

INSPIRATION bulletin

CL:AIRE's INSPIRATION bulletins describe practical aspects of research which have direct application to the management of contaminated soil or groundwater in an agricultural context. This bulletin provides an introduction to the topic of data assimilation and sequential parameter optimisation in the context of groundwater modelling.

Copyright © CL:AIRE.

Geological consistency in self-optimising groundwater models using nested particle filters

1. Introduction

In an age of increasing access to environmental data, be it through wireless sensor networks or remote sensing from orbit, promising new opportunities arise for the modelling of environmental systems. The interest surrounding machine-learning technologies designed to capitalise on this data wealth often focuses on 'black-box' techniques like neural networks. While exceedingly good at replicating observed patterns (e.g., Learn: here are some water levels and corresponding rainfall. Ask: What could be the water level corresponding to this rainfall?), such techniques are of little use in hydrogeology, where the objectives in questions generally are not – and often cannot – be directly observed, and instead have to be inferred indirectly through system understanding (e.g., Learn: here are some water levels and ERT profiles. Ask: How high can my pumping rate be before the 30 day buffer zone cuts into the nearby industrial area?). This translation of knowledge between observed quantities and the subject of interest is generally facilitated by numerical, physically-based models. Traditionally, such models are calibrated to a given data set, most commonly water levels, and may then – if validated – enable their parameters to be used to investigate the actual objective.

Data assimilation (DA) techniques have found use in projects where (a) a steady stream of new data is available, and (b) an interest in steadily updating the model simulation exists. Applications in hydrogeology have, for example, been the regulation of irrigation water extraction rates (Kinzelbach, 2018) or petroleum engineering (Aanonsen *et al.*, 2009). In its most primitive state, DA merely extracts the model's boundary conditions from an incoming data stream and simulates autonomously forward, correcting its own predictive shortcomings in the process. In more sophisticated applications, it may also use these data to optimise the model's parameters. This bulletin considers the latter case.

Such algorithms harbour particular promise in settings dominated by manifold sources of uncertainties, be it from uncertain boundaries, scarce observations, or large spatial scales. In agricultural catchments, for example, where such uncertainties abound and interests of multiple stakeholders collide, it is imperative that any model consulted in the decision-making process is both objective and honest about its uncertainties. It is our hope that DA techniques may pose a promising first step towards fully-automated, large-scale

monitoring and early-warning systems for groundwater contamination. Towards this end – and for transport problems in particular – the ability to optimise under complex geological priors is required. As such, a primitive example of how nested particle filters could be used will be explored.

2. Bayesian Statistics

Most DA techniques are based on a stochastic (Bayesian) representation of the system. A probabilistic perspective is often adopted to reflect uncertainty about aspects such as the model's states (hydraulic heads, contaminant concentrations, etc.), parameters (hydraulic conductivities, specific yield, etc.), forcings (boundary conditions), or even the model's fidelity itself (forecast errors). In such a system, belief about uncertain variables is reflected by a probability density function (pdf) defined over parameter or state space. These spaces are defined by taking all unknown variables of a certain type (states and/or parameters) required for a model, arranging them in a vector, and interpreting any realisation (i.e., specific values) thereof as coordinates of a point in a high-dimensional space. The pdf defined over this space then allows degrees of understanding about all different possible narratives to be expressed.

Unfortunately, it is generally impossible to formulate this pdf analytically. Instead, techniques like the Ensemble Kalman Filter (EnKF) or the particle filter (PF) approximate it through an ensemble of Monte Carlo samples, which are in essence a selection of different model parameterisations considered in parallel. The reader is referred to Doucet & Johanson (2009) for information on the topic of particle filters. In this bulletin, a slight variation of the nested particle filter setup known as SMC² is employed (Chopin *et al.*, 2013). It is noted that a variety of other approaches such as MCMC with likelihoods derived from a classic error model or non-sequential approaches might be equally feasible.

