Mathematical Geosciences manuscript No.
(will be inserted by the editor)

Ensemble smoother with multiple data assimilation as a tool for curve fitting and parameter uncertainty characterization: Example applications to fit non-linear sorption isotherms

Vanessa A. Godoy · Gian F. Napa-García · J. Jaime Gómez-Hernández

Received: date / Accepted: date

Abstract The ensemble smoother with multiple data assimilation (ES-MDA) coupled to a normal-score transformation is used to fit a Langmuir isotherm curve to estimate its parameters (S_m and b) and their uncertainty. The highlights of this work are three: i) the ES-MDA can be used as a curve fitting procedure, ii) the ES-MDA provides also a full uncertainty quantification about the fitted parameters and iii) for the specific case of the Langmuir isotherm, parameter S_m is well identified with little uncertainty, while parameter b is well identified with a larger uncertainty, indicative that solute concentrations are more sensitive to S_m than to b. As a by-product, the number of samples required to characterize the joint uncertainty of Langmuir isotherm parameters is also investigated; it can be concluded that the minimum number of samples to use is six, with best results

Vanessa A. Godoy ✉
Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain
E-mail: godoyalmeida@gmail.com

Gian F. Napa-García
Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain

J. Jaime Gómez-Hernández
Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain

obtained with eight samples, a value larger than the number recommend in the literature.

Keywords Tracer tests · Inverse modeling · Solute transport · Batch test · Bayesian methods

Acknowledgements The first author acknowledges the financial support from the Schlumberger Foundation by means of the program Faculty for the Future. The last author wishes to acknowledge the financial contribution of the Spanish Ministry of Science and Innovation through project number PID2019-109131RB-I00.

Declarations

Funding

This research has been supported by the Spanish Ministry of Science and Innovation through project number PID2019-109131RB-I00 and by the Schlumberger Foundation by means of the program Faculty for the Future.

Conflicts of interest/Competing interests

Not applicable

Availability of data and material

Not applicable

Code availability

Not applicable

1 Introduction

The retention of chemical constituents through the transfer of ions from the aqueous phase (sorptives) to the solid phase (sorbent) is widely known as sorption. To quantify the sorption characteristics of a soil, batch experiments are usually performed in which the relationship between the equilibrium concentrations of the sorptive and the sorbate is measured for a variety of sorptive concentrations while holding temperature constant; the outcome of this experiment results in what is known as a sorption isotherm, an example of which is shown in Fig. 1. The most common sorption isotherms are linear, Freundlich, and Langmuir isotherms.

The transport of disolved contaminants in soils mostly depends on the sorption capacity of soils and rocks (Pathak and Sharma 2018). The sorption processes are important in a variety of applications in the field of geosciences, including remediation systems design (Luo et al. 2014), the selection of a waste site (Bouchelaghem 2018), the design of groundwater pump-and-treat systems or the evaluation of natural attenuation and salinity (Park et al. 2007; Tavakoli-Kivi et al. 2019). Most of these applications involve at some point the use of numerical models, in which the parameters defining the isotherm equation are required as input data (Guo et al. 2019; Masood and Abd Ali 2020). An accurate determination of these parameters is crucial for good transport predictions, especially when these models are used to perform quantitative risk analysis (Capilla et al. 1998; Franssen and Gómez-Hernández 2002; Gómez-Hernández and Wen 1994; Hinz et al. 1994; Li et al. 2011; Zhou et al. 2011; Fairey and Wahman 2013). Hinz et al. (1994) studied the sensitivity of solute transport predictions on sorption isotherm parameters by quantifying the ratio of the input concentration to the maximum sorption capacity; the results show that the retardation of contaminants are highly influenced by the Langmuir isotherm parameters.

Determining sorption isotherm parameters based on data from laboratory experiments is always difficult and involves uncertainties. The chemical composition of the solute, the physical structure of the sorbent, the effect of temperature,

measurement errors, and natural variations in soils and rocks are examples of sources of uncertainty in laboratory experiments. This uncertainty, then, propagates through the isotherm fitting process onto the isotherm parameters. It is also still unclear how many experiments should be run to get a good fit of the isotherm curve with small estimate uncertainties. Roy et al. (1991), in his thorough analysis of batch-type adsorption procedures for estimating soil attenuation of chemicals, recommends a minimum of five experiments to fit the isotherm parameters.

Although there exist a vast number of papers related to sorption isotherms, few deal with the uncertainty associated with parameter estimation from batch experiments. Fairey and Wahman (2013) compared weighted regression with Markov chain Monte Carlo (McMC) to determine the joint uncertainty of Freundlich sorption isotherm fitting parameters, and found that both frequentist and Bayesian analyses reduced the uncertainty in parameters compared with treating the uncertainty independently, and that the difference between the two techniques becomes more pronounced as the degree of non-linearity in the isotherm increases. In a recent study, a hierarchical Bayesian model combined with McMC was satisfactorily used to estimate parameters from multiple sources of experimental data of sorption and to characterize the uncertainty of Langmuir isotherm parameters (Shih et al. 2020).

In this work, several numerical examples were performed to propose a procedure to fit non-linear isotherm curves, and to analyze how the number of experimental pairs used affects the estimated values and their uncertainty. The example is applied to the fitting of the two parameters that define the Langmuir isotherm and the two parameters that define the Freundlich isotherm, the expressions of which will be introduced below. Additionally, we investigate the influence of the ensemble size, the measurement-error magnitude, and the number of ES-MDA iterations on the uncertainty estimation.

The good results obtained in this specific context allows us to make a recommendation that the ES-MDA be used for curve-fitting in the general sense, not

limited to fitting isotherm curves. Traditionally this type of fitting is done using least-square approaches, yielding reasonable results; however least-squares will not provide an estimation of the uncertainty of the estimates, something that the ES-MDA will do independently of the prior distributions adopted for the parameters to be fitted. Also, the analysis of the stabilization of the statistics of the posterior distributions of the fitted parameters can be used as a tool to determine how many samples are needed to obtain reliable parameter estimates, as will be shown.

The proposed procedure is based on the ensemble smoother with multiple data assimilation (ES-MDA) (Emerick and Reynolds 2013; Evensen 2018), a method widely used in geosciences (Emerick 2017; Chen and Oliver 2012; Todaro et al. 2019; 2021; Silva et al. 2021a), and it includes a normal-score transformation to deal with the possible non-Gaussianity of both prior and posterior uncertainties about the parameters (Capilla et al. 1999; Zhou et al. 2011; Li et al. 2012; Xu and Gómez-Hernández 2015; 2016; 2018). Recently, the ES-MDA was satisfactorily used to fit the parameters of a B-Spline curve conditioned to well-test data (Silva et al. 2021b); however, this application is a common approach to solve a history matching exercise in which the parameters to be identified are not material parameters, such as permeability or porosity, but the geometrical parameters (widths and lengths) that define the turbidite lobes in the reservoir. The fact that the authors define a B-Spline curve as a function of the parameters identified does not imply that they are performing curve fitting in its traditional sense as used in this paper. To the best of our knowledge, this is the first paper in which the normal-score ES-MDA is used for curve fitting and estimation uncertainty quantification of an isotherm curve, and that investigates how many samples are necessary for a proper characterization.

The remainder of this paper is organized as follows. After a description of the ES-MDA in Sect. 2, materials and methods are reported in Sect. 3, results are presented and discussed in Sect. 4, conclusions are drawn in Sect. 5, and finally additional synthetic examples are presented in the Appendix.

The work described in this paper was presented as a poster at the 46th Annual Congress of the International Association of Hydrogeologists but it was never published (Gómez-Hernández et al. 2019).

2 The ensemble smoother with multiple data assimilation (ES-MDA)

The ES-MDA algorithm is based on the ensemble smoother (ES) (Burgers et al. 1998), described by Emerick and Reynolds (2013) and Evensen (2018) as an alternative to the ensemble Kalman filter (Xu et al. 2013; Zhou et al. 2012). The ES-MDA is an iterative data assimilation method that updates parameters (in our case, the isotherm parameters) making use of a set of system states (in our case, equilibrium concentrations) and the deviations between the predictions resulting from the current parameter values with respect to the experimental observations. The relationship between parameters and observations must be known and a forward model relating parameters and state variables must be available (in our case, the forward model is simply the isotherm equation).

