A variational ensemble Kalman filtering method for data assimilation using 2D and 3D version of COHERENS model

Idrissa Amour* and Tuomo Kauranne
Lappeenranta University of Technology, P.O. Box 21 53851, Lappeenranta, Finland

SUMMARY

Decoupled implementation of data assimilation methods has been rarely studied. The variational ensemble Kalman filter has been implemented such that it needs not to communicate directly with the model, but only through input and output devices. In this work, an open multi-functional three-dimensional (3D) model, the coupled hydrodynamical-ecological model for regional and shelf seas (COHERENS), has been used. Assimilation of the total suspended matter (TSM) is carried out in 154 km² lake Säkylän Pyhäjärvi. Observations of TSM were derived from high-resolution satellite images of turbidity and chlorophyll-a. For demonstrating the method, we have used a low-resolution model grid of 1 km. The model was run for a period from May 16 to September 14. We have run the COHERENS model with two-dimensional (2D) mode time steps and 3D mode time steps. This allows COHERENS to switch between 2D and 3D modes in a single run for computational efficiency. We have noticed that there is not much difference between these runs. This is because satellite images depict the derived TSM for the surface layer only. The use of additional 3D data might change this conclusion and improve the results. We have found that in this study, the use of a large ensemble size does not guarantee higher performance. The successful implementation of decoupled variational ensemble Kalman filter method opens the way for other methods and evolution models to enjoy the benefits without having to spend substantial effort in merging the model and assimilation codes together, which can be a difficult task. © 2016 The Authors.

1. INTRODUCTION

In numerical models of continuous physical processes, errors emerge because of either an incorrect model formulation or incomplete physics applied in the model. These errors should be accounted for and, when possible, corrected with observations. This process is known as data assimilation and can also be seen as a systematic way of combining a model forecast with measurements, which in many cases are sparse in space and limited in time.

Data assimilation has been applied in hydrology, for example, in an error forecasting procedure [1–3]. A common way to apply data assimilation to hydrological models is to update the initial conditions based on the history of the model (the prior) and the measurements (the data), as can be seen, for example, in [4, 5]. Data assimilation in hydrology has proven to give a better forecast when compared with open-loop simulation, as can be seen in [4]. However, the task of determining well-functioning parameters for the model and observations remains a challenge and may lead to even deterioration of results [2]. A precise representation of the internal oscillations of a
three-dimensional (3D) hydrological model is critical to successful simulations [6], but with extra information from measurements, this behavior can be captured [4, 7].

1.1. Assimilation methods

In linear data assimilation, the following two relations define the task at hand:

\[ \xi_k = M(\xi_{k-1}) + \xi_p \]
\[ \Gamma_k = K(\xi_k) + \xi_o \]

where \( \xi_k \) is the state vector estimate of the model at time \( k \), \( \Gamma_k \) is the observation vector, \( M \) is the evolution model, \( K \) is the observation operator, \( \xi_p \) is the model error, and \( \xi_o \) is the observation error.

The Kalman filter (KF) is among the first methods in data assimilation to be studied, but it is strictly applicable to linear models only. The extended Kalman filter (EKF) fills the gap left open by the KF by addressing nonlinearity by a locally linear approximation (linearization of the nonlinear observation and model operators). EKF has been, and still is, popular in many data assimilation problems and has been used as a benchmark to test new assimilation methods. However, it is computationally demanding, and this makes it unfeasible for high-dimensional problems, like the ones we have in hydrodynamics. The algorithms of the KF and EKF differ only in their linearization of the nonlinear observation and model operators whose details can be found, for example, in [8].

The most challenging part of such an assimilation scheme is how the error of the state estimates is estimated and propagated over time. In both KF and EKF, this has been performed by the so-called Kalman gain matrix. The ensemble Kalman filter (EnKF) came along with its own fashion of estimating this error by propagating several model trajectories in time [9]. The quality of the estimate of the error covariance generated in EnKF is not guaranteed to be sustained, however, and may suffer a loss that is known as covariance leakage [10], with the consequence that subsequent state estimates are no longer of good quality.

