DART-PFLOTRAN: An Ensemble-based Data Assimilation System for Estimating Subsurface Flow and Transport Model Parameters

Peishi Jiang[a], Xingyuan Chen[a,*], Kewei Chen[a], Jeffrey Anderson[b], Nancy Collins[b], Tim Hoar[b] and Moha EL Gharamti[b]

[a]*Pacific Northwest National Laboratory*
[b]*NCAR Data Assimilation Research Section*

ARTICLE INFO

Keywords:
DART
PFLOTRAN
ensemble-based data assimilation
inverse modeling

ABSTRACT

Ensemble-based Data Assimilation (EDA), based on the Monte Carlo approach, has been effectively applied to estimate model parameters through inverse modeling in subsurface flow and transport problems. However, implementation of EDA approach involves a complicated workflow that include setting up and executing ensemble forward model simulations, processing observations and model simulation results for parameter updates, and repeat for sequential or iterative EDA. To facilitate the management of such workflow and lower the barriers for adopting EDA-based parameter estimation in subsurface science, we develop a generic software framework linking the Data Assimilation Research Testbed (DART) with a massively parallel subsurface FLOw and TRANsport code PFLOTRAN. The new DART-PFLOTRAN leverages both the core data assimilation engines in DART and the computational power afforded by PFLOTRAN. In addition to the standard smoother and filtering options, DART-PFLOTRAN enables an iterative EDA workflow based on the Ensemble Smoother for Multiple Data Assimilation method (ES-MDA) to improve estimation accuracy for nonlinear forward problems. We verify the implementation of ES-MDA in DART-PFLOTRAN using two synthetic cases designed to estimate static permeability and dynamic exchange fluxes across the riverbed, respectively, from continuous temperature measurements made across a depth profile. One-dimensional hydro-thermal simulations are performed in both cases to relate temperature responses with the parameters of interest. In the case of estimating dynamic parameters, we demonstrate the flexibility of DART-PFLOTRAN in automating sequential ES-MDA workflow, which will significantly reduce the time researchers spend on managing complex workflows in similar applications. Both studies yield accurate estimations of the parameters compared to their synthetic truth, while ES-MDA leads to more accurate estimation when a high level of nonlinearity exist between observed responses and unknown parameters. With a code base in Python and Fortran, DART-PFLOTRAN paves the way for applications in large-scale subsurface inverse modeling by automating the complex workflow of sequential ES-MDA that can be executed on various computing platforms.

1. Introduction

Ensemble-based Data Assimilation (EDA) methods, including Ensemble Kalman Filter (EnKF) and Ensemble Smoother, have been extensively used to update model state vectors or estimate model parameters in various Earth science domains [1, 2, 3, 4]. Using a Monte Carlo-based ensemble representation of the joint probability for model states or parameters, EDA not only relaxes the constraint of Gaussian states required by the classic Kalman Filter, but also enables the nonlinear evolution of system states through physics-based process models. In subsurface hydrology and petroleum engineering, EDA has been widely used for parameter estimation (e.g., hydraulic conductivity or permeability) or "history matching" using field observations, such as hydraulic head, soil moisture, and tracer concentrations [5, 6, 7, 8, 9]. Iterative EDA methods, including ensemble smoother with multiple data assimilation (ES-MDA) [10, 11] and ensemble randomized maximum likelihood (EnRML) [12, 13, 7], have been developed to alleviate the accuracy deterioration caused by the nonlinear relationship in the evolution of model states or between model parameters and model states, similar to the Gauss-Newton or Levenberg-Marquardt approaches for nonlinear optimization. ES-MDA has been applied to delineate distinct geological facies and estimating permeability and exchange fluxes through riverbed through inverse modeling for its computational efficiency and estimation accuracy [8, 14].

*Corresponding author: Xingyuan.Chen@pnnl.gov
ORCID(s):

Although the iterative ES methods significantly reduce the number of forward simulation restarts required to conserve physical laws [7], the implementation of ES-MDA for large-scale inverse modeling is not trivial besides the complex workflow in launching multi-physics, parallel forward simulations, which is often required for managing the computational challenges [7, 8, 15]. Therefore, a user-friendly software framework for performing EDA associated with computationally intensive forward models and heterogeneous observational data will significantly increase scientific productivity. There exist multiple community-supported data assimilation tools, including PEST++ and Data Assimilation Research Testbed (DART). PEST++ , developed by U.S. Geological Survey for both parameter estimation and uncertainty analysis, adopts an iterative ensemble smoother for solving Gauss-Levenberg-Marquardt algorithm in model calibration [16]. DART, developed by the National Center for Atmospheric Research, provides a variety of EDA tools, including different filter techniques as well as various localization and inflation options [17]. Here, we employ DART as the core assimilation due to its modular structure that allows integration with various forward simulators by customizing a model-specific interface while keeping the data assimilation engine of DART and the forward simulator intact. DART has been successfully linked with a number of community codes for Earth system research, such as the Weather Research and Forecasting Model [18], the Community Atmosphere Model [19], and the Community Land Model [20], to facilitate model-data integration and consequently improve model accuracy.

The objective of this study is to develop a generic EDA software framework for improving subsurface flow and transport models by linking DART with PFLOTRAN [21], an open-source parallel subsurface flow and reactive transport model. Sequential ES-MDA, which performs ES-MDA in a sequence of assimilation time windows, is considered as a generic EDA approach that is flexible to be configured for performing traditional EnKF and ES-MDA. One key feature of PFLOTRAN is its embedded ensemble simulation capability, which greatly facilitates the implementation of EDA workflow for subsurface permeability estimation as demonstrated in multiple applications [7, 8]. We will allow flexible data subsetting in space and time to reduce the data dimension by sequentially assimilating those data subsets using ES-MDA. We implemented the DART-PFLOTRAN in a combination of Python, C-shell, and Fortran scripting. We also provide a Jupyter notebook [22] template as an alternative to python scripting for users to configure the data assimilation options, set up forward simulation models, and eventually execute the sequential ES-MDA workflow using the C-shell script. Jupyter notebook not only provides a straightforward way of documentation using the Markdown language, but also serves as an interactive coding and visualization platform. To verify the performance of DART-PFLOTRAN software framework, we employ the proposed framework to conduct sequential ES-MDA through two synthetic case studies that aim to estimate the exchange fluxes across sediment-water interface from continuous temperature measurements.

In the remaining of the paper, Section 2 provides an overview of sequential ES-MDA workflow and the detailed design of the DART-PFLOTRAN framework. Then, in Section 3, we verify the implementation of inverse modeling framework using two synthetic test cases. The estimated static and dynamic parameters are compared against the synthetic true values to assess implementation success. A brief conclusion is drawn in Section 4.

2. Methodology

In this section, we first describe the general workflow of sequential ES-MDA. Then, we introduce the detailed software design of the DART-PFLOTRAN framework, which includes enabling ensemble smoother in DART, the integrated workflow for performing sequential ES-MDA in DART-PFLOTRAN, and utilities used for coupling DART with PFLOTRAN.

