydrologic models (Doctoral disserta-
821	tion, University of Arizona.) Retrieved from http://hdl.handle.net/10150/
822	290657
823	Braun, M. (1993). Differential Equations and Their Applications: An Introduction to
824	Applied Mathematics (Vol. 11). New York, NY: Springer New York. Retrieved
825	2024-02-21, from http://link.springer.com/10.1007/978-1-4612-4360-1
826	doi: 10.1007/978-1-4612-4360-1
827	Brunetti, G., Šimunek, J., Wöhling, T., & Stumpp, C. (2023, September). An
828	in-depth analysis of Markov-Chain Monte Carlo ensemble samplers for in-
829	verse vadose zone modeling. Journal of Hydrology, 624, 129822. Retrieved
830	2023-10-04, from https://linkinghub.elsevier.com/retrieve/pii/
831	S0022169423007643 doi: 10.1016/j.jhydrol.2023.129822
832	Burrows, W., & Doherty, J. E. (2015, July). Efficient Calibration/Uncertainty
833	Analysis Using Paired Complex/Surrogate Models. Groundwater, $53(4)$ , $531-$

-34-

834	541. Retrieved 2022-05-16, from https://onlinelibrary.wiley.com/doi/
835	10.1111/gwat.12257 doi: 10.1111/gwat.12257
836	Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., & Slooten, L. J. (2005, March).
837	Inverse problem in hydrogeology. $Hydrogeology Journal, 13(1), 206-$
838	222. Retrieved 2023-05-22, from http://link.springer.com/10.1007/
839	s10040-004-0404-7 doi: 10.1007/s10040-004-0404-7
840	Christen, J. A., & Fox, C. (2005). Markov Chain Monte Carlo Using an Approxima-
841	tion. Journal of Computational and Graphical Statistics, $14(4)$ . doi: 10.1198/
842	106186005 X76983
843	Cliffe, K. A., Giles, M. B., Scheichl, R., & Teckentrup, A. L. (2011, January). Mul-
844	tilevel Monte Carlo methods and applications to elliptic PDEs with random
845	coefficients. Computing and Visualization in Science, $14(1)$ , 3–15. Retrieved
846	2022-11-23, from http://link.springer.com/10.1007/s00791-011-0160-x
847	doi: $10.1007/s00791-011-0160-x$
848	Colecchio, I., Boschan, A., Otero, A. D., & Noetinger, B. (2020, June). On
849	the multiscale characterization of effective hydraulic conductivity in ran-
850	dom heterogeneous media: A historical survey and some new perspectives.
851	Advances in Water Resources, 140, 103594. Retrieved 2023-05-26, from
852	https://linkinghub.elsevier.com/retrieve/pii/S0309170819310681
853	doi: 10.1016/j.advwatres.2020.103594
854	Cui, T., Detommaso, G., & Scheichl, R. (2024, March). Multilevel dimension-
855	independent likelihood-informed MCMC for large-scale inverse prob-
856	lems. Inverse Problems, $40(3)$ , 035005. Retrieved 2024-02-28, from
857	https://iopscience.iop.org/article/10.1088/1361-6420/ad1e2c doi:
858	10.1088/1361-6420/ad1e2c
859	Dodwell, T. J., Ketelsen, C., Scheichl, R., & Teckentrup, A. L. (2019). Multilevel
860	Markov Chain Monte Carlo. SIAM / ASA Journal of Uncertainty Quantifica-
861	tion.
862	Doherty, J. E. (2003, March). Ground Water Model Calibration Using Pilot Points
863	and Regularization. Groundwater, $41(2)$ , 170–177. Retrieved 2022-07-11, from
864	https://doi.org/10.1111/j.1745-6584.2003.tb02580.x (Publisher: John
865	Wiley & Sons, Ltd) doi: 10.1111/j.1745-6584.2003.tb02580.x
866	Doherty, J. E. (2015). Calibration and Uncertainty Analysis for Complex Environ-

