SAND2017-2881J

Delineating Facies Spatial Distribution by Integrating Ensemble Data Assimilation and Indicator Geostatistics with Level Set Transformation

Xuehang Song[1,2], Ming Ye[1], Zhenxue Dai[3,4], Glenn Hammond[5], John M. Zachara[2], and Xingyuan Chen[2,*]

[1]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.

[2]Pacific Northwest National Laboratory, Richland, Washington, USA.

[3]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[4]College of Construction Engineering, Jilin University, Changchun, China.

[5]Applied Systems Analysis and Research, Sandia National Laboratories, Albuquerque, New Mexico, USA.

*Corresponding author

Email: Xingyuan.Chen@pnnl.gov; Phone: (509) 371-7510; Fax: (509) 375-2999

Key Points

- Ensemble-based data assimilation is integrated with indicator geostatistics and level set transformation for facies delineation.
- The spatial distribution and permeability of two distinct facies are estimated simultaneously from transient head data induced by pumping tests.
- Imposing spatial continuity by adaptively selecting conditioning points used by indicator models is proven essential for facies delineation.

Abstract

A new approach is developed to delineate the spatial distribution of discrete facies (geological units that have unique distributions of hydraulic, physical, and/or chemical properties) conditioned not only on direct data (measurements directly related to facies properties, e.g., grain size distribution obtained from borehole samples) but also on indirect data (observations indirectly related to facies distribution, e.g., hydraulic head and tracer concentration). Our method integrates for the first time ensemble data assimilation with traditional transition probability-based geostatistics. The concept of level set is introduced to build shape parameterization that allows transformation between discrete facies indicators and continuous random variables. The spatial structure of different facies is simulated by indicator models using conditioning points selected adaptively during the iterative process of data assimilation. To evaluate the new method, a two-dimensional semi-synthetic example is designed to estimate the spatial distribution and permeability of two distinct facies from transient head data induced by pumping tests. The example demonstrates that our new method adequately captures the spatial pattern of facies distribution by imposing spatial continuity through conditioning points. The new method also reproduces the overall response in hydraulic head field with better accuracy compared to data assimilation with no constraints on spatial continuity on facies.

1. Introduction

Characterizing spatial heterogeneity and connectivity within the physical, chemical, and ecological systems is a daunting challenge facing the modeling community in the various domains of Earth system sciences [*Clark et al*., 2015; *Harvey and Gooseff*, 2015]. The facies-based approach, which divides the system into a finite number of relatively homogenous units, is commonly used to reduce the dimensionality and complexity in parameterizing a complex system [e.g., *Sassen et al.*, 2012]. It is particularly well suited for systems that contain subunits with sharp contrast in properties. Facies have been defined in different contexts of subsurface characterization (e.g., lithofacies for lithologic features, hydrofacies for hydraulic properties, chemofacies for chemical attributes, thermofacies for thermal properties, and reactive facies for reaction potential) [*Bayer et al.,* 2015; Dai et al., 2009; *Yabusaki et al.*, 2011]. It is important to develop a mathematically general framework for delineating the spatial distribution of facies.

As direct data on facies attributes are usually scarce due to cost constraints, they are often insufficient to adequately delineate the spatial distribution of facies with reasonable uncertainty. Therefore, indirect data have been used to augment the limited direct data for estimating facies distribution and properties through inverse modeling or data assimilation techniques [e.g., *Ye and Khaleel*, 2008; *Harp et al.*, 2008]. The ensemble data assimilation (EDA) methods have been widely applied to incorporate direct and indirect data to inform process-based numerical models owing to the EDA methods' computational efficiency (compared to full Bayesian approaches, such as the Markov chain Monte Carlo (MCMC) methods applied to data assimilation by Wainwright et al. [2014]) and flexibility to handle uncertainty arising from multiple sources [*Aanonsen et al.*, 2009; *Oliver and Chen*, 2011; *Chen et al.*, 2013]. As the original EDA methods were developed for estimating single-modal continuous variables (such as a Gaussian random

86 field of permeability), new developments are necessary to implement EDA methods for discrete
87 variables, such as the indicators for different facies types.

88 To enable facies-based EDA, it is necessary to map the discrete facies distribution to space
89 function of continuous random variables is needed. This can be achieved by using various
90 parameterization methods, such as the truncated pluri-Gaussian method [*Agbalaka and Oliver*,
91 2008; *Liu and Oliver*, 2005], the Gaussian mixture method [*Dovera and Della Rossa*, 2010],
92 discrete cosine transform method [*Jafarpour and McLaughlin*, 2008], and the level set method
93 [*Chang et al.*, 2010; *Moreno and Aanonsen*, 2014]. These methods typically involve the
94 following steps: (1) estimating a prior ensemble of facies distribution using indicator
95 geostatistical models, (2) translating the ensemble facies distribution to shape parameters that
96 describe the facies boundaries, and (3) performing EDA on the shape parameters to improve the
97 match between model predictions and observations. While such parameterization is effective in
98 general, it imposes no spatial structure constraints on facies (e.g., facies volume proportions,
99 correlation lengths, and juxtapositional tendencies) in the process of data assimilation. Thus, it
100 might lead to an unrealistic discontinuity in facies distribution.

101 In this paper, we propose a new framework that extends the EDA methods developed for
102 continuous variable to discrete facies delineation by assimilating indirect data while imposing
103 spatial continuity constraints on resulting facies distribution. In the new framework, we first
104 adopt a parameterization method based on level set functions [*Chang et al.*, 2010] to describe the
105 occurrence of facies at given locations through facies probability, and then use the EDA methods
106 to update the facies probability. At each data assimilation step, additional conditioning points of
107 facies are selected based on the changes between prior and posterior facies probability and
108 combined with the prior direct data to update the facies spatial structure. The combined set of

conditioning points and the updated spatial structure are then used in the conditional simulation of the facies field by using the Transition Probability Geostatistical Software (T-PROGS) [*Carle*, 1999] for the next step of data assimilation. The new EDA framework for spatial facies delineation is evaluated with a synthetic two-dimensional (2-D) groundwater modeling with two facies, in which transient head data induced by pumping tests is assimilated as indirect data to delineate the spatial distribution of hydrofacies with contrasting high- and low-permeability. The synthetic case of groundwater modeling specifically investigates the importance of imposing geostatistical constraints for facies delineation during the data assimilation.

