Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Advances in Water Resources 27 (2004) 737–750

Advances in Water Resources

www.elsevier.com/locate/advwatres

A modified Levenberg–Marquardt algorithm for quasi-linear geostatistical inversing

Wolfgang Nowak [*], Olaf A. Cirpka

Institut für Wasserbau, Lehrstuhl für Hydromechanik und Hydrosystemmodellierung, Universität Stuttgart, Pfaffenwaldring 61, D 70569 Stuttgart, Germany

Received 20 March 2003; received in revised form 18 March 2004; accepted 25 March 2004
Available online 18 May 2004

Abstract

The Quasi-Linear Geostatistical Approach is a method of inverse modeling to identify parameter fields, such as the hydraulic conductivity in heterogeneous aquifers, given observations of related quantities like hydraulic heads or arrival times of tracers. Derived in the Bayesian framework, it allows to rigorously quantify the uncertainty of the identified parameter field. Since inverse modeling in subsurface flow is in general a non-linear problem, the Quasi-Linear Geostatistical Approach employs an algorithm for non-linear optimization. Up to presence, this algorithm has been similar to the Gauss–Newton algorithm for least-squares fitting and fails in cases of strong non-linearity. In this study, we present and discuss a new modified Levenberg–Marquardt algorithm for the Quasi-Linear Geostatistical Approach. Compared to the original method, the new algorithm offers increased stability and is more robust, allowing for stronger non-linearity and higher variability of the parameter field to be identified. We demonstrate its efficiency and improved convergence compared to the original version in several test cases. The new algorithm is designed for the general case of an uncertain mean of the parameter field, which includes the cases of completely known and entirely unknown mean as special cases.

Keywords: Geostatistical inverse modeling; Conditioning; Cokriging; Levenberg; Marquardt

1. Introduction

Cokriging is a geostatistically based technique to identify an unknown spatial parameter field given observations of a correlated quantity. Originally developed for mining exploration, cokriging has successfully been used as tool for inverse modeling in other fields such as hydrogeology (e.g. [15]). In geostatistical inverse modeling, we consider the unknown parameters, such as the hydraulic conductivity field of a porous formation, as a random space function which is conditioned on observations of dependent quantities, such as the hydraulic head or the travel time of a solute [33]. The cross-covariance between the random space functions, i.e., the unknown parameters and the dependent state variables, is fully determined by the functional relationship of the parameters and observations and the auto-covariance of the parameter field. In many cases, this functional relationship is a partial differential equation. It may be non-linear with respect to the unknown parameters, so that inverse modeling in hydrogeology is in general a non-linear problem [11,18,29].

As has been shown by Kitanidis [9], both kriging and cokriging are identical to a Bayesian analysis for the estimation of the conditional mean. The rigorous Bayesian context allows an accurate quantification of the parameter uncertainty while imposing a minimum of structural assumptions onto the unknowns. Parameter uncertainty is expressed by the posterior auto-covariance of the parameter field. A disadvantage of traditional cokriging techniques lies in the computational costs involved in handling the auto-covariances and cross-covariances. This has led to the development of alternative geostatistical methods of inversing in which the cross-covariance matrices are not fully determined [7,22,33], or in which only a certain neighborhood around each observation point is considered.

While these methods reduce the computational costs dramatically, they sacrifice the rigor in determining the parameter uncertainty. Exploiting specific properties of

[*] Corresponding author. Fax: +49-711-685-7020.
E-mail address: wolfgang.nowak@iws.uni-stuttgart.de (W. Nowak).

0309-1708/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2004.03.004

the auto-covariance matrix of the unknowns [32] and using periodic embedding and spectral methods for matrix–matrix multiplications involving this matrix [20], however, the computational costs of cokriging can be drastically reduced. These spectral methods make cokriging techniques competitive in terms of computational efficiency while outrivaling the alternative methods in quantifying the parameter uncertainty. They are applicable if the unknown parameter field is statistically second-order stationary or at least intrinsic, and defined on a regular grid. The flow conditions and the parameter fields are non-periodic prior to the embedding. In contrast to spectral methods that put linear perturbation theory into the Fourier framework (e.g. [8]), they do not require the perturbations of the dependent quantities to be second-order stationary or intrinsic. This allows to apply these spectral methods to inverse modeling under non-stationary flow conditions in bounded domains [3]. We provide an assessment of the reduced computational costs of cokriging in Section 2.

In case the forward problem is linear with respect to the parameters, linear cokriging returns the parameter estimate after a single computational step. For non-linear relations, cokriging-like procedures are applied in an iterative manner. Basically, three differing concepts have been published. The first group of methods, such as the Iterative Cokriging-Like Technique [28], defines an approximate linearization of the forward problem once. Then, cokriging is applied repeatedly while allowing both the linearization and the auto-covariance of the parameter field to remain constant. Other methods conceptualize the iterative procedure as a sequence of Bayesian updating steps and update both the linearization and the auto-covariance of the parameter field during their iteration algorithm, like the Successive Linear Estimator (SLE) by Yeh and coworkers [29]. The third group of methods successively linearizes the forward problem about the current estimate while keeping the covariances. Into this group fall the Quasi-Linear Geostatistical Approach (Geostatistical Approach) by Kitanidis [11] and the Maximum a Posteriori (MAP) method by McLaughlin and Townley [18]. The advantages of linearizing about the current estimate are dealt with by Carrera and Glorioso [1].

Among other differences discussed elsewhere [12,14,19], the Quasi-Linear Geostatistical Approach and the MAP method differ as follows. The former defines the solution in a parameterized form based on a rigorous Bayesian analysis. The sole purpose of the iteration procedure is to optimize the subspace in which the parameterization is defined. In each iteration step, the previous trial solution is projected onto the current subspace. Only at the last iteration step, when the optimal subspace has been found, the conditioning is carried out and the conditional covariance is evaluated based on this optimal subspace. The form of the solution used in this method is discussed in depth in Section 3. In contrast to this, the MAP method seeks a solution that is a sum of different parameterizations encountered during the course of iteration, and the conditional covariance is computed in the last step, based on the final parameterization.

Among the above methods, the Successive Linear Estimator and the Quasi-Linear Geostatistical Approach strictly follow the Baysian concept. In this study, we aim at large problems where the number of unknown parameters may easily rise above $n = 10^5$, e.g. $n = 10^6$ for well-resolved 2-D or 3-D applications. Given this problem size, super-fast FFT-based methods to compute auto-covariances and cross-covariances are indispensable. The updated covariances occurring in the SLE, or in sequential kriging and cokriging [27] for the linear case, do not allow to apply these FFT-based methods directly, since they are no more second-order stationary or intrinsic. Cirpka and Nowak [3] discuss how to handle conditional covariance matrices in the FFT framework. The concept of sequential conditioning that is employed in the SLE and in sequential kriging and cokriging saves computational effort which is associated with the number of observations. However, for large numbers of conditioning steps, the computational costs to include conditional covariance matrices in the FFT framework rise dramatically and make sequential methods unfeasable for large numbers of unknowns.

The iteration algorithm underlying the Geostatistical Approach is in some respects formally similar to the Gauss–Newton algorithm [21] for least-squares fitting. Initially, the problem statement of inverse modeling seems to be under-determined because the number of unknown discrete values of the parameter field typically exceeds the number of observations by several orders of magnitude. In order to reduce the number of degrees of freedom, the Geostatistical Approach employs the Bayesian concept and thereby obtains a unique parameterized form of the solution. The resulting problem is well-posed, given that the data structure of the observations obey certain requirements. A more detailed discussion on the well-posedness or ill-posedness of inverse problems and lucid examples are provided elsewhere [24,30]. Due to the parameterization of the solution, the iteration algorithm used in the Geostatistical Approach differs from the standard Gauss–Newton algorithm in certain terms that will be analyzed in this study.

For least-squares fitting problems, the Gauss–Newton algorithm is well-known to be efficient for mildly non-linear problems, but to fail for strongly non-linear problems. The Levenberg–Marquardt algorithm [16,17] is a modification of the Gauss–Newton method that, in a self-adaptive manner, navigates between Gauss–Newton and the method of steepest descent [21]. Combining the robustness of the method of steepest descent with the computational efficiency of the Gauss–Newton method, the Levenberg–Marquardt algorithm has become a highly valued optimization tool for non-linear tasks of least-squares fitting in many engineering fields.

In hydrogeological applications, increased variability of the parameters leads to higher degrees of non-linearity for the inverse problem, thus decreasing the convergence radius and increasing the number of necessary iterations in the Geostatistical Approach. Above a certain extent of non-linearity, this method fails. Observations such as the arrival time of tracers are especially non-linear with respect to hydraulic conductivity since their sensitivity pattern is strongly influenced by the streamline pattern which is distorted or may even oscillate in the iterative procedure. This gives a clear motivation to improve existing optimization algorithms for geostatistical inversing in terms of stability and robustness, reducing the number of iteration steps and increasing the convergence radius for application to strongly non-linear problems.

The basic idea of the Levenberg–Marquardt algorithm has already been applied to geostatistical inverse modeling by other authors. Dietrich and Newsam [4] discussed how increasing the measurement error in the cokriging procedure can help to improve ill-conditioned matrices and suppress artefacts of numerical error in the estimated parameter fields, but induces a loss of information. The Successive Linear Estimator introduced by Yeh et al. [29] uses an adaptively amplified measurement error term for the auto-covariance of measurements and a relaxation for the cross-covariances to stabilize the algorithm.

If one desires to rigorously quantify the parameter uncertainty through the Bayesian concept, the choice is between the Geostatistical Approach and Successive Linear Estimator (or similar methods). The latter has successfully been applied to non-linear problems not only for the saturated zone, but even to the vadose zone [30,31]. However, as mentioned above, the FFT-based methods to compute the covariance matrices involved are not applicable to this method at an acceptable level of computational costs.

In this study, motivated by the success of the Levenberg–Marquardt algorithm in other areas of engineering and in the Successive Linear Estimator, we present a modified Levenberg–Marquardt algorithm for the Quasi-Linear Geostatistical Approach. The original form of the Levenberg–Marquardt algorithm has to be modified to account for the parameterized form of the solution used in the Geostatistical Approach. Further modifications are necessary to avoid violations of the strict Bayesian concept. The Geostatistical Approach consists of two parts: (1) identifying the unknown parameters for a given auto-covariance function of the unknown parameters, and (2) optimizing the so-called structural parameters used to define the auto-covariance. The Levenberg–Marquardt modification presented here is aimed at improving the first part.

