(Proceedings ModelCARE2011 held at Leipzig, Germany, in September 2011) (IAHS Publ. 355, 2012).

Optimization of pilot points location for geostatistical inversion of groundwater flow

MARCO PANZERI, ALBERTO GUADAGNINI & MONICA RIVA
DIIAR, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
marco1.panzeri@mail.polimi.it

Abstract We investigate the influence of pilot points location on our ability to characterize key parameters describing a randomly heterogeneous porous medium via geostatistical inverse modelling. Our methodology is framed in a Maximum Likelihood (ML) context. We estimate the optimal location of pilot points through a differential evolution method (DEM) which we embed in the inversion of moment equations of groundwater flow. The DEM allows investigating a large number of candidate solutions to select those leading to the minimization of a given objective function through an algorithm that mimics the process of natural evolution. We explore the strength of the methodology by way of a synthetic example and we investigate the effect of the parameters embedded in the algorithm and the ability of model quality criteria such as negative log likelihood, the Bayesian criteria *BIC* and *KIC* and information criteria *AIC*, *AICc* and *HIC* to estimate the optimal pilot points locations.

Key words stochastic inverse model; geostatistics; pilot points; model selection criteria; Differential Evolution Method

INTRODUCTION AND THEORETICAL BACKGROUND

Heterogeneous groundwater systems are often characterized through inverse modelling techniques. The explicit goal of an inverse method is to provide images of the spatial distribution of log-conductivity, Y, minimizing the misfit between observed and calculated quantities (such as hydraulic heads, h). Parameterization techniques are often employed to describe the spatial distribution of Y in terms of a relatively small number of parameters.

The pilot points approach was introduced by de Marsily (1984). The parameters of the system are Y values at measurement locations (if available) and at additional selected (pilot) points. The choice of the number, N_P, and spatial distribution of pilot points is a main step in the approach. In the context of geostatistical inversion of deterministic groundwater flow equations, Alcolea *et al.* (2006) investigate the influence of N_P on the quality of model estimates. By way of a synthetic example, these authors note that the mean absolute estimation error of Y (defined as the mean absolute difference between the true and estimated values of Y at each grid node) decreases as N_P increases. During geostatistical inversion of groundwater flow Moment Equations, Riva *et al.* (2010) find that the mean absolute error of Y estimates tends to decrease with N_P until it attains a stable value.

The analysis of the optimal spatial distribution of pilot points has received little attention. LaVenue & Pickens (1992) and Ramarao *et al.* (1995) employ adjoint sensitivity analysis and estimate optimal pilot points locations on the basis of their potential to minimize a selected objective function. Wen *et al.* (2005, 2006) couple Sequential Self Calibration (SSC) with a genetic algorithm during inversion for a synthetic reservoir model. The genetic algorithm is employed to: (a) update reservoir properties at pilot points locations, and (b) select an optimal spatial distribution of pilot points. An optimal set of pilot points locations is not detected for their example. Christensen & Doherty (2008) deal with the same problem by using a singular value decomposition of the sensitivity matrix of the pilot points-based inverse model. Pilot points are distributed uniformly over the simulation domain, and the model is calibrated with a small number of super parameters, corresponding to the largest eigenvalues of the sensitivity matrix. The authors recommend adopting a dense pilot points network and a large number of super parameters.

Here, we address this issue in the context of stochastic Moment Equations of flow. Our approach is based on the work of Hernandez *et al.* (2006) and Riva *et al.* (2009). These authors developed nonlinear stochastic inverse methods and algorithms to condition estimates of steady state and transient hydraulic heads, fluxes and their associated uncertainty on information about