The assimilation procedure used by the ES-MDA includes an initialization step, to generate N_e parameter realizations through statistical or geostatistical methods, a forecast step, and an update step. In the forecast step, the forward model is solved for each realization i, to obtain model predictions of the system state. Then, the vector $\mathbf{P}$ of model parameters used for the forecast is updated based on the discrepancy between observations and their model predictions. The updated parameter vector $\mathbf{P}^{\mathrm{u}}$ is given, for each realization, by

$$\mathbf{P}_i^{\mathrm{u}} = \mathbf{P}_i + \mathbf{K}\left[\mathbf{Y}^{\mathrm{ob}} + \varepsilon_i^{\mathrm{ob}} - \mathbf{Y}_i\right], \quad \{i = 1, \ldots, N_e\}, \tag{1}$$

where the subscript i refers to a specific realization, $\mathbf{Y}_i$ is the vector of model predictions for realization i, $\mathbf{Y}^{\mathrm{ob}}$ is the vector of state observations, $\varepsilon_i^{\mathrm{ob}}$ is the vector of observation errors for realization i (the observations errors have zero mean and a covariance given by matrix $\mathbf{R}$) and $\mathbf{K}$ is the Kalman gain, given by

$$\mathbf{K} = \mathbf{C}_{P,Y}\left(\mathbf{C}_{Y,Y} + \mathbf{R}\right)^{-1}, \tag{2}$$

where $\mathbf{C}_{Y,Y}$ is the auto-covariance of the state variables and $\mathbf{C}_{P,Y}$ is the cross-covariance between all parameters and state variables, which are computed from the ensemble of realizations as

$$\mathbf{C}_{P,Y} = \frac{1}{N_\mathrm{e} - 1}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\left(\mathbf{P}_\mathrm{i} - \overline{\mathbf{P}}\right)\left(\mathbf{Y}_\mathrm{i} - \overline{\mathbf{Y}}\right)^\mathrm{T}, \tag{3}$$

$$\mathbf{C}_{Y,Y} = \frac{1}{N_\mathrm{e} - 1}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\left(\mathbf{Y}_\mathrm{i} - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_\mathrm{i} - \overline{\mathbf{Y}}\right)^\mathrm{T}, \tag{4}$$

with $\overline{\mathbf{P}}$ and $\overline{\mathbf{Y}}$ being the ensemble means of parameters and predictions, respectively.

In a problem where there are n_p parameters (in our case, n_p will be two, since there are only two parameters in the Langmuir isotherm equation) and n_o observations (in our case, n_o varies between four and sixteen), vectors $\mathbf{P}_i^\mathrm{u}$ and $\mathbf{P}_i$ have sizes $n_p \times 1$, vectors $\mathbf{Y}_i^\mathrm{ob}$, $\varepsilon_i^\mathrm{ob}$, and $\mathbf{Y}$ have sizes $n_o \times 1$, the Kalman gain $\mathbf{K}$ and the covariance $\mathbf{C}_{P,Y}$ are matrices of size n_p x n_o, and the matrices $\mathbf{C}_{Y,Y}$ and $\mathbf{R}$ are of size $n_o \times n_o$. When the observation errors are modeled as uncorrelated, $\mathbf{R}$ is a diagonal matrix. In the covariance matrix calculation, $\overline{\mathbf{P}}$ is a column vector of size $n_p \times 1$ with the average values of each parameter computed through the realizations, $\overline{\mathbf{P}} = \frac{1}{N_e}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\mathbf{P}_i$, and, similarly $\overline{\mathbf{Y}}$ is a column vector of size $n_o \times 1$ with the average values of each parameter computed through the ensemble of realizations, $\overline{\mathbf{Y}} = \frac{1}{N_e}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\mathbf{Y}_i$.

2.1 Dealing with nonlinear state equations

The resulting updated parameters from Eq. (1) will be optimal estimates if, and only if, the state equation is linear. The ES-MDA was proposed to deal with nonlinear systems by iteratively applying this process of forecasting and updating

using the last updated parameters to make the next forecast. This iteration implies that the same data will be assimilated multiple times; for this reason, there is a need to inflate the covariance matrix by a coefficient α_j, at each iteration j, satisfying the following equation Evensen (2018)

$$\sum_{j=1}^{N_j} \frac{1}{\alpha_j} = 1, \tag{5}$$

where N_j is the total number of iterations.

According to the approach proposed by Evensen (2018), to compute α_j, first, it is necessary to select any nonzero value for α'_0, then, the remaining α'_j are calculated as

$$\alpha'_j = \frac{\alpha'_{j-1}}{\alpha_{geo}}, \tag{6}$$

where α_{geo} is a constant that controls the extent of the changes of α'_j from one iteration to the next. Finally, the values from Eq. (6) are normalized to obtain the final coefficients as

$$\alpha_j = \alpha'_j \left(\sum_{j=1}^{N_j} \frac{1}{\alpha'_j} \right). \tag{7}$$

We refer to Evensen (2018) for more details on the computation of the α_j. A modification in the update step is also required to consider the α_j coefficients. The update equation for the ES-MDA results then

$$\mathbf{P}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathbf{P}_i + \mathbf{C}_{P,Y} \left(\mathbf{C}_{Y,Y} + \alpha_j \mathbf{R} \right)^{-1} \left[\mathbf{Y}^{\mathbf{ob}} + \sqrt{\alpha_j} \varepsilon_i^{\mathbf{ob}} - \mathbf{Y}_i \right]. \tag{8}$$

The updating step as presented in Eq. (8) has the limitation of being suboptimal for parameters displaying a non-Gaussian distribution. To take advantage of the fact that the ES-MDA formulation is optimal when dealing with Gaussian parameters, a normal-score transformation can be performed. The advantage of using such transformation is that it can be applied to any prior distribution. After the ES-MDA formulation is applied in Gaussian space, a back transformation recovers the physical meaning of the parameters into the original non-Gaussian

space. In this work, a normal-score transformation is used at each iteration following the work by Zhou *et al.* (2011) in their proposal of the normal-score ensemble Kalman filter (NS-EnKF). The method consists in assuming a non-Gaussian prior for the parameters $\mathbf{P}$, which are transformed into Gaussian parameters $\mathbf{G}$ after applying a Gaussian anamorphosis $\mathrm{T}(\cdot)$, also known as Nataf transformation or normal-score transform (Nataf 1962)

$$\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathrm{T}\left(\mathbf{P}^{\mathrm{u}}_{\mathrm{MDA},i}\right). \tag{9}$$

The forecasting step is performed using $\mathbf{P}$ as input to the state equation, but the updating is performed on the Gaussian parameter vector $\mathbf{G}$ computed after the Gaussian transform of $\mathbf{P}$

$$\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathbf{G}_i + \mathbf{C}_{G,Y}\left(\mathbf{C}_{Y,Y} + \alpha_j \mathbf{R}\right)^{-1}\left[\mathbf{Y}^{\mathbf{ob}} + \sqrt{\alpha_j}\varepsilon_i^{\mathbf{ob}} - \mathbf{Y}_i\right]; \tag{10}$$

notice that the cross-covariance between parameters and state is computed on the Gaussian transform of the parameters. Finally, the updated or posterior distribution is recovered by applying the inverse transformation $\mathrm{T}^{-1}(\cdot)$ of the Eq. (9) as

$$\mathbf{P}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathrm{T}^{-1}\left(\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i}\right). \tag{11}$$

The different steps of the algorithm are summarized in the Algorithm 1 insert.