2.1. Sequential ES-MDA

Figure 1 illustrates the steps we take in sequential ES-MDA to assimilate different subsets of data. The observation data contain N time steps t_j $(1 \leq j \leq N)$, which are divided into M sub-domains or assimilation windows, i.e., $\boldsymbol{t}^i$ with $i = [1, ..., M]$. Each window $\boldsymbol{t}^i$ contains one or multiple consecutive observations. The workflow starts with a model spin-up to ensure that the model reaches a reasonable initial state for forward simulations that generate state predictions to confront observation data. After the spin-up, ES-MDA is sequentially performed on each $\boldsymbol{t}^i$ following a prescribed order. Within each assimilation time window, ES-MDA assimilates all the observations taken within that time window, which may include more than one observation step. Note that EnKF is a special case of this general workflow when observations from one time step are assimilated for every $\boldsymbol{t}^i$ (i.e., $M = N$) without the multiple/iterative data assimilation. Similarly, ES is another special case of this general workflow when all the observations are included in one single assimilation time window without iterations.

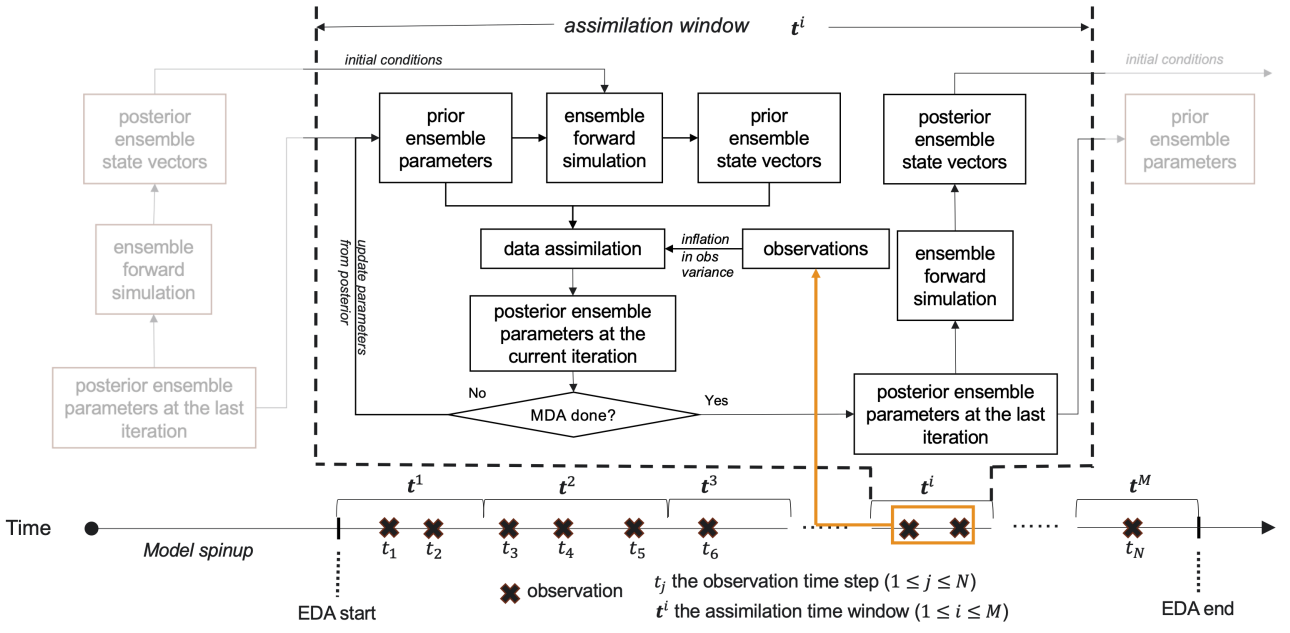

Figure 1: Illustration of the workflow for sequential ES-MDA. The posterior ensemble states are obtained from rerunning the forward simulation using updated parameters. The posterior end states from a previous assimilation time window serve as the initial states for the next assimilation time window.

At the first assimilation window $\boldsymbol{t}^1$, the prior ensemble of the model parameters are sampled from their prior distributions. In the remaining windows $\boldsymbol{t}^i$, the prior ensemble for static parameters are directly adopted from their updated posterior ensemble at the preceding assimilation window $\boldsymbol{t}^{i-1}$; the prior ensemble for dynamic parameters, if assuming continuity in time, can be sampled from distributions that preserve their mean values computed from the posterior ensemble of the preceding assimilation window with the same variance or lower and upper bounds used in the first assimilation time window; otherwise, the prior ensemble used in the first assimilation time window can be adopted in all the remaining time windows as well.

During each iteration of ES-MDA, ensemble forward simulations are performed to generate the prior ensemble of the state vectors within the assimilation time window. Model parameters are updated using the following equation [11]:

$$\boldsymbol{m}^u_{k,l} = \boldsymbol{m}^f_{k,l} + \boldsymbol{C}^f_{MD,l}(\boldsymbol{C}^f_{DD,l} + \alpha_l \boldsymbol{C}_D)^{-1}(\boldsymbol{d}_{obs} + \sqrt{\alpha_l}\boldsymbol{C}_D^{1/2}\boldsymbol{z}_k - \boldsymbol{d}^f_{k,l}), \quad k = 1, ..., N_e \quad \text{and} \quad l = 1, ..., L, \tag{1}$$

where the superscripts u and f refer to updated and forecast, respectively; the subscripts k and l are the indices of the ensemble member and the iteration, respectively; N_e is the ensemble size; L is the total number of iterations in ES-MDA; $\boldsymbol{m}^u_{k,l}$ and $\boldsymbol{m}^f_{k,l}$ are the kth ensemble member of the updated (i.e., posterior) and forecast (i.e., prior) parameters, respectively, at the lth iteration; $\boldsymbol{d}_{obs}$ is the observation data; $\boldsymbol{z}_k$ is the corresponding observation noise vector sampled from independent standard normal distributions for the kth ensemble member; $\boldsymbol{d}^f_{k,l}$ is the kth ensemble member of the predicted observation variables by the forward model driven by the prior ensemble of the parameters at the lth iteration; $\boldsymbol{C}^f_{MD,l}$ is the cross-covariance matrix between the prior parameters and the predicted observation variables; $\boldsymbol{C}^f_{DD,l}$ is the auto-covariance matrix of the predicted observation variables based on all the ensemble members $\boldsymbol{d}^f_{k,l}$; $\boldsymbol{C}_D$ is the auto-covariance matrix of the observation errors; and α_l is the inflation coefficient at the lth iteration, satisfying $\sum_{l=1}^{L} 1/\alpha_l = 1$.

Once all the iterations within the assimilation time window are completed, the posterior ensemble of the parameters are fed to the forward simulator to generate the posterior ensemble of the state vector at the end of the assimilation window $\boldsymbol{t}^i$, which will then serve as the ensemble of the initial conditions for the next assimilation time window.

2.2. Design of the DART-PFLOTRAN framework

We developed a new software framework to perform the sequential ES-MDA illustrated in Figure 1 by linking DART and PFLOTRAN. We first enabled ensemble smoother capability in DART (see Section 2.3) to leverage all the assimilation options in its core data assimilation engine. Then, we developed multiple utility functions (see Section 2.5) to modularize the execution of DART-PFLOTRAN in four primary steps: (1) configuring data assimilation specifics; (2) PFLOTRAN forward model configuration and preparation; (3) DART preparation; and (4) performing sequential ES-MDA. Scripts were developed in Python and C-shell to conduct the first three steps and the last step, respectively. A Jupyter notebook is also provided to integrate the entire workflow, which can be used as a tool to learn the framework and as a template for adapting to other applications.