2. Methodology

Two primary building blocks of our framework include the data assimilation method used for parameter estimation (described in Section 2.1) and the level set method that enables transformations between discrete facies indicators and Gaussian random variables (described in Section 2.2). One unique contribution of our framework is on imposing spatial continuity when updating facies field using T-PROGS, which is also explained in Section 2.2.

2.1 Ensemble data assimilation methods

While EDA methods were originated from the Ensemble Kalman Filter (EnKF) developed by *Evensen* [1994] and *Burgers et al.* [1998], iterative EDA approaches [*Gu and Oliver*, 2007; *Chen et al.*, 2013], similar to the Gauss-Newton algorithm for solving nonlinear problems, have been developed for nonlinear systems by controlling the adverse effect of nonlinearity through reducing the increment vector in the updating formula by a fraction and iterating the procedure multiple times. Ensemble Smoother with Multiple Data Assimilation (ES-MDA) [*Emerick and*

Reynolds, 2013] is one of such approaches, and it is adopted in this study for its straightforward implementation and proven efficiency in dealing with nonlinear systems.

The ES-MDA method assimilates all observations simultaneously in each iteration, while the reduction in the increment vector resulted from data assimilation is achieved by inflating the observational variance, as in the following analysis scheme for updating **m** (system states and parameters):

$$\mathbf{m}^{i,l+1} = \mathbf{m}^{i,l} + \mathbf{C}_{MD}\left(\mathbf{C}_{DD} + \alpha_l \mathbf{R}\right)^{-1}\left(\mathbf{d}_{\text{obs}} + \sqrt{\alpha_l}\mathbf{e}^i - \mathbf{d}^{i,f}\right), \tag{1}$$

where $\mathbf{m}^{i,l}$ is the *i*-th realization of the ensemble of **m** during the *l*-th iteration, $\mathbf{d}_{\text{obs}}$ is the observational data to be assimilated, $\mathbf{d}^{i,f} = f(\mathbf{m}^i)$ is the model prediction counterpart of $\mathbf{d}_{\text{obs}}$ simulated by a forward model denoted as $f(\cdot)$, matrix $\mathbf{C}_{MD}$ is the cross-covariance between the parameter vector and the model predictions, $\mathbf{C}_{DD}$ is the auto-covariance of model predictions, $\mathbf{e}^i$ is the *i*th realization of measurement errors that are assumed to follow a Gaussian distribution with zero mean and covariance matrix $\mathbf{R}$, and α_l is the iteration coefficient of each iteration. While as many iterations as needed can be performed to approach the optimal solution, the multiple or iterative updating of Eq. (1) is only valid when the coefficients α_l satisfy $\sum_{l=1}^{N_a}\frac{1}{\alpha_l} = 1$, with *l* and N_a being the iteration index and the total iteration number, respectively. The covariance matrices $\mathbf{C}_{MD}$ and $\mathbf{C}_{DD}$ are approximated by their ensemble statistics:

$$\mathbf{C}_{MD} \approx \frac{1}{N_e - 1}\sum_{i=1}^{N_e}\left[\left(\mathbf{m}^{i,l} - \left\langle \mathbf{m}^{i,l}\right\rangle\right)\left(\mathbf{d}^{i,f} - \left\langle \mathbf{d}^{i,f}\right\rangle\right)^{\mathrm{T}}\right], \tag{2}$$

$$\mathbf{C}_{DD} \approx \frac{1}{N_e - 1}\sum_{i=1}^{N_e}\left[\left(\mathbf{d}^{i,f} - \left\langle \mathbf{d}^{i,f}\right\rangle\right)\left(\mathbf{d}^{i,f} - \left\langle \mathbf{d}^{i,f}\right\rangle\right)^{\mathrm{T}}\right], \tag{3}$$

150 where N_e is the number of realizations in the ensemble, and T is the transpose of matrix.

151 Example parameter vector **m** can be the permeability **p** for each facies and the Gaussian
152 variables **y** related to facies through the level set transformation as described in the next section,
153 i.e., the *i*-th realization of the parameter vector at each iteration of ES-MDA is
154 $\mathbf{m}^{i,l} = \left[\left(\mathbf{y}^{i,l}\right)^{\mathrm{T}}, \left(\mathbf{p}^{i,l}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$. The observations, $\mathbf{d}_{\mathrm{obs}}$, are indirect data, such as the transient head data
155 used in the numerical example of section 3.

156

157 **2.2 Facies-based ES-MDA with level set transformation and transition probability-based**
158 **geostatistics**
159
160 We adopted the concept of level set [*Chang et al.*, 2010] for the transformation between
161 discrete facies types and continuous random variables in order to implement ES-MDA for facies
162 delineation. Taking a system of two facies types as an example, the facies types at a given
163 computational cell within a domain can be related to the signs of a group of Gaussian variables
164 via level set transformation as follows:

165
$$\begin{cases} r_j^i = 1, & \text{if } y_j^i > 0 \\ r_j^i = 2, & \text{if } y_j^i \le 0 \end{cases}, \tag{4}$$

166 where r_j^i is the *i*-th realization of facies type on the *j*-th cell of the computational domain and y_j^i
167 is the transformed continuous variable. By assuming that y_j^i follows a Gaussian distribution
168 with mean u_j and standard deviation σ_j, the probability of the facies being type 1 or type 2 at
169 the *j*-th node can be computed as

$$\begin{cases} p_{j,1} = \text{Prob}(y_j^i > 0) = 1 - \Phi\left(\dfrac{0-\mu_j}{\sigma_j}\right) \\ p_{j,2} = \text{Prob}(y_j^i \le 0) = \Phi\left(\dfrac{0-\mu_j}{\sigma_j}\right) \end{cases}, \quad (5)$$

where Φ is the Cumulative Distribution Function (CDF) of the standard Gaussian distribution.