-35-

867	mental Models. Brisbane: Watermark Numerical Computing.
868	Doherty, J. E., & Christensen, S. (2011, December). Use of paired simple and
869	complex models to reduce predictive bias and quantify uncertainty. Water $Re$ -
870	sources Research, 47(12). Retrieved 2022-02-28, from http://doi.wiley.com/
871	10.1029/2011WR010763 doi: 10.1029/2011WR010763
872	Erdal, D., & Cirpka, O. A. (2020, September). Technical Note: Improved sampling
873	of behavioral subsurface flow model parameters using active subspaces. $Hydrol$ -
874	ogy and Earth System Sciences, 24(9), 4567–4574. Retrieved 2023-10-04, from
875	$\tt https://hess.copernicus.org/articles/24/4567/2020/ ~doi:~10.5194/hess$
876	-24-4567-2020
877	Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., & Stephenson, J. (2009,
878	April). Markov chain Monte Carlo (MCMC) sampling methods to deter-
879	mine optimal models, model resolution and model choice for Earth Science
880	problems. Marine and Petroleum Geology, 26(4), 525–535. Retrieved
881	2023-10-04, from https://linkinghub.elsevier.com/retrieve/pii/
882	S0264817209000075 doi: 10.1016/j.marpetgeo.2009.01.003
883	Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2020,
884	October). Data for "Rainfall-Runoff Prediction at Multiple Timescales with
885	a Single Long Short-Term Memory Network". Zenodo. Retrieved from
886	https://doi.org/10.5281/zenodo.4072701 doi: 10.5281/zenodo.4072701
887	Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021,
888	April). Rainfall–runoff prediction at multiple timescales with a single Long
889	Short-Term Memory network. Hydrology and Earth System Sciences, 25(4),
890	2045-2062. Retrieved 2024-02-20, from https://hess.copernicus.org/
891	articles/25/2045/2021/ doi: 10.5194/hess-25-2045-2021
892	Gelman, A., & Rubin, D. B. (1992, November). Inference from Iterative Simulation
893	Using Multiple Sequences. Statistical Science, $7(4)$ . Retrieved 2022-11-08, from
894	https://projecteuclid.org/journals/statistical-science/volume-7/
895	issue-4/Inference-from-Iterative-Simulation-Using-Multiple
896	-Sequences/10.1214/ss/1177011136.full doi: 10.1214/ss/1177011136
897	Geyer, C. J. (1992, November). Practical Markov Chain Monte Carlo. Statistical
898	Science, 7(4). Retrieved 2024-04-02, from https://projecteuclid.org/
899	journals/statistical-science/volume-7/issue-4/Practical-Markov

-36-

900	-Chain-Monte-Carlo/10.1214/ss/1177011137.full doi: 10.1214/ss/
901	1177011137
902	Geyer, C. J. (2011, May). Introduction to Markov Chain Monte Carlo. In Handbook
903	of Markov Chain Monte Carlo (1st ed., pp. 3–48). New York: Chapman and
904	Hall/CRC. Retrieved 2024-04-02, from https://www.taylorfrancis.com/
905	books/9780429138508/chapters/10.1201/b10905-2 doi: $10.1201/b10905$ -2
906	Giles, M. B. (2008, June). Multilevel Monte Carlo Path Simulation. Opera-
907	tions Research, 56(3), 607-617. Retrieved 2023-05-02, from https://doi.org/
908	10.1287/opre.1070.0496 (Publisher: INFORMS) doi: 10.1287/opre.1070
909	.0496
910	Giles, M. B. (2015, May). Multilevel Monte Carlo methods. Acta Numerica,
911	24, 259-328. Retrieved 2022-11-24, from https://www.cambridge.org/core/
912	product/identifier/S096249291500001X/type/journal_article doi: 10
913	.1017/S096249291500001X
914	Gosses, M., & Wöhling, T. (2019, March). Simplification error analysis for ground-
915	water predictions with reduced order models. Advances in Water Resources,
916	125, 41-56. Retrieved 2022-02-17, from https://linkinghub.elsevier.com/
917	retrieve/pii/S030917081830575X doi: 10.1016/j.advwatres.2019.01.006
918	Gosses, M., & Wöhling, T. (2021, September). Robust Data Worth Analysis with
919	Surrogate Models. Groundwater, 59(5), 728–744. Retrieved 2022-05-12,
920	from https://onlinelibrary.wiley.com/doi/10.1111/gwat.13098 doi:
921	10.1111/gwat.13098
922	Harbaugh, A. W. (2005). MODFLOW-2005, The U.S. Geological Survey Modular
923	Ground-Water Model—the Ground-Water Flow Process (Tech. Rep. No. U.S.
924	Geological Survey Techniques and Methods 6–A16). Reston, VA: USGS.
925	Heinrich, S. (2001). Multilevel Monte Carlo Methods. In S. Margenov,
926	J. Waśniewski, & P. Yalamov (Eds.), Large-Scale Scientific Computing (pp.
927	58–67). Berlin, Heidelberg: Springer Berlin Heidelberg.
928	Herman, J. D., Reed, P. M., & Wagener, T. (2013, March). Time-varying sensitivity
929	analysis clarifies the effects of watershed model formulation on model behav-
930	ior. Water Resources Research, $49(3)$ , 1400–1414. Retrieved 2023-12-16, from
931	https://doi.org/10.1002/wrcr.20124 (Publisher: John Wiley & Sons,
932	Ltd) doi: 10.1002/wrcr.20124