2.3. Enabling ensemble smoother in DART

Designed for updating model states using the filter approach, the original DART adopts a local least-square optimization algorithm, referred to as the Anderson & Collins algorithm hereafter [23, 24]. The algorithm obtains the posterior state and parameter ensembles by adding up all individual increment ensembles which result from assimilating a single dimension of the multi-dimensional observation at a given time step. Although the original DART does not directly allow the assimilation of observations from multiple time steps within a given time window, the underlying Anderson & Collins algorithm offers the flexibility to be readily extended for ensemble smoother-based approaches. It is noted that the Anderson & Collins algorithm is theoretically equivalent to using Eq. (1) for non-iterative updating (i.e., setting l=1 in Eq. (1)) when assimilating one observation dimension in each time window [24], and is also practically equivalent to the ES using Eq. (1) when several multi-dimensional observations are assimilated within the window. Therefore, we enabled the smoother option in DART by modifying how the original DART maps the observations with model simulated states in a given assimilation time step. Observed or modeled states from different time steps are treated as multiple dimensions of data from a single time step. In doing so, both filter and smoother options are now available in the assimilation engine of DART.

2.4. Integrated workflow of performing sequential ES-MDA in DART-PFLOTRAN

With the smoother option enabled in DART, we run DART-PFLOTRAN following the integrated workflow shown in Figure 2:

Step 1: Data assimilation configuration. The following information is required to configure a DART-PFLOTRAN application: (1) the path to relevant files/folders, such as the PFLOTRAN executable, the application folder storing PFLOTRAN files and DART prior/posterior ensemble, and the DART-PFLOTRAN framework folder storing the utility files needed to execute the workflow (e.g., those described in Section 2.5); (2) information about the observation data, such as the observation variables to be assimilated, the spatio-temporal domains of the observations, and the folder location of the observation file; (3) information about the model parameters to be updated, such as the list of parameter names and their corresponding prior distributions for generating the prior ensemble; and (4) the data assimilation setting, such as the ensemble size, the assimilation window size, number of iterations along with the inflation coefficient for each iteration [11], and any other data assimilation options supported by DART.

Step 2: PFLOTRAN preparation. At this step, PFLOTRAN input files are prepared and model spin-up is performed. PFLOTRAN inputs are composed of a PFLOTRAN input deck, which users have to provide to configure a PFLOTRAN model conforming to the conceptual model of a specific application, and an HDF5 [25] file that contains the prior ensemble of the parameters generated from their prior distributions defined at Step 1. Once the PFLOTRAN inputs are ready, model spin-up will be performed for a selected period of time prior to the beginning of the first assimilation time window to ensure reasonable initial conditions for the forward simulations.

Step 3: DART preparation. To configure DART for the ES-MDA task for a specific application, we use utility tools (see Section 2.5) to automatically generate the following files based on the user input and preparation done at the previous two steps: (1) the Fortran namelist file that records a variety of DART configurations using the data assimilation setting specified at Step 1; (2) the updated DART variable library that includes new PFLOTRAN parameter and state names at Step 2; and (3) a utility function file *convert_nc.f90* to convert the observations at each time window from the netCDF file into a DART sequence file (see Section 2.5). Finally, this step will generate all the executables for running DART.

Step 4: Performing the sequential ES-MDA. The assimilation process performs ES-MDA in each time window to update model parameters, as illustrated in Figure 3, until all observations are assimilated. In each assimilation time window, the ES-MDA starts with updating the prior ensemble of model parameters in HDF5 file, which is then

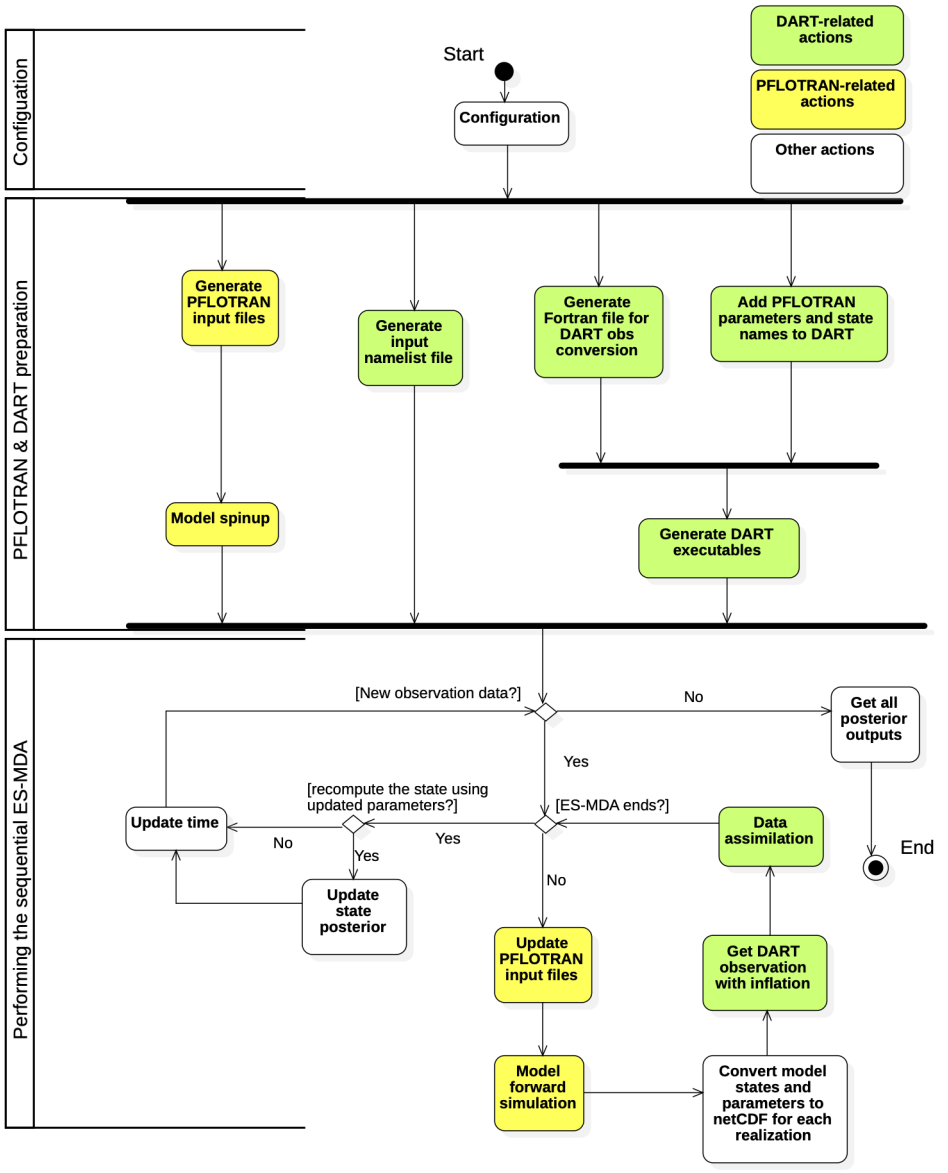

Figure 2: Integrated workflow of running the DART-PFLOTRAN software framework, including the initial configuration, preparing PFLOTRAN and DART files, and conducting sequential Ensemble Smoother for Multiple Data Assimilation (ES-MDA).

provided to launch ensemble PFLOTRAN simulations that produce the prior ensemble of simulated state vectors. After the forward simulations, the prior ensembles of parameters and model states are combined and saved into a netCDF file for each ensemble member. Then, the data in the DART observation sequence file are assimilated through DART to generate the posterior ensemble of parameters, which become the prior ensemble of parameters for the next iteration. After all ES-MDA iterations are completed within an assimilation time window, the posterior ensemble of parameters are used to run PFLOTRAN simulations to generate the posterior state vectors at the current window, which are used as the initial conditions for the subsequent assimilation time window.