Other varieties of data assimilation schemes are the ones that use the principle of optimization. Such schemes have to find the optimum state that minimizes a cost function. There are a number of variational methods in data assimilation that belong to this category. In the case of data assimilation in hydrology, stochastic optimal control approach was proposed in [5]. This scheme finds the model trajectory that best fits to the observed data series and takes into account the covariance of the measurements. A detailed description of the method can be found in [5].

The variational Kalman filter (VKF) falls into this latter category whose cost function is defined by Eq. (1), as derived in [8]

\[ J(\xi) = \frac{1}{2} (\xi - \xi_p)^T C_{p}^{-1} (\xi - \xi_p) + \frac{1}{2} (\Gamma - K(\xi))^T C_{o}^{-1} (\Gamma - K(\xi)) \]  

The cost function \( J \) is minimized with respect to the model state vector \( \xi \), with parameters \( \xi_p \) as the prior, \( \Gamma \) as the observations, \( C_p \) as the prior covariance, \( C_o \) as the observation error covariance, and \( K \) as the observation operator. The cost function (1) is similar to the one used in three-dimensional variational data assimilation (3D-VAR), which, along with its four-dimensional extension 4D-VAR, has been an operational method in many weather centers since its adoption at the European Centre for Medium-Range Weather Forecasts [11]. In VKF, the optimization is performed with the help of the limited memory quasi-Newton method by Broyden–Fletcher–Goldfarb–Shanno (LBFGS), which is explained in depth in [12], and this method gives both the state estimate \( \xi \) and an estimate of the error covariance, which is used to estimate error covariance \( C_p \). More details can be found in [8]. A good feature of VKF over EKF is that it reduces substantially the time consumption in the linearization of the model [8].

The variational ensemble Kalman filter (VEnKF) is a hybrid scheme that utilizes the benefits of both ensemble methods and variational methods. VEnKF uses an ensemble technique to estimate the error covariance matrix. The basic assumption in VEnKF is that the model error and model response are independent [13]. Dropping the time indices for simplicity, the prior covariance to be used in Eq. (1) by VEnKF is estimated as

\[ C_p = \text{Cov} (M(\xi) + \xi_p) = \text{Cov} (M(\xi)) + \text{Cov}(\xi_p) \approx XX^T + C_{\xi_p} \]
where $C_p$ is the model error covariance and $X$ is defined in the same way as in EnKF, with the exception of using the state estimate, instead of the ensemble mean; thus,

$$X = \frac{1}{n} \left( (\psi_1 - \xi), (\psi_2 - \xi), \ldots, (\psi_n - \xi) \right)$$

where $n$ is the total number of ensemble members, $\psi_i$ is the $i$th ensemble member, and $\xi$ is the state estimate of the model.

The result in Eq. (2) is substituted into the cost function (1), with the following form:

$$C_p^{-1} = (XX^T + C_p)^{-1}$$

(3)

With the use of Sherman–Morrison–Woodbury (SMW) formula [14], we end up with an attractive expression

$$C_p^{-1} = C_{\text{est}}^{-1} - C_{\text{est}}^{-1} X (I + X^T C_{\text{est}}^{-1} X)^{-1} X^T C_{\text{est}}^{-1}$$

(4)