This paper is organized as follows: in Section 2, we discuss the derivation of linear cokriging in the Bayesian framework and discuss certain properties which are helpful in dealing with optimization algorithms. In Section 3, we discuss the conventional form of the Geostatistical Approach as introduced by Kitanidis [11], and point out the formal differences between the Gauss–Newton algorithm and the iteration algorithm used in the Geostatistical Approach. In Section 4, we present the modified Levenberg–Marquardt algorithm for the Geostatistical Approach. Section 5 is a performance test in which we apply both the conventional and the new algorithm to a typical non-linear inverse modeling problem taken from Cirpka and Kitanidis [2].

2. Bayesian framework for linear cokriging

2.1. Prior distributions

Consider a random $n \times 1$ multi-Gaussian vector of unknowns $\mathbf{s}$ with expectation $E[\mathbf{s}] = \mathbf{X}\boldsymbol{\beta}$ and covariance $\mathbf{Q}_{ss} : \mathbf{s} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Q}_{ss})$. In hydrogeological applications, $\mathbf{s}$ may be the vector of unknown log-conductivity values in all grid cells, and $\mathbf{Q}_{ss}$ the corresponding covariance matrix sized $n \times n$. $\mathbf{X}$ is an $n \times p$ matrix of known deterministic base functions, and $\boldsymbol{\beta}$ is $p \times 1$ vector of uncertain drift coefficients. The probability density function (pdf) of $\mathbf{s}$ for given $\boldsymbol{\beta}$ is

$$p(\mathbf{s}\,|\,\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{Q}_{ss}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})\right], \tag{1}$$

and the uncertainty of the drift coefficients $\boldsymbol{\beta}$ is quantified by a (multi-)Gaussian distribution with mean $\boldsymbol{\beta}^*$ and covariance $\mathbf{Q}_{\beta\beta}$: $\boldsymbol{\beta} \sim \mathbf{N}(\boldsymbol{\beta}^*, \mathbf{Q}_{\beta\beta})$. From Bayesian analysis, we obtain that the distribution of $\boldsymbol{\beta}$ for given $\mathbf{s}$ is again (multi-)Gaussian with $\boldsymbol{\beta}\,|\,\mathbf{s} \sim \mathbf{N}(\widehat{\boldsymbol{\beta}}, \mathbf{Q}_{\beta\beta\,|\,s})$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + (\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{Q}_{ss}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}_{ss}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^*), \tag{2}$$

$$\mathbf{Q}_{\beta\beta\,|\,s} = (\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{Q}_{ss}^{-1}\mathbf{X})^{-1}. \tag{3}$$

The distribution of $\mathbf{s}$ regardless of the drift coefficients $\boldsymbol{\beta}$ can be obtained by marginalizing $p(\mathbf{s}\,|\,\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, yielding (compare Kitanidis, 1986, for the case of unknown mean [9])

$$p(\mathbf{s}) \propto \exp\left[-\frac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{G}_{ss}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^*)\right], \tag{4}$$

$$\begin{aligned}\mathbf{G}_{ss} &= (\mathbf{Q}_{ss}^{-1} - \mathbf{Q}_{ss}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{Q}_{ss}^{-1}\mathbf{X} + \mathbf{Q}_{\beta\beta}^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}_{ss}^{-1})^{-1} \\ &= \mathbf{Q}_{ss} + \mathbf{X}\mathbf{Q}_{\beta\beta}\mathbf{X}^{\mathrm{T}}.\end{aligned} \tag{5}$$

From Eq. (2) follows after application of some matrix algebra

$$\mathbf{G}_{\mathrm{ss}}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^{*}) = \mathbf{Q}_{\mathrm{ss}}^{-1}(\mathbf{s}-\mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{6}$$

which will be used below for simplifications.

2.2. Observations

Now consider the $m \times 1$ vector of observations $\mathbf{y}$ related to $\mathbf{s}$ via a linear transfer function $\mathbf{f}$

$$\mathbf{y} = \mathbf{f}(\mathbf{s}) + \mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{r}. \tag{7}$$

where $\mathbf{r}$ is the $m \times 1$ vector of observation error with zero mean and $m \times m$ covariance matrix $\mathbf{R}$, and $\mathbf{H}$ is the so-called sensitivity matrix which does not depend on $\mathbf{s}$ in the linear case. The likelihood of the measurements is

$$p(\mathbf{y}\,|\,\mathbf{s}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{s})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{s})\right]. \tag{8}$$

In hydrogeological applications, $\mathbf{y}$ may be a vector of head measurements, $\mathbf{f}(\mathbf{s})$ the modeled head values at the locations of the measurements for a given conductivity field, and $\mathbf{r}$ the errors when measuring hydraulic heads. Error propagation yields the expected value of $\mathbf{y}$ for given $\boldsymbol{\beta}$, the auto-covariance matrix $\mathbf{Q}_{\mathrm{yy}}$ of the observations $\mathbf{y}$, and the cross-covariance matrix $\mathbf{Q}_{\mathrm{sy}}$ between $\mathbf{s}$ and $\mathbf{y}$ [23]

$$\begin{aligned} E[\mathbf{y}\,|\,\widehat{\boldsymbol{\beta}}] &= \mathbf{H}\mathbf{X}\widehat{\boldsymbol{\beta}}, \\ \mathbf{Q}_{\mathrm{yy}} &= \mathbf{H}\mathbf{Q}_{\mathrm{ss}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}, \\ \mathbf{Q}_{\mathrm{sy}} &= \mathbf{Q}_{\mathrm{ss}}\mathbf{H}^{\mathrm{T}}. \end{aligned} \tag{9}$$

We multiply Eq. (8) by Eq. (4) and marginalize to find that $\mathbf{y} \sim \mathbf{N}(\mathbf{H}\mathbf{X}\boldsymbol{\beta}^{*}, \mathbf{G}_{\mathrm{yy}})$, with $\mathbf{G}_{\mathrm{yy}}$ defined by

$$\begin{aligned} \mathbf{G}_{\mathrm{yy}} &= (\mathbf{Q}_{\mathrm{yy}}^{-1} - \mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{H}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{H}\mathbf{X} + \mathbf{Q}_{\beta\beta}^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1})^{-1} \\ &= \mathbf{Q}_{\mathrm{yy}} + \mathbf{H}\mathbf{X}\mathbf{Q}_{\beta\beta}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}. \end{aligned} \tag{10}$$

Again, a useful identity similar to Eq. (6) holds

$$\mathbf{G}_{\mathrm{yy}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{X}\boldsymbol{\beta}^{*}) = \mathbf{Q}_{\mathrm{yy}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{X}\hat{\boldsymbol{\beta}}). \tag{11}$$

Bayesian analysis for $p(\boldsymbol{\beta}\,|\,\mathbf{y})$ yields that $\boldsymbol{\beta} \sim \mathbf{N}(\widehat{\boldsymbol{\beta}}, \mathbf{Q}_{\beta\beta\,|\,\mathrm{y}})$ with

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{*} + (\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{X}\boldsymbol{\beta}^{*}), \tag{12}$$

$$\mathbf{Q}_{\beta\beta\,|\,\mathrm{y}} = (\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{H}\mathbf{X})^{-1}. \tag{13}$$

2.3. Posterior distribution

The cokriging estimate $\hat{\mathbf{s}}$ for the unknowns $\mathbf{s}$ given the observations $\mathbf{y}$ is identical to the mean value of the posterior distribution of $\mathbf{s}$ given $\mathbf{y}$. The posterior distribution is obtained from Bayesian analysis (compare [9])

$$\begin{aligned} p(\mathbf{s}\,|\,\mathbf{y}) \propto \exp\Bigg[&-\frac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^{*})^{\mathrm{T}}\mathbf{G}_{\mathrm{ss}}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^{*}) \\ &-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{s})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{s})\Bigg]. \end{aligned} \tag{14}$$

The posterior mean value is identified as the value of $\mathbf{s}$ that maximizes $p(\mathbf{s}\,|\,\mathbf{y})$, which is identical to

$$\begin{aligned} \hat{\mathbf{s}} = \min\Bigg[&\frac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^{*})^{\mathrm{T}}\mathbf{G}_{\mathrm{ss}}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^{*}) \\ &+\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{s})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{s})\Bigg]. \end{aligned} \tag{15}$$

The function to be minimized is referred to as the objective function $L(\mathbf{s})$. Now, we set the first derivative of $L(\mathbf{s})$ to zero in order to obtain the normal equations. After several rearrangements, we obtain:

$$\hat{\mathbf{s}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Q}_{\mathrm{sy}}\mathbf{Q}_{\mathrm{yy}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{16}$$

showing that $\hat{\mathbf{s}}$ is comprised of the prior mean $\mathbf{X}\widehat{\boldsymbol{\beta}}$ plus an innovation term, which depends on the deviations of the observations from their expectation. This innovation term represents those random fluctuations of $\mathbf{s}$ about its mean value that are relevant for the values of the observations. Defining the $m \times 1$ vector ξ allows to express the posterior mean $\hat{\mathbf{s}}$ in a parameterized form [13]:

$$\xi = \mathbf{Q}_{\mathrm{yy}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{17}$$

$$\hat{\mathbf{s}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Q}_{\mathrm{sy}}\xi. \tag{18}$$

To obtain $\widehat{\boldsymbol{\beta}}$, we insert Eq. (18) into Eq. (2) and simplify

$$\mathbf{Q}_{\beta\beta}^{-1}\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^{*} + \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\xi. \tag{19}$$

Enforcing the constraint (19) while solving Eq. (17) is accomplished by solving the $(m+p) \times (m+p)$ system (compare with the results of Kitanidis [13] for the case of known and unknown mean)

$$\begin{bmatrix} \mathbf{Q}_{\mathrm{yy}} & \mathbf{H}\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} & -\mathbf{Q}_{\beta\beta}^{-1} \end{bmatrix} \begin{bmatrix} \xi \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ -\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^{*} \end{bmatrix}. \tag{20}$$

For the special case of $\mathbf{Q}_{\beta\beta}^{-1} = \mathbf{0}$, Eq. (18) is also known as the function estimate form of ordinary cokriging, Eq. (20) is the system of cokriging equations with the matrix therein being the cokriging matrix, and Eq. (19) is known as the unbiasedness constraint. In standard geostatistical literature, these equations are derived by finding an unbiased linear estimator with minimum estimation variance (Best Linear Unbiased Estimator, BLUE). For the sake of easy reading, we adapt this nomenclature.