3. Nested Particle Filters for Parameter Optimisation

In hydrogeological model optimisation attempts are often made to infer unknown subsurface parameters from a set of state observations (e.g., hydraulic heads). In a Bayesian context, this can be formulated as an attempt to derive the posterior pdf of the parameters θ with respect to the state observations $y_{1:t}$, where the

INSPIRATION bulletin

subscript denotes a time-sequence of observations:

$$p(\theta|y_{1:t}) = \frac{p(\theta)p(y_{1:t}|\theta)}{p(y_{1:t})} \quad \text{Equation 1}$$

Where $p(\theta)$ is the prior parameter pdf (our belief about the parameters before the assimilation of data), $p(y_{1:t}|\theta)$ the likelihood (essentially an evaluation of the performance of different parameters), and $p(y_{1:t})$ a normalisation factor. Unfortunately, it is generally difficult to obtain this likelihood directly. Instead, Equation 1 is expanded in Equation 2 to include the sequence of simulated states $x_{1:t}$:

$$p(\theta|y_{1:t}) = \frac{p(\theta) \int p(x_{1:t})p(y_{1:t}|x_{1:t}) dx_{1:t}}{p(y_{1:t})} \quad \text{Equation 2}$$

It is possible to quantify the observational likelihood $p(y|x)$ directly, which allows a nested particle filter to be formulated:

$$p(\theta|y_{1:t}) \propto \sum_{i=1}^{N_\theta} \left(\delta(\theta_i)p(\theta_i) \sum_{j=1}^{N_x} \left(\delta(x_{j,1})p(x_{j,1}) \prod_{s=1}^t p(y_s|x_{j,s}) \prod_{s=2}^t p(x_{j,s}|x_{j,s-1}, \theta_i) \right) \right) \quad \text{Equation 3}$$

where $\delta(A)$ denotes the Dirac delta measure centred on A , \propto denotes proportionality, and the subscripts i and j denote parameter and state particle indices, respectively. Equation 3 illustrates the eponymous nested structure: An outer particle filter (outer sum) of N_θ particles for inference about the parameters, and N_θ inner particle filters (inner sum) of N_x particles each for inference about the states, conditional on a specific parameter set, from which the likelihood $p(y_{1:t}|\theta)$ is ultimately inferred by integration. This likelihood then serves as the basis for the unnormalised particle weights. The term $p(x_{j,s}|x_{j,s-1}, \theta_i)$ is a stochastic forecast, which is normally represented by a deterministic simulation with an arbitrary model plus a small random Gaussian error. Its purpose is to represent the information loss from using an imperfect model (Figure 1): since your model is not reality, its prediction will be overconfident; this can be remedied by explicitly increasing the uncertainty of its predictions¹.

The weight-based forecast error implemented here consists of an exponentiation of the unnormalised particle weights $W_{j,s}$ with a factor α between 0 (*complete loss of information*) and 1 (*deterministic forecast*).

$$W_{j,s} = W_{j,s-1}^\alpha \quad \text{Equation 4}$$

The subsequent normalisation then results in partially equalised weights (Figure 1), thereby increasing the uncertainty of x . This has the advantage that it preserves the deterministic simulation (thus preserving its internal properties such as mass balances), but conversely does not allow the particles to take on states the model cannot predict. It is also worth noting that this exponentiation is not restricted to the states: it also carries over to the parameters, technically constituting a covariance inflation, thus allowing the

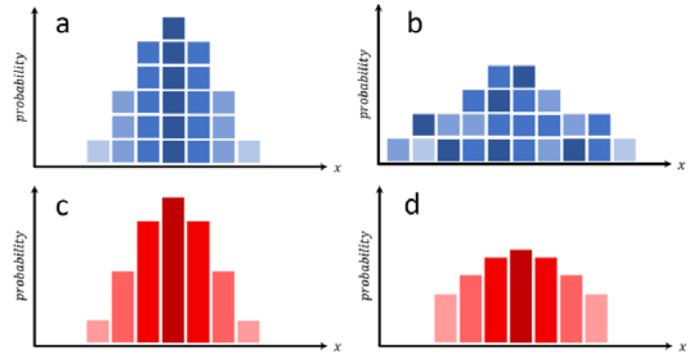


Figure 1: Difference between stochastic (a→b) and weight-based (c→d) forecast error in a histogram representation of the particle approximation: boxes represent individual particles, their height represent their respective normalised *weights* (corresponding to discrete probabilities). The wider the distribution (pile of boxes), the higher the entropy of the approximated probability distribution, and the less information is available about arbitrary variable x . The stochastic forecast error (a,b) achieves this by adding a random error to each (*diffusing* the particles through state space), the weight-based error achieves a similar effect by equalising the particles' weights.