Algorithm 1: Iterative data assimilation

Set: N_j = The number of iterations of ES-MDA

Set: $\mathbf{Y}^{\mathrm{ob}}$ = Observation data (here, concentrations at the solid phase)

Set: N_e = The number of parameter realizations

Set: $\alpha_0^{'}$ = Initial inflation coefficient

Set: α_{geo} = Constant that controls the extent of the change of α_j between iterations

begin

Generate an ensemble of initial parameters $\mathbf{P}$ (here, these parameters are drawn from their prior uniform distributions)

Calculate all α_j coefficients such that $\alpha_j^{'} = \frac{\alpha_{j-1}^{'}}{\alpha_{geo}}$ and $\alpha_j = \alpha_j^{'}\left(\sum_{j=1}^{N_j}\frac{1}{\alpha_j^{'}}\right)$

for $j \leftarrow 1$ **to** N_j **do**

for $i \leftarrow 1$ **to** N_e **do**

Perturb the observations: $\mathbf{Y^{ob}} + \sqrt{\alpha_j}\varepsilon_i^{\mathbf{ob}}$

Run forward model (here, evaluate the Langmuir sorption isotherm for the different liquid phase equilibrium solute concentrations) using $\mathbf{P}_i$ as input parameters to obtain $\mathbf{Y}_i$

end for

Apply a Gaussian anamorphosis: $\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathrm{T}\left(\mathbf{P}^{\mathrm{u}}_{\mathrm{MDA},i}\right)$

Calculate: $\mathbf{C}_{G,Y} = \frac{1}{N_e-1}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\left(\mathbf{G}_{\mathrm{i}} - \overline{\mathbf{G}}\right)\left(\mathbf{Y}_{\mathrm{i}} - \overline{\mathbf{Y}}\right)^{\mathrm{T}}$

Calculate: $\mathbf{C}_{Y,Y} = \frac{1}{N_e-1}\sum_{\mathrm{i}=1}^{\mathrm{N_e}}\left(\mathbf{Y}_{\mathrm{i}} - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_{\mathrm{i}} - \overline{\mathbf{Y}}\right)^{\mathrm{T}}$

Update:

$$\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathbf{G}_i + \mathbf{C}_{G,Y}\left(\mathbf{C}_{Y,Y} + \alpha_j\mathbf{R}\right)^{-1}\left[\mathbf{Y^{ob}} + \sqrt{\alpha_j}\varepsilon_i^{\mathbf{ob}} - \mathbf{Y}_i\right]$$

Back transform: $\mathbf{P}^{\mathrm{u}}_{\mathrm{MDA},i} = \mathrm{T}^{-1}\left(\mathbf{G}^{\mathrm{u}}_{\mathrm{MDA},i}\right)$

end for

end

3 Materials and methods

The Langmuir isotherm is one of the most common models used for sorption in relation with transport in porous media. It explains how a solute distributes between the solid and liquid phases once equilibrium is reached. This isotherm considers that the surface at which the solute can adsorb onto the solid phase is finite and, therefore, there is a maximum adsorbed concentration possible; its expression is

$$S = \frac{S_m b C_e}{1 + b C_e}, \tag{12}$$

where S is the solid-phase equilibrium concentration $\left[\mathrm{M}_{\mathrm{chemical}}\ \mathrm{M}_{\mathrm{sorbent}}^{-1}\right]$, C_e is the liquid-phase equilibrium concentration $\left[\mathrm{M}_{\mathrm{chemical}}\ \mathrm{L}_{\mathrm{water}}^{-3}\right]$, S_m represents the maximum concentration of soil-adsorbed solute $\left[\mathrm{M}\ \mathrm{M}^{-1}\right]$, and b is an adsorption constant related to binding energy $\left[\mathrm{L}^3\ \mathrm{M}^{-1}\right]$. Figure 1 shows a typical Langmuir isotherm.

In our case, in the context of the ES-MDA algorithm, the forward model is the Langmuir sorption isotherm function, Eq. (12), S_m and b are the model parameters, whereas S is the system state. The solute concentrations at equilibrium, C_e for the liquid phase will be the forcing terms of the forward model and they are known. The parameter vector $\mathbf{P}_i$ for a given realization i is

$$\mathbf{P}_i = \begin{bmatrix} S_{m,i} \\ b_i \end{bmatrix}. \tag{13}$$

The system state vector $\mathbf{Y}_i$ is the set of predicted solid-phase concentrations at equilibrium corresponding to the set of liquid-phase concentrations at equilibrium for which the corresponding laboratory tests have been performed

$$\mathbf{Y}_i = \begin{bmatrix} S_{1,i} \\ S_{2,i} \\ \cdots \\ S_{no,i} \end{bmatrix}. \tag{14}$$

From those laboratory tests, there will be n_o observed solid-phase equilibrium concentrations resulting from the experiments

$$\mathbf{Y^{ob}} = \begin{bmatrix} S_1^{ob} \\ S_2^{ob} \\ \cdots \\ S_{no}^{ob} \end{bmatrix}. \tag{15}$$

Based on a range of previous numerical experiments, several scenarios were considered in order to analyze the impact of different parameters in the estimation process of the Langmuir coefficients. More precisely, the number of realizations of the ensemble took the values of 30, 100 and 300; the number of observation n_o took all integer values between four and sixteen, the observation error standard deviation took the values of 10^{-2}, $5 \cdot 10^{-3}$ and 10^{-4} mg·g^{-1}; and the number of iterations of the ES-MDA, N_j, took all integer values between 1 and 6. The steps followed for any given scenario are described next.

The first step is to generate a set of data pairs (C_e, S) consistent with a Langmuir isotherm as they could have obtained in the laboratory. These would be the value pairs that have to be curve fitted by the isotherm function. An ideal soil for which sorption follows the Langmuir isotherm and with realistic parameter values $S_m = 0.1$ mg·g^{-1} and $b = 100$ l·mg^{-1} (Godoy et al. 2018) is considered. Sixteen experiments are mimicked with equally spaced values of C_e in the interval between 0 and 0.1 mg·l^{-1}. The corresponding observations are computed by applying the Langmuir equation and perturbing the resulting value with an error $\varepsilon^{\mathrm{ob}}$ drawn

from a Gaussian distribution with zero mean and standard deviation $\sigma_{\varepsilon^{\mathrm{ob}}}$. From these sixteen data pairs, the necessary n_o observations will be chosen.

Second, an ensemble of initial values for the two parameters is generated. The initial parameters are drawn from the uniform distributions $S_m \sim U[0,3]$ mg·g^{-1} and $b \sim U[0,300]$ l·mg^{-1}.

Third, the ES-MDA with a normal-score transformation, as described previously, was applied for each scenario.

Fourth, the moments (mean, standard deviation, kurtosis, and skewness) of the ES-MDA final estimates were analyzed to investigate how many isotherm samples are necessary to reach stable statistics and, consequently, acceptable estimates. Specifically, the optimal number of samples will be determined by visual analysis of the moments of the final probability distributions of the different parameters; this optimal number will be achieved when the moments stabilize.

The results presented next correspond only to synthetic experiments, no real experimental data have been used. The principal reason of this choice is that it is the only way in which a comparison between the estimates and the "true" values can be made and to perform an effective evaluation of the methodology. Including an additional example with experimental data will not serve to verify or increase the reliability of the synthetic results, as long as the soil being analyzed does display an adsorption behavior suitable to be modeled by a Langmuir isotherm. The method proposed does not pretend to be a method to discriminate between isotherm curves, and therefore, it is important to note that including an additional case with laboratory data will not improve the validity of the approach. In any case, for the sake of completeness, two additional synthetic cases have been run, which are discussed in the Appendix: one with a synthetic soil with a different Langmuir isotherm, and another one with a Freundlich isotherm. These two additional examples prove that the method is general enough for curve fitting and it could be used with a different isotherm and even in a different curve-fitting context.