The posterior covariance $\mathbf{Q}_{\mathrm{ss}\,|\,\mathrm{y}}$ of $\mathbf{s}$ given $\mathbf{y}$ quantifies the amount of uncertainty remaining after the conditioning procedure. It is defined by the inverse Hessian of the objective function, which is after application of some matrix algebra (again, compare [13], for the case of known and unknown mean)

$$\mathbf{Q}_{\mathrm{ss}\,|\,\mathrm{y}} = \mathbf{G}_{\mathrm{ss}} - \mathbf{G}_{\mathrm{ss}}\mathbf{H}^{\mathrm{T}}\mathbf{G}_{\mathrm{yy}}^{-1}\mathbf{H}\mathbf{G}_{\mathrm{ss}}. \tag{21}$$

2.4. Partitioned form

In this section, we introduce a new form of the BLUE that will prove convenient in later sections. According to the rules for partitioned matrices (see, e.g. [23]), the inverse of the cokriging matrix is

$$\begin{bmatrix} \mathbf{Q}_{\mathrm{yy}} & \mathbf{HX} \\ \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} & -\mathbf{Q}_{\beta\beta}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}_{\mathrm{yy}} & \mathbf{P}_{\mathrm{yb}} \\ \mathbf{P}_{\mathrm{by}} & \mathbf{P}_{\mathrm{bb}} \end{bmatrix} \tag{22}$$

in which the submatrices are

$$\mathbf{P}_{\mathrm{yy}} = \mathbf{Q}_{\mathrm{yy}}^{-1} - \mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{HX}(\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{HX} + \mathbf{Q}_{\beta\beta}^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}, \tag{23}$$

$$\mathbf{P}_{\mathrm{by}} = (\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{HX})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1} = \mathbf{P}_{\mathrm{yb}}^{\mathrm{T}}, \tag{24}$$

$$\mathbf{P}_{\mathrm{bb}} = -(\mathbf{Q}_{\beta\beta}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{Q}_{\mathrm{yy}}^{-1}\mathbf{HX})^{-1}. \tag{25}$$

Obviously, $\mathbf{P}_{\mathrm{yy}} = \mathbf{G}_{\mathrm{yy}}^{-1}$ and $\mathbf{P}_{\mathrm{bb}} = -\mathbf{Q}_{\beta\beta\,|\,\mathrm{y}}$. For the case of unknown mean, this partitioning has been used previously to derive analytical solutions for the conditional covariance of $\hat{\mathbf{s}}$ in matrix notation [13]. We will employ this partitioning to separate ξ from $\widehat{\boldsymbol{\beta}}$ in the cokriging system (Eq. (20)) and the BLUE (Eq. (18))

$$\xi = \mathbf{P}_{\mathrm{yy}}\mathbf{y} - \mathbf{P}_{\mathrm{yb}}\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^{*}, \tag{26}$$

$$\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\mathrm{by}}\mathbf{y} - \mathbf{P}_{\mathrm{bb}}\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^{*}, \tag{27}$$

$$\hat{\mathbf{s}} = (\mathbf{X}\mathbf{P}_{\mathrm{by}} + \mathbf{Q}_{\mathrm{sy}}\mathbf{P}_{\mathrm{yy}})\mathbf{y} - (\mathbf{X}\mathbf{P}_{\mathrm{bb}} + \mathbf{Q}_{\mathrm{sy}}\mathbf{P}_{\mathrm{yb}})\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^{*}. \tag{28}$$

2.5. Simplified objective function

We now present a new computationally most efficient form of the objective function. Consider the a priori criterion of Eq. (15)

$$L_p = \frac{1}{2}(\mathbf{s} - \mathbf{X}\boldsymbol{\beta}^{*})^{\mathrm{T}}\mathbf{G}_{\mathrm{ss}}^{-1}(\mathbf{s} - \mathbf{X}\boldsymbol{\beta}^{*}). \tag{29}$$

Inserting Eqs. (6), (18) and (19) into Eq. (29), L_p becomes after simplification

$$L_p = \frac{1}{2}\xi^{\mathrm{T}}(\mathbf{H}\mathbf{Q}_{\mathrm{ss}}\mathbf{H}^{\mathrm{T}} + \mathbf{HX}\mathbf{Q}_{\beta\beta}\mathbf{X}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})\xi. \tag{30}$$

Likewise, the likelihood term L_m can be simplified to

$$L_m = \frac{1}{2}\xi^{\mathrm{T}}(\mathbf{R})\xi. \tag{31}$$

Thus, the objective function in total is:

$$L = \frac{1}{2}\xi^{\mathrm{T}}\mathbf{G}_{\mathrm{yy}}\xi. \tag{32}$$

2.6. Properties of cokriging

2.6.1. Uniqueness and well-posedness

The matrix of base functions $\mathbf{X}$ and the cross-covariance matrix $\mathbf{Q}_{\mathrm{sy}}$ evidently are used as a geostatistically based parameterization of $\hat{\mathbf{s}}$ in the BLUE (Eq. (18)). $\mathbf{X}$ and $\mathbf{Q}_{\mathrm{sy}}$ span a $(m+p)$-dimensional subspace for $\hat{\mathbf{s}}$, reducing the degrees of freedom from n to $(m+p)$. As a consequence, only $(m+p)$ parameters $\widehat{\boldsymbol{\beta}}$ and ξ have to be solved in the cokriging system of equations (Eq. (20)). Although the task of parameter identification is initially underdetermined in the sense that there are less observations than unknown values $(m < n)$, the geostatistical approach converts the problem into a well-determined problem with $(m+p)$ equations for $(m+p)$ unknown parameters. Hence, provided that the cokriging matrix is not rank deficient, parameter estimation and inverse modeling based on cokriging-like procedures is a well-posed problem and yields unique solutions. More detailed discussions and lucid examples on the well-posedness or ill-posedness of inverse modeling problems for saturated and unsaturated flow are provided elsewhere [24,30].

It is important to understand that the subspace spanned by $\mathbf{X}$ and $\mathbf{Q}_{\mathrm{sy}}$ is not an arbitrary choice, but that it stems from strict Bayesian analysis. Both in the parameterization of the solution and in the definition of the posterior covariance (Eq. (21)), the sensitivity matrix $\mathbf{H}$ plays a central role. In case the sensitivity matrix is inaccurate for any reason, both the cokriging estimate and the estimation variance are subject to error and may significantly differ from the result of a rigorous Bayesian analysis.

2.6.2. Computational costs

The discussion of computational efficiency becomes relevant for large problems, since the computational costs of cokriging typically increase with the square or cube of the problem size. In cases where the transfer function $\mathbf{f}(\mathbf{s})$ is a partial differential equation, the discretization of the unknowns is dictated by stability criteria of the numerical schemes applied for evaluating $\mathbf{f}(\mathbf{s})$. Hence, a number of unknowns in the order of $n = 10^6$ is not unusual. The number of observations m is in most cases much smaller than n, e.g. in the order of 10 or 100, and the number p of base functions for the unknown mean $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is typically one for the case of a constant unknown or uncertain mean. The cokriging equations are rapidly solved as the cokriging matrix is sized $(m+p) \times (m+p)$. The potentially expensive tasks are (1) computing $\mathbf{H}$, (2) computing $\mathbf{Q}_{\mathrm{sy}}$ and $\mathbf{Q}_{\mathrm{yy}}$, and (3) evaluating the value of the objective function. However, means to reduce these costs have been found:

(1) Applying standard numerical differentiation, computing $\mathbf{H}$ takes $(n+1)$ evaluations of $\mathbf{f}(\mathbf{s})$. A highly

efficient way to compute $\mathbf{H}$ is the adjoint-state method, which requires only $(m+1)$ solutions of problems that are formally similar to $\mathbf{f}(\mathbf{s})$ [24,25].

(2) Computing $\mathbf{Q}_{sy}$ and $\mathbf{Q}_{yy}$ by standard methods is strictly impossible for large n, since storage of $\mathbf{Q}_{ss}$ requires memory $\mathcal{O}(n^2)$, e.g. 8.000 GByte for $n = 10^6$, exceeding the capacity of all present-day HDD devices. However, by exploiting certain properties of $\mathbf{Q}_{ss}$, storage costs can be reduced to $\mathcal{O}(n)$ [32], and the CPU time for computing $\mathbf{Q}_{sy}$ and $\mathbf{Q}_{yy}$ can be reduced from $\mathcal{O}(mn^2)$ to $\mathcal{O}(mn\log_2 n)$ via FFT-based methods [6,20,26].

(3) The objective function in the form of Eq. (15) requires storage of $\mathbf{Q}_{ss}$ and the solution of $\mathbf{G}_{ss}^{-1}(\mathbf{s}-\boldsymbol{\beta}^*)$ (Eq. (5)), associated with costs $\mathcal{O}(n^2)$. The simplified form presented in this study (Eq. (32)) reduces these costs to $\mathcal{O}(m^2)$. Using $\mathbf{G}_{yy}$ instead of $\mathbf{Q}_{ss}$, this removes the last reason to store $\mathbf{Q}_{ss}$ in an explicit and full form.