933	Herrera, P. A., Marazuela, M. A., & Hofmann, T. (2022, January). Parameter
934	estimation and uncertainty analysis in hydrological modeling. $WIREs Wa$ -
935	ter, $9(1)$ , e1569. Retrieved 2023-10-04, from https://wires.onlinelibrary
936	.wiley.com/doi/10.1002/wat2.1569 doi: 10.1002/wat2.1569
937	Kavetski, D., Kuczera, G., & Franks, S. W. (2006, March). Bayesian analysis
938	of input uncertainty in hydrological modeling: 1. Theory: INPUT UNCER-
939	TAINTY IN HYDROLOGY, 1. Water Resources Research, $42(3)$ . Retrieved
940	2022-11-30, from http://doi.wiley.com/10.1029/2005WR004368 doi:
941	10.1029/2005 WR004368
942	Kennedy, M. C., & O'Hagan, A. (2001, September). Bayesian Calibration
943	of Computer Models. Journal of the Royal Statistical Society Series B:
944	Statistical Methodology, 63(3), 425–464. Retrieved 2024-03-01, from
945	https://academic.oup.com/jrsssb/article/63/3/425/7083367 doi:
946	10.1111/1467-9868.00294
947	Kitanidis, P. K., & Vomvoris, E. G. (1983, June). A geostatistical approach to the
948	inverse problem in groundwater modeling (steady state) and one-dimensional
949	simulations. Water Resources Research, 19(3), 677–690. Retrieved 2023-
950	05-25, from http://doi.wiley.com/10.1029/WR019i003p00677 doi:
951	10.1029/WR019i003p00677
952	Knoben, W. J. M., Freer, J., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2019,
953	June). Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT)
954	v1.2: an open-source, extendable framework providing implementations of
955	46 conceptual hydrologic models as continuous state-space formulations.
956	Geoscientific Model Development, 12(6), 2463–2480. Retrieved 2024-02-
957	21, from https://gmd.copernicus.org/articles/12/2463/2019/ doi:
958	10.5194/gmd-12-2463-2019
959	Kuffour, B. N. O., Engdahl, N. B., Woodward, C. S., Condon, L. E., Kollet, S.,
960	& Maxwell, R. M. (2020, March). Simulating coupled surface–subsurface
961	flows with ParFlow v3.5.0: capabilities, applications, and ongoing develop- $% \left( {{\left[ {{{\rm{A}}} \right]}_{{\rm{A}}}}_{{\rm{A}}}} \right)$
962	ment of an open-source, massively parallel, integrated hydrologic model.
963	Geoscientific Model Development, 13(3), 1373–1397. Retrieved 2023-10-
964	04, from https://gmd.copernicus.org/articles/13/1373/2020/ doi:
965	10.5194/gmd-13-1373-2020

966	Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified li-
967	brary for exploratory analysis of Bayesian models in Python. Journal of Open
968	Source Software, 4(33), 1143. Retrieved from https://doi.org/10.21105/
969	$\tt joss.01143$ (Publisher: The Open Journal) doi: 10.21105/joss.01143
970	Kumar, R., Samaniego, L., & Attinger, S. (2013, January). Implications of dis-
971	tributed hydrologic model parameterization on water fluxes at multiple
972	scales and locations: DISTRIBUTED HYDROLOGIC MODEL PARAM-
973	ETERIZATIONS. Water Resources Research, $49(1)$ , 360–379. Retrieved
974	2023-05-23, from http://doi.wiley.com/10.1029/2012WR012195 doi:
975	10.1029/2012WR012195
976	Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D., & Jacques, D. (2013, May). Ef-
977	ficient posterior exploration of a high-dimensional groundwater model from
978	two-stage Markov chain Monte Carlo simulation and polynomial chaos ex-
979	pansion: Speeding up MCMC Simulation of a Groundwater Model. Wa-
980	ter Resources Research, $49(5)$ , 2664–2682. Retrieved 2023-09-15, from
981	http://doi.wiley.com/10.1002/wrcr.20226 doi: 10.1002/wrcr.20226
982	Laloy, E., & Vrugt, J. A. (2012, January). High-dimensional posterior exploration of
983	hydrologic models using multiple-try $\mathrm{DREAM}(\mathrm{ZS})$ and high-performance com-
984	puting: EFFICIENT MCMC FOR HIGH-DIMENSIONAL PROBLEMS.
985	Water Resources Research, 48(1). Retrieved 2022-07-25, from http://
986	doi.wiley.com/10.1029/2011WR010608 doi: 10.1029/2011WR010608
987	Leopoldina, G. N. A. o. S. (Ed.). (2022). Earth system science: Dis-
988	covery, diagnosis, and solutions in times of global change : Re-
989	port on tomorrow's science. Retrieved from https://levana
990	.leopoldina.org/receive/leopoldina_mods_00591 (ISBN: 978-
991	3-8047-4256-7 Url: https://doi.org/10.26164/leopoldina_03_00590 Url:
992	$https://doi.org/10.26164/leopoldina_03_00591) \qquad doi: 10.26164/leopoldina_03_00591)$
993	_00591
994	Linde, N., Ginsbourger, D., Irving, J., Nobile, F., & Doucet, A. (2017, Decem-
995	ber). On uncertainty quantification in hydrogeology and hydrogeophysics.
996	Advances in Water Resources, 110, 166–181. Retrieved 2023-05-16, from
997	https://linkinghub.elsevier.com/retrieve/pii/S0309170817304608
998	doi: 10.1016/j.advwatres.2017.10.014