optimisation process to 'forget' the past. This may or may not be desirable, depending on the objective. Furthermore, particle filters unfortunately suffer from the so-called *curse of dimensionality* (e.g., Bengtsson *et al.*, 2008) which renders them virtually ineffective in high-dimensional systems (i.e., those with many unknowns). A possible way to alleviate this issue is dimension reduction, for example via hyperparameterisation: instead of treating each cell parameter as an independent unknown variable, auxiliary parameters (ideally of lower number than the model parameters) are introduced instead. These so-called hyperparameters can be used to construct a full parameter field. Conveniently, hyperparameterisation can also be used to enforce conformity with a prescribed geological characterisation. This is demonstrated below for nested particle filters in a hydrogeological application with a synthetic example.

4. Synthetic Example

For illustration, the case of a two-dimensional unconfined groundwater model implemented in MODFLOW-USG is considered (Panday *et al.*, 2013). The parameter field consists of a high-conductive paleo-riverbed extending from north to south. The northern edge features a time-varying prescribed sinusoidal head boundary with an amplitude of 1 m and an offset of 3 m from the aquifer. The southern edge consists of a constant fixed-head boundary of 1 m. The western and eastern boundaries are prescribed zero head gradient. The aquifer bottom is at an elevation of zero. Head observations are extracted from three observation wells along a north-south axis in the centre of the model domain (Figure 2) with a standard deviation of 0.25 m. The true hydraulic conductivity of the two facies (paleo-riverbed and background) is $10^{-1.5}$ and 10^{-4} m·s⁻¹, respectively.

¹ A more practical example: Assume you bring your dog to a park. Your *variable of interest* is the dog's position, and you have formulated a mental *model* of your dog's behaviour. You close your eyes, let the dog off the leash, and make a prediction. You predict that the dog takes a beeline to its favourite spot on the river (*deterministic forecast*). Acknowledging the imperfection of your model, e.g. from a diversion in form of a squirrel along the way, you soften your prediction: the dog will probably end up somewhere nearby its favourite spot (deterministic forecast + *error*). Then you open your eyes, realise that your dog ran precisely the other way and is currently enthusiastically digging head-first through a toppled rubbish bin (*data assimilation*). You sigh heavily and adjust your model for next time (*optimisation*).

