XIX International Conference on Water Resources
CMWR 2012
University of Illinois at Urbana-Champaign
June 17-22,2012

A DUAL ENSEMBLE KALMAN FILTERING FOR ASSIMILATION INTO A COUPLED CONTAMINANT MODEL

Mohamad El Gharamti[†], **Ibrahim Hoteit**[⋆], **Johan Valstar**[◇]

[†]PhD, King Abdullah University of Science and Technology, Earth Sciences and Engineering, Thuwal 23955-6900 - Saudi Arabia
[⋆]Assistant Professor, King Abdullah University of Science and Technology, Applied Mathematics and Computational Science, Thuwal 23955-6900 - Saudi Arabia
[◇] Geohydrologist, Deltares, Subsurface and groundwater systems, Utrecht, Netherlands.
Email: [†]mohamad.elgharamti@kaust.edu.sa, [⋆]ibrahim.hoteit@kaust.edu.sa, [◇]johan.valstar@deltares.nl
[†⋆]Webpage: http://assimilation.kaust.edu.sa/

Key words: Data assimilation, Coupled models, Ensemble Kalman filtering.

Summary. Modeling contaminant evolution in geologic aquifers requires coupling a subsurface flow model with a contaminant transport model. Assuming perfect flow model, one can directly apply an ensemble Kalman filter on the contaminant transport model. This is however a very crude assumption as flow models can be subject to many sources of uncertainties. If the flow is not accurately simulated, contaminant predictions will likely be inaccurate even after successive Kalman updates of the contaminant with the data. In this study, we propose a dual strategy for this coupled system by treating the flow and the contaminant models separately while intertwining a pair of distinct Kalman filters; one on each model. Preliminary results suggest that on top of simplifying the implementation of the filtering system, the dual approach provide more stable and accurate solutions than the standard joint approach.

1 INTRODUCTION

Our knowledge of the physical parameters driving dynamic models, play an essential role in the quality of the resulting estimates. Inaccurate parameters, for instance reflects some uncertainty in the model and this in return makes the model outputs less reliable. An intuitive way to solve this, is to estimate those parameters alongside with the states of the model using sequential data assimilation. This has been carried out in several applications including hydrology[8], oceanography[3], ... using the well-known and widely used Kalman filter and its nonlinear ensemble extensions[1,4,5,7]. The idea is simple and is based on building one single vector where both, the states and the parameters are concatenated. This is known as the joint state-parameter estimation problem and has been tested in several studies before.

One drawback of such approach is when the system unknowns (states and parameters) increase, the degree of freedom in the system increases making the estimation problem unstable and somehow intractable especially in nonlinear dynamics[6]. Another approach to tackle this problem is the dual state-parameter estimation[6,9] where two mutually interactive filters are utilized; one for the parameters and the other for the states. The parameters are set constant in time with some random perturbations, whereas the state is integrated via the available dynamic model. This approach has a nice feature that allows information from state adjustment to be fed-back and incorporated into the parameter estimates.

In coupled models such as, flow and transport, wind and wave, ... where more than one physical state is involved, one can consider that the output state of the first model becomes input *"parameters"* for the second. So if this input state is inaccurate, the final result is likely inaccurate as well. A good strategy is then to estimate both states relying on some observations using either joint or the dual approach.

We introduce the dual-states estimation technique for coupled models using ensemble-based Kalman filters. The method generalizes the dual state-parameter estimation approach[6] to systems where the parameters are observed and evolve according to a dynamic model. This method is illustrated and tested with a coupled subsurface flow and transport system in an unconfined contaminated aquifer.

2 FILTERING SCHEMES

Consider the following state-space discrete system:

$$X_k = M_{k,k-1}X_{k-1} + W_{k-1} + \eta_k, \quad (1)$$
$$Y_k^o = H_k X_k + \varepsilon_k, \quad (2)$$

where X_k is the state vector at time t_k, $M_{k,k-1}$ is an operator integrating the system from time t_{k-1} to t_k, W_{k-1} is a forcing term, H_k is the observational operator and Y_k^o is the observation given at t_k. η_k and ε_k are independent system and observational noise, respectively.