999	Liu, J. S. (Ed.). (2008). Monte Carlo Strategies in Scientific Computing. New York:
1000	Springer.
1001	Liu, Y., Li, J., Sun, S., & Yu, B. (2019, October). Advances in Gaussian ran-
1002	dom field generation: a review. Computational Geosciences, $23(5)$ , 1011–
1003	1047. Retrieved 2022-08-31, from http://link.springer.com/10.1007/
1004	<b>s10596-019-09867-у</b> doi: 10.1007/s10596-019-09867-у
1005	Lykkegaard, M. B. (2022, November). tinyDA. Retrieved from https://pypi.org/
1006	project/tinyda/
1007	Lykkegaard, M. B., & Dodwell, T. J. (2022, June). Where to drill next? A
1008	dual-weighted approach to adaptive optimal design of groundwater surveys.
1009	Advances in Water Resources, 164, 104219. Retrieved 2023-01-16, from
1010	https://linkinghub.elsevier.com/retrieve/pii/S0309170822000914
1011	doi: 10.1016/j.advwatres.2022.104219
1012	Lykkegaard, M. B., Dodwell, T. J., Fox, C., Mingas, G., & Scheichl, R. (2023,
1013	March). Multilevel Delayed Acceptance MCMC. SIAM/ASA Journal on Un-
1014	certainty Quantification, 11(1), 1–30. Retrieved 2023-03-16, from https://
1015	epubs.siam.org/doi/10.1137/22M1476770 doi: 10.1137/22M1476770
1016	Mai, J. (2023, May). Ten strategies towards successful calibration of environmen-
1017	tal models. Journal of Hydrology, 620, 129414. Retrieved 2023-06-01, from
1018	https://linkinghub.elsevier.com/retrieve/pii/S0022169423003566
1019	doi: 10.1016/j.jhydrol.2023.129414
1020	Mirzaei, M., Huang, Y. F., El-Shafie, A., & Shatirah, A. (2015, July). Applica-
1021	tion of the generalized likelihood uncertainty estimation (GLUE) approach
1022	for assessing uncertainty in hydrological models: a review. Stochastic En-
1023	vironmental Research and Risk Assessment, 29(5), 1265–1273. Retrieved
1024	2023-05-03, from http://link.springer.com/10.1007/s00477-014-1000-6
1025	doi: 10.1007/s00477-014-1000-6
1026	Montanari, A. (2007, March). What do we mean by 'uncertainty'? The need
1027	for a consistent wording about uncertainty assessment in hydrology. $Hy$ -
1028	drological Processes, 21(6), 841-845. Retrieved 2022-11-30, from https://
1029	onlinelibrary.wiley.com/doi/10.1002/hyp.6623 doi: 10.1002/hyp.6623
1030	Moore, C., & Doherty, J. E. (2006, April). The cost of uniqueness in groundwa-
1031	ter model calibration. Advances in Water Resources, 29(4), 605–623. Re-

1032	trieved 2022-07-13, from https://linkinghub.elsevier.com/retrieve/pii/
1033	<b>S0309170805001752</b> doi: 10.1016/j.advwatres.2005.07.003
1034	Moore, C., Wöhling, T., & Doherty, J. (2010, August). Efficient regularization and
1035	uncertainty analysis using a global optimization methodology: REGULAR-
1036	IZATION, UNCERTAINTY AND GLOBAL OPTIMIZATION. Water Re-
1037	sources Research, 46(8). Retrieved 2023-10-16, from http://doi.wiley.com/
1038	10.1029/2009WR008627 doi: 10.1029/2009WR008627
1039	Nash, J., & Sutcliffe, J. (1970, April). River flow forecasting through conceptual
1040	models part I — A discussion of principles. Journal of Hydrology, $10(3)$ , 282–
1041	290. Retrieved from https://www.sciencedirect.com/science/article/
1042	pii/0022169470902556 doi: 10.1016/0022-1694(70)90255-6
1043	Niswonger, R. G., Panday, S., & Ibaraki, M. (2011). MODFLOW-NWT, A Newton
1044	formulation for MODFLOW-2005 (Tech. Rep. No. U.S. Geological Survey
1045	Techniques and Methods 6–A37).
1046	Nott, D. J., Marshall, L., & Brown, J. (2012, December). Generalized likelihood un-
1047	certainty estimation (GLUE) and approximate Bayesian computation: What's
1048	the connection?: TECHNICAL NOTE. Water Resources Research, $48(12)$ .
1049	Retrieved 2023-05-03, from http://doi.wiley.com/10.1029/2011WR011128
1050	doi: 10.1029/2011WR011128
1051	Page, T., Smith, P., Beven, K., Pianosi, F., Sarrazin, F., Almeida, S., Wagener,
1052	T. (2023, July). Technical note: The CREDIBLE Uncertainty Estimation
1053	(CURE) toolbox: facilitating the communication of epistemic uncertainty.
1054	Hydrology and Earth System Sciences, 27(13), 2523–2534. Retrieved 2023-
1055	10-04, from https://hess.copernicus.org/articles/27/2523/2023/ doi:
1056	10.5194/hess-27-2523-2023
1057	Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,
1058	Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of
1059	Machine Learning Research, 12, 2825–2830.
1060	Plumlee, M. (2017, July). Bayesian Calibration of Inexact Computer Models. Jour-
1061	nal of the American Statistical Association, 112(519), 1274–1285. Retrieved
1062	from https://doi.org/10.1080/01621459.2016.1211016 (Publisher: Taylor
1063	& Francis) doi: 10.1080/01621459.2016.1211016

<sup>1064</sup> Pokhrel, P., Gupta, H. V., & Wagener, T. (2008, December). A spatial regulariza-