Expression (4) is attractive because it will never be computed as a whole matrix, but in the form of a vector product in cost function (1), that makes computations less expensive even for high-dimensional problems. Optimization in VEnKF is carried out with the help of the LBFGS method, which gives the state estimate $\xi_{\text{est}}$ as the minimizer of (1), and the estimate of the covariance $C_{\text{est}}$, which is a by-product of the LBFGS as inverse Hessian of (1) stored in vector format. A new ensemble is generated from the normal distribution with mean $\xi_{\text{est}}$ and covariance $C_{\text{est}}$. More detail on the sampling technique from low-memory storage of error covariance $C_{\text{est}}$ can be found in the Appendix. A close relative to the VEnKF method is the maximum likelihood ensemble filter (MLEF) of [15]. As in VEnKF, the analysis in MLEF is obtained as the maximizer of the likelihood posterior probability distribution, similar to the minimization of the 3D-VAR cost function. In a similar fashion to VEnKF, the analysis error covariance is obtained as the inverse Hessian as by-product of the minimization [15]. A clear difference between VEnKF and MLEF is that MLEF assumes a perfect model and that in VEnKF, the ensemble members are resampled at every data assimilation cycle, that extra feature makes VEnKF robust against filter in-breeding problems.

1.2. Data assimilation in hydrology and limnology

The need of data assimilation in hydrology has become apparent in parallel with the traditional approach of parameter estimation of the model. Several assimilation methods have been applied in the fields of hydrology and limnology. However, the performance of assimilation methods is very dependent on the method employed. Despite coupled hydrodynamical-ecological model for regional and shelf seas (COHERENS) being used by many researchers and institutions in the field of hydrology and limnology, data assimilation is not often a part of the simulations carried out. One of the few such studies was carried out with the simplified 1D-COHERENS model of a North sea, whose results with EnKF were promising, but depend much on ensemble initialization [7].

The EKF has a long history in hydrology studies [16]. However, its practical implementation has been a major obstacle with high-dimensional models. EKF has to invert and operate huge matrices of the order $10^7$ by $10^7$, which is a setback on current computing platforms. Nonlinearity in the model has also been another challenge for EKF, as one has to produce the tangent linear and adjoint codes, which are not always available for the model and are tedious to construct and maintain.

The EnKF has been more popular in hydrology, because of its nonlinear error covariance propagation in time (see publications [17–19]). In [16], streamflow observations have been used to update model state in a distributed hydrological model using EnKF.

Water quality is an important phenomenon to be studied with both modeling and measurements, but it has not been covered much in data assimilation literature [18, 20]. In [18], a study of algal bloom dynamics in a river with the EnKF was performed. The model had to be run with a small ensemble size because of the heavy computation in 3D hydrodynamics models. Small ensembles make the quality of covariance error propagation uncertain [18, 19]. Soil moisture has been also an active field of study with the use of EnKF (for example, [19, 21]). In publication [21],...
the EnKF underestimated the forecast error covariance for small ensemble sizes but gave reasonable results with bigger ensembles, which suggests significance of dynamic error covariance propagation [21].

Variational ensemble Kalman filter is still a new candidate in data assimilation, but it has proven to be competitive in many aspects. One attractive feature of VEnKF is that it needs neither tangent linear nor adjoint codes [13]. Therefore, it can be easily implemented into hydrological models. In [4], VEnKF has demonstrated its ability to forecast water heights in a dam-break experiment of Martin and Gorelick [22], which was first experimented by Bellos et al. [23]. Results were convincing enough to carry out more complex assimilation tasks with VEnKF, taking into account another advantage, namely, the fact that VEnKF can be easily parallelized.

Algorithm 1 VEnKF Algorithm

1: procedure INITIALIZATION
2: Initialize the prior $\xi_0$
3: Initialize the ensemble members $\psi_{0,i}$, $i = 1, 2, ..., n$
4: $k \leftarrow 1$
5: Time loop:
6: procedure PROPAGATION
7: Propagate the prior $\xi_k = M(\xi_{k-1})$
8: Propagate the ensemble members $\psi_{k,i} = M(\psi_{k-1,i})$, $i = 1, 2, ..., n$
9: procedure SUBPROBLEM
10: Use SMW formula (4) to get $C_p^{-1}$
11: procedure OPTIMIZATION
12: Use LBFGS for cost function (1) to obtain $\xi^{est}$ and $C^{est}$
13: $\xi \leftarrow \xi^{est}$
14: procedure SAMPLING
15: Generate new ensemble members, $\psi_{k,i} \sim N(\xi^{est}, C^{est})$, $i = 1, 2, ..., n$
16: Update time
17: GoTo Time loop