2.6.3. Reproduction of measurements

To clarify the influence of the measurement error $\mathbf{R}$ on $\hat{\mathbf{s}}$, we analyze how $\hat{\mathbf{s}}$ reproduces the measurements. We define the measurements values returned by the BLUE

$$\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{s}}. \tag{33}$$

The residuals of the BLUE are defined as

$$\hat{\mathbf{r}} = \mathbf{y} - \mathbf{H}\hat{\mathbf{s}} = \mathbf{y} - \mathbf{H}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Q}_{sy}\boldsymbol{\xi}) = \mathbf{R}\boldsymbol{\xi} \tag{34}$$

in which Eqs. (17) and (9) were used. Eq. (34) shows that $\hat{\mathbf{s}}$ never perfectly reproduces the observations for non-zero $\mathbf{R}$ (unless in the unlikely case that $\boldsymbol{\xi}$ vanishes). It is of great importance to understand that the residuals $\hat{\mathbf{r}}$ are not a lack of accuracy in the BLUE, but a consequence of meeting the measurements under the smoothness condition implied by the prior statistics of $\mathbf{s}$. After ortho-normalization, these residuals can be used for model criticism [10]. For later derivations, it is convenient to replace $\boldsymbol{\xi}$ in Eq. (34) by Eq. (26)

$$\hat{\mathbf{r}} = \mathbf{R}\mathbf{P}_{yy}\mathbf{y} - \mathbf{R}\mathbf{P}_{yb}\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^*. \tag{35}$$

Using Eq. (34), it is easy to obtain the properties of $\hat{\mathbf{r}}$ and $\hat{\mathbf{y}}$ for zero measurement error

$$\begin{aligned}\lim_{\mathbf{R}\to\mathbf{0}} \hat{\mathbf{y}} &= \mathbf{y},\\ \lim_{\mathbf{R}\to\mathbf{0}} \hat{\mathbf{r}} &= \mathbf{0}.\end{aligned} \tag{36}$$

At the limit of infinite measurement error, inserting $\mathbf{R}^{-1} = \mathbf{0}$ into the posterior pdf (Eq. (14)) yields the prior pdf (Eq. (4)), so that

$$\begin{aligned}\lim_{\mathbf{R}^{-1}\to\mathbf{0}} \hat{\mathbf{s}} &= \mathbf{X}\boldsymbol{\beta}^*,\\ \lim_{\mathbf{R}^{-1}\to\mathbf{0}} \hat{\mathbf{y}} &= \mathbf{H}\mathbf{X}\boldsymbol{\beta}^*,\\ \lim_{\mathbf{R}^{-1}\to\mathbf{0}} \hat{\mathbf{r}} &= \mathbf{y} - \mathbf{H}\mathbf{X}\boldsymbol{\beta}^*.\end{aligned} \tag{37}$$

If the prior mean is absolutely unknown, i.e. $\mathbf{Q}_{\beta\beta}^{-1} = \mathbf{0}$, the analysis of Eq. (12) yields

$$\lim_{\mathbf{Q}_{\beta\beta}^{-1}\to 0} \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{H}^T\mathbf{Q}_{yy}^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^T\mathbf{Q}_{yy}^{-1}\mathbf{y}, \tag{38}$$

so that the value of $\widehat{\boldsymbol{\beta}}$ in Eq. (18) is solely based on the observations. For the limiting case of known prior mean, with $\mathbf{Q}_{\beta\beta} = \mathbf{0}$, the second term in Eq. (12) vanishes

$$\lim_{\mathbf{Q}_{\beta\beta}\to 0} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* \tag{39}$$

and the value of $\widehat{\boldsymbol{\beta}}$ is fully determined by the prior values $\boldsymbol{\beta}^*$ of the trend parameters.

3. Quasi-linear Geostatistical Approach

In case the transfer function $\mathbf{f}(\mathbf{s})$ is non-linear, minimizing the objective function (Eq. (15)) becomes a matter of non-linear optimization. The Quasi-Linear Geostatistical Approach successively linearizes the transfer function about the current estimate $\mathbf{s}_k$:

$$\begin{aligned}\mathbf{f}(\mathbf{s}) &\approx \mathbf{f}(\mathbf{s}_k) + \widetilde{\mathbf{H}}_k(\mathbf{s}-\mathbf{s}_k),\\ \widetilde{\mathbf{H}}_k &= \left.\frac{\partial \mathbf{f}(\mathbf{s})}{\partial \mathbf{s}}\right|_{\mathbf{s}_k},\end{aligned} \tag{40}$$

in which $\widetilde{\mathbf{H}}_k$ is the $m \times n$ sensitivity matrix linearized about $\mathbf{s}_k$. Eq. (40) is exact at the limit of $\mathbf{s} \to \mathbf{s}_k$. In the following, all quantities marked with a tilde are quantities that depend on the current linearization. We introduce a modified vector of observations

$$\tilde{\mathbf{y}}_k = \mathbf{y} - \mathbf{f}(\mathbf{s}_k) + \widetilde{\mathbf{H}}_k\mathbf{s}_k \tag{41}$$

and obtain by linearized error propagation (see, e.g. [23])

$$\begin{aligned}E[\tilde{\mathbf{y}}_k \mid \widehat{\boldsymbol{\beta}}] &= \widetilde{\mathbf{H}}_k\mathbf{X}\widehat{\boldsymbol{\beta}},\\ \widetilde{\mathbf{Q}}_{yy,k} &= \widetilde{\mathbf{H}}_k\mathbf{Q}_{ss}\widetilde{\mathbf{H}}_k^T + \mathbf{R},\\ \widetilde{\mathbf{Q}}_{sy,k} &= \mathbf{Q}_{ss}\widetilde{\mathbf{H}}_k^T.\end{aligned} \tag{42}$$

Using the modified vector of observations produces a linearized objective function that is formally identical to Eq. (15)

$$\begin{aligned}\mathbf{s}_{k+1} = \min\Bigg[&\frac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^*)^T\mathbf{G}_{ss}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta}^*)\\ &+\frac{1}{2}(\tilde{\mathbf{y}}_k-\widetilde{\mathbf{H}}_k\mathbf{s})^T\mathbf{R}^{-1}(\tilde{\mathbf{y}}_k-\widetilde{\mathbf{H}}_k\mathbf{s})\Bigg],\end{aligned} \tag{43}$$

so that all subsequent derivations are identical. However, when evaluating the objective function, only the simplification of the prior term (Eq. (30)), but not Eq. (31) can be applied, since Eq. (31) is only exact for $\mathbf{s} = \mathbf{s}_k$.

3.1. Gauss–Newton optimization

The formalism used in the Quasi-Linear Geostatistical Approach is formally similar to the Gauss–Newton method [21] with a non-linear constraint. For comparison, both algorithms are given below.

Algorithm 1 (*Gauss–Newton method with constraint*). An unknown $n_g \times 1$ vector of parameters $\boldsymbol{\xi}$ is related to the $m_g \times 1$ vector of measurements $\mathbf{y}$, $m_g > n_g$, by the relation $\mathbf{y} = \mathbf{f}(\boldsymbol{\xi}) + \mathbf{r}$. The objective function to be minimized is $\chi^2 = (\mathbf{y} - \mathbf{f}(\boldsymbol{\xi}))^{\mathrm{T}}\mathbf{W}_{\xi\xi}(\mathbf{y} - \mathbf{f}(\boldsymbol{\xi}))$, in which $\mathbf{W}_{\xi\xi}$ is a $m_g \times m_g$ weighting matrix, while fulfilling a constraint of the form $\mathbf{G}(\boldsymbol{\xi}) = \mathbf{G}_0$. Deviations of $\mathbf{G}(\boldsymbol{\xi})$ from $\mathbf{G}_0$ are punished by the weighting matrix $\mathbf{W}_{\nu\nu}$. Define an initial guess $\boldsymbol{\xi}_0$. Then

(1) Compute $\widetilde{\mathbf{H}}_k = \frac{\partial \mathbf{f}(\xi)}{\partial \xi}\big|_{\xi_k}$ and $\tilde{\mathbf{g}}_k = \frac{\partial \mathbf{G}(\xi)}{\partial \xi}\big|_{\xi_k}$.

(2) Find $\boldsymbol{\xi}_{k+1}$ by solving

$$\boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k + \Delta\boldsymbol{\xi}, \qquad (44)$$

$$\begin{bmatrix} \widetilde{\mathbf{H}}_k^{\mathrm{T}}\mathbf{W}_{\xi\xi}\widetilde{\mathbf{H}}_k & \tilde{\mathbf{g}}_k^{\mathrm{T}} \\ \tilde{\mathbf{g}}_k & -\mathbf{W}_{\nu\nu}^{-1} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\xi} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} -\widetilde{\mathbf{H}}_k^{\mathrm{T}}\mathbf{W}_{\xi\xi}(\mathbf{y} - \mathbf{f}(\boldsymbol{\xi}_k)) \\ (\mathbf{G}_0 - \mathbf{G}(\boldsymbol{\xi}_k)) \end{bmatrix}. \qquad (45)$$

(3) Increase k by one and repeat until convergence.

The algorithm introduced by Kitanidis [11] covers the case of unknown mean. The version discussed in the following is an extension of the Geostatistical Approach to the general case of uncertain mean.

Algorithm 2 (*Quasi-Linear Geostatistical Approach*). Define an initial guess $\mathbf{s}_0$. Then

(1) Compute $\widetilde{\mathbf{H}}_k$.

(2) Find $\mathbf{s}_{k+1}$ by solving

$$\mathbf{s}_{k+1} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{k+1} + \widetilde{\mathbf{Q}}_{\mathrm{sy},k}\boldsymbol{\xi}_{k+1}, \qquad (46)$$

$$\begin{bmatrix} \widetilde{\mathbf{Q}}_{\mathrm{yy},k} & \widetilde{\mathbf{H}}_k\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{H}}_k^{\mathrm{T}} & -\mathbf{Q}_{\beta\beta}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{k+1} \\ \widehat{\boldsymbol{\beta}}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{y} - \mathbf{f}(\mathbf{s}_k) + \widetilde{\mathbf{H}}_k\mathbf{s}_k \\ -\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^* \end{bmatrix}. \qquad (47)$$

(3) Increase k by one and repeat until convergence.

Algorithms 1 and 2 have in common, that both $\widetilde{\mathbf{H}}_k^{\mathrm{T}}\mathbf{W}\widetilde{\mathbf{H}}_k$ and $\widetilde{\mathbf{Q}}_{\mathrm{yy},k}$, are derived from the Hessian matrices of the corresponding linearized objective functions (compare Eq. (32)). Further, the second line in both cases follows from the constraints, with $\widetilde{\mathbf{H}}_k\mathbf{X}$ and $\tilde{\mathbf{g}}_k$ being the corresponding derivatives. Finally, the right-hand side vectors of both Eqs. (45) and (47) contain the residuals from the previous step in one form or another.