Sequential data assimilation aims at estimating the state of the system at each observation time using all available observations up to this time. For linear models, the Kalman filter (KF) can be utilized but it might get very expensive depending on the system dimension. To overcome this, the singular evolutive Kalman filter (SEKF)[2], which is based on low rank approximation of the system covariance, can be used. In nonlinear cases, linearization of the model is possible as in the extended Kalman filter (EKF)[4] and it's low rank formulation given by the singular evolutive extended Kalman (SEEK)[5,7]. The SEKF and the SEEK filters proceed in two main steps apart from an initialization step as follows:

- *Initialization*: Run the model and generate a long sequence of state vectors from which one often takes the mean as the initial state X_0^a and a low-rank r approximation of the sample covariance matrix P_0^a obtained through EOF analysis[7].

(i) *Forecast Stage*: At time t_{k-1}, an estimate of he system X_{k-1}^a and its covariance P_{k-1}^a in the form $L_{k-1}U_{k-1}L_{k-1}^T$ are available. The model in (1) is used to forecast the mean and the covariance:

$$X_k^f = M_{k,k-1}\left(X_{k-1}^a\right), \tag{3}$$

$$L_k = \mathcal{M}_{k,k-1}L_{k-1}, \tag{4}$$

where $\mathcal{M}_{k,k-1}$ is $M_{k,k-1}$ in the SEKF and $\mathbf{M}_{k,k-1}$ in the SEEK filter ($\mathbf{M}$ here is the gradient of M evaluated at X^a).

(ii) Analysis Stage: When new observations becomes available from (2), the predicted state and covariance are corrected as below

$$U_k^{-1} = \left[\alpha U_{k-1} + \left(L_k^T L_k\right)^{-1} L_k^T Q_k L_k \left(L_k^T L_k\right)^{-1}\right]^{-1} + L_k^T H_k^T R_k^{-1} H_k L_k, \tag{5}$$

$$X_k^a = X_k^f + L_k U_k L_k^T H_k^T R_k^{-1}\left[Y_k^o - H_k X_k^f\right], \tag{6}$$

where α is an inflation factor and Q_k and R_k are the model and observational error covariances, respectively.

The analysis and predicted error covariance matrices are then always decomposed into LUL^T.

In strong nonlinear dynamics, several studies[4] have shown that the linearization of the system might cause some instabilities and even divergence. The ensemble formulation of the KF for nonlinear models was then proposed[1] namely, EnKF. A similar ensemble-based variant of the SEEK filter is the singular evolutive interpolated Kalman filter (SEIK)[5,7]. In this filter, the evolution of the matrix L_k in (4) is carried out using an ensemble of states, $X_{k-1}^1, X_{k-1}^2, \ldots, X_{k-1}^N$ which are integrated forward with the model as follows

$$X_k^{f,i} = M_{k,k-1}\left(X_{k-1}^{a,i}\right). \tag{7}$$

The forecast error covariance matrix is then $P_k^f = \mathrm{cov}(X_k^{f,i}) = L_k V_k L_k^T$ with

$$L_k = \left[\left(X_k^{f,1} - X_k^f\right), \left(X_k^{f,2} - X_k^f\right), \ldots \left(X_k^{f,r+1} - X_k^f\right)\right] T, \tag{8}$$

where T is $(r+1) \times r$ full rank matrix with zero column sums and $V_k = \left[(r+1)T^T T\right]^{-1}$. The analysis step is then identical to the one in the SEEK filter. After analysis, a resampling step is needed to sample new members $X_k^{a,i}$ given the most recent estimate X_k^a as the mean and P_k^a as the covariance.

3 DUAL STATES ESTIMATION

The dual states estimation for coupled models consists of two interactive filters exchanging information to produce states estimates for both models. Each filter has its own separate forecast and analysis steps. Because the states from the second model are function of the driving states of the first model, their observations should help improving their respective estimates.

To illustrate, assume at time t_k an observation from the second model is available (any variable with $\widetilde{\cdot}$ sign belongs to the second model).

1. Integration of the first model states ensemble takes place as in (7).

2. If there are available observations for this first state, then use it to update its ensemble to $X_k^{a,i}$. If not, we use the forecast ensemble, $X_k^{f,i}$. To generalize both cases, use $X_k^{c,i}$ for the current ensemble of the first model.