2.5. Utility functions to link DART and PFLOTRAN

To facilitate the four-step integrated workflow, the following utility functions are developed to link DART and PFLOTRAN (shown as blue lines and grey texts in Figure 3):

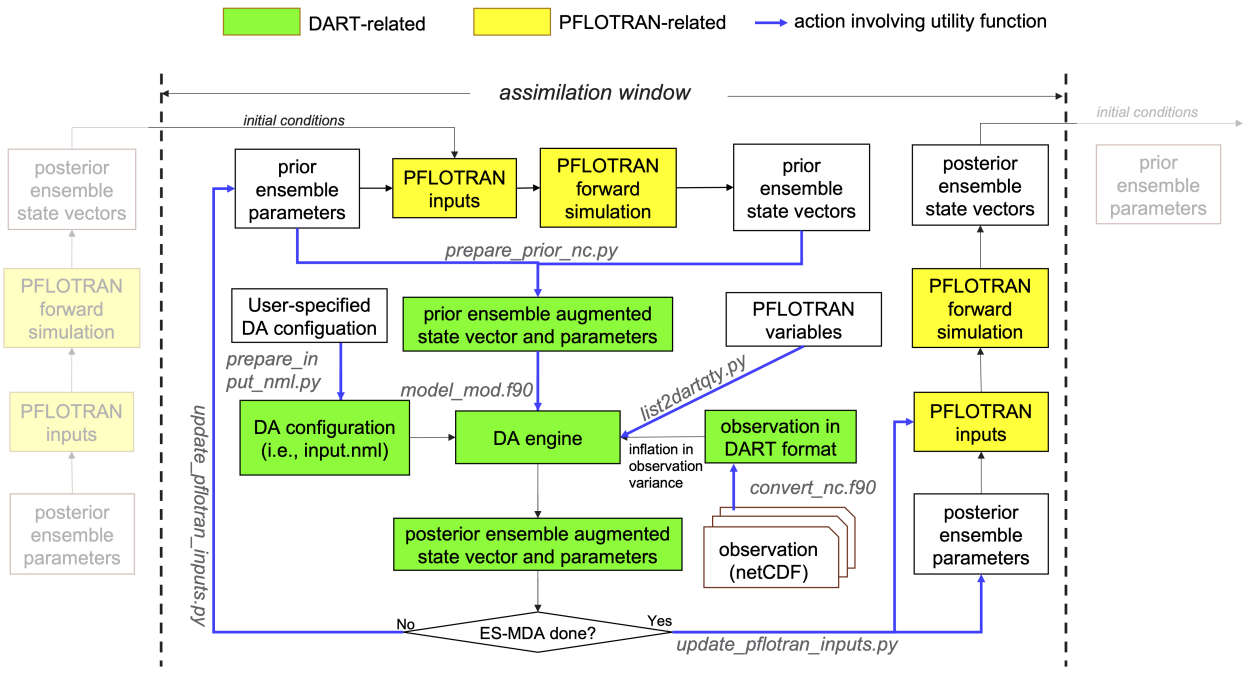

Figure 3: Design of the DART-PFLOTRAN software framework in one assimilation window of the sequential ES-MDA workflow in Figure. 1. (Green and yellow boxes are files associated with DART and PFLOTRAN, respectively; and blue arrows represent actions involving using utility functions, with the corresponding utility file name in gray text)

- *prepare_input_nml.py* is used in Step 3 to generate a Fortran namelist file from user-specified data assimilation configurations, such as the number of ensemble members, paths to prior and posterior files, temporal range of observations. A detailed description on the namelist file is available at DART's official website [26].

- *list2dartqty.py* is used in Step 3 to modify DART Fortran files to register a list of PFLOTRAN parameter and state variable names in DART variable library, so that DART can extract the prior ensemble from in the netCDF files and map the ensemble model state vectors with the data in the observation file.

- *prepare_prior_nc.py* is used to prepare an individual netCDF file for each prior ensemble member of PFLOTRAN parameters and state vectors as well as their spatial locations and time steps.

- *convert_nc.f90* is used to generate a DART observation sequence file at each assimilation window by extracting the associated data from the observation netCDF file. To use this utility function, users need to provide a standardized netCDF file in Step 1, which includes the times when the observations were taken, the spatial locations of observations, values of each observation variable in a two-dimensional matrix (i.e., temporal and spatial dimensions), and an additional two-dimensional matrix for observation errors with one-to-one correspondence to all the observation values.

- *model_mod.f90* contains a set of Fortran subroutines that allows DART to (1) define the spatial and temporal domains of observations at a data assimilation step and (2) compile the ensemble members of model simulated state variables at the same locations and times of the observations.

- *update_pflotran_inputs.py* is used to update PFLOTRAN input files after an ES-MDA iteration. The realizations of PFLOTRAN parameters in the HDF5 file will be updated using the posterior ensemble of parameters. The model simulation time window in PFLOTRAN input deck will be updated if the data assimilation proceeds to the next assimilation time window.

3. Verification of data assimilation implementation

We verified the implementation of DART-PFLOTRAN using two test cases that aim to estimate static and dynamic parameters, respectively. We attempted to estimate the permeability field as well as the dynamic exchange fluxes across the riverbed from temperature depth profiles monitored beneath the riverbed over time, as illustrated in Figure 4(a). The groundwater temperatures below the riverbed at different depths are related to the dynamic exchange fluxes through one-dimensional (1-D) flow and heat transport processes simulated by PFLOTRAN:

$$q = \frac{k\rho_w g}{\mu_w}\frac{dh}{dl}, \tag{2a}$$

$$Q_e = \frac{\partial}{\partial t}[\phi\eta U + (1-\phi)\rho_r c_p T] + \nabla\cdot(\eta q H - \kappa\nabla T), \tag{2b}$$

where q is the groundwater exchange flux [m/s]; k is the soil permeability [m^2]; ρ_w is water density [kg/m^3]; g is the gravitational acceleration [m/s^2]; μ_w is water viscosity [kg/ms]; dh is the difference between two hydraulic heads [m]; dl is the flow path length between two points [m]; ϕ is the porosity of soil matrix; Q_e is source/sink terms for energy transport [J/(m^{3}K)]; η is molar water density [kmol/m^3]; U is internal energy of the fluid [J/kg]; T is the groundwater temperature in Kelvin K; H is enthalpy [J/kg]; ρ_r is rock density [kg/m^3]; c_p is specific heat capacity [J/(kgK)]; and κ is thermal conductivity [J/(mKs)] of the porous media. For the flow process, the Darcy's law is used to compute the exchange flux (Eq.(2a)), which is coupled to the heat transport process in groundwater governed by the energy balance (Eq.(2b)).