1065	tion approach to parameter estimation for a distributed watershed model. $Wa$ -
1066	ter Resources Research, $44(12)$ , 2007WR006615. Retrieved 2023-10-16, from
1067	https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2007WR006615
1068	doi: $10.1029/2007WR006615$
1069	Reinecke, R., Wachholz, A., Mehl, S., Foglia, L., Niemann, C., & Döll, P.
1070	(2020, May). Importance of Spatial Resolution in Global Groundwater
1071	Modeling. $Groundwater, 58(3), 363-376.$ Retrieved 2022-06-08, from
1072	https://onlinelibrary.wiley.com/doi/10.1111/gwat.12996 doi:
1073	10.1111/gwat.12996
1074	Rudolph, M. G., Wöhling, T., Wagener, T., & Hartmann, A. (2024, April). Extend-
1075	$ing \ GLUE \ with \ multilevel \ methods \ to \ accelerate \ statistical \ inversion \ of \ hydro-$
1076	logical models - code and data. Zenodo. Retrieved from https://doi.org/
1077	10.5281/zenodo.10963983 doi: 10.5281/zenodo.10963983
1078	Sadegh, M., & Vrugt, J. A. (2013, December). Bridging the gap between GLUE
1079	and formal statistical approaches: approximate Bayesian computation. $Hydrol$
1080	ogy and Earth System Sciences, 17(12), 4831–4850. Retrieved 2023-05-03, from
1081	$\verb+https://hess.copernicus.org/articles/17/4831/2013/ doi: 10.5194/hess.copernicus.org/articles/17/4831/2013/ doi: 10.5194/hess.copernicus.org/articles/18/ doi: 18/ doi: 18/$
1082	-17-4831-2013
1083	Samaniego, L., Kumar, R., & Attinger, S. (2010, May). Multiscale parameter region-
1084	alization of a grid-based hydrologic model at the mesoscale: MULTISCALE
1085	PARAMETER REGIONALIZATION. Water Resources Research, $46(5)$ . Re-
1086	trieved 2023-05-23, from http://doi.wiley.com/10.1029/2008WR007327 doi:
1087	10.1029/2008WR007327
1088	Savage, J. T. S., Pianosi, F., Bates, P., Freer, J., & Wagener, T. (2016, Novem-
1089	ber). Quantifying the importance of spatial resolution and other fac-
1090	tors through global sensitivity analysis of a flood inundation model. $Wa$ -
1091	ter Resources Research, $52(11)$ , 9146–9163. Retrieved 2023-10-04, from
1092	https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2015WR018198
1093	doi: $10.1002/2015$ WR018198
1094	Schoups, G., & Vrugt, J. A. (2010, October). A formal likelihood function for
1095	parameter and predictive inference of hydrologic models with correlated, het-
1096	eroscedastic, and non-Gaussian errors. $Water Resources Research, 46(10),$
1097	2009WR008933. Retrieved 2023-01-09, from https://onlinelibrary.wiley

-42-

1098	.com/doi/abs/10.1029/2009WR008933 doi: $10.1029/2009WR008933$
1099	Team, R. (2022). Ray. Retrieved from https://pypi.org/project/ray/2.2.0/
1100	Tonkin, M. J., & Doherty, J. E. (2005, October). A hybrid regularized inversion
1101	methodology for highly parameterized environmental models: HYBRID REG-
1102	ULARIZATION METHODOLOGY. Water Resources Research, 41(10).
1103	Retrieved 2022-05-16, from http://doi.wiley.com/10.1029/2005WR003995
1104	doi: $10.1029/2005$ WR003995
1105	Trotter, L., & Knoben, W. J. M. (2022, April). MARRMoT v2.1. Zen-
1106	odo. Retrieved from https://doi.org/10.5281/zenodo.6484372 doi:
1107	10.5281/zenodo. $6484372$
1108	Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., & Peel, M. C. (2022, Au-
1109	gust). Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT)
1110	v2.1: an object-oriented implementation of $47$ established hydrological models
1111	for improved speed and readability. $Geoscientific Model Development, 15(16),$
1112	6359-6369. Retrieved 2024-02-21, from https://gmd.copernicus.org/
1113	articles/15/6359/2022/ doi: 10.5194/gmd-15-6359-2022
1114	Vidal-Codina, F., Nguyen, N., Giles, M., & Peraire, J. (2015, September). A model
1115	and variance reduction method for computing statistical outputs of stochastic
1116	elliptic partial differential equations. Journal of Computational Physics, 297,
1117	700-720. Retrieved 2024-02-29, from https://linkinghub.elsevier.com/
1118	retrieve/pii/S0021999115003757 doi: 10.1016/j.jcp.2015.05.041
1119	von Gunten, D., Wöhling, T., Haslauer, C., Merchán, D., Causapé, J., & Cirpka,
1120	O. A. (2014, November). Efficient calibration of a distributed pde -based hy-
1121	drological model using grid coarsening. Journal of Hydrology, 519, 3290–3304.
1122	Retrieved 2023-03-03, from https://linkinghub.elsevier.com/retrieve/
1123	pii/S0022169414008191 doi: 10.1016/j.jhydrol.2014.10.025
1124	Vrugt, J. A. (2016, January). Markov chain Monte Carlo simulation using the
1125	DREAM software package: Theory, concepts, and MATLAB implementation.
1126	Environmental Modelling & Software, 75, 273–316. Retrieved 2022-07-19, from
1127	https://linkinghub.elsevier.com/retrieve/pii/S1364815215300396
1128	doi: 10.1016/j.envsoft.2015.08.013
1129	Vrugt, J. A., & Beven, K. J. (2018, April). Embracing equifinality with effi-
1130	ciency: Limits of Acceptability sampling using the DREAM(LOA) algo-