The main difference originates from the parameterized form used in the Geostatistical Approach. The estimator in Algorithm 2 (Eq. (46)) is based on an $(m + p)$ dimensional subspace spanned by $\mathbf{X}$ and $\widetilde{\mathbf{Q}}_{\mathrm{sy},k}$. In each iteration step, when $\widetilde{\mathbf{H}}_k$ is updated, the subspace spanned by $\widetilde{\mathbf{Q}}_{\mathrm{sy},k}$ is updated simultaneously. Then, the old subspace spanned by $\widetilde{\mathbf{Q}}_{\mathrm{sy},k-1}$, is outdated and no more valid. Hence, unlike in Algorithm 1, Eq. (44), the updated solution is not given by the previous solution plus a modification. Instead, Algorithm 2 projects the previous solution onto the new subspace by including the term $\widetilde{\mathbf{H}}_k\mathbf{s}_k$ in the right-hand side vector of Eq. (47).

3.2. Form of the solution

The following is a graphical and instructive example to discuss the form of the solution and the iteration algorithm chosen in the Geostatistical Approach (Algorithm 2). Consider that two subspaces are available: an outdated subspace spanned by $\widetilde{\mathbf{Q}}_{\mathrm{sy},k-1}$ and the current subspace spanned by $\widetilde{\mathbf{Q}}_{\mathrm{sy},k}$, $\widetilde{\mathbf{Q}}_{\mathrm{sy},k} \neq \widetilde{\mathbf{Q}}_{\mathrm{sy},k-1}$. The subspace spanned by $\widetilde{\mathbf{Q}}_{\mathrm{sy},k-1}$ is not a linear combination of the components of $\widetilde{\mathbf{Q}}_{\mathrm{sy},k}$. The current sensitivity matrix $\widetilde{\mathbf{H}}_k$ is more accurate for the following estimation, since it has been linearized about a value of $\mathbf{s}$ that is closer to $\hat{\mathbf{s}}$ than the value which $\widetilde{\mathbf{H}}_{k-1}$ has been linearized about. Let us clearly regard $\widetilde{\mathbf{H}}_{k-1}$ as a poor linearization. Assume we defined a solution to the inverse problem in the following form:

$$\hat{\mathbf{s}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_k + \mathbf{X}\widehat{\boldsymbol{\beta}}_{k-1} + \widetilde{\mathbf{Q}}_{\mathrm{sy},k}\boldsymbol{\xi}_k + \widetilde{\mathbf{Q}}_{\mathrm{sy},k-1}\boldsymbol{\xi}_{k-1} \qquad (48)$$

and then fitted the parameters $\widehat{\boldsymbol{\beta}}_k$, $\widehat{\boldsymbol{\beta}}_{k-1}$, $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k-1}$ so that $\hat{\mathbf{s}}$ minimizes the objective function (Eq. (43)). It is clear that, if neglecting the smoothness condition implied by the a priori term, we are free to choose any combination of the parameters that lead to a perfect fit with the measurements. Since we have $(2m + 2p)$ parameters to fit while the observations and the conditions for the trend parameters result in no more than $(m + p)$ equations, the solution for $\mathbf{s}$ would not be unique. Now, we take into account the contribution of the prior term in the objective function. The current subspace is based on the more accurate linearization. This fact makes it more "efficient" for meeting the measurements in the sense that smaller perturbations lead to the same satisfaction of the measurements while allowing for a smaller value of the a priori term. Hence, the previous subspace is completely discarded by finding that only with $\boldsymbol{\xi}_{k-1} = \mathbf{0}$ the objective function is minimized. The same analysis holds for any number of available subspaces.

At this point, let us discuss several consequences:

(1) As mentioned in previous sections, the subspaces spanned by $\mathbf{X}$ and $\mathbf{Q}_{\mathbf{sy}}$ reduce the degrees of freedom and hence allow to define a unique solution in the case of linear cokriging. If several subspaces were available, the property of uniqueness would be lost.

(2) The iteration procedure in Algorithm 2 finds the optimal subspace for the estimator, in which the optimum subspace is defined such that the measurements are satisfied by minimum perturbations. By defining the unique optimal subspace for the non-linear case, the Quasi-Linear Geostatistical Approach maintains the uniqueness of the solution.

(3) Algorithm 2 does not adhere to previous trial solutions, but projects them onto the current subspace in each iteration step. The final estimate is entirely based on the final (optimal) subspace, and the conditional covariance is defined in the very same subspace, using Eq. (21) with $\mathbf{H} = \widetilde{\mathbf{H}}_k$. The single iteration steps are not a process of conditioning, but merely of finding the optimal subspace. Hence, it is not necessary to update the prior covariance during the iteration procedure as it would be the case in a sequential or successive Bayesian updating procedure. In the Quasi-Linear Geostatistical Approach, seen from the Bayesian point of view, the act of conditioning is entirely carried out in the final step.

(4) Line search algorithms find solutions of the form $\mathbf{s}_{k+1} = \mathbf{s}_k + \Delta\mathbf{s}$, which is in our case equivalent to the form given in Eq. (48). If a line search modification was applied to Algorithm 2, the outdated subspaces would not be discarded, and the Bayesian concept would be violated.

(5) Solutions of the form $\mathbf{s}_{k+1} = \mathbf{s}_k + \Delta\mathbf{s}$ do not violate the Bayesian concept if, and only if, they are used in the context of Bayesian updating procedures such as the Successive Linear Estimator [29] or sequential kriging and cokriging [27]. In these methods, each iteration step is defined as a process of conditioning. The covariances are successively updated so that the prior covariance of each iteration step is given by the conditional covariance of the preceding step.

3.3. Drawbacks of the conventional algorithm

For strongly non-linear problems, the Gauss–Newton method (Algorithm 1) in general and the Quasi-Linear Geostatistical Approach (Algorithm 2) in particular are known to diverge due to overshooting and oscillations. In comparison to Algorithm 1, Algorithm 2 has an additional disadvantage based on the changing subspace of the solution, as we will show in the following analysis.

3.3.1. Deterioration of the solution

We split the entries of the right-hand side vector in Eq. (20) into an innovative and a projecting part:

$$\begin{bmatrix} \tilde{\mathbf{y}} \\ -\mathbf{Q}_{\beta\beta}^{-1}\boldsymbol{\beta}^* \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{y} - \mathbf{f}(\mathbf{s}_k) \\ -\mathbf{Q}_{\beta\beta}^{-1}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_k) \end{bmatrix}}_{\text{innovative}} + \underbrace{\begin{bmatrix} \widetilde{\mathbf{H}}_k\mathbf{s}_k \\ -\mathbf{Q}_{\beta\beta}^{-1}\widehat{\boldsymbol{\beta}}_k \end{bmatrix}}_{\text{projecting}}. \tag{49}$$

When inserting these parts into the linearized cokriging equations (Eq. (47)) separately, one obtains a projecting part and an innovative part of the parameter vector:

$$\begin{bmatrix} \boldsymbol{\xi}_{k+1} \\ \widehat{\boldsymbol{\beta}}_{k+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}_{\mathrm{pr}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{pr}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_{\mathrm{in}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{in}}, \end{bmatrix} \tag{50}$$

which in turn can be inserted into the estimator (Eq. (46))

$$\mathbf{s}_{k+1} = \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{pr}} + \widetilde{\mathbf{Q}}_{\mathbf{sy},k}\boldsymbol{\xi}_{\mathrm{pr}}}_{\mathbf{s}_{\mathrm{pr}}} + \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{in}} + \widetilde{\mathbf{Q}}_{\mathbf{sy},k}\boldsymbol{\xi}_{\mathrm{in}}}_{\mathbf{s}_{\mathrm{in}}}. \tag{51}$$

The innovative part generates new innovations based on the residuals $(\mathbf{y} - \mathbf{f}(\mathbf{s}_k))$ and $(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_k)$ of the previous trial solution, while the projecting part projects $\mathbf{s}_k$ onto the new subspace spanned by $\widetilde{\mathbf{Q}}_{\mathbf{sy},k}$. Finally, the splitting affects the measurements $\tilde{\mathbf{y}}$ returned by the BLUE

$$\hat{\mathbf{y}}_{k+1} = \underbrace{\widetilde{\mathbf{H}}_k\mathbf{s}_{\mathrm{pr}}}_{\hat{\mathbf{y}}_{\mathrm{pr}}} + \underbrace{\widetilde{\mathbf{H}}_k\mathbf{s}_{\mathrm{in}}}_{\hat{\mathbf{y}}_{\mathrm{in}}}. \tag{52}$$

Now, we substitute the projecting part $\hat{\mathbf{y}}_{\mathrm{pr}} = \widetilde{\mathbf{H}}_k\mathbf{s}_{\mathrm{pr}}$ for $\mathbf{y}$ in Eq. (35) to analyze how the projection reproduces the observations

$$\hat{\mathbf{r}}_{\mathrm{pr}} = \mathbf{R}\widetilde{\mathbf{P}}_{\mathbf{yy},k}\widetilde{\mathbf{H}}_k\mathbf{s}_k - \mathbf{R}\widetilde{\mathbf{P}}_{\mathbf{yb},k}\mathbf{Q}_{\beta\beta}^{-1}\hat{\boldsymbol{\beta}}_k. \tag{53}$$

Then it becomes evident that $\hat{\mathbf{y}}_{\mathrm{pr}} \neq \hat{\mathbf{y}}_k$ unless $\mathbf{R} = \mathbf{0}$

$$\lim_{\mathbf{R}\to\mathbf{0}} \hat{\mathbf{y}}_{\mathrm{pr}} = \hat{\mathbf{y}}_k. \tag{54}$$

This means that, for non-zero $\mathbf{R}$, the projecting part $\mathbf{s}_{\mathrm{pr}}$ never satisfies the observations to the same extent as the previous trial solution $\mathbf{s}_k$. The extreme case of infinite $\mathbf{R}$ may serve as an illustrative example: Inserting the projecting part into Eq. (37) yields

$$\lim_{\mathbf{R}^{-1}\to\mathbf{0}} \mathbf{s}_{\mathrm{pr}} = \mathbf{X}\boldsymbol{\beta}^*. \tag{55}$$

For $\widetilde{\mathbf{H}}_k = \widetilde{\mathbf{H}}_{k-1}$, it can be shown that $\mathbf{s}_{\mathrm{pr}}$ is equal to $\mathbf{s}_k$ if, and only if, $\mathbf{R} = \mathbf{0}$. This is easy to see since the subspace for $\hat{\mathbf{s}}$ does not change, i.e., $\widetilde{\mathbf{Q}}_{\mathbf{sy},k} = \widetilde{\mathbf{Q}}_{\mathbf{sy},k-1}$, and the projection from the previous subspace onto the current one is an identity operation. For non-linear $\mathbf{f}(\mathbf{s})$, i.e. $\widetilde{\mathbf{H}}_k \neq \widetilde{\mathbf{H}}_{k-1}$, the projection is not an identity operation, and $\mathbf{R} = \mathbf{0}$ is no more sufficient to ensure that $\mathbf{s}_{\mathrm{pr}} = \mathbf{s}_k$, and we will find in general that $\mathbf{s}_{\mathrm{pr}} \neq \mathbf{s}_k$. Then, because $\mathbf{f}(\mathbf{s})$ is non-linear and Eq. (40) is only approximate, it is not necessarily true that $\mathbf{f}(\mathbf{s}_{\mathrm{pr}}) = \mathbf{f}(\mathbf{s}_k)$ even for $\mathbf{R} = \mathbf{0}$ and $\widetilde{\mathbf{H}}_k\mathbf{s}_{\mathrm{pr}} = \widetilde{\mathbf{H}}_k\mathbf{s}_k$. Thus, the projecting part of the solution deteriorates in any case. The larger the extent of non-linearity or the larger the step sizes occurring during iteration, the less accurate is the linearization. This, in turn, leads to a higher degree of deterioration in the projection, with the potential to prevent the entire algorithm from converging.