In the first test case, we assumed that the hydraulic heads at the top and bottom boundaries were measured continuously along with temperature. Therefore, the exchange fluxes can be estimated using the Darcy's Law if the permeability of the porous media is known. In this case, we implemented ES-MDA to estimate the permeability of the soil column that does not change over time (see Figure. 4(b)). In contrast, in the second test case, we assumed that no hydraulic heads were measured. As a result, the exchange flux within a time window has to be directly estimated from the temperature responses below the riverbed. Furthermore, the exchange flux could vary over time driven by the stage fluctuations in the river, which was reflected as the Neumann type boundary condition for exchange flux based on the synthetic data. Thus, we implemented the sequential ES-MDA in the second test case to sequentially estimate the exchange fluxes in a set of predefined time windows (see Figure. 4(c)).

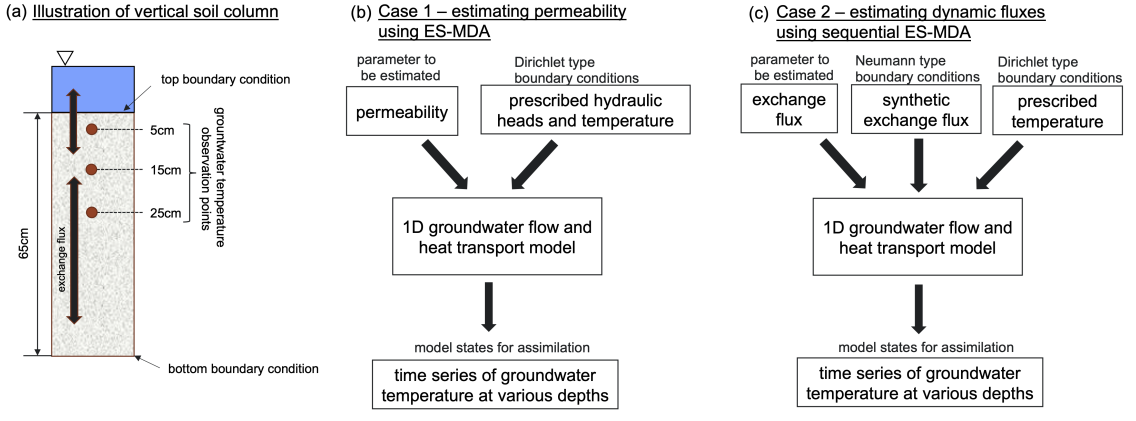

Figure 4: Illustration of the one-dimensional groundwater flow and heat transport model and the two data assimilation test case in Section 3. (a) depicts the vertical soil column where the model generates the synthetic groundwater exchange flux and temperature at 5/15/25cm below the riverbed. (b) shows the diagram of Case 1, where permeability is estimated using ensemble smoother for multiple data assimilation (ES-MDA) by assimilating groundwater temperature observation at multiple depths given the hydraulic head. (c) shows the diagram of Case 2 where the temporal dynamics of exchange flux are estimated using sequential ES-MDA assuming hydraulic heads are not known at the boundaries.

3.1. Generation of synthetic observation data

In order to evaluate the accuracy of the estimated parameters in both test cases, we generated synthetic observation data of hydraulic heads and temperature responses with known permeability and dynamic exchange fluxes, which serve as the ground truth of the parameters estimated by DART-PFLOTRAN for performance assessment. PFLOTRAN was used to generate the temperature responses in a 65cm soil column shown in Figure 4(a). The model domain was discretized into 1cm vertical grid cells with a homogeneous soil permeability value across the entire depth. The time-varying hydraulic heads and temperature at the top and bottom boundaries, which can be obtained from monitoring data in practice, were used as the boundary conditions. The PFLOTRAN simulation generated riverbed exchange flux at a 30-min resolution and groundwater temperature at the center of each grid cell at a 5-min resolution for 3 months. We assumed that the temperature observations were available at the depths of 5cm, 15cm and 25cm, in addition to the top and bottom of the boundaries. The synthetic temperature observation data were obtained by adding observation errors generated from a Gaussian distribution with a mean of 0 and a standard deviation of 0.05/3 Celsius, mimicking an observation error of 0.05 Celsius. We used the same observation error during the data assimilation process.

3.2. Case 1: estimating static permeability using ES-MDA

In this test case, the prior ensemble of log10 transformed permeability was generated by sampling 100 realizations from a log-normal distribution with mean and standard deviation being -11 and 1 ($\log_{10}(m^2)$), respectively. We first performed a two-day spin-up data assimilation to constrain the initial temperature profile conformed to the point observations at the observation depths. Then, the permeability ensemble was updated from the aforementioned prior ensemble by assimilating observations at the depths of 5cm, 15cm and 25cm over 50 time steps using the ES-MDA in DART-PFLOTRAN with a single assimilation time window.

We assessed the impact of the number of iterations, i.e., $L = [1,2,3]$, on estimating the unknown permeability. In Figure 5, the prior and updated posterior distributions of log10 transformed permeability resulted from different number of iterations are shown in violin plots and compared with its ground truth (i.e., -10.41 $\log_{10}(m^2)$) represented by the red dashed line. It can be observed that the mean of the posterior ensemble is improved to approach the true permeability with the increasing number of iterations, while the spread in the posterior ensemble shrinks significantly with more iterations. Two iterations appear to be adequate in this test case as there is negligible improvement in the estimation by increasing to three iterations. The convergence of posterior permeability estimation to its true value verifies our implementation of DART-PFLOTRAN in using ES-MDA for a single assimilation time window by compiling data taken in different locations and at different times.

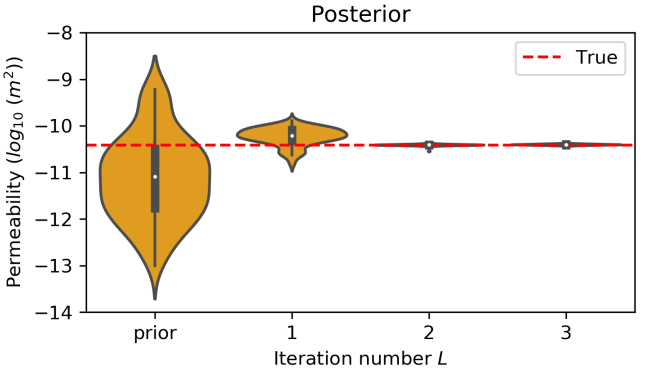

Figure 5: Violin plots of the prior and posterior ensemble of the permeability updated using different iteration numbers $L = [1, 2, 3]$ of ES-MDA for Case 1 in Section 3. The red dashed line is the true permeability value, i.e., -10.41 $\log_{10}(m^2)$.