The level set transformation offers one way to handle discrete facies when implementing the ES-MDA method. As shown in Figure 1, the implementation starts with generating geostatistical realizations of facies distribution from prior information on probability of facies occurrence, which may include borehole data and expert opinions. Subsequently, the level set method is used to transform the realizations of discrete facies to realizations of continuous Gaussian variable at each grid cell by preserving the probability of facies occurrence. One key step in this transformation is to determine the mean and variance of the transformed Gaussian random variable. Since the results of data assimilation are not impacted by the value of variance (see the proof given in the supplementary information), we use a fixed variance of 1 for the transformed continuous random variable, which simplifies the transformation to find the mean of the transformed Gaussian random variable at each grid cell.

The transformed Gaussian variables at all grid cells are taken as the state variables in the ES-MDA formulation and their ensembles are updated at each iteration. The final posterior ensemble of the transformed variables is transformed back to facies through the probability of occurrence using equation (5). While this is a typical procedure used in existing facies-based EDA methods, its main drawback is that the facies at each grid cell changes independently from its neighbors without considering spatial continuity. The absence of spatial continuity constraint can lead to an unrealistic discontinuous distribution of facies. In this study, an additional procedure is introduced at each data assimilation step or each iteration within a step to impose a spatial

191 structure in generating updated facies field using geostatistical conditional simulation (e.g., T-
192 PROGS). The spatial structure is informed by a group of conditioning points selected using a set
193 of criteria discussed below.

194 The level-set-based transformation and T-PROGS based conditional simulation of facies are
195 two building blocks that enable us to delineate facies distribution using the ES-MDA method
196 developed for the continuous random variables. The important steps in the analysis procedure are
197 summarized in Figure 1 and described as follows:

198 (1) Generate prior ensembles of facies field and associated facies property (e.g., permeability)
199 based on prior information (e.g., borehole data and expert knowledge).

200 (2) Transform the realizations of facies into the realizations of the Gaussian variable at each
201 grid cell using mean values calculated from Eq. (A1) in the supplementary information
202 with standard deviation fixed at 1. This step is referred to as the level set (LS)
203 Transformation marked in Figure 1.

204 (3) Run forward simulations with the ensemble of facies fields and their associated facies
205 properties to produce the modeled counterparts of observations.

206 (4) Update the ensemble of the transformed Gaussian variables and property parameter of
207 each facies by assimilating observation data using ES-MDA;

208 (5) Update the facies probabilities at each grid cell using the mean of transformed variables
209 (Eq. 5) based on the ensemble updated at step (4). This step is referred to as the level set
210 (LS) back transformation marked in Figure 1.

211 (6) Select additional conditioning points using the criteria described below.

212 (7) Combine the conditioning points selected at step (6) with the original set of conditioning
213 points from the previous iteration to update facies spatial structure information needed by

214 T-PROGS, i.e., discrete transition probability and facies volumetric proportions. Then
215 generate a new ensemble of geostatistical realizations of facies field using T-PROGS.
216 (8) Repeat (2) to (7) until convergence or prescribed iteration number is reached.
217 It should be noted that, while T-PROGS is used in the procedure above, the ES-MDA method is
218 compatible with other geostatistical simulators for generating random fields of facies.
219 Steps (5) – (7) (shown in the gray shaded boxes in Figure 1) are the unique contribution of
220 our framework, which are not included in any existing facies-based EDA approaches. We term
221 these steps as a "reconditioning" procedure. For selecting the additional conditioning points in
222 step (6), the absolute changes of facies probability before and after an iteration at all the grid
223 cells are first ranked in the ascending order, then the locations with the top 1% change are
224 selected as the candidate points because they are more sensitive to the observation data in the
225 given iteration. The pool of candidate points accumulates over iterations. We then down select
226 from all these candidate points to a subset, based on the changes of the updated facies probability
227 values from their prior estimates. The down selection step is to ensure that the selected points
228 represent the "new" information content assimilated from the observational data using the prior
229 as the baseline. To avoid overfitting, we keep the number of additional conditioning points the
230 same as the number of observation data points. More research on the number of additional
231 conditioning points and how to control the size of candidate pool is warranted in a future study.
232

233 **3. Synthetic Example**

234 To demonstrate and evaluate our ES-MDA method for facies-based data assimilation, a
235 synthetic study of groundwater flow modeling in a domain with two facies was developed by
236 revising that of *Harp et al.* [2008] with a reduced amount of direct data, which are borehole logs

of sediment texture. The study domain (Figure 2a) is a 2-D confined aquifer with the size of 1000m × 200m in 10m × 20m resolution. The left and right boundaries at x = 0 m and x = 1000 m are set with constant hydraulic heads of 100 m and 95 m, respectively. The top and bottom boundaries at z = 200 m and z = 0 m are set as no-flow boundaries. The domain contains two facies with contrasting permeability, 10^{-9} m^2 for the facies that is more permeable and 10^{-12} m^2 for the less permeable one, which are comparable with the measured values for two contrasting geologic layers at the DOE Hanford site [*Chen et al.*, 2012; 2013]. We refer to these two facies as Hanford and Ringold hereafter. The true facies distribution was generated by *Harp et al.* [2008] with the volumetric proportions of 0.7 and 0.3 for Hanford and Ringold, respectively. The mean lengths of Ringold in the x (length) and z (thickness) directions are 300 m and 20 m, respectively. Borehole geological data (i.e., direct data) are assumed to be available at each grid cell along three wells located at x = 0 m, 250 m, and 500 m (marked by the vertical black lines in Figure 2a). Given that *Harp et al.* [2008] used direct data at five wells (the other two located at x = 750m and 1000m), the synthetic example of this study with smaller number of observation data is more challenging for data assimilation. Groundwater pumping with a constant rate of Q = 10.2 L/s was imposed at the domain center (x = 500 m). Transient head data (indirect data) were collected at seven discrete times until a steady state was reached at eight observation locations (marked with green dots in Figure 2a). The data were corrupted by measurement errors, which were modeled as white noise with the standard deviation of 1 cm.