-43-

1131	rithm. Journal of Hydrology, 559, 954–971. Retrieved 2023-10-04, from
1132	https://linkinghub.elsevier.com/retrieve/pii/S0022169418301021
1133	doi: 10.1016/j.jhydrol.2018.02.026
1134	Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005,
1135	January). Improved treatment of uncertainty in hydrologic modeling: Com-
1136	bining the strengths of global optimization and data assimilation. Water
1137	Resources Research, 41(1), 2004WR003059. Retrieved 2024-02-14, from
1138	https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2004WR003059
1139	doi: 10.1029/2004WR003059
1140	Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S. (2003, August). A
1141	Shuffled Complex Evolution Metropolis algorithm for optimization and uncer-
1142	tainty assessment of hydrologic model parameters: EFFICIENT METHOD
1143	FOR ESTIMATING PARAMETER UNCERTAINTY. Water Resources Re-
1144	search, 39(8). Retrieved 2023-03-03, from http://doi.wiley.com/10.1029/
1145	2002WR001642 doi: 10.1029/2002WR001642
1146	Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A.
1147	(2008, December). Treatment of input uncertainty in hydrologic model-
1148	ing: Doing hydrology backward with Markov chain Monte Carlo simulation.
1149	Water Resources Research, 44(12). Retrieved 2022-07-25, from http://
1150	doi.wiley.com/10.1029/2007WR006720 doi: 10.1029/2007WR006720
1151	Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., & Robinson, B. A. (2009, Oc-
1152	tober). Equifinality of formal (DREAM) and informal (GLUE) Bayesian
1153	approaches in hydrologic modeling? Stochastic Environmental Re-
1154	search and Risk Assessment, 23(7), 1011–1026. Retrieved 2022-03-07,
1155	from http://link.springer.com/10.1007/s00477-008-0274-y doi:
1156	10.1007/s00477-008-0274-y
1157	Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian,
1158	S. (2001, March). A framework for development and application of hydrolog-
1159	ical models. Hydrology and Earth System Sciences, 5(1), 13–26. Retrieved
1160	2024-01-03, from https://hess.copernicus.org/articles/5/13/2001/ doi:
1161	10.5194/hess-5-13-2001
1162	Wagener, T., & Gupta, H. V. (2005, December). Model identification for hy-

Stochastic Environmental Re-

drological forecasting under uncertainty.

1163

1164	search and Risk Assessment, 19(6), 378–387. Retrieved 2023-05-08,						
1165	from http://link.springer.com/10.1007/s00477-005-0006-5 doi:						
1166	10.1007/s00477-005-0006-5						
1167	White, J. T. (2018, November). A model-independent iterative ensemble smoother						
1168	for efficient history-matching and uncertainty quantification in very high						
1169	dimensions. Environmental Modelling & Software, 109, 191–201. Re-						
1170	trieved from https://www.sciencedirect.com/science/article/pii/						
1171	S1364815218302676 doi: 10.1016/j.envsoft.2018.06.009						
1172	White, J. T., Hunt, R. J., Fienen, M. N., & Doherty, J. E. (2020). Approaches						
1173	to highly parameterized inversion: $PEST++$ Version 5, a software suite						
1174	for parameter estimation, uncertainty analysis, management optimization						
1175	and sensitivity analysis (Report No. 7-C26). Reston, VA. Retrieved from						
1176	http://pubs.er.usgs.gov/publication/tm7C26 doi: $10.3133/tm7C26$						
1177	White, J. T., Knowling, M. J., & Moore, C. R. (2020, September). Consequences						
1178	of Groundwater-Model Vertical Discretization in Risk-Based Decision-						
1179	Making. $Groundwater, 58(5), 695-709.$ Retrieved 2022-07-06, from						
1180	https://doi.org/10.1111/gwat.12957 (Publisher: John Wiley & Sons,						
1181	Ltd) doi: 10.1111/gwat.12957						
1182	Wildemeersch, S., Goderniaux, P., Orban, P., Brouyère, S., & Dassargues, A. (2014,						
1183	March). Assessing the effects of spatial discretization on large-scale flow model						
1184	performance and prediction uncertainty. Journal of Hydrology, 510, 10–25.						
1185	Retrieved 2023-03-17, from https://linkinghub.elsevier.com/retrieve/						
1186	pii/S0022169413009177 doi: 10.1016/j.jhydrol.2013.12.020						
1187	Zhou, H., Gómez-Hernández, J. J., & Li, L. (2014, January). Inverse methods in						
1188	hydrogeology: Evolution and recent trends. Advances in Water Resources,						
1189	63, 22-37. Retrieved 2022-04-22, from https://linkinghub.elsevier.com/						
1190	retrieve/pii/S0309170813002017 doi: 10.1016/j.advwatres.2013.10.014						
1191	Zimmerman, D. A., de Marsily, G., Gotway, C. A., Marietta, M. G., Axness, C. L.,						
1192	Beauheim, R. L., Rubin, Y. (1998, June). A comparison of seven geosta-						
1193	tistically based inverse approaches to estimate transmissivities for modeling						
1194	advective transport by groundwater flow. Water Resources Research, $34(6)$ ,						
1195	1373-1413. Retrieved 2022-05-25, from http://doi.wiley.com/10.1029/						
1196	98WR00003 doi: 10.1029/98WR00003						

-45-

Figure 1.