Copyright © 2012 IAHS Press

measured values of *Y* and hydraulic head. Their methodology is based on zero- and second-order approximations of exact nonlocal stochastic first and second moment equations of groundwater flow (Guadagnini & Neuman, 1999). Estimates of *Y* at measurement and pilot points locations is framed in the context of Maximum Likelihood (ML) theory. The negative log likelihood criterion (Carrera & Neuman, 1986):

$$NLL = \left(\hat{\mathbf{h}} - \mathbf{h}^*\right)^T \mathbf{C}_\mathbf{h}^{-1} \left(\hat{\mathbf{h}} - \mathbf{h}^*\right) + \left(\hat{\mathbf{Y}} - \mathbf{Y}^*\right)^T \mathbf{C}_\mathbf{Y}^{-1} \left(\hat{\mathbf{Y}} - \mathbf{Y}^*\right) + \ln\left|\mathbf{C}_\mathbf{Y}\right| + \ln\left|\mathbf{C}_\mathbf{h}\right| + N \ln 2\pi \tag{1}$$

is minimized with respect to the model parameters. In equation (1), $\hat{\mathbf{h}}$ is a vector of conditional mean heads predicted at N_h observation points, $\mathbf{h}^*$ is the vector of head measurements, $\hat{\mathbf{Y}}$ is an inverse estimate of log-conductivity at N_M measurement points and N_P pilot points, $\mathbf{Y}^*$ represents measured values of Y at measurement and pilot points locations. Prior values of *Y* at pilot points are evaluated via kriging. $\mathbf{C}_h = \sigma_{hE}^2 \mathbf{V}_h$ is the covariance matrix of head measurement errors, $\mathbf{C}_Y = \sigma_{YE}^2 \mathbf{V}_Y$ is the covariance matrix of *Y* measurement errors, σ_{hE}^2 and σ_{YE}^2 are (typically unknown and hence to be estimated) statistical scaling parameters; $\mathbf{V}_h$ and $\mathbf{V}_Y$ are known symmetric positive-definite matrices, $N_Y = N_M + N_P$ and $N = N_h + N_Y$. The covariance matrix $\mathbf{C}_Y$ is expressed as:

$$\mathbf{C}_\mathbf{Y} = \begin{bmatrix} \mathbf{C}_\mathbf{YM} & 0 \\ 0 & \mathbf{C}_\mathbf{YP} \end{bmatrix} \tag{2}$$

where $\mathbf{C}_{YM}$ is the diagonal covariance matrix of *Y* measurement errors and $\mathbf{C}_{YP}$ is the (generally) non-diagonal covariance matrix of *Y* estimation errors at pilot points.

If the variogram of *Y*, σ_{hE}^2 and σ_{YE}^2 are fixed, minimizing equation (1) is equivalent to minimization of the regularized general least squares criterion:

$$J = \left(\hat{\mathbf{h}} - \mathbf{h}^*\right)^T \mathbf{V}_\mathbf{h}^{-1} \left(\hat{\mathbf{h}} - \mathbf{h}^*\right) + \lambda \left(\hat{\mathbf{Y}} - \mathbf{Y}^*\right)^T \mathbf{V}_\mathbf{Y}^{-1} \left(\hat{\mathbf{Y}} - \mathbf{Y}^*\right) \tag{3}$$

where $\lambda = \sigma_{hE}^2 / \sigma_{YE}^2$ is a regularization term.

The Bayesian criterion *BIC* (Schwarz, 1978) and information criteria *AIC* (Akaike, 1974), *AICc* (Hurvich & Tsai, 1989) and *HIC* (Hannan, 1980) differ from *NLL* by constant values that depend on N_Y, N_h and *N*. Riva *et al.* (2010, 2011) find that measurement error variances and key parameters of the variogram of *Y* can be estimated properly for a fixed set of pilot points locations by minimizing the Bayesian model discrimination criterion *KIC* (Kashyap, 1982):

$$KIC = NLL + N_Y \ln\left(N/2\pi\right) - \ln\left|\mathbf{Q}\right| \tag{4}$$

where $\mathbf{Q}$ is the estimation errors covariance matrix. In the following sections we explore the ability of a zero-order approximation of moment equations of flow to estimate the optimal locations of a prescribed number of pilot points.