INSPIRATION bulletin

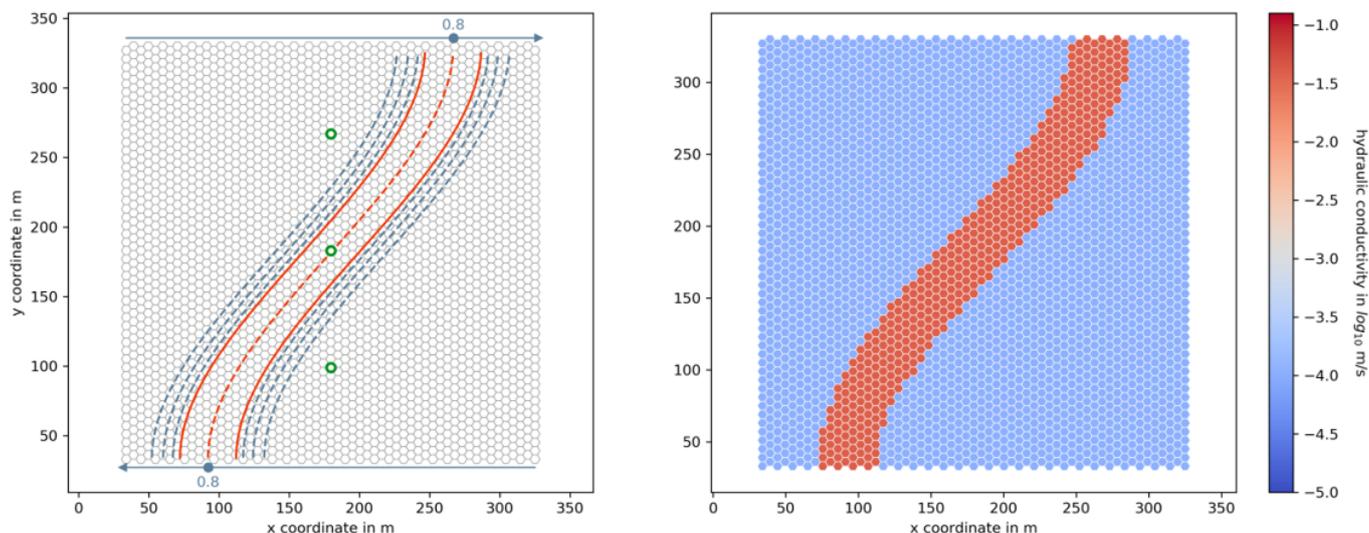


Figure 2: Synthetic reference for the nested particle filter demonstration. The left side shows a conceptual setup: the hyperparameter a is defined in opposite directions along the northern and southern boundary, so one value defines both start and end point. The riverbed spline connects both (dashed orange line), with the full lines illustrating the 20 m offset. The blue dashed lines signify the second facies outside of this riverbed. The green rings show the location of the observation wells. The right side shows the resulting reference parameter field.

The full parameter field has 2850 unknown values of hydraulic conductivity. This field is hyperparameterised with only three variables: the first is a , ranging from 0 to 1, which defines relative positions along the northern and southern boundaries. Based on these start and end points, a spline with orthogonal derivatives is drawn to the boundaries and defines points within a distance of 20 m as part of the paleo-riverbed. The second and third hyperparameters are the hydraulic conductivities of the facies, which are initially assumed to be unknown within the range of $10^{-0.5}$ and 10^{-5} $\text{m}\cdot\text{s}^{-1}$. The full parameter field is then assembled by assigning hydraulic conductivities for facies based on the facies map provided through the factor a . In doing so, a 2850-dimensional parameter space is simplified to a 3-dimensional hyperspace. The proposal function for the hydraulic conductivity hyperparameters is a zero-centred Gaussian with a standard deviation of $0.1 \log \text{m}\cdot\text{s}^{-1}$, and for a is a uniform distribution between 0 to 1.

The simulation covers a period of 100 hours where data is assimilated every two hours. [NB. We commit slight sacrifice by introducing a weight-based forecast error for performance-related reasons despite working in a synthetic (and thus theoretically error-free) setting. We further limit ourselves to a particle size of $N_{\theta}=250$ for the outer and $N_x=1$ for the inner particle filters for similar reasons.]

The system is designed to be symmetrical relative to the central axis along which the observation wells are aligned. This results in at least two functionally equivalent parameter fields: the parameter field depicted on the right side of Figure 2, and the same field mirrored along the x-direction. For cells which have ambivalent facies adherence, a bi-modal marginal pdf along the corresponding parameter space dimension would thus be expected. This feature is used to test whether the particle filter can capture this uncertainty.

5. Results

The results of the nested particle filter simulation are depicted in Figure 3. The expected parameter field does express the desired symmetry and the pdfs of the parameter space dimensions of the off-centre meander cells are indeed bi-modal. However, there is also a difference between the obtained results and the reference case: the hydraulic conductivity of the meander and background is markedly lower than in the synthetic reference. This difference seems to be compensated by placing the meander start and end-points closer towards the western and eastern edges, since the identified parameter fields appear to yield equivalent state predictions at the observation points (white and black isolines).

A possible interpretation of this phenomenon could be that these settings provide more stability than the optimal values around the true reference: the blurred-out features of the meanders at the edges and the spread of the pdf suggest there is some leeway for deviations from the identified optima.

6. Discussion

In this bulletin, a hydrogeological implementation of the nested particle filter is presented for a very simple synthetic example. Hyperparameterisation was employed, and successfully approximated non-Gaussian features of the parameter pdf. While the hyperparameters obtained differ slightly from the synthetic reference, the state predictions at the observation wells are equivalent to the reference. The ability of sequential, self-optimising groundwater model frameworks to honour prescribed but uncertain geological structures during parameter estimation will be important for their future use in many applications, especially transport problems. Particle filters are not strictly required for this, other sequential optimisers based around Markov Chain Monte Carlo (MCMC) updates may cope well with the often highly nonlinear state responses to complex parameter updates.