4 Results and discussion

Figures 2, 3 and 4 serve to illustrate how the ES-MDA works. An initial ensemble of 100 realizations of parameters (Fig. 2a) is generated using the above mentioned uniform distributions, the pairs (S_m, b) are distributed randomly within the domain $U[0,3] \times U[0,300]$. Each of the points in Fig. 2a corresponds to one of the isotherm curves plotted in Fig. 3a. We can see a wide scatter of potential isotherm curves all of them quite far from the "true" curve. In Fig. 3 the points corresponding to the observations are also displayed. The discrepancies between the S values for the different curves and the observed ones, the intrinsic variability of the parameters values as measured by its covariance, and the cross-correlation between parameters and predicted values serve to compute the different elements in the ES-MDA equations and to update each one of the points in Fig. 2a into a new pair that gives a new isotherm curve closer to the real one. After the first update, the new pairs of parameters are shown in Fig. 2b, the dashed lines correspond to the values of the isotherm curve used to generate the observations. It is quite evident how, after one iteration, the range of variability of the S_m parameter is much narrower than the initial range and quite close to the true value, whereas the range of the b parameters is still scattered over the entire initial range. The reason for this fast convergence of the S_m parameter is due to the higher sensitivity that the shape of the isotherm curve has to the S_m parameter than to the b parameter. We can see how the cloud of pairs keeps reducing its spread after each assimilation iteration, and at iteration #4 the cloud of parameters has collapsed onto the true values with little spread. The remaining spread is a measure of the residual uncertainty in their estimation. The results obtained for iterations #5 and #6 are not shown since they are almost identical to those of iteration #4. Fig. 4 shows the histograms of the initial ensembles and after four iterations. The initial histograms correspond to the starting uniform distributions and the histograms after four iterations show a spike at the true value for S_m and a histogram with some spread and a little bias for the b parameter. The smaller sensitivity of the

isotherm curves to this parameter makes it impossible to identify it more precisely. This small sensitivity translates in that the set of isotherm curves corresponding to this range of b values result in almost superposing curves in Fig. 3d.

The previous figures, which have been obtained for a specific scenario, demonstrate the power of the ES-MDA to identify the Langmuir parameters, together with a measure of their uncertainty. This exercise has been repeated for several other scenarios with the objective to determine how many experiments should be run, that is, how many observation pairs are needed to find the parameters that provide the best fitting curve. One way to analyze this aspect is to seek when the estimate of the statistics of the final set of parameters stabilizes with the number of samples, if at all. This analysis will be performed on the results after four iterations of the ES-MDA; similarly as for the scenario displayed in Figures 2, 3 and 4, four iterations were enough for the parameter estimates to stabilize in all scenarios.

Figure 5 shows results for the final ensemble of updated values for parameters S_m. The values of the mean, standard deviation, skewness and kurtosis are shown for all combinations of number of ensemble realizations, observation error and observation samples. It is evident that the best results are obtained when using an ensemble of 300 realizations, but good estimates of the true value (as given by the ensemble mean) with little uncertainty (as given by the ensemble standard deviation) can be obtained for all scenarios as soon as six observations are used. The estimates stabilizes at six observations when using 300 realizations for all statistics. A stable estimate for the means and standard deviations when using a smaller number of realizations requires between eight and ten samples. The estimated skewness and kurtosis vary more erratically for the scenarios with 30 and 100 realizations, in part due to the smaller number of elements to compute these statistics, and in part due to the narrowness of the final distribution which make these values very sensitive to small departures from the mean. It is very interesting to note that the magnitude of the observation error has little or no

effect in the estimates of the means and standard deviations in all scenarios as long as at least six observations are used in the estimation.

Figure 6 shows results for the final ensemble of updated values for parameter b. The values of the mean, standard deviation, skewness and kurtosis are shown for all combinations of number of ensemble realizations, observation error and observation samples. Contrary to Fig. 5, there is not a striking difference on the curves as a function of the number of realizations. The main reason for this results is the already-mentioned fact that the Langmuir curve is less sensitive to parameter b than to parameter S_m; for this reason, as soon as an estimated value is relatively close to the real one, the estimated Langmuir isotherm is almost indistinguishable with the true one, and there is no need to update the parameter anymore. This behavior is particularly noticeable in the values of the standard deviations; they do not get as close to zero as for S_m but remain with non-zero values throughout and with larger values when the observation errors are larger. Skewness and kurtosis estimates behave as for S_m. Due to the smaller sensitivity of the isotherm to the b value, an estimate based only on four samples would be enough.

The ES-MDA performed remarkably well for the purpose of estimating the fitting parameters of a Langmuir isotherm in a wide range of scenarios, with the additional benefit of providing also an estimate of their uncertainty. This uncertainty estimate is a confidence measurement about the estimated value and it is also a measurement of the sensitivity of the fitting to the parameter. After all, at the end of the exercise, there is an experimental histogram showing the full distribution of potential values for the parameters, from which the mean or the median could be selected as best estimates, but from which an analysis of the parameter values which are consistent with the observations can also be performed. Such an analysis of the uncertainty about the estimates cannot be performed with standard fitting procedures, such as least-squares, that, at most, provide an estimated value and an estimation error.

In summary and with regard to the recommendation by Roy et al. (1991) that a minimum of five observations should be used to estimate the isotherm parameters, we conclude that such a number would be enough for the estimation of the b coefficient, but not for the estimation of S_m. Our recommendation would be to increase that minimum number to six and, preferably, to eight.

The successful results in the application of the ES-MDA for curve fitting for the three non-linear isotherm curves analyzed in the paper, makes us postulate the use of the ES-MDA for general curve fitting when characterization of the uncertainty about the final estimates is important.

5 Conclusions

In this paper, we proposed a procedure to fit sorption isotherm curves using an ensemble smoother with multiple data assimilation (ES-MDA) coupled to a normal-score transform. The main advantage of this approach is not only that the parameter is good, but also that a characterization of the parameter estimation uncertainty is obtained. In order to evaluate the proposed procedure, we performed numerical examples with a variety of scenarios to additionally investigate the influence of the number of experimental pairs, the ensemble size, the measurement-error magnitude, and the number of ES-MDA iterations on the uncertainty estimation. Our results show that, since the shape of the Langmuir isotherm is much more sensitive to the S_m parameter than to the b parameter, the precision in their identification is not the same. After four ES-MDA iterations, the cloud of experimental pairs has collapsed onto the reference value for the S_m parameter and presents some spread and a little bias for the b parameter. By investigating other scenarios with several combinations of number of ensemble realizations, observation error and observation samples we find that, for the S_m parameter, the best results are obtained when using an ensemble of 300 realizations, and the use of at least six observations can be enough to produce relatively good estimates of the true value regardless of the scenario. An interesting finding is that when at least six obser-

vations are used, the magnitude of the observation error has almost no effect in the estimates of the mean and standard deviation of S_m. For the b parameter, there is not a clear difference on the curves as a function of the scenarios because the Langmuir curve is little sensitive to this parameter. These results demonstrate the power of the ES-MDA to identify the Langmuir parameters together with an estimate of their uncertainty. We conclude that the actual recommendation that a minimum of five observations should be used to estimate the Langmuir parameters would be enough for the estimation of the b coefficient, but not for the estimation of S_m. In order to correctly estimate Langmuir parameters together with their uncertainty a minimum of six and, preferably, eight samples should be used. The results of the application in the two additional cases included in the Appendix are similar and reinforce our belief that the ES-MDA could be applied for standard curve fitting when an uncertainty characterization about parameter estimates is needed.

References

Bouchelaghem, F. (2018). Multi-scale study of pollutant transport and uptake in compacted bentonite. *Mathematical Geosciences*, 50:495–523.

Burgers, G., Van Leeuwen, P. J., and Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724.

Capilla, J. E., Gömez-Hernández, J. J., and Sahuquillo, A. (1998). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric head data—3. application to the culebra formation at the waste isolation pilot plan (wipp), new mexico, usa. *Journal of Hydrology*, 207:254–269.

Capilla, J. E., Rodrigo, J., and Gómez-Hernández, J. J. (1999). Simulation of non-gaussian transmissivity fields honoring piezometric data and integrating soft and secondary information. *Math. Geology*, 31:907–927.

Chen, Y. and Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44:1–26.

Emerick, A. A. (2017). Investigation on principal component analysis parameterizations for history matching channelized facies models with ensemble-based data assimilation. *Mathematical Geosciences*, 49:85–120.

Emerick, A. A. and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15.

Evensen, G. (2018). Analysis of iterative ensemble smoothers for solving inverse problems. *Computational Geosciences*, 22:885–908.