3.3. Case 2: estimating dynamic groundwater exchange fluxes using sequential ES-MDA

In this test case, we sequentially estimated the hourly exchange fluxes over a month by assimilating the associated groundwater temperature observations within each assimilation time window (i.e., hourly, with 12 observed temperature data points at each depth). The initial prior ensemble of the exchange flux was generated by sampling 100 realizations from a Gaussian distribution with a mean of 0 m/s and standard deviation of 0.5 m/s (positive and

negative fluxes refer to the downwelling and upwelling fluxes, respectively). In each of the subsequent assimilation time windows, we generated the 100 realizations of the exchange flux for the prior ensemble by shifting its mean to the posterior mean resulting from the immediate preceding assimilation time window while maintaining 0.5 m/s as the standard deviation. The data assimilation was started two days earlier than the targeted estimation time window as the spin-up to minimize the impact of the initial conditions on the flux estimation accuracy.

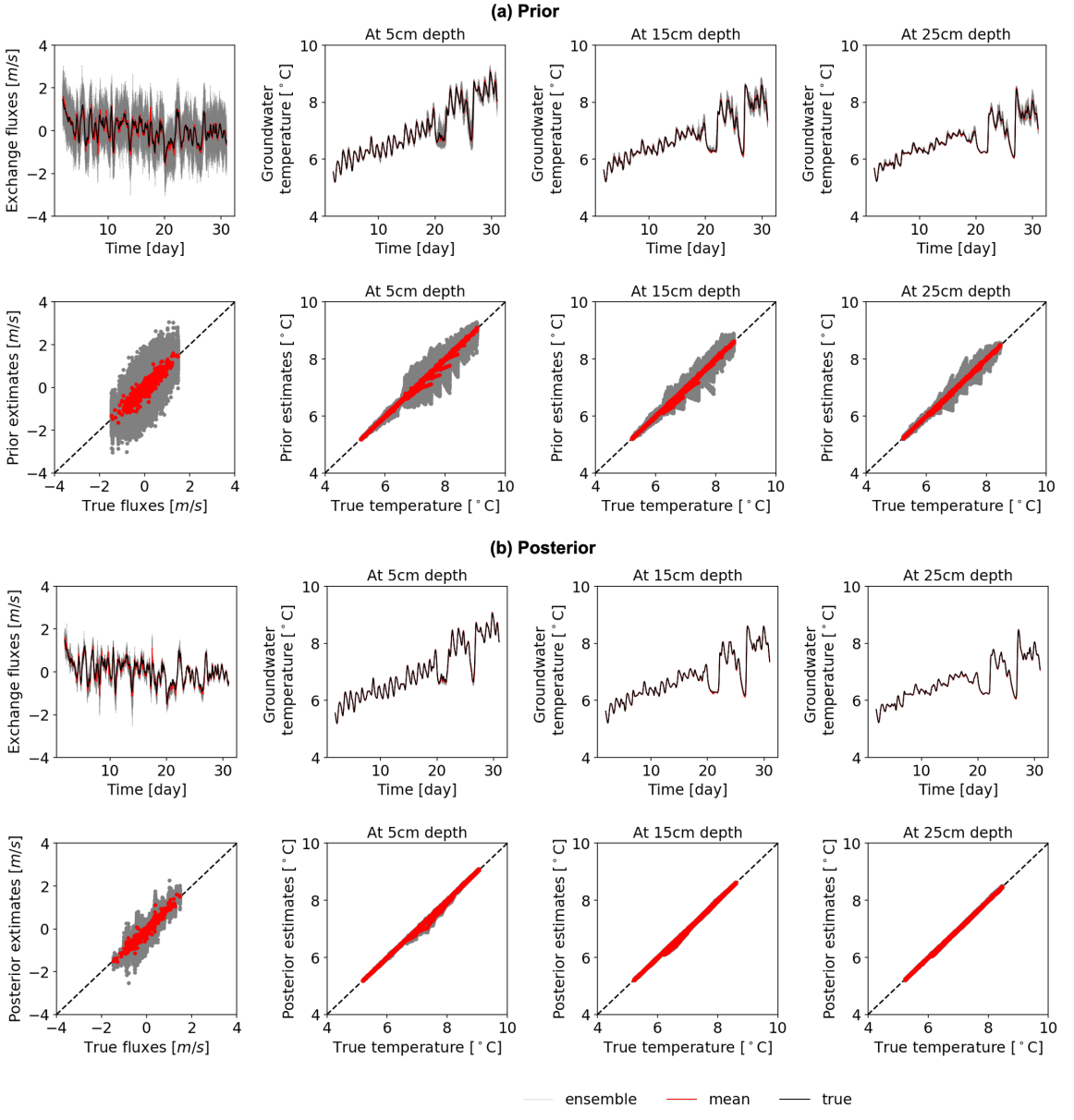

Figure 6: The assimilation results using DART-smoother for Case 2 in Section 3. The sequential ES is employed (i.e., the number of iterations $L = 1$ in sequential ES-MDA), with one-hour assimilation window. (a) plots the prior of the exchange flux and groundwater temperature at depths of 5/15/25cm, including all the ensemble (the dashed gray line), the ensemble mean (the red line), and the ground truth (the black line). (b) plots the corresponding posterior results.

We first tested the flux estimations using one iteration (i.e., $L = 1$) to verify the implementation of sequential ES in DART-PFLOTRAN. Figure 6 shows the ensembles of the hourly exchange fluxes and the simulated groundwater temperature before and after assimilating temperature responses, as compared against the synthetic observations and

ground truth in both the time series and scatter plots. All pairs of ensemble mean of the estimated hourly flux vs its ground truth are tightly distributed around the 1:1 line with substantial reduction in the ensemble uncertainty, which consequently improve the simulated temperature responses below the riverbed. The results in Figure 6 clearly demonstrate the effective dynamic parameter estimation using the sequential ES approach implemented in DART-PFLOTRAN.