The facies-based data assimilation framework was applied to estimate the spatial distribution of the two facies and their associated permeabilities. The initial ensemble of log-transformed permeability for Hanford and Ringold were generated from the Gaussian distributions with a variance of 0.5 and mean values of -8 ($\log_{10}$-m^2) for Hanford and -11 ($\log_{10}$-m^2) for Ringold.

Note that the initial guesses of mean permeability for the both facies were set one order of magnitude higher than their true values to test the robustness of our data assimilation framework. An ensemble size of 300 was used in our data assimilation process to ensure the convergence of ensemble approximation. Four iterations of ES-MDA were performed with the iteration coefficient of $\alpha_l = 4$ used in Eq. (1). The number of additional conditioning points used in the "reconditioning" procedure (Figure 1) was set to 50 (the number of head observations), and they were selected adaptively during the iterations to augment the spatial structure of the facies for the T-PROGS geostatistical simulations. The discrete lags for T-PROGS to generate a continuous-lag Markov chain model were 250m and 2m in the horizontal and vertical directions, respectively, consistent with the minimum lag distances of the borehole data. The flow simulation for each realization of the permeability field was performed using the high-performance reactive flow and transport code PFLOTRAN [*Hammond et al.*, 2014].

4. Results and Discussion

Upon the completion of ES-MDA with the reconditioning procedure, the probability of Ringold occurrence was calculated for each grid cell (by counting the occurrence frequency in the posterior ensemble of the facies field), and the results were compared with the true Ringold distribution (Figure 2a) to assess the accuracy of our estimation (Figures 2c, 2e, and 2g). Comparing Figures 2c and 2e of the first two assimilation steps with Figure 2b of the prior Ringold distribution shows that the facies distribution estimated using our method changed significantly towards the true field. Convergence was achieved after four iteration steps as there was negligible difference in estimations between the 3rd (results not shown) and 4th iterations (Figure 2g). The final estimate of facies distribution (Figure 2g) captured all the major features

283 of Ringold distribution, with remarkable improvements noted in the top-left corner and right side
284 of the domain highlighted in the green boxes. Most of the additional conditioning points (in
285 Figures 2c, 2e, and 2g) during the data assimilation steps occur within the highlighted regions,
286 where the initial uncertainty of facies distribution is high due to the lack of direct data.
287 We also compared the above results with those obtained without the reconditioning
288 procedure to evaluate the importance of imposing the facies spatial continuity in facies
289 delineation (Figures 2d, 2f, and 2h). It is evident that the removal of spatial structure
290 reconstruction at each data assimilation step decreased the overall accuracy of facies
291 reconstruction with much noisier spatial patterns. One posterior realization of facies field,
292 corresponding to the same randomly picked prior realization, is provided in Figures 2i and 2j
293 with and without reconditioning, respectively. The difference between the two realizations is
294 representative of all the realizations in the ensemble and is consistent with that between their
295 mean fields (e.g., the facies probability field) shown in Figures 2g and 2h. Imposing spatial
296 continuity through conditioning is effective in avoiding the unrealistic noisy structure of facies.
297 Thus this reconditioning step is essential when using data assimilation techniques for facies
298 delineation, especially when a smaller amount of direct data is available for the facies delineation.
299 The estimated volumetric proportion and mean length of each facies are compared to their
300 true values to evaluate the effectiveness of ES-MDA in capturing the primary parameters that
301 describe the spatial structure of facies distribution. Figures 3a-3c show that the constraint on
302 spatial continuity of facies improves the accuracy of ES-MDA in reproducing both the
303 volumetric portion and mean length for both facies, compared to ES-MDA without the
304 reconditioning procedure. The final estimates with the reconditioning procedure deviate less than
305 4% from their true values. On the other hand, the means of estimated permeabilities for both the

Hanford and Ringold facies are nearly identical to the true values regardless of the reconditioning, as shown in Figure 3d. The reconditioning only leads to marginal improvement in estimating the Ringold permeability.

Root mean square errors (RMSEs) between the estimated facies probabilities and the corresponding true values were calculated to assess the accuracy of facies estimation at each grid cell. The RMSEs represent the spatial average of goodness-of-fit over the entire domain, and they were evaluated for the estimated mean facies field and the individual realizations of facies field as:

$$\mathrm{RMSE}^{I}_{\mathrm{mean}} = \sqrt{\frac{1}{N_g}\sum_{j=1}^{N_g}\left(\frac{1}{N_e}\sum_{i=1}^{N_e} I_{i,j} - I^{\mathrm{ref}}_{j}\right)^2}$$
$$\mathrm{RMSE}^{I}_{i} = \sqrt{\frac{1}{N_g}\sum_{j=1}^{N_g}\left(I_{i,j} - I^{\mathrm{ref}}_{j}\right)^2}\ , \tag{6}$$

where N_g is the number of grid cells in the domain, N_e is the number of realizations, and I and I^{ref} are the facies indicators (1 for Ringold and 0 for Hanford). $\mathrm{RMSE}^{I}_{\mathrm{mean}}$ is calculated using mean field of facies, while RMSE^{I}_{i} is calculated for each realization of posterior facies field. The probability distribution of RMSE^{I}_{i} can be constructed from all the realizations for the range of goodness-of-fit among different realizations, and the distribution is plotted in Figure 3e.

The $\mathrm{RMSE}^{I}_{\mathrm{mean}}$ value of the mean facies field before ES-MDA is 0.349. After ES-MDA, it decreases to 0.316 (with reconditioning) and 0.348 (without reconditioning), confirming that the reconditioning led to an overall improvement in facies delineation across the domain. The probability density functions (PDFs) of RMSE^{I}_{i} plotted in Figure 3e also shows that the reconditioning procedure of our framework shifts the distribution of RMSEs towards smaller values, i.e., higher accuracy. To evaluate the $\mathrm{RMSE}^{I}_{\mathrm{mean}}$ in the horizontal and vertical directions,

the squared difference was averaged over the columns of the grid instead of over the entire domain. The pattern of $\mathrm{RMSE}^{I}_{\mathrm{mean}}$ along the horizontal direction (Figure 3g) shows a strong influence of the direct data on the facies estimation. Assimilating the indirect data has little effect in the region (e.g., that between x = 0 m and x = 500m) where a significant amount of direct data are available. However, a major improvement is observed in the region (e.g., the area with x > 500m) with little direct data, indicating the importance of the reconditioning procedure of our ES-MDA framework.