3.3.2. Local minima

For strongly non-linear problems, the objective function may have multiple minima. In such cases, the Geostatistical Approach (Algorithm 2) may find a local minimum that satisfies the measurements to an extent specified by the measurement error statistics, but its identity to the global minimum cannot be proved. Then, it is common practice to accept the solution if (1) it is acceptably smooth to the subjective satisfaction of the modeler, and (2a) the objective function does not exceed a prescribed value, typically derived from the χ^2-distribution for $(m+p)$ degrees of freedom or (2b) the orthonormalized residuals obey certain statistics [10]. According to our experience, most failures to find an acceptable solution originate from overshooting of iteration steps, which leads to solutions that are not sufficiently smooth in the sense of the prior distribution.

4. Modified Levenberg–Marquardt algorithm

In this section, we present and discuss a modified Levenberg–Marquardt Algorithm for the Quasi-Linear Geostatistical Approach. The choice of the Levenberg–Marquardt algorithm and the nature of the modifications is based on the following simple and perspicuous train of thought.

(1) The Geostatistical Approach (Algorithm 2) suffers from oscillations and over-shooting, leading to solutions that fail to comply with the smoothness constraint. Further, in addition to typical problems of successive linearization methods with strongly non-linear problems, the solution deteriorates whenever the step size is too large and the algorithm may fail. Applying a line search on top of Algorithm 2 would violate the required form of the solution.

(2) The Levenberg–Marquardt algorithm [16,17] suppresses oscillations and overshooting by controlling the step size and direction. It does so by amplifying the diagonal entries of Eq. (45) in the Gauss–Newton algorithm (Algorithm 1).

(3) This is similar to amplifying the measurement error $\mathbf{R}$ in Algorithm 2, Eq. (47). Using $\mathbf{R}$ to amplify the diagonal entries of the linearized cokriging system will put the step size control into a statistically based and well controllable framework.

(4) When exerting intelligent control over $\mathbf{R}$ during the course of iteration, the solution space can systematically be screened starting at the prior mean, which increases the chance that the solution complies with the smoothness condition.

(5) The role of the projecting and the innovative parts can be taken into account such that $\mathbf{R}$ is controlled separately for these two parts: to suppress the deterioration through the projection and to prevent overshooting in the innovative part. The measurement error $\mathbf{R}$ has to be decreased in the projection part and increased in the innovation part to reduce the step size.

(6) Error analysis of the linearization can be used to prescribe a certain maximum step size. Further, if an iteration step is very small and the linearization is still sufficiently accurate, it can be re-used for the next iteration step.

Again, we discuss a few properties of the standard Levenberg–Marquardt algorithm for least-squares fitting before introducing the counterpart for the Geostatistical Approach.

Algorithm 3 (*Levenberg–Marquardt algorithm with constraint*). The problem description is identical to Algorithm 1. Define an initial guess ξ_0 and initialize the Levenberg–Marquardt parameter λ with $\lambda > 0$. Then

(1) Compute $\widetilde{\mathbf{H}}_k$ and $\widetilde{\mathbf{g}}_k$.

(2) Find ξ_{k+1} by solving

$$\xi_{k+1} = \xi_k + \Delta\xi, \tag{56}$$

$$\begin{bmatrix} \widetilde{\mathbf{H}}_k^{\mathrm{T}}\mathbf{W}_{\xi\xi}\widetilde{\mathbf{H}}_k + \lambda\mathbf{D}_1 & \widetilde{\mathbf{g}}_k^{\mathrm{T}} \\ \widetilde{\mathbf{g}}_k & -\mathbf{W}_{\nu\nu}^{-1} + \lambda\mathbf{D}_2 \end{bmatrix} \begin{bmatrix} \Delta\xi \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} -\widetilde{\mathbf{H}}_k^{\mathrm{T}}\mathbf{W}_{\xi\xi}(\mathbf{y} - \mathbf{f}(\xi_k)) \\ \mathbf{G}_0 - \mathbf{G}(\xi_k) \end{bmatrix}. \tag{57}$$

If the objective function does not improve, increase λ and repeat step 2. Otherwise, decrease λ.

(3) Increase k by one and repeat until convergence.

The terms $\lambda\mathbf{D}_1$ and $\lambda\mathbf{D}_2$ amplify the diagonal entries of the Hessian matrix in Eq. (57). Initially, λ is assigned a low value, $\lambda > 0$. Whenever convergence is poor, λ is increased by a user-defined factor, and is again decreased whenever convergence is good. For $\lambda \to \infty$, the step size $|\Delta\xi|$ approaches zero, the search direction approaches the direction of steepest descent, and there is always an improvement of the objective function for sufficiently large λ unless ξ_k is a minimum. As ξ_k converges towards the solution, λ can be decreased to zero. Ideally, during the last iteration steps, the unmodified system of equations is used and the algorithm is identical to Algorithm 1.

Algorithm 4 (*Modified Levenberg–Marquardt Algorithm for the Quasi-Linear Geostatistical Approach*). The problem statement is as specified for Algorithm 2. Error analysis yields that the error of linearization is acceptable only for $|\mathbf{s}-\mathbf{s}_k| < \Delta\mathbf{s}_1$, and negligible for $|\mathbf{s}-\mathbf{s}_k| < \Delta\mathbf{s}_2$. Define an initial guess $\mathbf{s}_0 = \mathbf{X}\boldsymbol{\beta}^*$ and initialize λ with $\lambda > 0$.

(1) Compute $\widetilde{\mathbf{H}}_k$ unless $|\mathbf{s}_{k-1}-\mathbf{s}_k| < \Delta\mathbf{s}_2$.

(2) Find $\mathbf{s}_{k+1}$ by solving the following equations:

$$\mathbf{s}_{k+1} = \mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{pr}} + \hat{\boldsymbol{\beta}}_{\mathrm{in}}) + \widetilde{\mathbf{Q}}_{\mathrm{sy},k}(\boldsymbol{\xi}_{\mathrm{pr}} + \boldsymbol{\xi}_{\mathrm{in}}), \tag{58}$$

$$\begin{bmatrix} \widetilde{\mathbf{Q}}_{\mathrm{yy},k} + \lambda\mathbf{R} & \widetilde{\mathbf{H}}_k\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{H}}_k^{\mathrm{T}} & -(1+\lambda)\mathbf{Q}_{\beta\beta}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{\mathrm{in}} \\ \hat{\boldsymbol{\beta}}_{\mathrm{in}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} - \mathbf{f}(\mathbf{s}_k) \\ -\mathbf{Q}_{\beta\beta}^{-1}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_k) \end{bmatrix}, \tag{59}$$

$$\begin{bmatrix} \widetilde{\mathbf{Q}}_{\mathrm{yy},k} - \tau\mathbf{R} & \widetilde{\mathbf{H}}_k\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{H}}_k^{\mathrm{T}} & -(1+\lambda)\mathbf{Q}_{\beta\beta}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{\mathrm{pr}} \\ \hat{\boldsymbol{\beta}}_{\mathrm{pr}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{H}}_k\mathbf{s}_k \\ -(1+\lambda)\mathbf{Q}_{\beta\beta}^{-1}\hat{\boldsymbol{\beta}}_k \end{bmatrix}, \tag{60}$$

$$\tau = 1 - (1+\lambda)^{-\gamma}. \tag{61}$$

If $|\mathbf{s}_{k+1}-\mathbf{s}_k| \geqslant \Delta\mathbf{s}_1$ or if the objective function does not improve, increase λ and repeat step 2. Otherwise decrease λ and continue.

(3) Increase k by one and repeat until convergence.

4.1. Properties of the modified algorithm

Algorithm 4 has the following properties:

(1) The Levenberg–Marquardt parameter λ controls the step size. If the previous step is sufficiently small, the limit for $\lambda \to \infty$ is a step size of zero. This property is discussed below in more detail.

(2) By appropriate choice of $\gamma > 0$, the algorithm can be fine-tuned to the problem at hand. For large γ, the algorithm becomes more aggressive by suppressing the deterioration of the projecting part. We recommend $\gamma > 1$ to ensure that $\boldsymbol{\xi}_{\mathrm{rep}} \to \boldsymbol{\xi}_k$ is faster than $\boldsymbol{\xi}_{\mathrm{in}} \to \mathbf{0}$.

(3) As the algorithm converges, λ can be decreased towards zero. For $\lambda \to 0$, the algorithm is identical to the conventional form (Algorithm 2).

(4) The solution found by Algorithm 4 has the same properties as the solution found by Algorithm 2, following the strict Bayesian framework. Eq. (58) is not to be confused with the form of the solution that would violate the Bayesian concept as discussed above (Eq. (48)), since no outdated subspaces appear here.

(5) The solution space is screened in a controlled manner, starting at the prior mean. In case the uniqueness of the solution is questionable because the objective function is strongly non-linear and has multiple local minima, the solution found by Algorithm 4 has better chances to comply with the smoothness condition and to fulfill common statistical criteria for testing the solution.