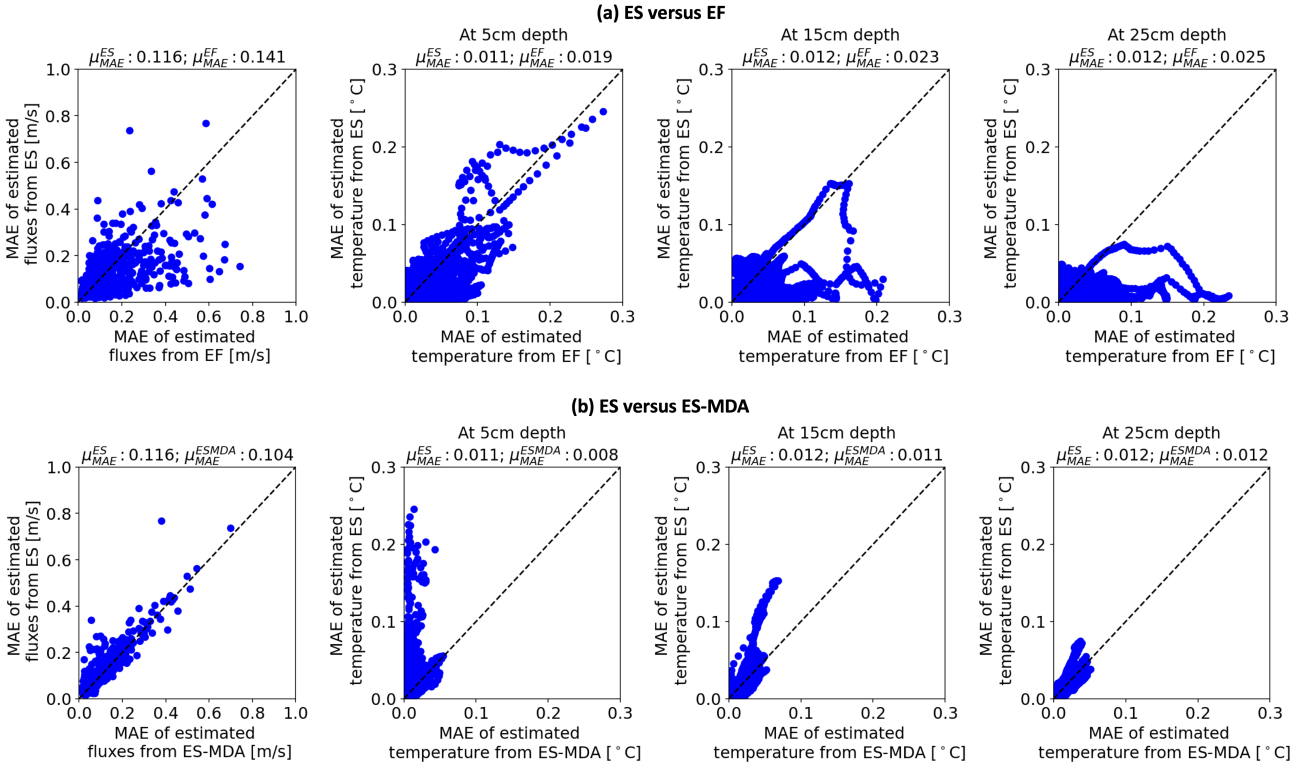

Figure 7: Mean absolute error (MAE) on the posterior of the exchange fluxes and groundwater temperature estimated using three different approaches in Case 2 in Section 3, including sequential ensemble smoother (ES), ensemble filter (EF), and sequential ensemble smoother for multiple data assimilation (ES-MDA) with three iterations $L = 3$. All the three approaches were performed using one-hour assimilation window. (a) shows the 1:1 plots between MAE from ES and EF in estimating exchange fluxes and groundwater temperature at 5/15/25cm depths. (b) shows the corresponding 1:1 plots between MAE from ES and ES-MDA. (Note that μ_{MAE} refers to the temporally averaged MAE in each subplot.)

We then compared the performance of the sequential ES with the original ensemble filter (EF) scheme in DART for estimating the dynamic exchange fluxes, which does not honor different observed time steps in each assimilation window. This is done by computing the mean absolute error (MAE) against the ground truth in posterior flux and temperature estimations at each time window, as plotted in Figure 7(a) showing the MAE comparisons between ES and EF, with their corresponding means over the entire estimation time window (μ_{MAE}) shown on the top of each subplot. The results show that MAE of estimated fluxes from the two approaches are comparable, with most of the MAEs smaller than 0.2 m/s and data pairs distributed nearly symmetrically around the black-dotted 1:1 line. There are more data pairs falling below the 1:1 line in the larger MAE regime (i.e., larger than 0.3 m/s), illustrating that EF tends to produce more higher absolute errors, which is also consistent with its higher average MAE than that of ES (i.e., 0.141 m/s vs 0.116 m/s). The more accurate estimations of exchange fluxes by ES result in smaller MAEs in the simulated groundwater temperature across all depths, more so at the depths of 15 and 25 cm as evidenced by more data pairs falling below the 1:1 line. While the exchange fluxes by both ES and EF yield highly accurate predictions of groundwater temperature, as illustrated by the small magnitude of their maximum and average MAEs, ES reduces the average temperature MAEs to approximately half compared to EF. Such gain in estimation accuracy demonstrates the potential advantage of ES in parameter estimation over the original EF scheme in DART.

Lastly, we assessed the performance gain of multiple iterations in sequential ES-MDA by comparing the MAEs of both exchange fluxes and simulated groundwater temperatures produced by $L = 3$ and $L = 1$ (i.e., the ES approach), as shown in Figure 7(b). The results show universal reductions in MAEs when the number of iterations in ES-MDA is increased from one to three. Although there is only slight decrease in average MAE of estimated exchange fluxes (i.e., from 0.116 m/s to 0.104 m/s) when increasing the iteration number from one to three, there are substantial reductions in a number of large flux MAEs with the iterative data assimilation as represented by the data pairs far above the 1:1 line. Such improvement in flux estimation through iterations leads to significant reductions in the MAEs of the simulated groundwater temperature, effectively eliminating all MAEs larger than 0.1 °C in ES. The performance gain through iterative ES in this test case not only verifies our implementation of the sequential ES-MDA in DART-PFLOTRAN, it also demonstrates the necessity of taking the iterative ES to improve data assimilation accuracy under nonlinearity.

4. Conclusion

In this study, we developed an open-source software framework, DART-PFLOTRAN, for conducting sequential ES-MDA to estimate static and dynamic parameters for subsurface flow and transport models. This new software framework links DART – a community facility for data assimilation with PFLOTRAN – a parallel simulation code for subsurface flow and reactive transport processes. We enabled the ensemble smoother option in DART and developed multiple utility functions to establish communications between DART and PFLOTRAN for sequential ES-MDA. We verified the implementation of DART-PFLOTRAN for both the static and dynamic parameter estimations using two synthetic cases, which demonstrated that we have successfully extended DART beyond its traditional applications in the atmospheric science for updating model state vectors.

We implemented the integrated workflow of performing ES-MDA in both Python and C-shell scripts. We also provide a user-friendly interface using Jupyter notebook. The scripts and the Jupyter notebook templates can be easily adapted to other applications to alleviate the burden in managing the complex data assimilation and parameter estimation workflow, especially when sequential and iterative assimilation is necessary to reduce the adverse effects of nonlinearity on estimation accuracy. With the added flexibility in subsetting observation data in space and time, DART-PFLOTRAN is poised for large-scale large scale applications. The workflow developed to link DART and PFLOTRAN can also be extended to link DART with other similar simulators such as the Advanced Terrestrial Simulator [27] and ParFlow [28]), which will greatly accelerate the integration of multi-scale and multi-type observations above and below ground with watershed models to improve the predictability of a wide variety of real systems.

Software availability

The source code of DART-PFLOTRAN is available at: gitlab.pnnl.gov/sbrsfa/dart-pflotran.

Acknowledgements

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). This contribution originates from the SBR Scientific Focus Area (SFA) at the Pacific Northwest National Laboratory (PNNL) and was supported by the partnership with the IDEAS-Watersheds. 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. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans 99 (C5) (1994) 10143–10162. doi:10.1029/94JC00572.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572

[2] G. Evensen, The ensemble kalman filter: theoretical formulation and practical implementation, Ocean Dynamics 53 (4) (2003) 343–367. doi:10.1007/s10236-003-0036-9.
URL https://doi.org/10.1007/s10236-003-0036-9

[3] G. Evensen, Data assimilation: the ensemble Kalman filter, Springer Science & Business Media, 2009.