The RMSEs were also calculated for the hydraulic head field simulated using the estimated facies distribution and their associated permeabilities before and after applying ES-MDA. The calculation is the same as that for the facies distribution, with the only difference being averaging the squared difference over the time steps before averaging over space as shown below:

$$\mathrm{RMSE}^{H}_{\mathrm{mean}} = \sqrt{\frac{1}{N_g}\sum_{j=1}^{N_g}\sum_{k=1}^{N_t}\left(\frac{1}{N_e}\sum_{i=1}^{N_e}H_{i,j,k} - H^{\mathrm{ref}}_{j,k}\right)^2}$$
$$\mathrm{RMSE}^{H}_{i} = \sqrt{\frac{1}{N_g}\sum_{j=1}^{N_g}\sum_{k=1}^{N_t}\left(H_{i,j,k} - H^{\mathrm{ref}}_{j,k}\right)^2}\ , \tag{7}$$

where H and H^{ref} are the simulated head from the estimated and true facies field, respectively. Our ES-MDA method with reconditioning yields a $\mathrm{RMSE}^{H}_{\mathrm{mean}}$ value of 0.165m, which is significantly smaller than the value of 0.673m for the prior and the value of 0.235m for the ES-MDA without reconditioning. The PDF of RMSE^{H}_{i} (Figure 3f) shows higher density in smaller values, confirming the improvement in prediction of hydraulic head field. The $\mathrm{RMSE}^{H}_{\mathrm{mean}}$ values calculated for the columns of the grid show that ES-MDA can effectively reduce the errors in simulated hydraulic head field, and the improvement is enhanced when the spatially continuity of facies distribution is properly accounted for by reconditioning (Figure 3h). The larger RMSEs

346 near the pumping well are likely due to the numerical error in solving the pressure discontinuity
347 imposed by pumping.
348
349 **4. Conclusions**
350 A new data assimilation framework (ES-MDA) was developed for delineating the spatial
351 distributions of facies and for estimating facies permeability based on different data types by
352 integrating ensemble data assimilation with indicator geostatistics. A parameterization method
353 based on the level set concept was used to transform between discrete facies distribution and
354 continuous Gaussian random variables to allow the application of the data assimilation methods
355 developed for continuous variables for discrete facies delineation. The unique feature of our ES-
356 MDA framework is that the delineated facies distribution is not only informed by the direct and
357 indirect information, but also constrained by spatial continuity through the reconditioning
358 procedure at each assimilation step. The results from a two-dimensional synthetic example
359 demonstrated that our framework can accurately characterize facies distribution and associated
360 permeability. The reconditioning procedure is unique and innovative to our framework, and the
361 procedure was shown to be essential in maintaining spatial continuity in the facies distribution,
362 which is especially important for systems known to have preferential flow path. It should be
363 noted that successful facies delineation also depends on the contrast in facies properties. Further
364 testing the proposed method in a system with mild contrast in facies, such as the Borden aquifer
365 [*Ritzi et al.*, 2013], could be worthwhile. Although ES-MDA was demonstrated with a two-facies
366 system, extending it to more than two facies is straightforward by introducing additional
367 parameters for level set transformation [*Chang et al.,* 2010; *Mannseth*, 2011]. ES-MDA is

mathematically general, and thus can also be readily extended for delineating other facies beyond the hydrofacies, such as reactive facies and thermofacies.

Acknowledgement

This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Subsurface Biogeochemical Research Program (SBR). The first author was jointly supported by a DOE Early Career Award (DE-SC0008272), and the SBR Scientific Focus Area (SFA) at the Pacific Northwest National Laboratory (PNNL). The second author was supported by project DE-SC0008272 and NSF-EAR grant 1552329, and the third, fifth and sixth authors were supported by the SBR SFA at PNNL. The authors thank Haibin Chang for helpful discussions on level set based parameterization. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Requests for data that are not explicitly provided in the manuscript could be made to the corresponding author.

References

Aanonsen, S., G. Nævdal, D. Oliver, A. Reynolds, and B. Vallès (2009), The Ensemble Kalman Filter in Reservoir Engineering--a Review, *SPE Journal*, *14*(3), 393-412.

Agbalaka, C. C., and D. S. Oliver (2008), Application of the EnKF and Localization to Automatic History Matching of Facies Distribution and Production Data, *Mathematical Geosciences*, *40*(4), 353-374.

Bayer, P., A. Comunian, D. Hoyng, and G. Mariethoz (2015), High resolution multi-facies realizations of sedimentary reservoir and aquifer analogs, *Sci Data*, *2*, 150033.

Burgers, G., P. Jan van Leeuwen, and G. Evensen (1998), Analysis Scheme in the Ensemble Kalman Filter, *Monthly Weather Review*, *126*(6), 1719-1724.

Carle, S. F. (1999), *T-PROGS: Transition probability geostatistical software*, version 2.1 ed., 84 pp., University of California, Davis, CA.

Carle, S. F., and G. E. Fogg (1997), Modeling Spatial Variability with One and Multidimensional Continuous-Lag Markov Chains, *Mathematical Geology*, *29*(7), 891-918.

Chang, H., D. Zhang, and Z. Lu (2010), History matching of facies distribution with the EnKF and level set parameterization, *Journal of Computational Physics*, *229*(20), 8011-8030.

Chen, X., G. E. Hammond, C. J. Murray, M. L. Rockhold, V. R. Vermeu, and J. M. Zachara (2013), Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at Hanford 300 area, *Water Resources Research*, *49*(10), 7064-7076.

Chen, X., H. Murakami, M. S. Hahn, G. E. Hammond, M. L. Rockhold, J. M. Zachara, and Y. Rubin (2012), Three-dimensional Bayesian geostatistical aquifer characterization at the Hanford 300 Area using tracer test data, Water Resour. Res. , 48,W06501, doi:10.1029/2011WR010675.