Fairey, J. L. and Wahman, D. G. (2013). Bayesian and Frequentist Methods for Estimating Joint Uncertainty of Freundlich Adsorption Isotherm Fitting Parameters. *Journal of Environmental Engineering*, 139:307–311.

Franssen, H. H. and Gómez-Hernández, J. (2002). 3d inverse modelling of groundwater flow at a fractured site using a stochastic continuum model with multiple statistical populations. *Stochastic Environmental Research and Risk Assessment*, 16:155–174.

Godoy, V. A., Zuquette, L. V., and Gómez-Hernández, J. J. (2018). Scale effect on hydraulic conductivity and solute transport: Small and large-scale laboratory experiments and field experiments. *Engineering Geology*, 243:196–205.

Gómez-Hernández, J. J., Napa-García, G. F., and Godoy, V. A. (2019). How to account for uncertainty in the estimation of adsorption isotherm parameters. In Gómez-Hernández, J. J. and Navarro, B. A., editors, *Groundwater Management and Governance: Coping with Uncertainty, proceedings of IAH2019, the 46th Annual Congress of the International Association of Hydrogeologists*, page 342. Spanish Chapter of the International Association of Mathematical Geosciences, Spanish Chapter of the International Association of Mathematical Geosciences.

Gómez-Hernández, J. J. and Wen, X.-H. (1994). Probabilistic assessment of travel times in groundwater modeling. *J. of Stochastic Hydrology and Hydraulics*, 8(1):19–56.

Guo, Z., Fogg, G. E., Brusseau, M. L., LaBolle, E. M., and Lopez, J. (2019). Modeling groundwater contaminant transport in the presence of large heterogeneity: a case study comparing MT3D and RWhet. *Hydrogeology Journal*, 27:1363–1371.

Hinz, C., Gaston, L., and Selim, H. (1994). Effect of sorption isotherm type on predictions of solute mobility in soil. *Water Resources Research*, 30:3013–3021.

Li, L., Zhou, H., and Gómez-Hernández, J. J. (2011). A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (made) site, columbus air force base, mississippi (usa). *Journal of Hydrology*, 404:278–293.

Li, L., Zhou, H., Hendricks Franssen, H.-J., and Gómez-Hernández, J. J. (2012). Modeling transient groundwater flow by coupling ensemble kalman filtering and upscaling. *Water Resources Research*, 48:W01537.

Luo, Q., Wu, J., Yang, Y., Qian, J., and Wu, J. (2014). Optimal design of groundwater remediation system using a probabilistic multi-objective fast harmony search algorithm under uncertainty. *Journal of Hydrology*, 519:3305–3315.

Masood, Z. B. and Abd Ali, Z. T. (2020). Numerical modeling of two-dimensional simulation of groundwater protection from lead using different sorbents in permeable barriers. *Environmental Engineering Research*, 25:605–613.

Nataf, A. (1962). Determination des distribution don t les marges sont donnees. *Comptes Rendus de l Academie des Sciences*, 225:42–43.

Park, D. K., Ko, N. Y., and Lee, K. K. (2007). Optimal groundwater remediation design considering effects of natural attenuation processes: Pumping strategy with enhanced-natural-attenuation. *Geosciences Journal*, 11:377–385.

Pathak, P. and Sharma, S. (2018). Sorption isotherms, kinetics, and thermodynamics of contaminants in indian soils. *Journal of Environmental Engineering*, 144:04018109.

Roy, W., Krapac, I., Chou, S., and Griffin, R. (1991). Batch-Type Adsorption Procedures for Estimating Soil Attenuation of Chemicals.

Shih, C., Park, J., Sholl, D. S., Realff, M. J., Yajima, T., and Kawajiri, Y. (2020). Hierarchical Bayesian estimation for adsorption isotherm parameter determination. *Chemical Engineering Science*, 214:115435.

Silva, T. M., Pesco, S., Barreto Jr, A., and Onur, M. (2021a). A new procedure for generating data covariance inflation factors for ensemble smoother with multiple data assimilation. *Computers & Geosciences*, page 104722.

Silva, T. M., Villalobos, R. S., Cardona, Y. A., Barreto, A., and Pesco, S. (2021b). Well-testing based turbidite lobes modeling using the ensemble smoother with multiple data assimilation. *Computational Geosciences*, 25:1139–1157.

Tavakoli-Kivi, S., Bailey, R. T., and Gates, T. K. (2019). A salinity reactive transport and equilibrium chemistry model for regional-scale agricultural groundwater systems. *Journal of Hydrology*, 572:274–293.

Todaro, V., D'Oria, M., Tanda, M. G., and Gómez-Hernández, J. J. (2019). Ensemble smoother with multiple data assimilation for reverse flow routing. *Computers & Geosciences*, 131:32–40.

Todaro, V., D'Oria, M., Tanda, M. G., and Gómez-Hernández, J. J. (2021). Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer. *Journal of Hydrology*, page 126215.

Xu, T. and Gómez-Hernández, J. J. (2015). Inverse sequential simulation: A new approach for the characterization of hydraulic conductivities demonstrated on a non-Gaussian field. *Water Resources Research*, 51:2227–2242.

Xu, T. and Gómez-Hernández, J. J. (2016). Characterization of non-Gaussian conductivities and porosities with hydraulic heads, solute concentrations, and water temperatures. *Water Resources Research*, 52:6111–6136.

Xu, T. and Gómez-Hernández, J. J. (2018). Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter. *Advances in Water Resources*, 112:106–123.

Xu, T., Jaime Gómez-Hernández, J., Zhou, H., and Li, L. (2013). The power of transient piezometric head data in inverse modeling: an application of the localized normal-score enkf with covariance inflation in a heterogenous bimodal hydraulic conductivity field. *Advances in Water Resources*, 54:100–118.

Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H.-J., and Li, L. (2011). An approach to handling non-Gaussianity of parameters and state variables in ensemble Kalman filtering. *Advances in Water Resources*, 34(7):844–864.

Zhou, H., Li, L., Franssen, H.-J. H., and Gómez-Hernández, J. J. (2012). Pattern recognition in a bimodal aquifer using the normal-score ensemble kalman filter. *Mathematical Geosciences*, 44:169–185.

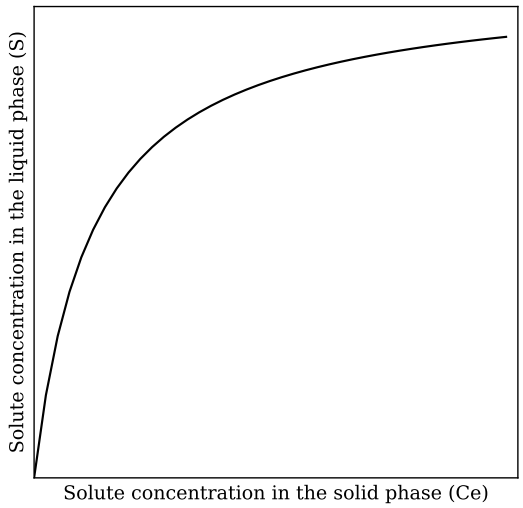

Fig. 1: Typical Langmuir isotherm.

A Additional Synthetic Examples

In the following, two additional examples are included to support the claims in the main body of the text; the first one is for a synthetic soil with adsorption characteristics given by a Langmuir isotherm with parameters $S_m = 7.2$ mg·g^{-1} and $b = 0.174$ l·mg^{-1} (**?**), and the second one for a soil characterized by a Freundlich isotherm

$$S = K_f C_e^{1/n} \tag{16}$$

with parameters $K_f = 1.5$ and $n^{-1} = 0.39$; S is given in mg·g^{-1} and C_e in mg·l^{-1} (**?**). Figure A.1 shows the three isotherms considered in the paper.

The procedure to fit the curves is the same as the one used in the body of the text. The initial sets of realizations are drawn from the following bivariate uncorrelated uniform distributions: $(S_m, b) \in U[0, 230] \times U[0, 0.8]$, and $(K_f, n^{-1}) \in U[0, 30] \times U[0.001, 0.99]$.