[4] P. J. van Leeuwen, G. Evensen, Data assimilation and inverse methods in terms of a probabilistic formulation, Monthly Weather Review 124 (12) (1996) 2898–2913. doi:10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2.
URL https://doi.org/10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2

[5] R. Bailey, D. Baù, Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution, Water Resources Research 46 (12). doi:10.1029/2010WR009147.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR009147

[6] V. E. J. Haugen, L.-J. Natvik, G. Evensen, A. M. Berg, K. M. Flornes, G. Naevdal, et al., History matching using the ensemble kalman filter on a north sea field case, in: SPE Annual Technical Conference and Exhibition, Vol. 13, Society of Petroleum Engineers, 2006, pp. 382–391. doi:10.2118/102430âĂŘPA.

[7] X. Chen, G. E. Hammond, C. J. Murray, M. L. Rockhold, V. R. Vermeul, J. M. Zachara, Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at hanford 300 area, Water Resources Research 49 (10) (2013) 7064–7076. doi:10.1002/2012WR013285.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2012WR013285

[8] X. Song, X. Chen, M. Ye, Z. Dai, G. Hammond, J. M. Zachara, Delineating facies spatial distribution by integrating ensemble data assimilationand indicator geostatistics with level-set transformation, Water Resources Research 55 (4) (2019) 2652–2671. doi:10.1029/2018WR023262.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023262

[9] P. Zhu, L. Shi, Y. Zhu, Q. Zhang, K. Huang, M. Williams, Data assimilation of soil water flow via ensemble kalman filter: Infusing soil moisture data at different scales, Journal of Hydrology 555 (2017) 912 – 925. doi:https://doi.org/10.1016/j.jhydrol.2017.10.078.
URL http://www.sciencedirect.com/science/article/pii/S0022169417307540

[10] A. A. Emerick, A. C. Reynolds, et al., History-matching production and seismic data in a real field case using the ensemble smoother with multiple data assimilation, in: SPE Reservoir Simulation Symposium, Society of Petroleum Engineers, 2013. doi:10.2118/163675-MS.

[11] A. A. Emerick, A. C. Reynolds, Ensemble smoother with multiple data assimilation, Computers and Geosciences 55 (2013) 3 – 15, ensemble Kalman filter for data assimilation. doi:https://doi.org/10.1016/j.cageo.2012.03.011.
URL http://www.sciencedirect.com/science/article/pii/S0098300412000994

[12] Y. Chen, D. S. Oliver, Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother, Mathematical Geosciences 44 (1) (2012) 1–26. doi:10.1007/s11004-011-9376-z.
URL https://doi.org/10.1007/s11004-011-9376-z

[13] Y. Chen, D. S. Oliver, Levenberg–marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification, Computational Geosciences 17 (2013) 689–703. doi:10.1007/s10596-013-9351-5.

[14] K. Chen, X. Chen, X. Song, M. A. Briggs, P. Jiang, P. Shuai, G. Hammon, H. Zhan, J. M. Zachara, Using ensemble data assimilation to estimate transient hydrologic exchange flow under highly dynamic flow conditions, Water Resources Research.

[15] P. Shuai, X. Chen, X. Song, G. E. Hammond, J. Zachara, P. Royer, H. Ren, W. A. Perkins, M. C. Richmond, M. Huang, Dam operations and subsurface hydrogeology control dynamics of hydrologic exchange flows in a regulated river reach, Water Resources Research 55 (4) (2019) 2593–2612. arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR024193, doi:10.1029/2018WR024193.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024193

[16] J. T. White, R. J. Hunt, J. E. Doherty, M. N. Fienen, Pest++ version 5, a parameter estimation and uncertainty analysis software suite optimized for large environmental models, US Geological Survey Techniques and Methods Report.

[17] J. Anderson, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, A. Avellano, The data assimilation research testbed: A community facility, Bulletin of the American Meteorological Society 90 (9) (2009) 1283–1296. doi:10.1175/2009BAMS2618.1.
URL https://doi.org/10.1175/2009BAMS2618.1

[18] F. Kurzrock, H. Nguyen, J. Sauer, F. Chane Ming, S. Cros, W. L. Smith Jr., P. Minnis, R. Palikonda, T. A. Jones, C. Lallemand, L. Linguet, G. Lajoie, Evaluation of wrf-dart (arw v3.9.1.1 and dart manhattan release) multiphase cloud water path assimilation for short-term solar irradiance forecasting in a tropical environment, Geoscientific Model Development 12 (9) (2019) 3939–3954. doi:10.5194/gmd-12-3939-2019.
URL https://www.geosci-model-dev.net/12/3939/2019/

[19] K. Raeder, J. L. Anderson, N. Collins, T. J. Hoar, J. E. Kay, P. H. Lauritzen, R. Pincus, Dart/cam: An ensemble data assimilation system for cesm atmospheric models, Journal of Climate 25 (18) (2012) 6304–6317. arXiv:https://doi.org/10.1175/JCLI-D-11-00395.1, doi:10.1175/JCLI-D-11-00395.1.
URL https://doi.org/10.1175/JCLI-D-11-00395.1

[20] A. M. Fox, T. J. Hoar, J. L. Anderson, A. F. Arellano, W. K. Smith, M. E. Litvak, N. MacBean, D. S. Schimel, D. J. P. Moore, Evaluation of a data assimilation system for land surface models using clm4.5, Journal of Advances in Modeling Earth Systems 10 (10) (2018) 2471–2494. doi:10.1029/2018MS001362.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001362

[21] G. E. Hammond, P. C. Lichtner, R. T. Mills, Evaluating the performance of parallel subsurface simulators: An illustrative example with pflotran, Water Resources Research 50 (2014) 208–228. doi:10.1002/2012WR013483.

[22] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, Jupyter notebooks – a publishing format for reproducible computational workflows, in: F. Loizides, B. Schmidt (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas, IOS Press, 2016, pp. 87 – 90.

[23] J. L. Anderson, A local least squares framework for ensemble filtering, Monthly Weather Review 131 (4) (2003) 634–642. doi:10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2.
URL https://doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2

[24] J. L. Anderson, N. Collins, Scalable implementations of ensemble filter algorithms for data assimilation, Journal of Atmospheric and Oceanic Technology 24 (8) (2007) 1452–1463. doi:10.1175/JTECH2049.1.

URL https://doi.org/10.1175/JTECH2049.1
[25] The HDF Group, Hierarchical data format version 5.
URL http://www.hdfgroup.org/HDF5
[26] T. Hoar, Dart manhattan release notes (2017).
URL https://www.image.ucar.edu/DAReS/DART/Manhattan/documentation/html/Manhattan_getting_started.html
[27] The ATS Group, Advanced Terrestrial Simulator (ATS) development.
URL https://github.com/amanzi/ats
[28] S. J. Kollet, R. M. Maxwell, Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model, Advances in Water Resources 29 (7) (2006) 945 – 958. doi:https://doi.org/10.1016/j.advwatres.2005.08.006.
URL http://www.sciencedirect.com/science/article/pii/S0309170805002101