(6) The costs of searching for an adequate value of λ and the for computing new linearizations are minimized through error analysis.

4.2. Step size control

In Eq. (59), $\widetilde{\mathbf{Q}}_{\mathrm{yy},k}$ is negligible for very large λ. Then, approximate $\mathbf{Q}_{\mathrm{yy}} \approx \lambda\mathbf{R}$ and substitute the modified cokriging matrix from Eq. (59) in Eqs. (23)–(25) to obtain the limit of the $\mathbf{P}$ submatrices for $\lambda \to \infty$. Insert the resulting expressions and the right-hand side vector from Eq. (59) into Eqs. (26) and (27) to give

$$\begin{aligned} &\lim_{\lambda\to\infty}(\boldsymbol{\xi}_{\mathrm{in}}) = \mathbf{0}, \\ &\lim_{\lambda\to\infty}(\hat{\boldsymbol{\beta}}_{\mathrm{in}}) = \mathbf{0}. \end{aligned} \tag{62}$$

This shows that increasing λ can be used to restrict the step size for the innovative part.

Similarly, we can show that $\mathbf{s}_{\mathrm{pr}} = \mathbf{s}_k$ for $\lambda \to \infty$. Considering that, according to Eq. (42), $\widetilde{\mathbf{Q}}_{\mathrm{yy},k} - \tau\mathbf{R} = \widetilde{\mathbf{H}}_k\mathbf{Q}_{\mathrm{ss}}\widetilde{\mathbf{H}}_k^{\mathrm{T}}$ for infinite λ, $\mathbf{R}$ vanishes from Eq. (53), and we obtain

$$\begin{aligned} &\lim_{\lambda\to\infty}\hat{\mathbf{r}}_{\mathrm{pr}} = \mathbf{0}, \\ &\lim_{\lambda\to\infty}\widetilde{\mathbf{H}}_k\mathbf{s}_{\mathrm{pr}} = \widetilde{\mathbf{H}}_k\mathbf{s}_k. \end{aligned} \tag{63}$$

Combining Eqs. (58)–(63) yields for the linear case with $\widetilde{\mathbf{H}}_k = \widetilde{\mathbf{H}}_{k-1}$

$$\lim_{\lambda\to\infty}\mathbf{s}_{k+1} = \mathbf{s}_k. \tag{64}$$

Still, the algorithm is subject to the deterioration of the solution, since Eq. (64) holds for the linear case with $\widetilde{\mathbf{H}}_k = \widetilde{\mathbf{H}}_{k-1}$. For $\widetilde{\mathbf{H}}_k \neq \widetilde{\mathbf{H}}_{k-1}$, these identities are only approximate. In most situations, the approximate character of Eq. (64) does not cause problems. In case problems should occur, the step size restriction defined by $\Delta\mathbf{s}_1$ can be chosen more drastically to ensure that $\widetilde{\mathbf{H}}_k \approx \widetilde{\mathbf{H}}_{k-1}$.

4.3. Application to known and unknown mean

Algorithm 4 is designed for the case of uncertain mean. The cases of known and unknown mean are merely special cases of this general case. To obtain an algorithm for the case of unknown mean, set $\mathbf{Q}_{\beta\beta}^{-1} = \mathbf{0}$ in all places. A more stable version for the unknown mean case can be obtained by substituting $(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_k) = \mathbf{0}$ in Eq. (59).

In this case, no bias is exerted onto $\hat{\boldsymbol{\beta}}$ so that effectively the algorithm behaves like in the unknown mean case, but the step size control over $\boldsymbol{\beta}$ is still active.

For the case of known mean, the entire derivations simplify, and the additional terms, rows and columns for $\hat{\boldsymbol{\beta}}$ disappear in all equations. Instead, the known mean value $\mathbf{X}\boldsymbol{\beta}$ is added to $\mathbf{s}_{k+1}$ in Eq. (58) and subtracted from $\mathbf{s}_k$ in Eq. (60).

5. Performance test

To compare the performance of the new algorithm compared to the conventional algorithm, we apply both of them to a problem described by Cirpka and Kitanidis [2] with several simplifications: We seek for the hydraulic conductivity distribution K with unknown mean in a 2-D locally isotropic aquifer, considering measurements of hydraulic head ϕ and arrival time t_{50} of a conservative tracer. Since a full mathematical description of the underlying problem is given in the original publication, we only provide a brief summary.

The unknown quantity is the log-conductivity $Y = \log K$, discretized as an elementwise constant function on a regular grid with n elements. The unknowns are second-order stationary with uncertain constant mean, making $\mathbf{X}$ an $n \times 1$ vector with unit entries. The covariance matrix $\mathbf{Q}_{ss}$ is given by the exponential model with structural parameters that are assumed to be known for simplicity. Since the grid is regular, $\mathbf{Q}_{ss}$ is block Toeplitz with Toeplitz blocks, allowing to apply FFT-based methods for multiplication of $\mathbf{Q}_{ss}$ [20].

The transformation $Y = \ln K$ is linearized about $\widetilde{Y}_k$ by

$$\exp(\widetilde{Y}_k + Y') \approx \widetilde{K}_k + \widetilde{K}_k Y' \tag{65}$$

with $\widetilde{K}_k = \exp(\widetilde{Y}_k)$. The observed hydraulic head ϕ is related to Y through the steady-state groundwater flow equation without sources and sinks

$$\nabla \cdot (\exp(Y)\nabla\phi) = 0 \quad \text{in } \Omega. \tag{66}$$

The flow domain Ω with the boundary Γ is rectangular: Γ_1 is the west section of Γ, Γ_2 is east, and Γ_3 and Γ_4 are north and south. Boundary conditions are

$$\begin{aligned} &\phi = \hat{\phi}_1 \quad \text{on } \Gamma_1, \\ &\phi = \hat{\phi}_2 \quad \text{on } \Gamma_2, \\ &(\exp(\widetilde{Y}_k)\nabla\phi) \cdot \mathbf{n} = 0 \quad \text{on} \quad \Gamma_{3,4}. \end{aligned} \tag{67}$$

The boundary conditions for the tracer are an instantaneous release of the tracer at Γ_1 and $t = 0$, with zero-flux conditions at $\Gamma_{3,4}$ and no diffusion at $\Gamma_{1,2}$.

The sensitivities of the head and arrival time measurements with respect to Y are computed by the adjoint-state method [24]. A major contribution to the non-linearity of this problem originates from the linearization of $K = \exp(Y)$. Error analysis yields

$$\varepsilon(Y') = \exp(Y') - (1 + Y'). \tag{68}$$

We decide the error to be acceptable for $Y' < 0.4$, $\varepsilon \approx 0.1$, and negligible for $Y' < 0.01$, $\varepsilon \approx 5\mathrm{e} - 5$, which will be taken into account in Algorithm 4. Further sources of non-linearity stem from the changing streamline pattern during the course of the iteration process and other effects of evolving heterogeneity during the iteration algorithm.

To set up test cases, we generate unconditional realizations of Y using the spectral approach of Dietrich and Newsam [5], solve Eqs. (66) and the transport problem, pick values of ϕ and t_{50} at the measurement locations, and add white noise to obtain artificial measurement data. An example of an unconditional realization of log K together with the corresponding head and arrival time distribution and measurement locations is displayed in Fig. 1(a)–(c).

Subsequently, we 'forget' the generated distributions of log K and proceed with the Quasi-Linear Geostatistical Approach to determine the 'unknown' spatial distribution of Y, using both algorithms for comparison. For illustration, the result of a specific test case is displayed in Fig. 1(d)–(f). The conductivity distribution recovered through the Geostatistical Approach is smoother than the original field. However, the dependent quantities used for conditioning meet the measurements at the measurement locations.

We will discuss five test cases. In case one, the variance of Y is chosen sufficiently small so that the problem is effectively linear. In all other test cases, we have increased the variance of Y by simply scaling the realization used in case one, making the problem non-linear. The chosen parameter values are listed in Table 1. The conventional form of the Geostatistical Approach has been reported to be applicable up to higher values for the variance of log-conductivity σ_Y^2 if only head measurements are available. The relation between measurements of arrival time, however, is non-linear to a higher degree even for smaller values of σ_Y^2.

The results of the test cases are shown in Table 2. In the almost linear setup, case one, both algorithms find the solution within few steps. The solutions are identical but for differences below chosen breakoff criteria of the iteration procedure, and have the same geostatistical properties. In case two, the conventional algorithm (Algorithm 2) begins to oscillate, but finally finds the solution after 15 steps, while the modified Levenberg–Marquardt version (Algorithm 4) still performs well. A comparison of the solution yields that both solutions

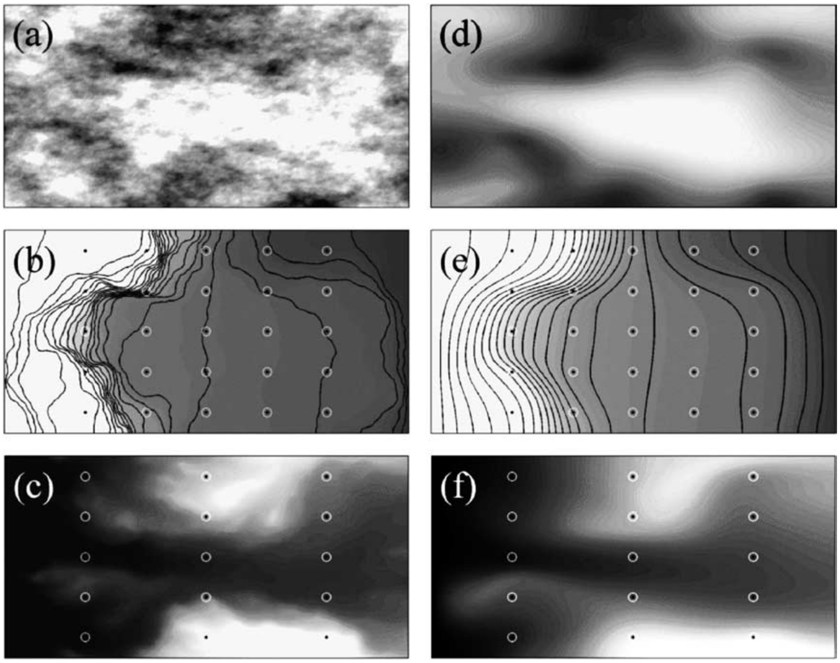

Fig. 1. Problem setup: (a) realization of $\ln K$ (here: variance 6.4 for test case 5); (b) modeled distribution of hydraulic heads; (c) modeled arrival time distribution; (d) distribution of $\ln K$ obtained from Quasi-Linear Geostatistical Approach using the modified Levenberg–Marquardt algorithm; (e) hydraulic heads for inverse problem and (f) arrival time for inverse problem. Dots represent locations of measurements. Greyscale normalized for direct comparison of true distribution (a–c) with cokriging results (d–f). Brighter colors correspond to higher values.