2. HYDRODYNAMICS MODEL: COHERENS

Coupled hydrodynamical-ecological model for regional and shelf seas is a 3D multipurpose model for coastal and shelf seas that resolves mesoscale to seasonal scale processes [24]. COHERENS is a well-documented open source code that is now available in Fortran 90. The model has four main compartments:

1. physical part for solving advection diffusion equations;
2. microbiological module for solving dynamics of organic matter;
3. sediment compartment that solves the deposition and resuspension of inorganic as well as organic matter; and
4. transport model for contaminant distribution.

The discretization of the model is performed with an Arakawa C-grid using either Cartesian or spherical coordinates [24].

The COHERENS equations for the 3D mode feature a continuity equation, momentum equations, and temperature and salinity equations; see the assumptions used in their derivation in [24]

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$$ (5)
\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} - f v = - \frac{1}{\rho_0} \frac{\partial p}{\partial x} + F_x^I + \frac{\partial}{\partial z} \left( v_T \frac{\partial u}{\partial z} \right) + \frac{\partial}{\partial x} \tau_{xx} + \frac{\partial}{\partial y} \tau_{xy} \]  

\[ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} + f u = - \frac{1}{\rho_0} \frac{\partial p}{\partial y} + F_y^I + \frac{\partial}{\partial z} \left( v_T \frac{\partial v}{\partial z} \right) + \frac{\partial}{\partial x} \tau_{yx} + \frac{\partial}{\partial y} \tau_{yy} \]  

\[ \frac{\partial p}{\partial z} = -\rho g \]  

\[ \frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} + w \frac{\partial T}{\partial z} = - \frac{1}{\rho_0 c_p} \frac{\partial I}{\partial z} + \frac{\partial}{\partial z} \left( \lambda_T \frac{\partial T}{\partial z} \right) + \frac{\partial}{\partial x} \left( \lambda_H \frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left( \lambda_H \frac{\partial T}{\partial y} \right) \]  

\[ \frac{\partial S}{\partial t} + u \frac{\partial S}{\partial x} + v \frac{\partial S}{\partial y} + w \frac{\partial S}{\partial z} = \frac{\partial}{\partial z} \left( \lambda_T \frac{\partial S}{\partial z} \right) + \frac{\partial}{\partial x} \left( \lambda_H \frac{\partial S}{\partial x} \right) + \frac{\partial}{\partial y} \left( \lambda_H \frac{\partial S}{\partial y} \right) \]  

where \((u, v)\) are the horizontal components of the current, \(w\) is the vertical current, \(f = 2\Omega \sin(\phi)\) is the Coriolis frequency, where \(\Omega = \pi/43082\) radians/sec is the frequency of the earth’s rotation, \(p\) is pressure, \(\rho_0\) is density, \(\rho_0\) is a uniform reference density, \(g\) is the acceleration due to gravity, \((F_x^I, F_y^I)\) are the components of the astronomical tidal force, \(v_T\) and \(\lambda_T\) are the vertical turbulent diffusion coefficients, \(\lambda_H\) and \(\lambda_H\) are the horizontal turbulent diffusion coefficients, \(\tau_{ij}\) is the horizontal friction tensor, \(T\) is potential temperature, \(I\) is the solar irradiance within the water column, \(c_p\) is the specific heat capacity of sea water at constant pressure, and \(S\) is the salinity, [24].