Figure 2.



Figure 3.



Figure 4.



![](_page_57_Figure_1.jpeg)

![](_page_57_Figure_2.jpeg)

Figure 5.

![](_page_59_Figure_0.jpeg)

Figure 6.

![](_page_61_Figure_0.jpeg)

![](_page_61_Figure_1.jpeg)

![](_page_61_Figure_2.jpeg)

Figure 7.

![](_page_63_Figure_0.jpeg)

![](_page_63_Figure_1.jpeg)

Figure 8.

![](_page_65_Figure_0.jpeg)

Figure 9.

## MLGLUE

![](_page_67_Figure_1.jpeg)

# GLUE

## MLDA

## MCMC

Figure 10.

![](_page_69_Figure_0.jpeg)

	MLGLU GLUE MLDA	JE		MCM True Nois
J.			JI.	
4	5 Obse	6 ervatic	7 on Po	int Nr

![](_page_69_Figure_3.jpeg)

## Supporting Information for "Extending GLUE with Multilevel Methods to Accelerate Statistical Inversion of Hydrological Models"

M. G. Rudolph<sup>1</sup>, T. Wöhling<sup>2,3</sup>, T. Wagener<sup>4</sup>, A. Hartmann<sup>1</sup>

 $^{1}$ Institute of Groundwater Management, Technische Universität Dresden, Dresden, Germany

<sup>2</sup>Chair of Hydrology, Institute of Hydrology and Meteorology, Technische Universität Dresden, Dresden, Germany

 $^{3}\mathrm{Lincoln}$  Agritech, Lincoln, New Zealand

<sup>4</sup>Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany

## Contents of this file

- 1. Text S1 to S4
- 2. Figures S1 to S4

## Introduction

This supporting information provides additional text and figures describing the results shown and discussed in the main article "Extending GLUE with Multilevel Methods to Accelerate Statistical Inversion of Hydrological Models". Texts S1 and S2 provide additional details on the derivation of MLGLUE. Texts S3 and S4 provide more detailed descriptions of results for the example inverse problems with a rainfall-runoff model and a groundwater flow model, respectively. Figures S1 and S2 illustrate results for the rainfall-

April 12, 2024, 2:39pm

runoff modelling example. Figures S3 and S4 illustrate results for the groundwater flow modelling example.

## Text S1 - Derivation of MLGLUE, the Wrong Combination of MLMC and GLUE

Assuming that likelihood thresholds are given on each level prior to sampling, a straightforward approach to combining MLMC and GLUE would be to use an MLMC algorithm (e.g., Giles, 2015) directly. Then, only model simulations would be considered that correspond to likelihoods that are above the level-dependent likelihood threshold. With that, as most MLMC samples come from lower levels, posterior parameter samples would mainly be comprised of samples from the posterior distribution corresponding to the coarser-level models. We aim, however, at generating samples that come from the posterior distribution on the finest level. This combination is therefore not purposeful. Otherwise we could directly use the model on level  $\ell = 0$  to perform statistical inversion on a single level, which contradicts the actual aim of the methodology.

Text S2 - Derivation of MLGLUE, Level-Dependent Likelihood Thresholds Using level-dependent likelihood thresholds instead of the highest-level threshold for all levels is motivated by the construction of the MLDA algorithm (Lykkegaard et al., 2023) as well as by the original delayed acceptance MCMC algorithm (Christen & Fox, 2005). In MLDA, different target densities are considered on each level because the likelihood function - seen as a (hyper-) surface in the parameter space - depends on the model used on a corresponding level. In the sense of Bayes' theorem, those densities can be considered to be Bayesian posterior densities. This is an intuitive construction; conside evaluations

April 12, 2024, 2:39pm
of the quantity of interest on different levels, made with the same parameter samples,  $\{\mathbf{Q}_{\ell}(\boldsymbol{\theta}^{(i)}), \mathbf{Q}_{\ell+1}(\boldsymbol{\theta}^{(i)})\}_{i=1}^{N}$ , as well as corresponding likelihods  $\{\widetilde{\mathcal{L}}_{\ell}(\boldsymbol{\theta}^{(i)}|\widetilde{\mathbf{Y}}), \widetilde{\mathcal{L}}_{\ell+1}(\boldsymbol{\theta}^{(i)}|\widetilde{\mathbf{Y}})\}_{i=1}^{N}$ :

$$\mathbf{Q}_{\ell}(\boldsymbol{\theta}_i) \neq \mathbf{Q}_{\ell+1}(\boldsymbol{\theta}_i) \tag{1}$$

$$\Rightarrow \widetilde{\mathcal{L}}_{\ell}(\boldsymbol{\theta}^{(j)}|\widetilde{\mathbf{Y}}) \neq \widetilde{\mathcal{L}}_{\ell+1}(\boldsymbol{\theta}^{(j)}|\widetilde{\mathbf{Y}})$$
<sup>(2)</sup>

$$\Rightarrow \widetilde{\mathcal{L}}_{T,\ell} \neq \widetilde{\mathcal{L}}_{T,\ell+1} \tag{3}$$

Therefore, level-dependent likelihood thresholds instead of a single highest-level threshold used on all levels need to be considered to accurately reflect the variations within the hierarchy of models.