Clark, Martyn P., Ying Fan, David M. Lawrence, Jennifer C. Adam, Diogo Bolster, David J. Gochis, Richard P. Hooper, Mukesh Kumar, L. Ruby Leung, D. Scott Mackay, Reed M. Maxwell, Chaopeng Shen, Sean C. Swenson, and Xubin Zeng (2015), Improving the representation of hydrologic processes in Earth System Models, *Water Resources Research,* 51 (8):5929-5956. doi: 10.1002/2015WR017096.

Dai, Z., A. Wolfsberg, Z. Lu, and H. Deng (2009), Scale dependence of sorption coefficients for contaminant transport in saturated fractured rock, *Geophysical Research Letters*, 36, L01403, doi:10.1029/2008GL036516.

Deutsch, C. V., and A. G. Journel (1992), *GSLIB: Geostatistical Software Library and User's Guide*, 340 pp., Oxford University Press, New York.

Dovera, L., and E. Della Rossa (2010), Multimodal ensemble Kalman filtering using Gaussian mixture models, *Computat Geosci*, *15*(2), 307-323.

Emerick, A., and A. C. Reynolds (2013), Ensemble smoother with multiple data assimilation, Computers & Geosciences, 55, 1-13.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, *99*(C5), 10143-10192.

Hammond, G.E., P.C. Lichtner, and R.T. Mills (2014) Evaluating the Performance of Parallel Subsurface Simulators: An Illustrative Example with PFLOTRAN, Water Resources Research, 50, doi:10.1002/2012WR013483.

Harp, D. R., Z. Dai, A. V. Wolfsberg, J. a. Vrugt, B. a. Robinson, and V. V. Vesselinov (2008), Aquifer structure identification using stochastic inversion, *Geophysical Research Letters*, *35*(8), L08404.

Harvey, Jud, and Michael Gooseff (2015), River corridor science: Hydrologic exchange and ecological consequences from bedforms to basins, *Water Resources Research,* 51 (9):6893-6922. doi: 10.1002/2015WR017617.

Gu, Y., and D. Oliver (2007), An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation, SPE Journal, 12(4), 438-446.

Jafarpour, B., and D. B. McLaughlin (2008), History matching with an ensemble Kalman filter and discrete cosine parameterization, *Computat Geosci*, *12*(2), 227-244.

Liu, G., Y. Chen, and D. Zhang (2008), Investigation of flow and transport processes at the MADE siste using ensemble Kalman filter, Advances in Water Resources, 31, 975-986.

Liu, N., and D. S. Oliver (2005), Ensemble Kalman filter for automatic history matching of geologic facies, *Journal of Petroleum Science and Engineering*, *47*(3-4), 147-161.

Moreno, D. L., and S. I. Aanonsen (2011), Continuous Facies Updating Using the Ensemble Kalman Filter and the Level Set Method, *Mathematical Geosciences*, *43*(8), 951-970.

Mannseth, T. (2014), Relation Between Level Set and Truncated Pluri-Gaussian Methodologies for Facies Representation, Mathematical Geosciences, 46(6), 711-731.

Oliver, D. S., and Y. Chen (2011), Recent progress on reservoir history matching: a review, *Computat Geosci*, *15*(1), 185-221.

Ritzi, R. W. Jr., L. Huang, R. Ramanathan, and R. M. Allen-King (2013), Horizontal spatial correlation of hydraulic and reactive transport parameters as related to hierarchical sedimentary architecture at the Borden research site, Water Resour. Res., 49, 1901–1913, doi:10.1002/wrcr.20165.

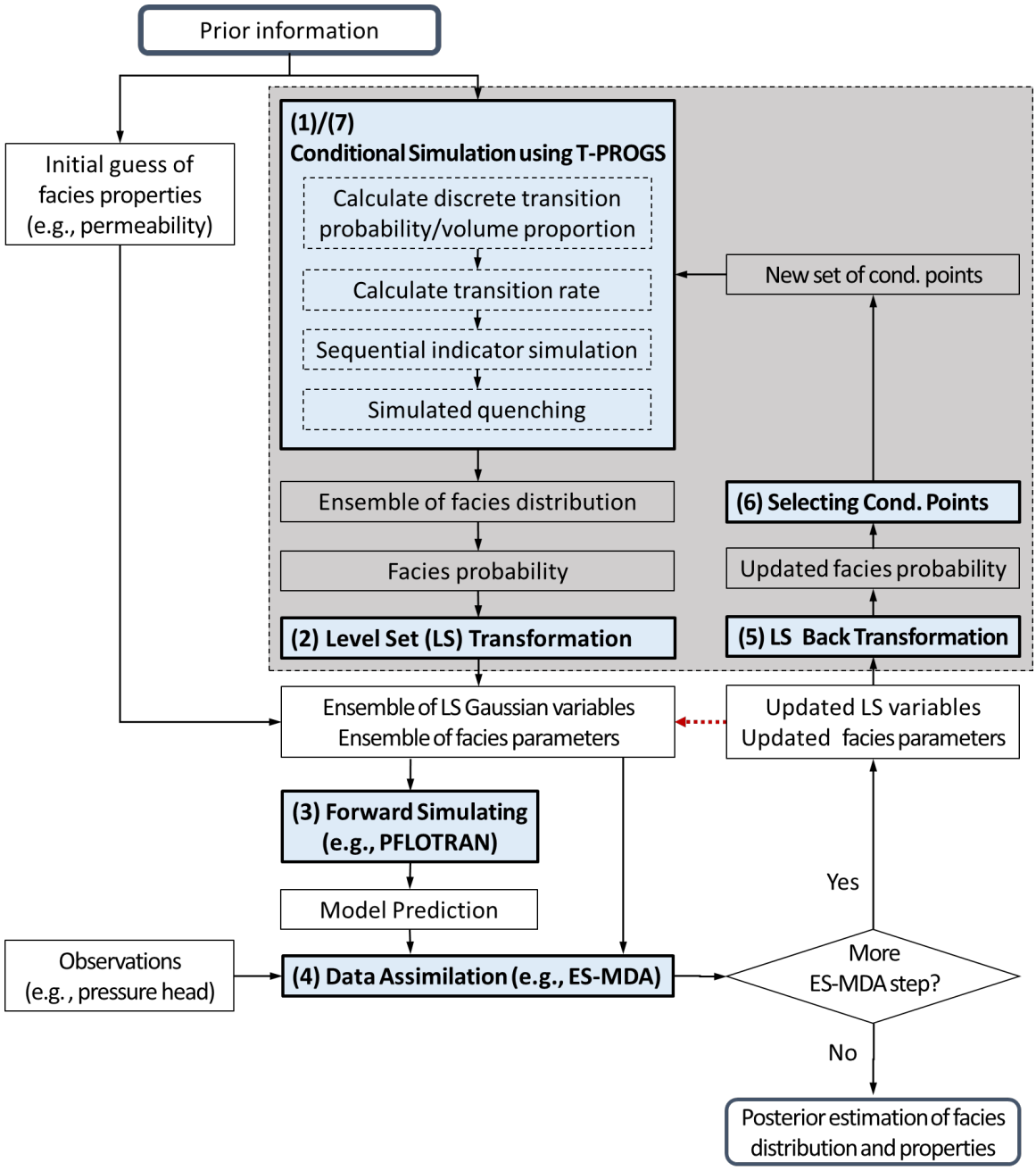
Sassen, D. S., S. S. Hubbard, S. A. Bea, J. Chen, N. Spycher, and M. E. Denham (2012), Reactive facies: An approach for parameterizing field-scale reactive transport models using geophysical methods, *WaterResour.Res.*, 48, W10526, doi:10.1029/2011WR011047.