The evolution with the number of samples of the best estimate as given by the mean of the ensemble of updated parameters for the two cases can be seen in Figure1 A.2. The conclusions that can be drawn from the analysis of these figures are the same as from the analysis of

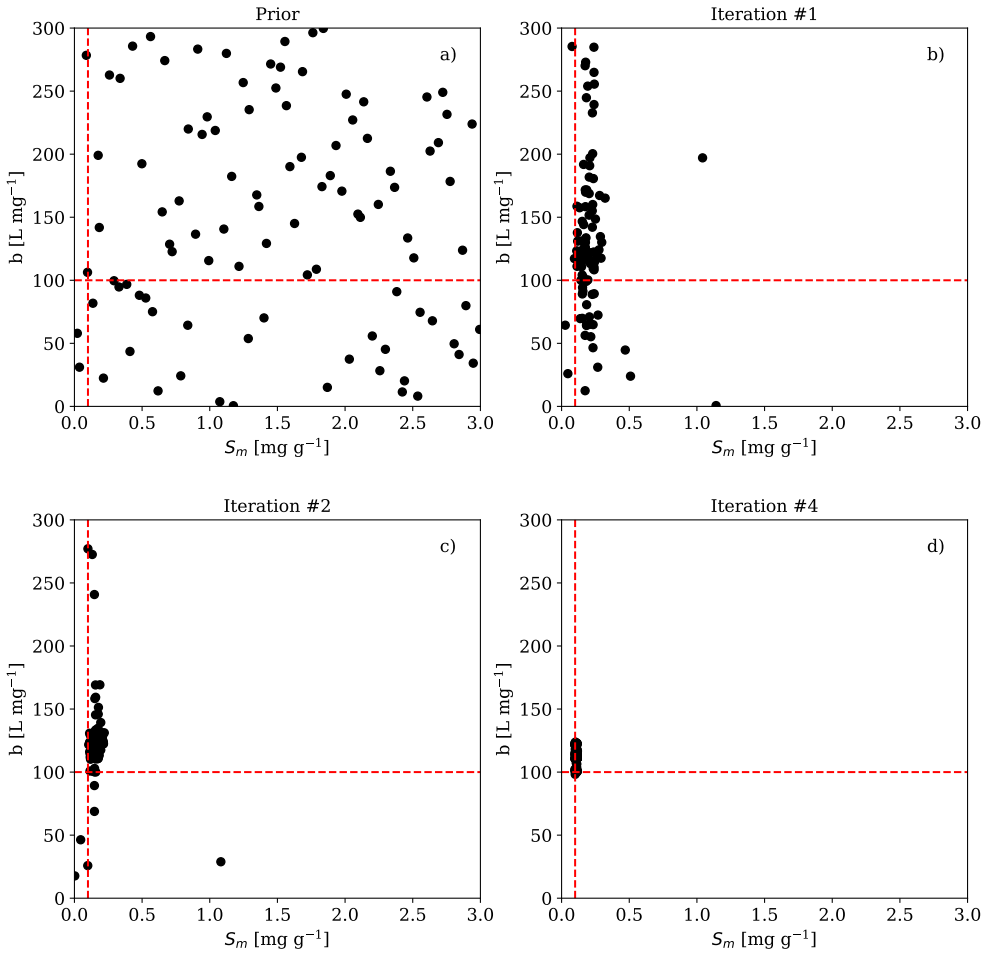

Fig. 2: Operation of the ES-MDA for: a prior pairs of S_m and b; parameters values obtained after ES-MDA #1, #2 and #4 iterations and 100 realizations.

the example in the main body. The estimated values are affected by the magnitude of the measurement errors, the larger the measurement errors, the larger the bias of the estimated value (as given by the mean of the ensemble results). When the error standard deviation is set to 1%, the estimates are quite close to the true value of the synthetic soil. For the Langmuir isotherm, the fluctuations of the mean $_m$ and mean b stabilize about six samples with a better stabilization the larger the number of realizations of the ensemble have been used. For the Freundlich isotherm, the estimation needs at least 13 samples and either 100 or 300 realizations to retrieve good estimates when the error standard deviation is above 1%; for the smaller error, seven samples are necessary before the mean estimate stabilizes close to the

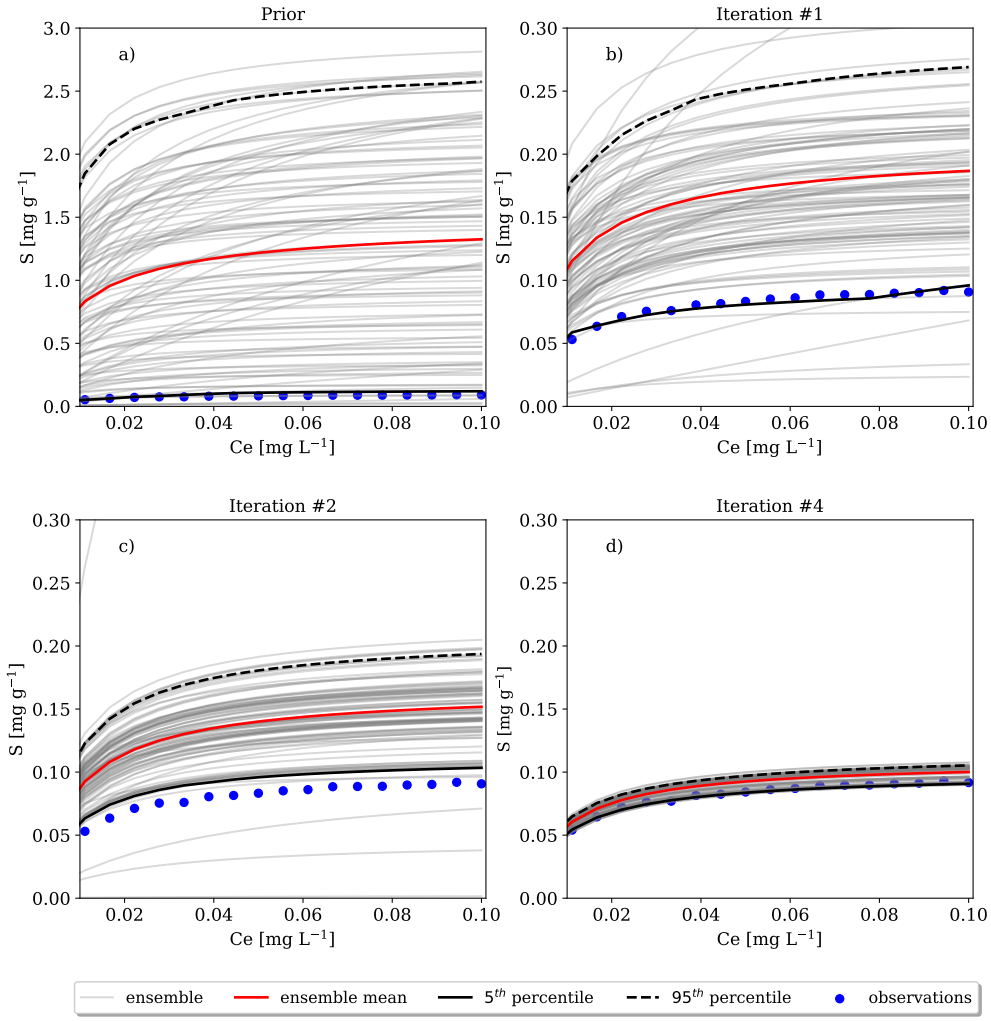

Fig. 3: Variability in the computation of the Langmuir isotherm using: **a** prior; and **b-d** updated parameters for 17 observations at iterations #1, #2 and #4, 100 realizations, and ε^{ob} of 10^{-3} mg g^{-1}.

452 reference value. It could be concluded that the Freundlich estimate may need more samples to
453 ensure a good estimation of its parameters.

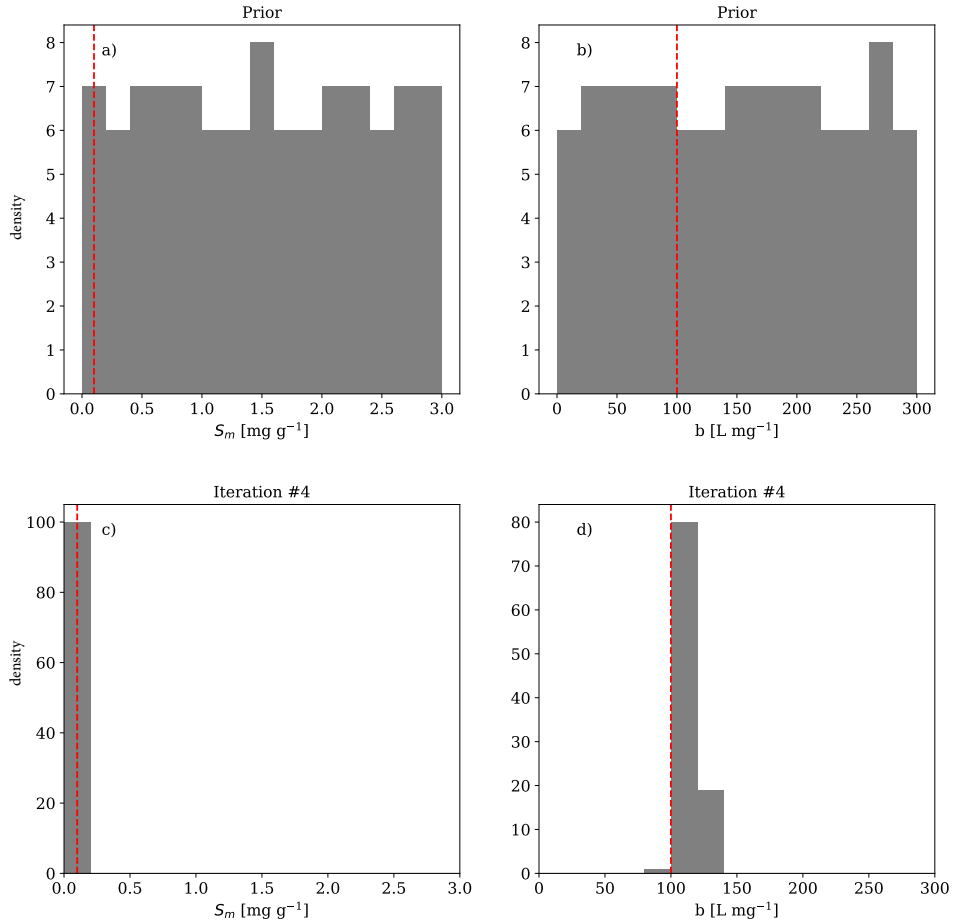

Fig. 4: The prior (**a-b**) and updated (**c-d**) histograms of the Langmuir isotherm parameters for 17 observations, 100 realizations, and $\varepsilon^{\mathrm{ob}}$ of 10^{-3} mg g^{-1}.

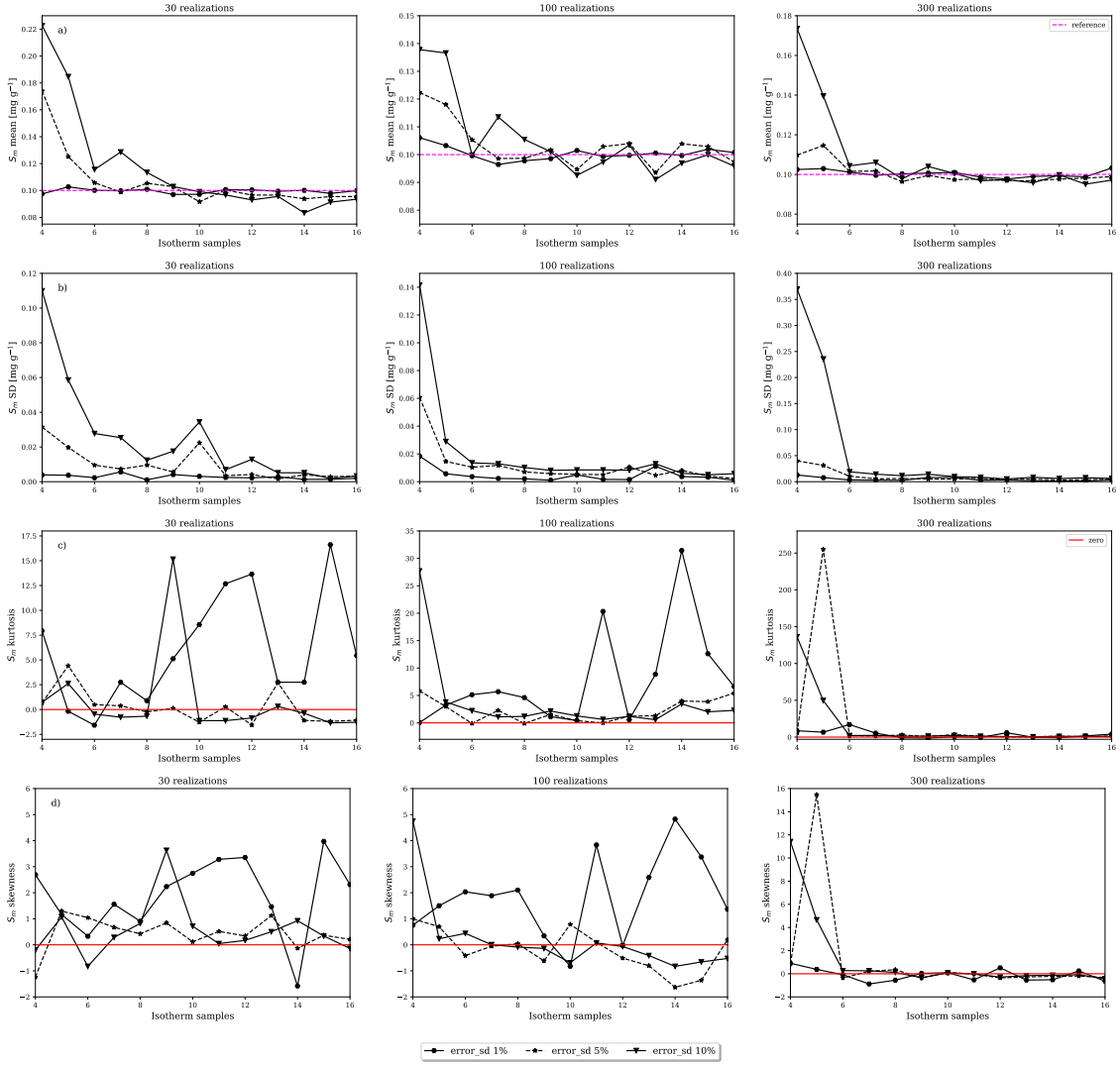

Fig. 5: From top to bottom, variation of the: a) mean, b) standard deviation, c) kurtosis, and d) skewness of S_m with the number of isotherm samples.