INSPIRATION bulletin

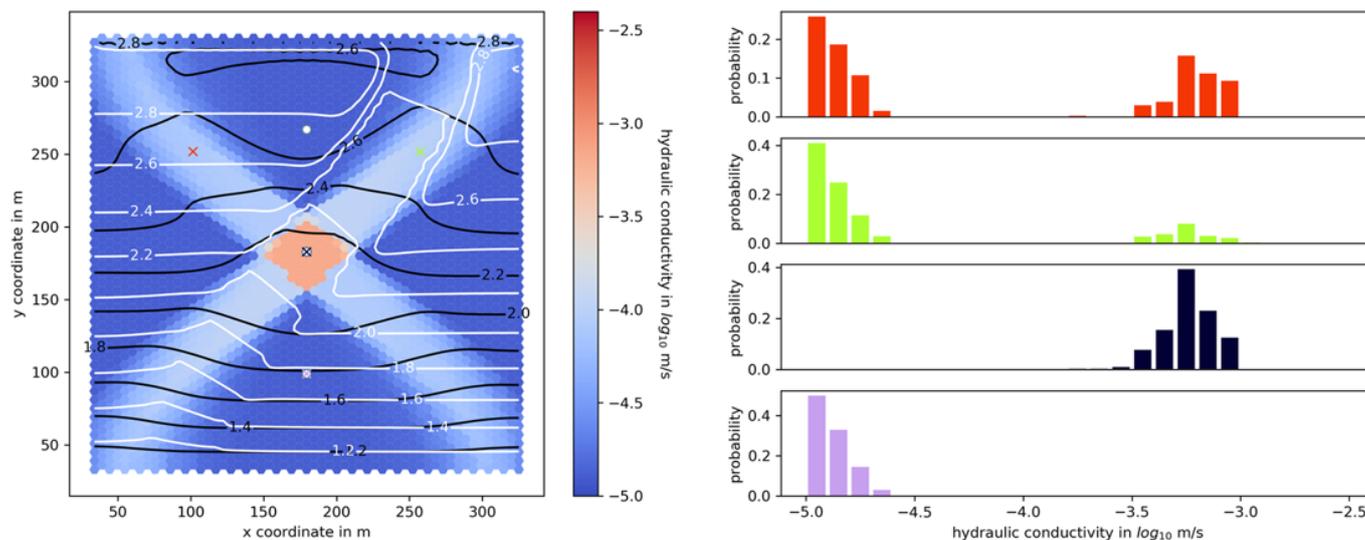


Figure 3: Results of the nested particle filter run. Left side: Expected parameter field (coloured hexagons), expected head distribution (black isolines), and true head distribution (white isolines) at the end of the simulation period. White circles mark observation wells and crosses mark the locations selected for investigation of the marginal parameter pdf. The correspondingly colour-coded histograms are plotted to the right.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675120 for the project entitled "Managing soil and groundwater impacts from agriculture for sustainable intensification – INSPIRATION".

References

- Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C., & Vallès, B. 2009. The ensemble Kalman filter in reservoir engineering--a review. *Spe Journal*, 14(03), 393-412.
- Bengtsson, Thomas, Peter Bickel, & Bo Li. 2008. "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems." *Probability and statistics: Essays in honor of David A. Freedman*. Institute of Mathematical Statistics. 316-334.
- Chopin, N., Jacob, P. E., & Papaspiliopoulos, O. 2013. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 397-426.

- Doucet, A., & Johansen, A. M. 2009. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704), 3.
- Kinzelbach, W. 2018. How to control overpumping of aquifers. https://www.ethz.ch/content/dam/ethz/special-interest/baug/ifu/hydrology-dam/documents/events/Seminars2018-2019/ifu_hyd_SeminarKinzelbach.pdf. [Accessed 31 Jan. 2019].
- Panday, S., Langevin, C. D., Niswonger, R. G., Ibaraki, M., & Hughes, J. D. 2013. MODFLOW – USG Version 1: An Unstructured Grid Version of MODFLOW for Simulating Groundwater Flow and Tightly Coupled Processes Using a Control Volume Finite-Difference Formulation. U.S. Geological Survey (Techniques and Methods 6-A45).

For more information on the INSPIRATION Project, please visit:
www.inspirationitn.co.uk

If you would like further information about other CL:AIRE publications please contact us at the Help Desk at www.claire.co.uk