In the current version of COHERENS (V2.9), the formation of ice is not allowed, therefore the temperature must be higher than the freezing point,

\[ T > \alpha_f S, \quad \alpha_f = -0.0575 \degree C/PSU \]  

where PSU is the salinity concentration unit standing for Practical Salinity Unity; PSU = \(10^{-3}\)kg/kg. More details of the model equations and the numerical methods can be seen in [24].

3. DECOUPLED IMPLEMENTATION OF VEnKF ON COHERENS

In many cases, the implementation of an ensemble type data assimilation scheme to a coupled model system is tedious and needs much technical effort. To circumvent this, we introduce a non-intrusive way of implementation of the data assimilation (VEnKF in this study) scheme to a coupled model system (COHERENS). In this approach, the model code remains untouched and the communication between the model (COHERENS) and the assimilation scheme (VEnKF) method is made possible through input and output devices (files). The two, the model and VEnKF operations are managed by an external program, for example, a shell script. At assimilation time, the prior model and ensemble members write their final states into final condition (IC) files. When the VEnKF is called, it reads the model prior state and ensemble states from these IC files. With the information from the measurements, VEnKF updates the model state and generates an updated estimate for the error covariance matrix using ensemble members. It then writes the analysis back into an IC file that now becomes an initial condition file, ready to be used by the model on the next interval. In addition, the Hessian update vectors of the VEnKF minimization are written into an updated error covariance matrix in vector form, from which the initial conditions for a new ensemble are drawn and likewise written into IC files. A summary of decoupled implementation is shown in Algorithm 2.
In this paper, we will study data assimilation of hydrodynamic flow in Lake Säkylän Pyhäjärvi, located in the south-western part of Finland. The lake has a total area of $154 \text{ km}^2$ and mean water depth of $5.4 \text{ m}$ [25], and its deepest point is $25 \text{ m}$ [26]. The lake Pyhäjärvi in Figure 1 has three openings. The first in the southern part; river Yläneenjoki, the second in the eastern part; river Pyhäjoki and the last one in the northern part; river Eurajoki. The direction of the flow is shown by the arrow in the map. For simplicity of calculations, we opted to use $1 \text{ km}$ grid resolution. The model was run with 20 vertical layers. The time step of the model was set to $20 \text{ s}$, and model fields were written after every 6 hrs from the start of simulation. In this work, we did not make a new COHERENS setup; instead, we have adopted the setup made in [27].

Figure 1. Map of lake Säkylän Pyhäjärvi. The green dot on the map shows the location of an automatic station, which is taking measurements records. The device was put at $1 \text{ m}$ depth from the lake surface [27].
4.1. Artificial temperature experiment

It is a common practice in data assimilation to start with an artificial experiment before the actual real task. The advantage is that it is simple to debug when something goes wrong, which is not the case for real data. In this experiment, we have created artificial temperature measurements by solving the COHERENS model for 5 days, from May 15, 2009 at 12:00 to May 20, 2009 at 12:00. The initial temperature was set to be 10.1°C as it has been recorded in the SYKE Hertta database as the vertical average at the deepest point between May 10 and May 20, 2009 [27]. To generate artificial measurements, we used normally distributed noise with mean 0 and variance $\sigma^2 = (0.15 \times 3.6414723)^2$ to add to the COHERENS temperature fields. Where $\sigma$ is the standard deviation used in climatological simulation. For the matter of testing, we have assumed data to be observed fully on the surface and in all 20 layers. The observation error covariance was set to be $10\sigma^2 I$, and the model covariance error was assumed to be the same as the observation covariance error.

The assimilation has been carried out after every 6 h. Fine tuning of line search parameters in the Wolfe conditions was assured for better convergence in the LBFGS. Maximum number of iterations in LBFGS was set to 100, with maximum number of stored vectors as 20. The VEnKF has been run with 10, 20, 50, and 100 ensemble members.