3. This current ensemble $X_k^{c,i}$ will then be substituted by an updated ensemble using observations of the other state. This is done by propagating the ensemble of the second model,

$$\widetilde{X}_k^{f,i} = \widetilde{M}_{k,k-1}\left(\widetilde{X}_{k-1}^{a,i}, X_k^{c,i}\right), \tag{9}$$

   and then use this information to update the state estimate of the first model

$$X_k^u = X_k^c + L_k\widetilde{U}_k\widetilde{L}_k^T\widetilde{H}_k^T\widetilde{R}_k^{-1}\left[\widetilde{Y}_k^o - \widetilde{H}_k\widetilde{X}_k^f\right], \tag{10}$$

   where $\widetilde{U}_k$ and $\widetilde{L}_k$ are the same as equations (4) and (5) evaluated using the second model's ensemble $\widetilde{X}_k^{f,i}$ and X_k^u here denotes the updated estimate of the first model.

4. Now, sample the ensemble of the first filter using the mean in (10) and the covariance P_k^a and use it in the second filter's forecast step for the integration of the ensemble

$$\widetilde{X}_k^{f,i} = \widetilde{M}_{k,k-1}\left(\widetilde{X}_{k-1}^{a,i}, X_k^{u,i}\right). \tag{11}$$

5. An analysis step is then applied for these predicted states

$$\widetilde{X}_k^a = \widetilde{X}_k^f + \widetilde{L}_k\widetilde{U}_k\widetilde{L}_k^T\widetilde{H}_k^T\widetilde{R}_k^{-1}\left[\widetilde{Y}_k^o - \widetilde{H}_k\widetilde{X}_k^f\right], \tag{12}$$

   to determine a new analyzed ensemble for the second filter, $\widetilde{X}_k^{a,i}$.

4 FLOW AND TRANSPORT COUPLED SYSTEM

This hydrologic coupled system consists of a flow model which provides forcing (velocity field) to a transport model leading for solute (contaminant) movement in the subsurface.

4.1 Flow Model

Consider a 2D flow in an unconfined aquifer coupled with a transport system. The unconfined aquifer is fully saturated with water and has some contaminated areas that moves under the effect of the flow. The flow is modeled using groundwater flow equation (Darcy flow + mass balance):

$$\frac{\partial}{\partial x}\left(K_x\varphi\frac{\partial\varphi}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\varphi\frac{\partial\varphi}{\partial y}\right)=S_y\frac{\partial\varphi}{\partial t}+G, \tag{13}$$

where K_x and K_y represent the hydraulic conductivity in x and y directions, respectively. S_y is the specific yield of the aquifer, G is the source/sink term and φ is the hydraulic head with two main components; the pressure and the elevation heads.
Assuming an isotropic heterogenous subsurface, we discretize the domain using finite volume (cell-centered finite differences, CCFD) and solve (13) using forward Euler;

$$\varphi_k = S_y^{-1}\left[\mathcal{A}(\varphi_{k-1})\Delta t - G\Delta t + S_y\varphi_{k-1}\right], \tag{14}$$

where Δt is the time stepping in the flow model and $\mathcal{A}$ is the transient nonlinear flow function. Using the resulting distribution of φ,the velocity field is obtained as follows

$$U_x = -\frac{k_x}{\mu}\frac{\partial P}{\partial x}, \qquad U_y = -\frac{k_y}{\mu}\frac{\partial P}{\partial y}. \tag{15}$$

where μ is the viscosity of water, P is the pressure head and k_x and k_y are the permeabilities in x and y directions, respectively.

4.2 Transport Model

Once the velocity field is known, we plug it in the transport system. For this, we use the general transport of species equation in 2D:

$$\frac{\partial}{\partial t}(\phi C)+\frac{\partial}{\partial x}\left(U_xC-D_x\frac{\partial C}{\partial x}\right)+\frac{\partial}{\partial y}\left(U_yC-D_y\frac{\partial C}{\partial y}\right)=r(C)+qC^*, \tag{16}$$

where ϕ is the porosity of the domain, C is the concentration of the solute (contaminant), D_x and D_y are the diffusion terms, $r(C)$ is the reaction term, q is the source and C^* is the upwind concentration.
Ignoring the reaction and the source terms, we discretize the system using the upwind scheme of CCFD to ensure stable solutions and solve using forward Euler again;

$$C_k = N^{-1}\left[-\mathcal{B}(C_{k-1})\Delta t' + \mathcal{F}(C_{k-1})\Delta t' + NC_{k-1}\right], \tag{17}$$

where N is a diagonal matrix holding information about the porosity in each cell of the domain, $\mathcal{B}$ is the advection coefficient matrix, $\mathcal{F}$ is the diffusion coefficient matrix and $\Delta t'$ is the time stepping in the transport model.