NUMERICAL MODEL SET-UP

We consider the synthetic test case presented by Hernandez *et al*. (2006). Convergent flow is superimposed to mean uniform flow in a rectangular domain of length 18 and width 8 (all quantities are given in arbitrary consistent units). The domain is discretized into N_e = 3600 square elements of uniform size equal to 0.2. The Sequential Gaussian simulator GCOSIM3D (Gómez-Hernandez & Journel, 1993) is employed to generate a single unconditional realization of *Y* at block centres having zero mean and an exponential isotropic variogram with sill σ_Y^2 = 4.0 and integral scale I_Y = 1.0. Deterministic head values of 10 and 0 are prescribed along the left and right boundaries, respectively. The top and bottom boundaries are considered to be impervious. A well is located at the domain centre (see Fig. 1) and pumps continuously at a constant unit rate. The corresponding forward steady-state flow problem is solved numerically to yield head values at all

grid nodes (see Fig. 1). These generated log-conductivities and heads constitute our reference values. We sample the head field at N_h = 36 measurement points and the Y field at 16 points. We then superimpose white Gaussian noise (measurement error) of unit variance to each set of measurements ($\sigma_{YE}^2 = \sigma_{hE}^2 = 1; \mathbf{V}_{YM} = \mathbf{I}; \mathbf{V}_h = \mathbf{I}$).

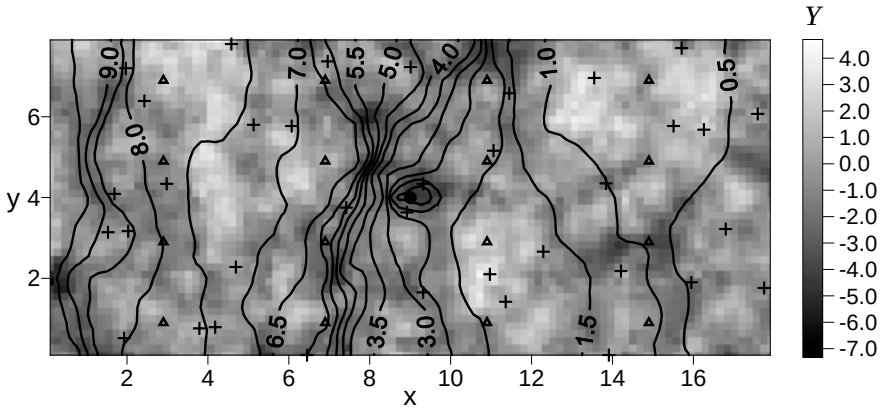

Fig. 1 Reference Y field, spatial distribution of Y (△) and h (+) conditioning measurements and pumping well location (●). Contour lines indicate the reference h field corresponding to a pumping rate Q = 1.

For demonstration purposes, we take the parameters of the variogram of Y to be given and pre-define the number of pilot points, N_P. Our goal is to estimate: (a) Y values at measurement and pilot points locations, (b) the statistical parameters σ_{YE}^2 and σ_{hE}^2, and (c) the optimal pilot points locations. These objectives are accomplished following the methodology proposed by Riva *et al.* (2010, 2011). The space of candidate locations of pilot points is searched by a differential evolution method (DEM).

RESULTS AND DISCUSSION

A critical point in the application of a DEM is the choice of the objective function to minimize for a given purpose. Here, we focus on the estimate of the optimal spatial distribution of pilot points for a given value of N_P. In our example we set N_P = 64. The total number of candidate spatial configurations is $N_e!/\left[N_P!\left(N_e - N_P\right)!\right] \approx 1.8 \times 10^{138}$ for our numerical example. It is noted that, when the location of pilot points changes between iterations of the inverse problem, $\mathbf{C_{YP}}$ can play an important role in the optimization process. When the spatial distribution of pilot points is such that their associated conductivity values are highly correlated, $\mathbf{C_{YP}}$ can become singular, and $\ln\left|\mathbf{C_Y}\right| = \ln\left|\mathbf{C_{YM}}\right| + \ln\left|\mathbf{C_{YP}}\right| \rightarrow -\infty$ (equation (1)). This implies that spatial arrangements where pilot points are clustered within a (relatively) small area tend to be favoured relative to scenarios where the pilot points are spread uniformly over the domain. These theoretical observations are supported by the numerical results (not shown) obtained minimizing *NLL* to detect the optimal pilot points locations. Considering that the Bayesian criterion *BIC* and the information criteria *AIC*, *AICc* and *HIC* differ from *NLL* by a constant, we conclude that none of these criteria is suitable to our purpose, at least in a situation where the number of pilot and (Y) measurement points is fixed.

In the following we explore the capability of *KIC* to identify the optimal pilot points arrangement. We do so by embedding the code INME (Riva *et al.*, 2011) into the genetic based algorithm DEM (Storn & Price, 1997). The convergence process of DEM relies on a heuristic search algorithm that mimics the process of natural evolution. This process is performed through the steps of initialization, mutation, cross-over and selection.

In the initialization step we generate a population of vectors $\mathbf{x}_{i,G=1}$, where G is the generation number, and $i = 1, \ldots, SP$. In our example we set $SP = 100$. Each vector contains the coordinates of the N_P pilot points and the values of the unknown statistical parameters. The locations of the pilot points are chosen randomly from the set of the grid element centroids, in such a way that the grid cells are associated with equal probability to be assigned a pilot point and that no more than one pilot point or one Y measurement can be found in the same grid element. With reference to the statistical parameters σ^2_{YE} and σ^2_{hE}, we extract two random values from a continuous uniform probability density function defined on the interval $\left[10^{-6};10^{4}\right]$. This allows use of the code INME to perform an inversion of the flow problem for each of the SP vectors and to calculate the corresponding value of KIC.

A set of SP mutant vectors is created in the mutation step by:

$$\mathbf{v}_{i,G+1} = \mathbf{x}_{best,G} + K\left(\mathbf{x}_{r1,G} + \mathbf{x}_{r2,G} - \mathbf{x}_{r3,G} - \mathbf{x}_{r4,G}\right) \tag{5}$$

where $\mathbf{x}_{best,G}$ is the vector of the set $\mathbf{x}_{i,G}$ with the smallest value of KIC; $r1$, $r2$, $r3$ and $r4 \in [1; SP]$ are random different integers. We set $K = 0.3$ in our simulation.

A third set of trial vectors $\mathbf{u}_{i,G+1}$, $i = 1,\ldots,SP$ is then created through the cross-over step. The value of the component of $\mathbf{x}_{i,G}$ or $\mathbf{v}_{i,G+1}$ is assigned to each corresponding component of a cross-over vector $\mathbf{u}_{i,G+1}$ with probability CR or $(1-CR)$, respectively. In our example we adopt a value of $CR = 0.5$. Computation of the value of KIC corresponding to each element $\mathbf{u}_{i,G+1}$ is performed after the population of trial vectors $\mathbf{u}_{i,G+1}$, $i = 1, \ldots, SP$ has been determined.

A new generation of vectors $\mathbf{x}_{i,G+1}$ is then created. This is accomplished for each $i = 1, \ldots, SP$ by setting $\mathbf{x}_{i,G+1} = \mathbf{x}_{i,G}$ or $\mathbf{x}_{i,G+1} = \mathbf{u}_{i,G+1}$, respectively, depending on whether the value of KIC corresponding to the component of the vector $\mathbf{x}_{i,G}$ is smaller or larger than the value of KIC associated with vector $\mathbf{u}_{i,G+1}$.

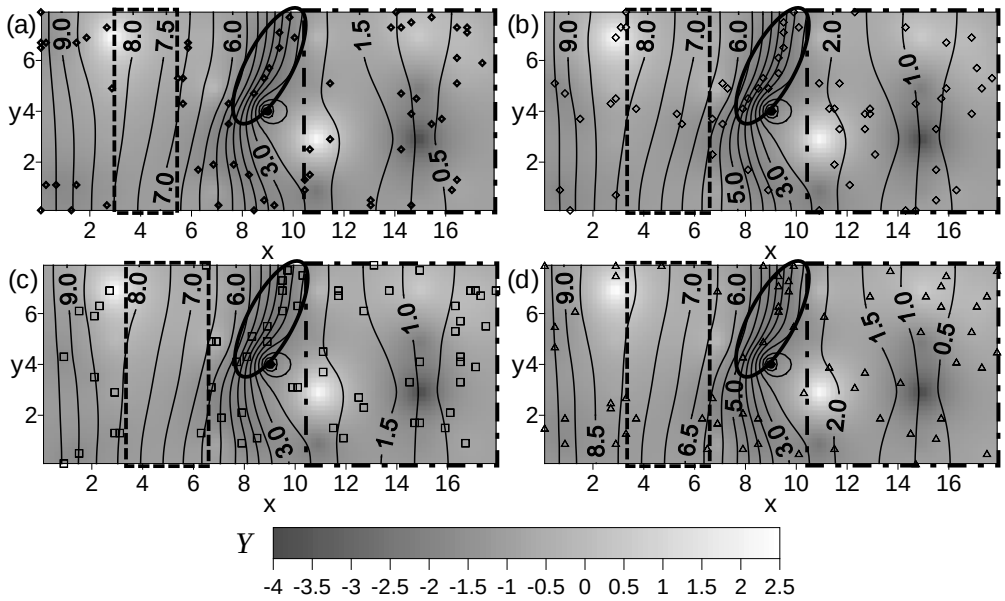

Fig. 2 Optimal pilot points configurations obtained by minimization of KIC and estimated Y fields, with corresponding colour scale. Contour lines indicate estimated h fields.

The mutation, cross-over and selection steps are iterated until convergence is reached (i.e. when $\mathbf{x}_{i,G+1} \approx \mathbf{x}_{i,G}$ for all $i = 1, \ldots, SP$). It is then clear that the convergence process of DEM is associated with some elements of randomness. These affect: (a) the distribution of the search parameters of the initial population vectors, and (b) the mutation and cross-over processes, which

drive the generation of the new trial vectors (offspring) from the current population. This is the reason why simulations performed starting from different seeds initializing a Random Number Generator (RNG) will employ different sequences of random numbers and will in general produce different outcomes. We explore this issue performing four different inversions, each associated with a different seed in the RNG. The common characteristics shared by the results of the simulations can be interpreted as the main-features of the (unknown) optimal spatial arrangement of the pilot points. Figure 2 depicts the optimal pilot points distribution obtained for each of the four simulations. Figure 3 reports the evolution of the values of the *KIC* criterion during the optimization process. Our results show that the algorithm converges to a stable value in less than 20 iterations for all the simulations.

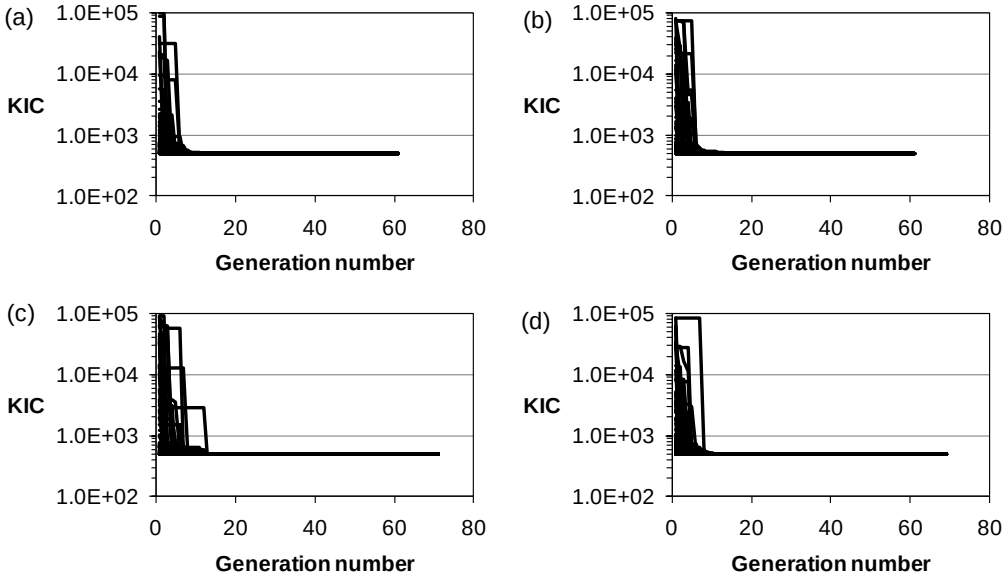

Fig. 3 Convergence of the *KIC* criterion during the optimization of pilot points locations.

Although the final pilot points configurations do not coincide for the four cases analysed, they all exhibit some common features (Fig. 2): (a) an area, identified with the (left) dashed rectangle, where the concentration of pilot points is small relative to the rest of the domain; (b) a zone (marked by the ellipse) with a large concentration of pilot points; and (c) a part of the domain (marked by the dashed-dotted rectangle) associated with a more-or-less uniform distribution of pilot points. These common elements appear to be correlated with the gradient of the reference (Fig. 1) and estimated (Fig. 2) hydraulic head fields, as the largest pilot points density occurs within regions of steep hydraulic head gradient.

We conclude our note with the analysis of the influence of the number of *Y* conditioning measurements, N_M, on the optimal distribution of pilot points. We set the seed of the RNG equal to the value adopted for the simulation depicted in Fig. 2(d) and perform the inversion after removing the four measurement of *Y* located at x = 6.9 from the conditioning data set (see Fig. 1), resulting in $N_M = 12$. Figure 4 displays the optimal pilot points setting obtained for this scenario, showing the effect of a decreased number of conditioning measurements on the distribution of pilot points in the region where the *Y* measurements have been disregarded. Comparing Figs 2(d) and 4 reveals that a significant number of pilot points have migrated from the left side of the left rectangular region (Fig. 2(d), $x \approx 1 \div 2$) towards the interior part of the area (at about $x \approx 4.5$, Fig. 4). This shift has been induced by the decreased number of measurement points adopted. We note that the density of pilot points remains large within the area delineated by the ellipse (Fig. 4) and pilot points tends to remain uniformly distributed for x > 10.

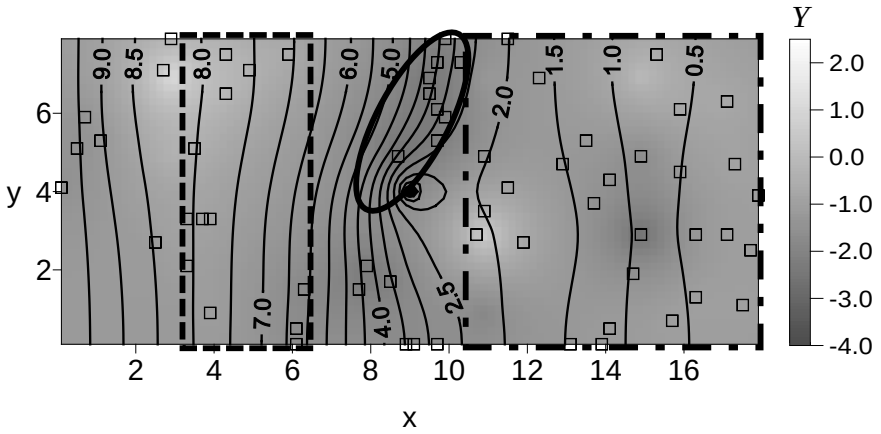

Fig. 4 Optimal pilot points configuration and estimated Y field, with corresponding colour scale. Contour lines indicate the estimated h field. Results are obtained for $N_M = 12$.

CONCLUSIONS

Our examples show that embedding the geostatistical inversion of groundwater flow Moment Equations within a genetic-based sampling algorithm allows identifying an optimal spatial setting for pilot points. These preliminary numerical results suggest that it is beneficial to concentrate the pilot points where the largest values of the hydraulic head gradient are expected, i.e. within regions where strong log-conductivity contrasts may occur or in the proximity of source/sink terms. On the other hand, the pilot points can be distributed uniformly in the areas of the domain where the spatial variation of the head field is relatively smooth. Another relevant aspect is related to the distribution of the available conditioning measurements of conductivity. Our simulations suggest locating the pilot points to compensate for the lack of measurements in given regions of the system.

REFERENCES

Alcolea, A., Carrera, J. & Medina, A. (2006) Pilot points method incorporating prior information for solving the groundwater flow inverse problem, *Adv. Water Resour.* 29(11), 1678–1689, doi:10.1016/j.advwatres.2005.12.009.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19(6), 716–723, doi:10.1109/TAC.1974.1100705.

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

Christensen, S. & Doherty, J. (2008) Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration. *Adv. Water Resour.* 31(4), 674–700, doi:10.1016/j.advwatres.2008.01.003.

de Marsily, G., Lavedan, C., Bouchere, M. & Fasanino, G. (1984) Interpretation of interference 614 tests in a well field using geostatistical techniques to fit the permeability distribution in a 615 reservoir model. In: *Geostatistics for Natural Resources Characterization*, Part 2. NATO ASI 616 Ser., Ser. C 182 (ed. by G. Verly *et al*.), 831–849. D. Reidel Publ. Co., Dordrecht, The Netherlands.

Gomez-Hernandez, J. J. & Journel, A. (1993) Joint sequential simulation of multiGaussian fields, In: *Geostatistics Tròia '92* (ed. by A. O. Soares), 85–94. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Guadagnini, A. & Neuman, S. P. (1999) Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains 1. Theory and computational approach, *Water Resour. Res.* 35(10), 2999–3018, doi:10.1029/1999WR900160.

Hannan, E. J. (1980) The estimation of the order of an ARMA process. *Ann. Stat.* 8(5), 1071–1081, doi: 10.1214/aos/1176345144.

Hernandez, A. F., Neuman, S. P., Guadagnini, A. & Carrera, J. (2006) Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media, *Water Resour. Res.* 42(5), W05425, doi:10.1029/2005WR004449.

Hurvich, C. M. & Tsai, C. L. (1989) Regression and time-series model selection in small samples. *Biometrika* 76(2), 297–307, doi:10.2307/2336663.

Kashyap, R. L. (1982) Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE T. Pattern Anal.* 4(2), 99–104, doi:10.1109/TPAMI.1982.4767213.

LaVenue, A. M. & Pickens, J. F. (1992) Application of a coupled adjoint sensitivity and kriging approach to calibrate a groundwater-flow model. *Water Resour. Res.* 28(6), 1543–1569, doi:10.1029/92WR00208.

Ramarao, B., LaVenue, A., De Marsily, G. & Marietta, M. (1995) Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields 1. Theory and Computational Experiments. *Water Resour. Res.* 31(3), 475-493, doi:10.1029/94WR02258.

Riva, M., Guadagnini, A., Neuman, S. P., Bianchi Janetti, E. & Malama, B. (2009) Inverse analysis of stochastic moment equations for transient flow in randomly heterogeneous media. *Adv. Water Resour.* 32(10), 1495–1507, doi:10.1016/j.advwatres.2009.07.003.

Riva, M., Guadagnini, A., De Gasperi, F. & Alcolea, A. (2010) Exact sensitivity matrix and influence of the number of pilot points in the geostatistical inversion of moment equations of groundwater flow, *Water Resour. Res.* 46, W11513, doi:10.1029/2009WR008476.

Riva, M., Panzeri, M., Guadagnini, A. & Neuman, S. P. (2011) Role of model selection criteria in geostatistical inverse estimation of statistical data- and model- parameters. *Water Resour. Res.* 47, W07502, doi:10.1029/2011WR010480.

Storn, R. & Price, K. (1997) Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, *J. Global Optim.* 11(4), 341–359, doi:10.1023/A:1008202821328.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Stat.*, 6(2), 461–464, doi:10.1214/aos/1176344136.

Wen, X.-H., Yu, T. & Lee, S. (2005) Coupling Sequential-Self calibration and Genetic Algorithms to Integrate Production Data in Geostatistical Reservoir Modeling. In: *Geostatistics Banff 2004* (ed. by O. Leuangthong & C. V. Deutsch), 691–701, Springer, The Netherlands.

Wen, X.-H., Lee, S. & Yu, T. (2006) Simultaneous integration of pressure, water cut, 1 and 4-D seismic data in geostatistical reservoir modeling. *Math. Geol.* 38(3), 301–325, doi:10.1007/s11004-005-9016-6.