4.1.1. Results. The aim of this experiment was to test the performance of the algorithm for a biased model case. If everything goes well in this experiment, then we have some hope for success also for the real assimilation task, where measurements and the prior are not correlated in any way. The results of the artificial experiment show that the decoupled VEnKF algorithm works well and gives the expected results. The convergence of the LBFGS optimization was monitored and found to be consistent for the whole period of assimilation. The LBFGS was converging between 10 and 60 iterations. For this test case, we compare the truth and the VEnKF estimates by plotting the root mean square error (RMSE) plots and relative error plots.

The temperature dynamics in Figure 2 is presented reasonably well by VEnKF. The average relative error plot in Figure 3 and average RMSE in Figure 4 are well behaved. The range bounds shown in Figure 3 and 4 by different ensemble members are reasonable.

4.2. Suspended particular matter assimilation

In this study, we assimilate with VEnKF the total suspended matter (TSM) in lake Säkylän Pyhijärvi. The TSM is derived from the turbidity and chlorophyll-a satellite images for the lake

![Figure 2. The true temperature values compared with the VEnKF estimates for 120 h of simulation from May 15 to May 20, using 100 ensemble members.](image-url)
Figure 3. The mean relative error plot for all layers, using 100 ensemble members. The vertical bars show the range of the relative errors.

Figure 4. The average of root mean square error plot for all layers, using 100 ensemble members. The vertical bars show the range of the mean square errors.

Säkylän Pyhäjärvi. The images are for 9 days, at 12:00, May 16, June 1, June 8, June 18, June 21, June 26, July 6, August 22, and September 14, 2009. The turbidity data contain the effect of organic substances, which have been eliminated by using the following formula [27]:

\[ TSM_{(no\ biomass)}(i, j) = \frac{Tur(i, j) - 0.0449}{1.09 - 0.1 \times Chl(i, j)} \]  

where \( TSM_{(no\ biomass)} \) stands for TSM concentration without biomass, Tur is turbidity concentration from the satellite image, and Chl is the chlorophyll-a concentration data from the satellite image [27]. TSM conversion and satellite image processing are described in [27]. The satellite data are usually taken to describe the top surface layer, which causes the assimilation process to have less or no information from the lower layers. Therefore, the observation operator \( K \) maps the satellite image data onto the top layer of the lake. The model has been initialized with TSM values of 0 mg/l, in May 15 at 12:00 and run for 24 h (spin-up phase), then the model was initialized with the TSM values of 2.5 mg/l in May 16 at 12:00 (same setting as in [27]), which is then run for as long as we have observations, that is, until September 14, 2009. The VEnKF has been run with 10, 30, and 50
Figure 5. Satellite TSM data in the first column with VEnKF estimates using 10, 30, and 50 ensembles in second, third, and fourth columns, respectively. The plots are for June 1, June 8, June 18, June 21, June 26, and July 6. The COHERENS setup is 2D mode. ens, ensemble members.

Figure 6. Satellite TSM data in the first column with VEnKF estimates using 10, 30 and 50 ensembles in second, third, and fourth columns, respectively. The plots are for June 1, June 8, June 18, June 21, June 26, and July 6. The COHERENS setup is 3D mode. ens, ensemble members.
ensemble members. The model observation covariance is set \((0.15 \times \sigma_0)^2 I\), while the model error covariance is set 10 times that of observation.

4.2.1. Results. The decoupling method presented in this paper suggests a stable and robust approach. The LBFGS minimization converges between 10 and 80 iterations with gradient tolerance norm between \(10^{-3}\) and \(10^{-5}\). The LBFGS convergence criterion was set as the function value tolerance of \(10^{-12}\) and gradient norm tolerance of \(10^{-8}\).

The VEnKF assimilation outcomes in Figures 5 and 6 are presented for 2D-COHERENS mode and 3D-COHERENS mode, respectively. The performance of the VEnKF is not obviously better with the use of larger ensemble sizes. One can see that the results, for example, for June 1 are almost identical for all ensemble sizes. However, we can see small discrepancies for the rest of the dates. This suggests that with this available measurements, the use of larger ensemble sizes is not effective. The discrepancies between the data and VEnKF estimates suggest that VEnKF output is a result of combining the prior and the measurements, as it also follows directly from the definition of a KF.