5 EXPERIMENTS AND RESULTS

5.1 Experimental Setup

We study the coupled system on a rectangular domain with length $h = 1.2$ km and width $\ell = 0.6$ km. The domain consists of two main rocks with different permeabilities where one is embedded in the other. The area of the embedded rock is 0.18 km^2 and is located in the center of the domain. The porosity of the subsurface is 25% with 20% specific yield. The viscosity of water is 1 cP.
The domain is descritized into 3600 cells (60 in each direction) which makes the area of each cell $\sim 0.028\%$ of the total area. The flow of water is perpendicular to z direction from west to east with impermeable north and south boundaries. Initially, the domain is fully pure except for one contaminated region close to the western boundary; $[60, 160] \times [50, 550]$ m. We carry out twin experiments based on perfect and perturbed flow conditions (table 1), in other words, we run the coupled flow-transport system using:

I. *Perfect flow* conditions and save the output as a reference solution.

II. *Perturbed flow* conditions and use the described estimation methods to recover the *true* solution in I.

	Perfect Flow	*Perturbed Flow*
West boundary φ	100 m-water	80 m-water
East boundary φ	10 m-water	40 m-water
Permeability, main rock	100 md	90 md
Permeability, embedded rock	20 md	30 md
Source (uniform)	4.32×10^{-3} m/day	8.64×10^{-4} m/day

Table 1: Parameters of the flow system for the two studied scenarios.

The time stepping in the flow model is 1 day and 1 month for the transport model. We observe both the pressure head (flow) and contaminant concentration (transport) using 50 wells spread inside the domain. The flow is observed every week, whereas, the contaminant is observed every 1 month and the total modeling time is 4 years.

5.2 Estimation Results

The reference contaminant and flow states at the end of the 4 years are shown fig.1(a),(e). Under the flow of water, the contaminant is transported in the domain from west to east boundaries where it tries to avoid the low-permeability region at the center. We apply both; the joint and the dual approaches in a way to recover those reference states.

The recovered flow from the joint and the dual is shown in fig.1(f),(g); clearly the flow produced by the dual is more comparable to the reference solution than the joint one. Similarly, the contaminant estimate in the dual is closer to the true state than the joint one as shown in fig.1(b),(c). The root mean square errors (RMSE) for the two states are also plotted in fig.1(d),(h). There are three important features, which make the dual more appropriate approach for such a problem:

⋆ The estimation error of the flow state by the dual was always on track and did not increase unlike the case with the joint. This goes back to the additional useful step provided whenever the contaminant is observed. The dual, when predicting the second state, uses the updated statistics from the first estimated states; however, the joint method make use of them at the next filtering step. But if the flow started to lose track it's very hard to bring it back even after using contaminant information.

⋆ Because of large model uncertainties, we use inflation in the flow to bring the predictions closer to the data. Inflating the flow covariance pushes the estimates more to the reference but this might cause instabilities in the contaminant because of rapid changes in the velocity field. The dual method do not suffer from this problem because once the flow is changed, the contaminant data adjusts this change in a physical sense to preserve stabilities in the contaminant state. For this, the dual method takes more inflation and gets more accurate estimates than the joint.

⋆ In the joint method, as seen in fig.1(d),(h), when the flow started to lose track (RMSE started to increase) the estimates of the contaminant started to get worse. This can be due to the nonlinearity of the model when increasing the degree of freedom of the state vector; nevertheless, in the dual approach the contaminant state is estimated independently from the flow state providing more degrees of freedom for efficient assimilation.

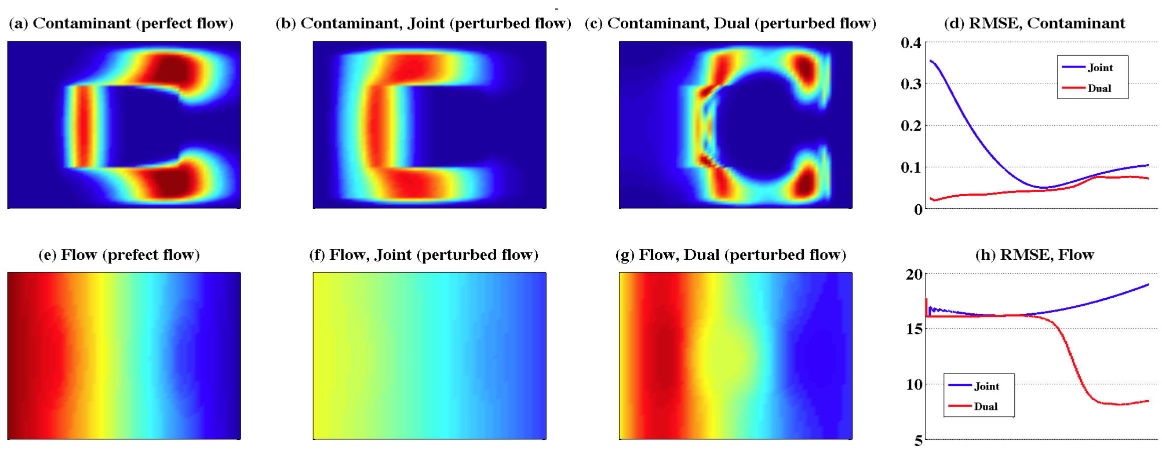

Figure 1: Joint and dual estimates for both flow and contaminant states after 4 years.

6 CONCLUSION

In this paper we introduced the dual states estimation technique for coupled models using the singular evolutive interpolated Kalman (SEIK) filter. Two interactive filters each with full dynamic model are combined to efficiently estimate the given physical states. An application to subsurface flow-transport coupled system is studied and assimilation experiments were ran with the goal of accurately locating the contaminant regions in a 2D domain using joint and dual techniques. Imposing perturbed flow conditions, numerical results suggest that in very uncertain flow settings, relying only on few pressure data for the flow is not enough to accurately recover the contaminant state. The dual approach uses contaminant data at every assimilation step to improve the flow and correct the contaminant state using the improved flow state. On the other hand, the joint approach uses the current flow state with no updates on its statistics when assimilating the contaminant causing some dynamical inconsistencies in this nonlinear setting.

REFERENCES

[1] G. Burgers, P.J. Van Leeuwen and G. Evensen (1998) *Analysis scheme in the ensemble Kalman filter*, Mon. Wea. Rev. 126, 1719-1724.

[2] M.E. Gharamti, I. Hoteit and S. Sun (2011) *Low-rank Kalman filtering for efficient state estimation of subsurface advective contaminant transport models*, J. of Env. Eng. 10.1061/(ASCE)EE.1943-7870.0000484.

[3] M. Ghil and P. Malanotte-Rizzoli (1991) *Data assimilation in meteorology and oceanography*, Adv. Geophys. 33, 141-266.

[4] I. Hoteit, G. Korres and G. Triantafyllou (2005) *Comparison of extended and ensemble based Kalman filters with low and high resolution primitive equation ocean models*, Nonlinear Proces. in Geophy. 12-5, 755-765.

[5] I. Hoteit, D.T. Pham and J. Blum (2002) *A simplified reduced order Kalman filtering and application to altimetric data assimilation in tropical Pacific*, J. Mar. Sys. 36, 101-127.

[6] H. Moradkhani, S. Sorooshian, H.V. Gupta and P.R. Houser (2004) *Dual state-parameter estimation of hydrological models using ensemble Kalman filter*, Adv. in Water Res. 28, 135-147.

[7] D.T. Pham, J. Verron and L. Gordeau (1998) *Singular evolutive Kalman filters for data assimilation in oceanography*, C. R. Acad. Sci. Paris 326, 255-260.

[8] R.H. Reichle, D. McLaughlin and D. Entekhabi (2002) *Hydrologic data assimilation with the ensemble Kalman filter*, Mon. Wea. Rev. 130, 103-14.

[9] E.A. Wan and T.A. Nelson (1997) *Dual Kalman filtering methods for nonlinear prediction, estimation and smoothing*, Adv. Neural Info. Pro. Syst. 9.