Wainwright, H. M., J. Chen, D. S. Sassen, and S. S. Hubbard (2014), Bayesian hierarchical approach and geophysical data sets for estimation of reactive facies over plume scales, *Water Resour. Res.*, 50, 4564–4584, doi:10.1002/2013WR013842.

Yabusaki, S. B., Y. Fang, K. H. Williams, C. J. Murray, A. L. Ward, R. D. Dayvault, S. R. Waichler, D. R. Newcomer, F. A. Spane, and P. E. Long (2011), Variably saturated flow and multicomponent biogeochemical reactive transport modeling of a uranium bioremediation field experiment, *J Contam Hydrol*, *126*(3-4), 271-290.

Ye, M., and R. Khaleel (2008), A Markov chain model for characterizing medium heterogeneity and sediment layering structure, *Water Resources Research*, *44*, W09427.

464 **LIST OF FIGURES**



465

466 **Figure 1**. Flow chart of integrating ensemble data assimilation methods (e.g., ES-MDA) and
467 indicator geostatistical methods (e.g., T-PROGS) based on level set parameterization. The part
468 with grey shadow is the integration of geostatistics described in Section 2.2. The red dashed line
469 shows the EDA procedure without conditional simulation of facies at each data assimilation step
470 as adopted by most of the existing facies-based EDA methods. The step number (1)~(7) in
471 Section 2.2 is highlighted.
472

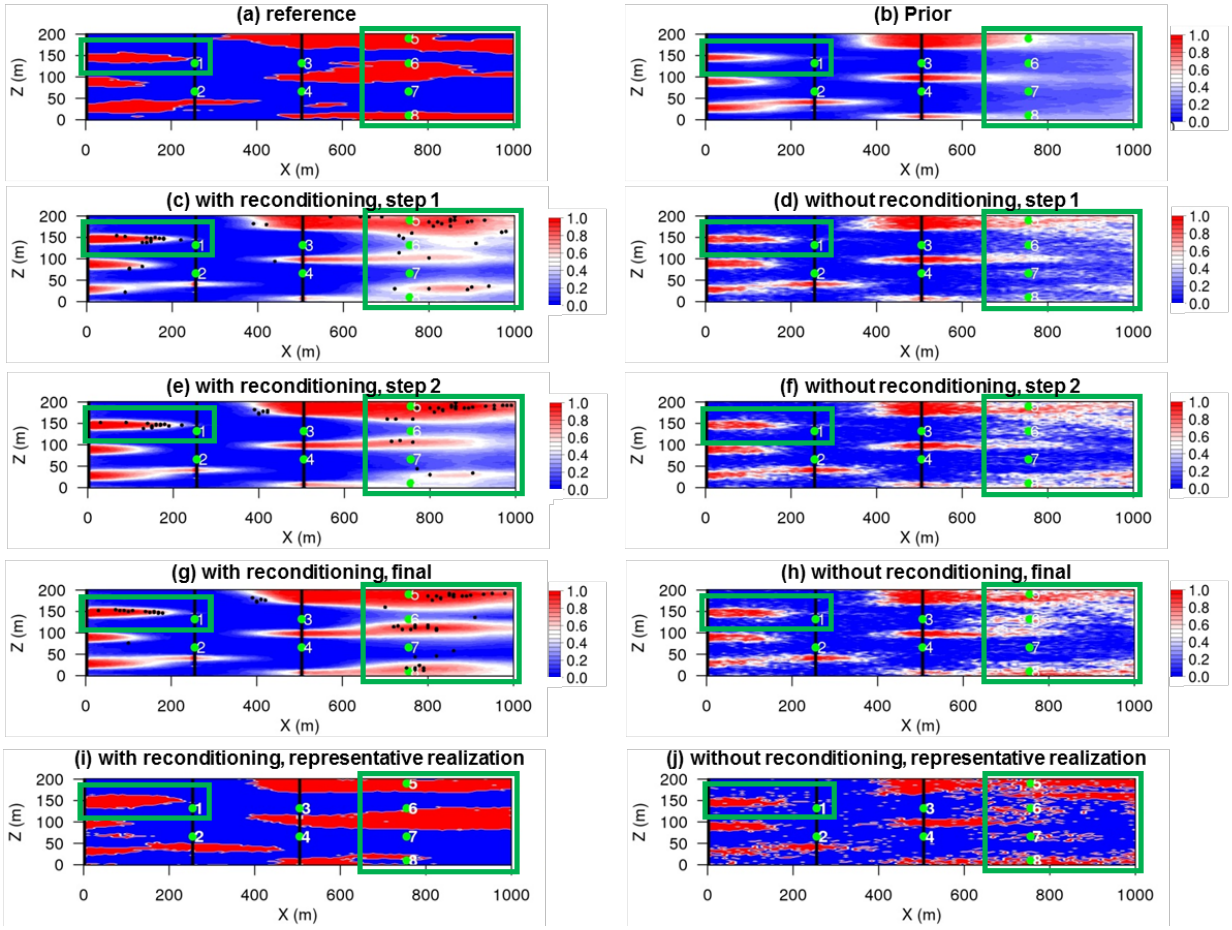


Figure 2. (a) Spatial distribution of two facies, Hanford (blue) and Ringold (red). Three black vertical lines (one coincides with the west boundary) indicate where facies indicator data are collected. The green dots indicate head observation locations. The pumping well is located on the central black line. (b) Prior probability field of Ringold estimated from the prior ensemble of facies field. (c, e, g) Posterior probability fields of Ringold after the 1st, 2nd and final data assimilation steps using our ES-MDA method with reconditioning (the black dots are additional conditioning points selected for conditional simulation of facies using T-PROGS). (d, f, h) Posterior probability fields of Ringold after 1st, 2nd and final data assimilation steps using ES-MDA without the reconditioning procedure. (i) A representative posterior realization of facies field using our ES-MDA method with reconditioing. (j) A representative posterior realization of facies field using the ES-MDA method without reconditioning.

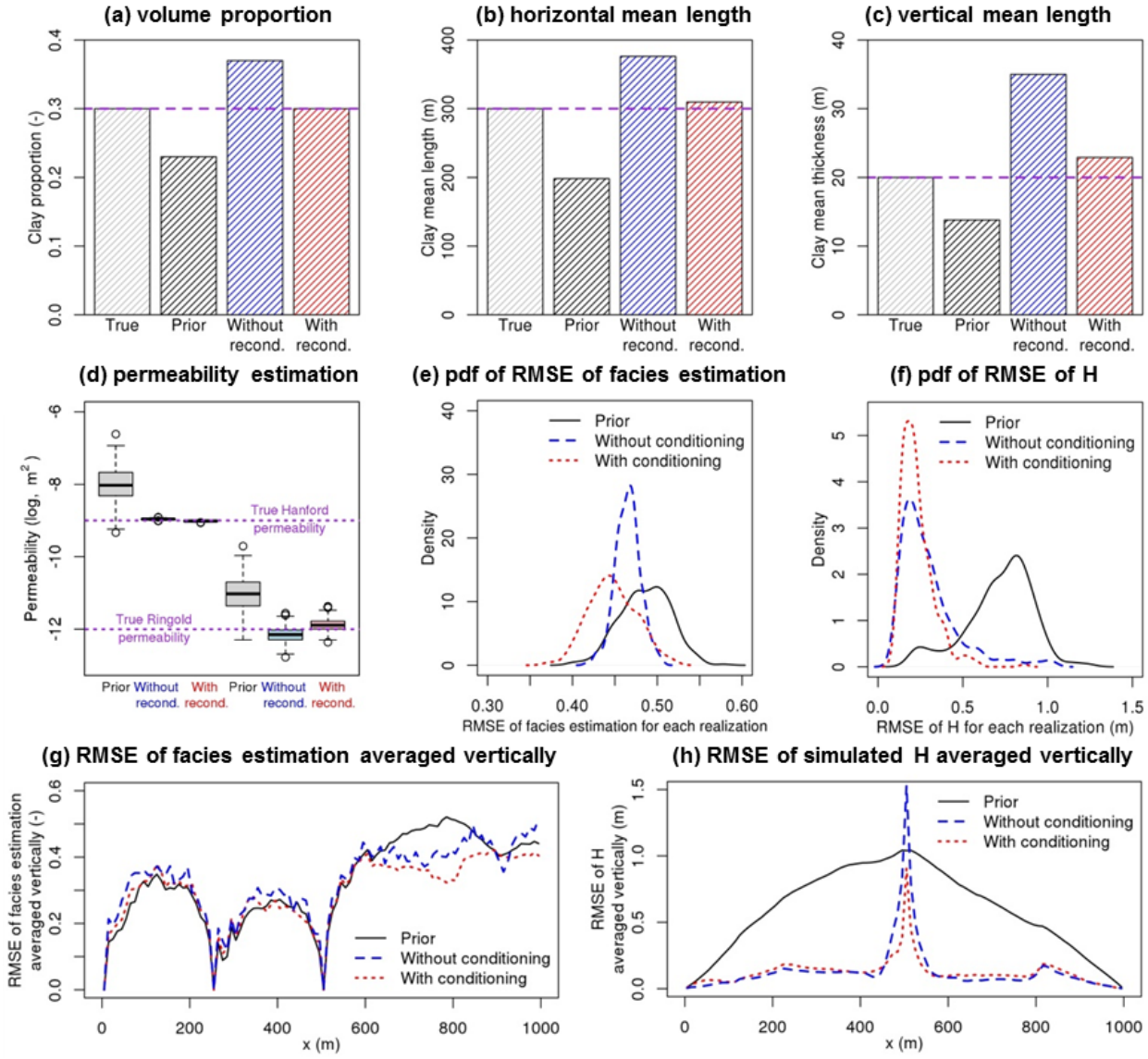


Figure 3. Comparisons between the prior and posterior estimates with or without reconditioning for Ringold volume proportion (a), horizontal mean length of Ringold (b), vertical mean length of Ringold (c), boxplots of prior and posterior estimations of permeabilities for Hanford and Ringold (d), pdfs of RMSE calculated on each realization of facies field (e), pdfs of RMSE calculated on each realization of hydraulic head (f), RMSE of facies probability averaged vertically (g), and RMSE of hydraulic head averaged vertically (h).