## Text S3 - Additional Description of Results, Rainfall-Runoff Modelling

With MLDA, a total number of  $N_{\ell=L} = 2,000$  samples were computed on the highest level using  $n_{chains} = 20$  and a subsampling rate of 5, resulting in a total of  $N_{MLDA} =$  $2,000\cdot5^2\cdot20 = 1,000,000$  samples from the prior distribution. No samples were burnt from the 20 MLDA chains on the highest level, resulting in  $\hat{R} = 1.0$  for all 5 parameters. Out of the 40,020 remaining samples (including randomly initialized samples on the highest level), only 8,204 effective samples could be used (mean effective sample size estimate for the bulk of the posterior). Therefore, a thinning of 5 was applied, resulting in 8,020 effective samples. With MCMC, a total number of  $N_{\ell=L} = 50,000$  samples were computed on the highest level using  $n_{chains} = 20$ , resulting in a total of  $N_{MCMC} = 50,000 \cdot 20 =$ 1,000,000 samples from the prior distribution. No samples are burnt from the 20 MCMC chains, resulting in  $\hat{R} = 1.00$  for all parameters. Out of the 1,000,020 remaining samples (including randomly initialized samples), only 16,353 effective samples could be used

(mean effective sample size estimate for the bulk of the posterior). Therefore, a thinning of 62 was applied, resulting in 16,140 effective samples.

## Text S4 - Additional Description of Results, Groundwater Flow

With MLDA, a total number of  $N_{\ell=L} = 250$  samples were computed on the highest level using  $n_{chains} = 32$ , resulting in a total of  $N_{MLDA} = 250 \cdot 5^3 \cdot 32 = 1,000,000$  samples from the prior distribution. No sample was burnt from the 32 MLDA chains, resulting in a mean Gelman-Rubin statistic of  $\overline{\hat{R}} = 1.02$  ( $\widehat{R}_{min} = 1.01$ ,  $\widehat{R}_{max} = 1.03$ ), averaged over all 51 parameters. Out of the 8,032 remaining samples (including randomly initialized samples on the highest level), only 1,982 effective samples could be used (mean effective sample size estimate for the bulk of the posterior). Therefore, a thinning of 4 was applied, resulting in 2,008 effective samples. With MCMC, a total number of  $N_{\ell=L} = 31,250$ samples were computed on the highest level using  $n_{chains} = 32$  and a subsampling rate of 5, resulting in a total of  $N_{MCMC} = 31,250 \cdot 32 = 1,000,000$  samples from the prior distribution. The initial sample was burnt from the 32 MCMC chains, resulting in a mean Gelman-Rubin statistic of  $\overline{\hat{R}} = 1.02$  ( $\widehat{R}_{min} = 1.01$ ,  $\widehat{R}_{max} = 1.03$ ), averaged over all 51 parameters. Out of the 1,000,000 remaining sample, only 2,080 effective samples can be used (mean effective sample size estimate for the bulk of the posterior). Therefore, a thinning of 481 was applied, resulting in 2,080 effective samples.

## References

Christen, J. A., & Fox, C. (2005). Markov Chain Monte Carlo Using an Approximation. Journal of Computational and Graphical Statistics, 14(4). doi: 10.1198/ 106186005X76983

Giles, M. B. (2015, May). Multilevel Monte Carlo methods. Acta Numerica,

24, 259-328. Retrieved 2022-11-24, from https://www.cambridge.org/core/ product/identifier/S096249291500001X/type/journal\_article doi: 10.1017/ S096249291500001X

Lykkegaard, M. B., Dodwell, T. J., Fox, C., Mingas, G., & Scheichl, R. (2023, March). Multilevel Delayed Acceptance MCMC. SIAM/ASA Journal on Uncertainty Quantification, 11(1), 1–30. Retrieved 2023-03-16, from https://epubs.siam.org/doi/ 10.1137/22M1476770 doi: 10.1137/22M1476770



Figure S1. Level-dependent likelihood thresholds for the rainfall-runoff modelling example, estimated with different numbers of tuning samples for threshold settings corresponding to the top 2% (left) and the top 7% (right)



Figure S2. Relations between levels for the linear regression example, using an informal likelihood





Figure S3. Level-dependent likelihood thresholds for the groundwater flow example, estimated with different numbers of tuning samples for threshold settings corresponding to the top 2 % (left) and the top 7 % (right)



**Figure S4.** Relations between levels for the groundwater flow example, using an informal definition of the likelihood

April 12, 2024, 2:39pm

: