

When it is not normal to be Gaussian

Quando non è normale essere gaussiano

J. Jaime Gómez-Hernández and Teng Xu

Abstract The parameters characterizing geological formations display spatial patterns that are very often not amenable to be modeled using a multiGaussian random function. Multimodal marginal distributions, spatially connected low or high values and curvilinear features are some of the observed characteristics that a multiGaussian random function will fail to reproduce. Hydraulic conductivity is one such parameter, of great importance for the prediction of the movement of groundwater. Many efforts have been carried out in the last decades to create algorithms capable to generate hydraulic conductivity realizations according to non-Gaussian random functions. And, in the last years, these efforts have shifted to creating algorithms for the stochastic inverse modeling of hydraulic conductivity in a non-Gaussian context. The ensemble Kalman filter is one of the methods which have been favored in stochastic inverse modeling; however, it fails when hydraulic conductivity is best modeled as non-Gaussian. This paper discusses why it fails and shows an alternative algorithm based on a simple modification of the standard formulation of the ensemble Kalman filter.

Abstract *I parametri che caratterizzano le formazioni geologiche presentano configurazioni spaziali che non sono spesso suscettibili di essere modellate mediante una funzione aleatoria Gaussiana multivariata. Distribuzioni marginali multimodali, valori bassi o elevati spazialmente collegati e configurazioni curvilinee sono alcune delle caratteristiche osservate che una funzione aleatoria Gaussiana multivariata non è in grado di riprodurre. La conducibilità idraulica appartiene a tali parametri ed è di grande importanza per la previsione del movimento delle acque sotterranee. Negli ultimi decenni sono stati effettuati molti sforzi al fine di produrre algoritmi in grado di generare realizzazioni della conducibilità idraulica*

J. Jaime Gómez-Hernández

Research Institute for Water and Environmental Engineering, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain e-mail: jgomez@upv.es

Teng Xu

Research Institute for Water and Environmental Engineering, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain e-mail: tenxu@posgrado.upv.es

basate su funzioni aleatorie non-Gaussiane. In particolare, negli ultimi anni, questi sforzi hanno riguardato la predisposizione di algoritmi per la modellazione stocastica inversa della conducibilità idraulica in un contesto non- Gaussiano. L'ensemble Kalman filter è uno dei metodi preferiti nella modellazione stocastica inversa; tuttavia tale metodo fallisce se la conducibilità idraulica è modellata come non-Gaussiana. In questo articolo, si chiarisce il motivo di tale performance e si illustra un algoritmo alternativo basato su una semplice modifica della formulazione standard dell'ensemble Kalman filter.

Key words: stochastic inverse modeling, ensemble Kalman filter, non-Gaussian, hydraulic conductivity

1 Introduction

Back in 1993, we had the opportunity to make a presentation at the Italian Statistical Society meeting warning about “The dangers of parsimony” associated to the use of multiGaussian models in the Earth Sciences [2]. That presentation gave rise to a short paper published in the Journal of the Italian Society [3] followed, five years later!, by a longer paper published in Advances in Water Resources [4] — after it being rejected for publication in another journal— in which we insisted on the importance of drifting away from the established hypothesis that earth science parameters had to be modeled as multiGaussian random functions. The latter paper, initially dismissed as non-academic, has received, today, more than 150 citations and is one of our most cited works. Few researchers deny today that the spatial variability patterns observed in geologic formations are seldom amenable to a multi-Gaussian model, that the observed channel-like structures, the continuity streaks, the impermeable barriers are best modeled with stochastic processes that are based on higher-order statistics, not just the two-point covariance that fully characterizes a multiGaussian probability distribution.

When multiGaussian models are still used is because there are no better alternatives from an analytical or numerical point of view. For example, in inverse modeling, the ensemble Kalman filter (EnKF) [1] has been proven to be a very efficient method for stochastic inverse modeling in many disciplines. The EnKF is a variant of the Kalman filter [5] aimed at overcoming the inefficiencies of the Kalman filter when the state equation is non-linear. It is well known that the Kalman filter is optimal for linear transfer functions and for state parameters that can be modeled as multi-Gaussian. The EnKF works around the non-linearity of the state function but does not work well when either the state variables or the parameters controlling the state equation are non-Gaussian.

Next, we present the standard formulation of the EnKF and show its performance in a non-Gaussian setting, and we follow with a simple alternative that solves the problem.

2 The Ensemble Kalman Filter

The Ensemble Kalman filter is an assimilation algorithm that incrementally updates the estimates of the parameters controlling a state equation, on the basis of the discrepancy between predictions and observations as new state data are acquired in time. Consider a phenomenon, such as groundwater flow, in which the state of the system y at time step k is determined as a function of the system state at time step $k - 1$ and the latest estimate of the parameters x controlling the state equation:

$$\mathbf{Y}_k = f(\mathbf{Y}_{k-1}, \mathbf{X}_{k-1}), \quad (1)$$

where f represents the numerical model approximating the state equation, and $\mathbf{Y}$ and $\mathbf{X}$ are vectors containing state and parameter values, respectively, on all the cells of the numerical model.

The Kalman filter needs the auto- and cross-covariances of $\mathbf{Y}$ and $\mathbf{X}$ to compute the updates of both $\mathbf{Y}$ and $\mathbf{X}$ once the state variable has been observed at a number of locations —generally much smaller than the number of cells in the model. When the state equation is linear, these covariances are easily obtainable after the covariance of x . The ensemble Kalman filter solves this problem for non-linear state equations by computing numerically these covariances at any time step k from an ensemble of realizations of the parameters $\mathbf{X}_k$ and their associated responses $\mathbf{Y}_k = f(\mathbf{X}_k, \mathbf{Y}_{k-1})$. The standard EnKF flowchart would be as follows:

1. Initialization of the ensemble. Generate a set of N realizations, $\mathbf{X}_0^i, i = 1, \dots, N$. At this point, no state data has been collected, therefore the generation is only conditional to whatever x data are available; no inverse conditioning yet. Set the state initial conditions $\mathbf{Y}_0^j, j = 1, \dots, N$ —generally the same for all $\mathbf{X}$ realizations.
2. Assimilation loop. Repeat for $k = 1, \dots, K$ with K being the total number of time steps for which state data will be collected.
 - a. Forecast. Predict the state of the system for each member of the ensemble at time step k , $\mathbf{Y}_k^i = f(\mathbf{Y}_{k-1}^j, \mathbf{X}_{k-1}^j), j = 1, \dots, N$.
 - b. Covariance calculation. Compute the experimental auto- and cross-covariances C_{yy}, C_{xy} from the ensemble of realizations, i.e.,

$$C_{yy, i_1, i_2} = \frac{1}{N} \sum_{j=1}^N (y_{i_1} - \bar{y}_{i_1})(y_{i_2} - \bar{y}_{i_2}), \quad (2)$$

where C_{yy, i_1, i_2} is the element of the state covariance matrix associated to cells i_1 and i_2 , and the overbar indicates average value computed through the ensemble, i.e.,

$$\bar{y}_{i_1} = \frac{1}{N} \sum_{j=1}^N y_{i_1}. \quad (3)$$

- c. Update. For each member of the ensemble, update the parameters based on the discrepancy between observed and predicted states

$$\mathbf{X}_k = \mathbf{X}_{k-1} + C_{xy}(C_{yy} + C_D)^{-1}(\mathbf{Y}_k^{\text{obs}} - \mathbf{Y}_k^{\text{for}}), \quad (4)$$

where $\mathbf{Y}_k^{\text{obs}}$ is a vector containing just the observed state at the measurement locations—we will assume that the measurement locations coincide with prediction locations by the numerical model—, $\mathbf{Y}_k^{\text{for}}$ is a vector containing the forecast values predicted by the model at measurement locations, and C_D is a prescribed measurement error covariance matrix—generally diagonal.

As we have already mentioned, the numerical calculation of the covariance in step 2b circumvents the problem of the non-linearity of the transfer function, as is the case in groundwater modeling, where x will be permeability, y will be piezometric head, and f will be a groundwater flow model. However, the fact that the update step relies exclusively in two-point statistics such as covariances and cross-covariances makes that the repeated application of this updating equation—as new state data are assimilated—will render the parameter estimates Gaussian, even if at the initialization step they were far from Gaussian.

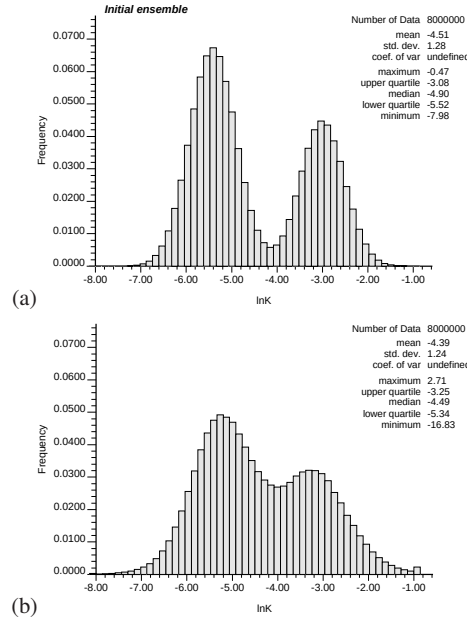


Fig. 1 Logpermeability histograms for (a) the initial ensemble of 500 realizations and (b) after 50 time steps of data assimilation.

Figure 1 top shows the histogram of the initial set of logpermeabilities displaying a clear bimodal distribution related to the two distinct facies associated to channel and non-channel materials in the aquifer. Figure 1 bottom shows the histogram of the updated logpermeabilities after repeatedly applying the update equations for 50 time steps. It is quite noticeable how the bimodal distribution is shifting towards a more Gaussian-like one. This departure of the marginal distribution from the initial one is not affecting the reproduction of the state variables at the measurement locations. Figure 2 shows how the initial ensemble of logpermeabilities fails to predict the state at a given location, whereas all realizations of the updated ensemble are capable of a good reproduction of the observed state.

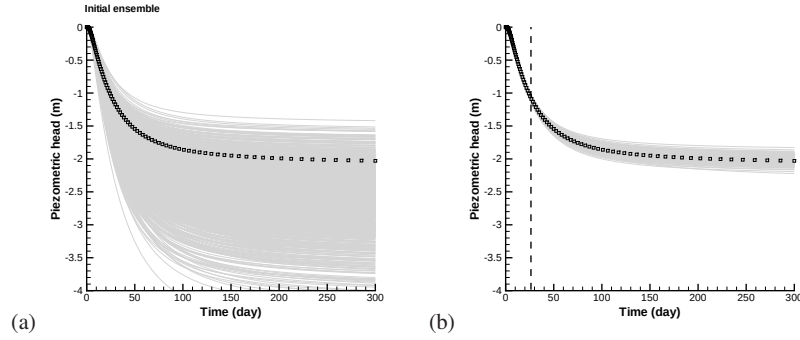


Fig. 2 Reproduction of the observed piezometric heads at a given location (a) for each member of the initial ensemble of 500 realizations and (b) for each of the updated realizations after 50 time steps of data assimilation.

3 A variant of the EnKF

Since, the non-linearity of the state function is not a problem for the EnKF to generate realizations matching the observed state data, Zhou *et al.* [6] have proposed to apply a univariate transformation to parameters and state variables so that the transformed variables are marginally Gaussian (not yet multiGaussian) maintaining a relationship through a new state equation that is even more non-linearly than the original one. If we apply normal-score transforms to both x and y , to yield new variables $x' = \Phi_x(x)$ and $y' = \Phi_y(y)$, we can envision applying the EnKF on these two new variables, which, now will be related by a new state equation:

$$\mathbf{Y}'_k = g(\mathbf{Y}'_{k-1}, \mathbf{X}'_{k-1}), \quad (5)$$

where g represents a convolution of the groundwater flow equation and the normal-score transform functions.

This simple transformation produces results that respect the bimodal distribution of logpermeability without compromising the reproduction of the observed state data, as can be seen in Figure 3

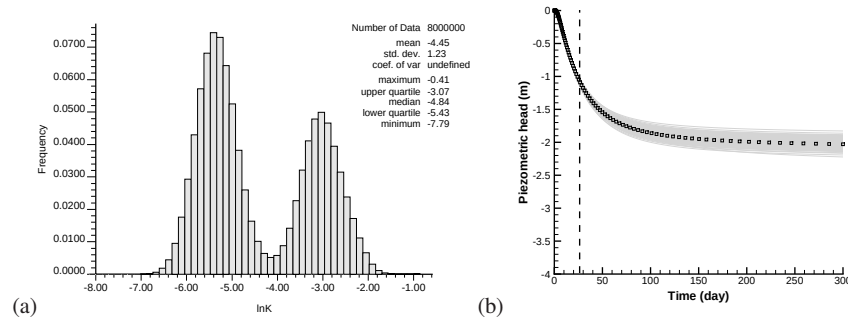


Fig. 3 Applying the EnKF to the normal-score transformed variables. Results after 50 time steps of data assimilation: (a) Logpermeability histogram after 50 assimilation steps, (b) reproduction of the observed piezometric heads at a given location for each member of the 500 updated realizations.

Acknowledgements The authors gratefully acknowledge the financial support by the Spanish Ministry of Economy and Competitiveness through grant CGL2011-23295.

References

1. G. Evensen. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.
2. J. J. Gómez-Hernández and X.-H. Wen. MultiGaussian models: The danger of parsimony. In V. Capasso, G. Girone, and D. Posa, editors, *Statistics of Spatial Processes: Theory and Applications*. IRMA, 1993.
3. J. J. Gómez-Hernández and X.-H. Wen. Multigaussian models: The danger of parsimony. *Journal of the Italian Statistical Society*, 4(2):167–181, 1995.
4. J. J. Gómez-Hernández and X.-H. Wen. To be or not to be multiGaussian: A reflection in stochastic hydrogeology. *Advances in Water Resources*, 21(1):47–61, 1998.
5. R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.
6. H. Zhou, J. J. Gómez-Hernández, H.-J. Hendricks Franssen, and L. Li. An approach to handling non-gaussianity of parameters and state variables in ensemble kalman filtering. *Advances in Water Resources*, 34(7):844–864, 2011.