Table 1
Parameters used for the test cases

Parameter	Units	Value
Domain length L_x	m	1000
Correl. length λ_x	m	4
Grid spacing d_x	m	4
Observations ϕ	–	25
Observations t_{50}	–	15
Error σ_h for h	m	0.01
Error σ_t for t_{50}	%	10
Domain length L_y	m	500
Correl. length λ_y	m	2
Grid spacing d_y	m	4
Porosity	–	0.3
Dispersivity a_l	m	10
Dispersivity α_t	m	1
Diffusion D_m	$\frac{m^2}{s}$	10^{-9}

For the variance of each test case, please refer to Table 2.

have similar values of $\hat{\mathbf{r}}$ that cannot be rejected based on χ^2 statistics using a 95% confidence level. However, the value of the prior term is smaller for the solution obtained from Algorithm 4, i.e., the latter solution is smoother than the one obtained from the conventional algorithm. In case three, the new algorithm proves stable and monotonically converges towards the solution, whereas the non-linearity of the problem causes the conventional algorithm to fail immediately. Case four represents the degree of non-linearity that the new algorithm can deal with, and case five demonstrates its limits. The solution obtained from the new algorithm for case five is shown in Fig. 1(d)–(f).

The contour-lines of the hydraulic head indicate that the streamlines are inclined up to almost 90° due to the heterogeneity of the flow field, resembling the extreme degree of non-linearity for all streamline-related pro-

Table 2
Performance of the conventional algorithm for the Quasi-Linear Geostatistical Approach (Algorithm 2) and the modified Levenberg–Marquardt algorithm (Algorithm 4), comparison for the test cases 1–5

Case no.	Variance σ_Y^2	Algorithm 2		Algorithm 4	
		No. of steps	Status	No. of steps	Status
1	0.1	4	Converged	3	Converged
2	0.4	15	Oscillations	5	Converged
3	1.6	–	Failed	15	Converged
4	3.2	–	Failed	19	Converged
5	6.4	–	Failed	20	Stagnated

cesses and quantities like, e.g. the transport of a tracer and hence its arrival time distribution.

6. Summary and conclusions

We have analyzed the Gauss–Newton algorithm used in the Quasi-Linear Geostatistical Approach for inverse modeling. A discussion in the Bayesian framework revealed that, and why, the solution has to obey a certain form. The solution is defined in a subspace that is obtained from the geostatistical approach and changes during the iteration procedure. To comply with the Bayesian concept and maintain the uniqueness of the solution, only the final and optimal subspace must be used, and the previous trial solution must be projected onto the current subspace in each iteration step.

Like the Gauss–Newton algorithm for least-squares fitting, the Gauss–Newton version of the Quasi-Linear Geostatistical Approach encounters problems in strongly non-linear optimization tasks. Overshooting and oscillations may occur, causing the algorithm to diverge. In case the inverse problem is strongly non-linear and the objective function has multiple local minima, excessively large steps lead to solutions that fail to obey the smoothness condition implied by the geostatistical approach. Additionally, the projection onto the current subspace introduces new instability problems.

To increase the stability of the iteration procedure, we introduced a modified Levenberg–Marquardt version of the Quasi-Linear Geostatistical Approach. This new algorithm splits each iteration step into two parts: projecting the previous trial solution onto the current subspace and reducing the residuals. Each part is provided with its own stabilization mechanism. The first stabilization is to reduce the deterioration of the trial solution during the projection onto the current subspace, while the second restricts the improvement of the residuals to prevent overshooting.

The resulting algorithm screens the solution space starting at the prior mean in a geostatistically controlled manner. Exerting control over the step size, it reduces the risk of oscillation or overshooting of the solution. In case of strong non-linearity, the objective functions may have multiple minima and the identity of the solution to the global minimum cannot be proved. Instead, decision criteria based on geostatistical considerations are used to reject or accept the solution. According to our experience, local minima do in most cases not fulfill these decision criteria since they fail to obey the smoothness condition implied by the geostatistical approach. By putting the step size control into the geostatistical framework, the new algorithm improves the chance to obtain solutions that comply with this smoothness condition. We demonstrated in test cases, that the new algorithm has an increased convergence radius and can cope with stronger non-linearity while requiring less iteration steps than its Gauss–Newton relative. This allows to apply the Quasi-Linear Geostatistical Approach to cases of higher variability and increased non-linearity.

Acknowledgements

This study has been funded by the Emmy Noether program of the Deutsche Forschungsgemeinschaft under the grant Ci 26/3-4. The authors like to thank two anonymous reviewers for their constructive remarks.

References

[1] Carrera J, Glorioso L. On geostatistical formulation of the groundwater flow inverse problem. Adv Water Resour 1991;14(5):273–83.

[2] Cirpka O, Kitanidis P. Sensitivity of temporal moments calculated by the adjoint-state method, and joint inversing of head and tracer data. Adv Water Resour 2001;24(1):89–103.

[3] Cirpka O, Nowak W. First-order variance of travel time in non-stationary formations. Water Resour. Res. (in press).

[4] Dietrich C, Newsam G. A stability analysis of the geostatistical approach to aquifer transmissivity identification. Stochast Hydrol Hydraul 1989;3:293–316.

[5] Dietrich C, Newsam G. A fast and exact method for multidimensional Gaussian stochastic simulations. Water Resour Res 1993;29(8):2861–9.

[6] Dietrich C, Newsam G. A fast and exact method for multidimensional Gaussian stochastic simulations: extension to realizations conditioned on direct and indirect measurements. Water Resour Res 1996;32(6):1643–52.

[7] Gomez-Hernandez J, Sahuquillo A, Capilla J. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data. 1. Theory. J Hydrol 1997;203:162–74.

[8] Harter T, Gutjahr A, Yeh T-J. Linearized cosimulation of hydraulic conductivity, pressure head, and flux in saturated and unsaturated, heterogenous porous media. J Hydrol 1999;183:169–90.

[9] Kitanidis P. Parameter uncertainty in estimation of spatial functions: Bayesian analysis. Water Resour Res 1986;22(4):499–507.

[10] Kitanidis P. Orthonormal residuals in geostatistics: model criticism and parameter estimation. Math Geol 1991;23(5):741–58.

[11] Kitanidis P. Quasi-linear geostatistical theory for inversing. Water Resour Res 1995;31(10):2411–9.

[12] Kitanidis P. On the geostatistical approach to the inverse problem. Adv Water Resour 1996;19(6):333–42.

[13] Kitanidis P. Analytical expressions of conditional mean, covariance, and sample functions in geostatistics. Stochast Hydrol Hydraul 1996;12:279–94.

[14] Kitanidis P. Comment on "a reassessment of the groundwater inverse problem" D. Mclaughlin and L.R. Townley. Water Resour Res 1997;33(9):2199–202.

[15] Kitanidis P, Vomvoris E. A geostatistical approach to the inverse problem in groundwater modeling (steady-state) and one-dimensional simulations. Water Resour Res 1983;19(3):677–90.

[16] Levenberg K. A method for the solution of certain nonlinear problems in least squares. Quart Appl Math 1944;2:164–8.

[17] Marquardt D. An algorithm for least squares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431–41.

[18] McLaughlin D, Townley L. A reassessment of the groundwater inverse problem. Water Resour Res 1996;32(5):1131–61.

[19] McLaughlin D, Townley L. Reply to comment by P. K. Kitanidis on "a reassessment of the groundwater inverse problem" by D. Mclaughlin and L.R. Townley. Water Resour Res 1997;33(9):2203.

[20] Nowak W, Tenkleve S, Cirpka O. Efficient computation of linearized cross-covariance and auto-covariance matrices of interdependent quantities. Math Geol 2003;35(1):53–66.

[21] Press W, Teukolsky BFS, Vetterling W. Numerical Recipes: The Art of Scientific Computing. second ed. Cambridge: Cambridge University Press; 1992.

[22] RamaRao B, La Venue A, de Marsily G, Marietta M. Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields. 1. Theory and computational experiments. Water Resour Res 1995;31(3):475–93.

[23] Schweppe F. Uncertain dynamic systems. Englewood Cliffs, NJ: Prentice-Hall; 1973.

[24] Sun N-Z. Inverse problems in groundwater modeling. In: Theory and applications of transport in porous media. Dordrecht: Kluwer Academic Publishers; 1994.

[25] Sun N-Z, Yeh W-G. Coupled inverse problems in groundwater modeling. 1. Sensitivity analysis and parameter identification. Water Resour Res 1990;26(10):2507–25.

[26] van Loan C. Computational frameworks for the fast Fourier transform. Philadelphia, PA: SIAM Publications; 1992.

[27] Vargas-Guzmán J, Yeh T-J. Sequential kriging and cokriging: two powerful geostatistical approaches. Stochast Environ Res Risk Assess 1999;13:416–35.

[28] Yeh T-C, Gutjahr A, Jin M. An iterative cokriging-like technique for ground-water flow modeling. Ground Water 1995;33(1):33–41.

[29] Yeh T-C, Jin M, Hanna S. An iterative stochastic inverse method: conditional effective transmissivity and hydraulic head fields. Water Resour Res 1996;32(1):85–92.

[30] Yeh T-C, Ŝimůnek J. Stochstic fusion of information for characterizing and monitoring the vadose zone. Vadose Zone J 2002;1:207–21.

[31] Zhang J, Yeh T-J. An iterative geostatistical inverse method for steady flow in the vadose zone. Water Resour Res 1997;33(1):63–71.

[32] Zimmerman D. Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. J Statist Computat Simulat 1989;32(1/2):1–15.

[33] Zimmerman D, de Marsily G, Gotway C, Marietta M, Axness C, Beauheim R, et al. A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resour Res 1998;34(6):1373–413.