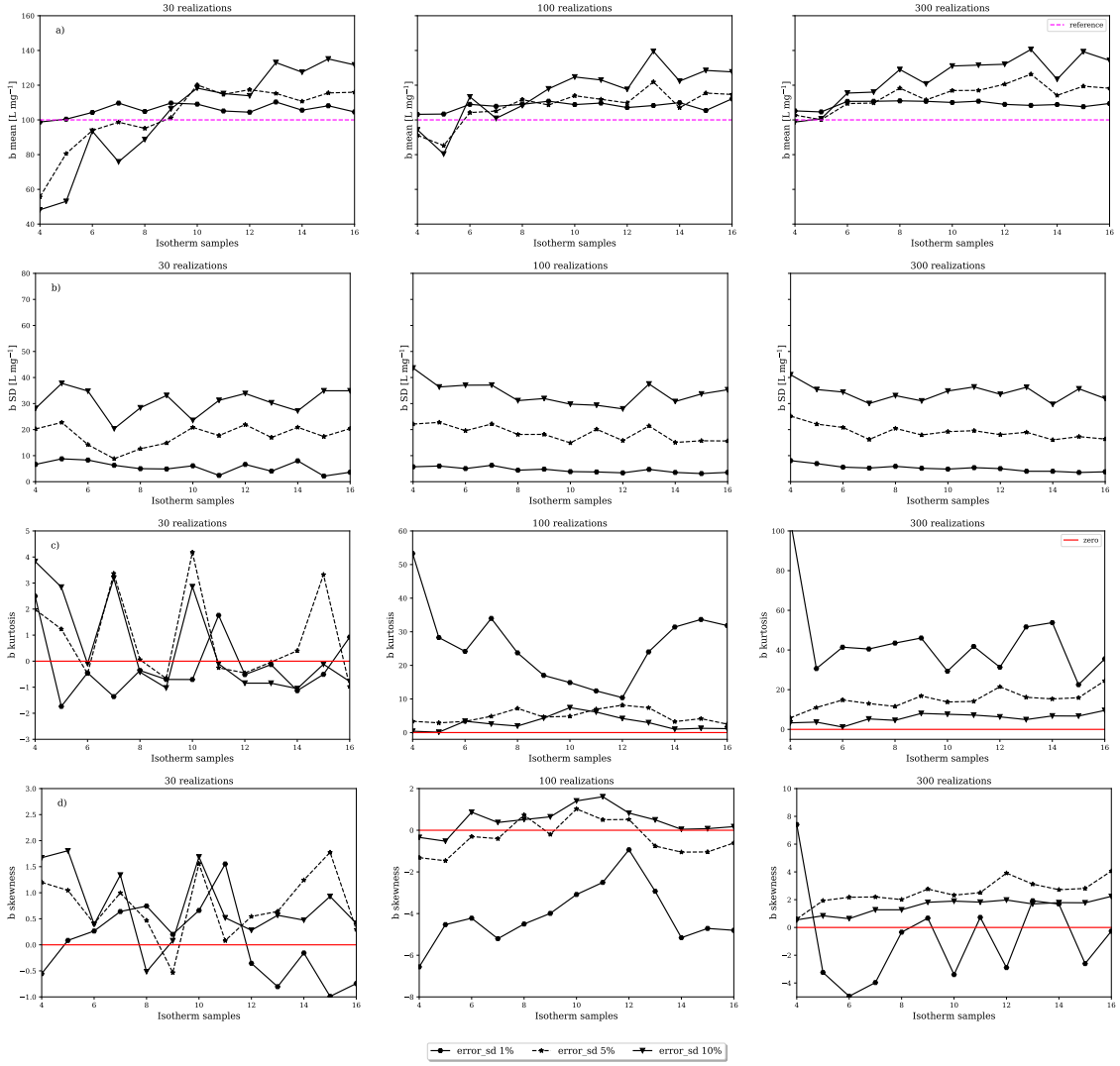

Fig. 6: From top to bottom, variation of the: a) mean, b) standard deviation, c) kurtosis, and d) skewness of b with the number of isotherm samples.

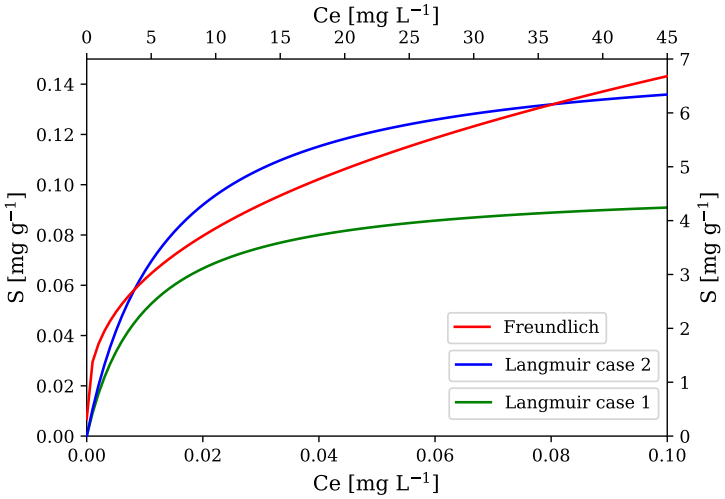

Fig. 7: The three isotherms analyzed in the paper. The Freundlich isotherm uses the left vertical axis, while the Langmuir isotherms use the right vertical axis.

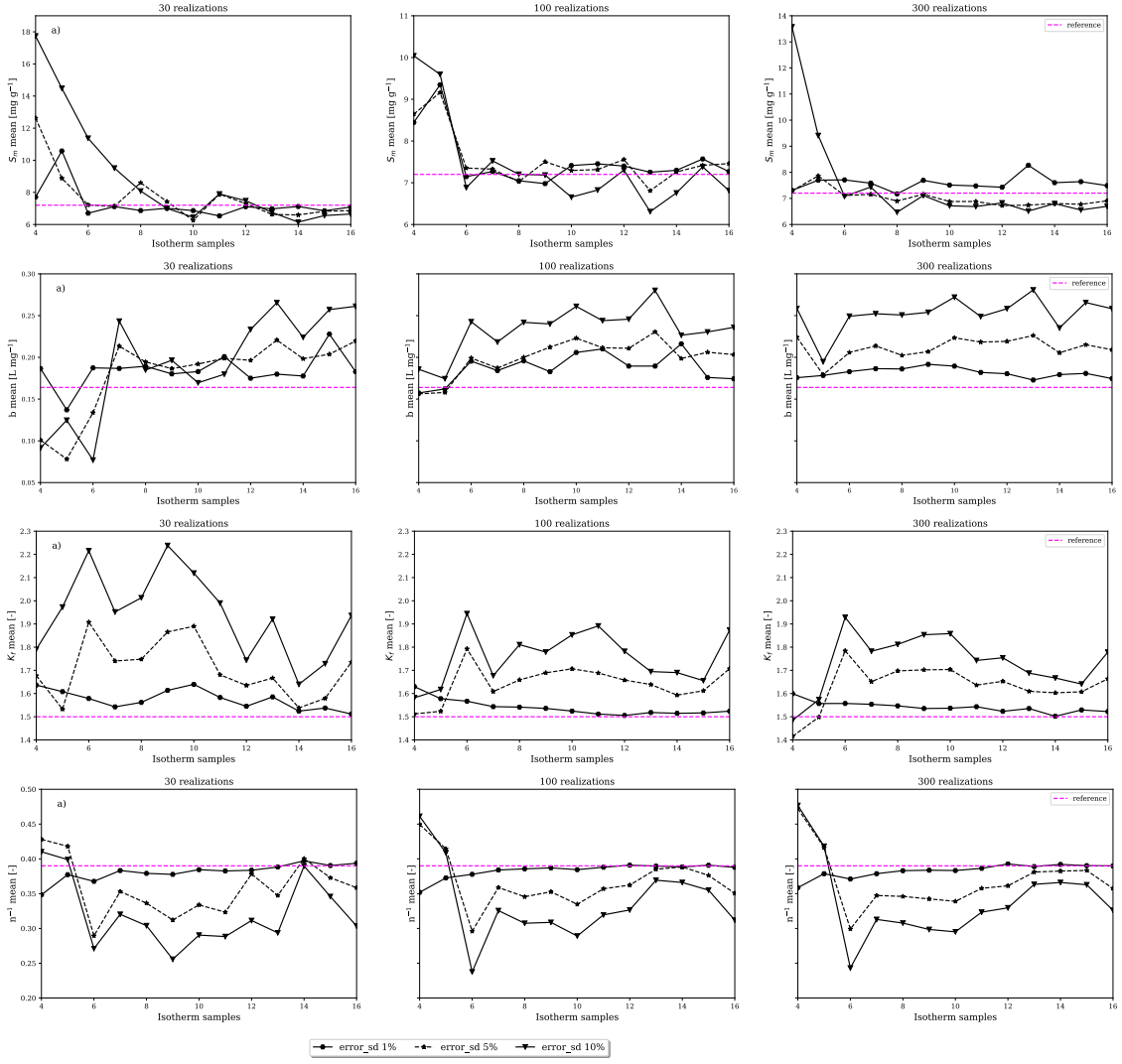

Fig. 8: From top to bottom, variation of the mean of parameter: S_m, b, K_f, and n^{-1} with the number of isotherm samples.