Figure 7. VEnKF assimilation of the TSM for surface layer from May 16 to July 6, 2009 at the automatic station using 10 ensemble members in 2D mode.

Figure 8. VEnKF assimilation of the TSM for surface layer from May 16 to July 6, 2009 at the automatic station using 50 ensemble members in 2D mode.
The COHERENS mode does not have influence on the results, as it can be seen in Figures 5 and 6. This can be explained by the fact that the satellite data are given for the lake surface only. Therefore, VEnKF does not get any information on the bottom layers of the lake.

The results of the VEnKF are compared with the records collected by the automatic station located at the green dot in the map Figure 1. The instrument was located at a depth of 1 m below the surface, the records were available from May 18 to July 7 2009 at an interval of 1 h. The results of the assimilation using VEnKF are shown at the surface layer because the instrument at the automatics station is located near the surface of the lake. COHERENS can be set to run with full 3D mode or with partial 3D mode. In this section, we present results for 3D mode and a partial 3D mode where COHERENS run with a defined number of time steps in 2D mode. In this work, that time step is set to 15. The VEnKF seems to capture reasonably well the evolution of TSM. In Figures 7–10, the VEnKF captured the evolution of turbidity in many parts of the assimilation period except at the beginning of the assimilation, as the first satellite data were available from June 1, 2009 only. The results show that the mode of the run is not very significant because the satellite images contain
Figure 11. VEnKF assimilation of the TSM for surface layer from May 16 to July 6, 2009 at the automatic station using 50 ensemble members in 3D mode.

Figure 12. Total root-mean-square error (RMSE) curves for TSM with respect to satellite observation and automatic station data using 10, 30, and 50 ensemble members, respectively. The size of the ensemble has little impact on total deviation of the analysis from measurements.

information on the surface of the lake only. The results of TSM VEnKF estimate with 50 ensemble members look a little better than those with 30 ensemble members (refer to Figures 7–11) in the sense of quality of estimates. Likewise, it suggests that even a small ensemble size is enough to capture the error propagation of the model state, as it can be seen in Figure 12.

5. CONCLUSIONS AND RECOMMENDATIONS

In this study, we have successfully implemented a decoupled version of the VEnKF algorithm. Its performance matches with the quality that we have expected from a grid resolution of 1 km that was used. Results for both 2D mode and 3D mode of the model do not differ much in quality, which can be explained by the fact that satellite images give measurements on the surface only and therefore pass no information from the bottom layers to VEnKF. The size of observed data in this case was 1/20 of the size of the VEnKF state variable, and so this assimilation task represents a sparse data case.
The VEnKF was run with various ensemble sizes, and we find in this study that the use of high ensemble size is not important, as extra members contribute little to performance. This observation may be important in the development of reduced-order hybrid and ensemble data assimilation methods. This is inspired in particular by the work of the authors in [28, 29]. In reduced-order modelling, a numerical unique representation of the model system is used to speed up the solution of the model. It has been noted that the speed is 10 times higher with hybrid proper orthogonal decomposition/discrete empirical interpolation method than with full space data assimilation system [28]. The approach is also significant in the field of reduced-order optimization [28]. The reduced-order hybrid Monte Carlo sampling smoother has been introduced in [29]; the scheme is found to be faster than its full order system counterpart while preserving most of the characteristics of the full space hybrid Monte Carlo smoother.

Moreover, this finding does not rule out the need for covariance localization, which is known to give better results for smaller ensemble than for a larger ensemble without localization [30]. The result may be caused by the fact that the error covariance is underestimated. This underestimation can be corrected by introducing covariance inflation, for example, [31], where the covariance matrix is scaled with a fixed or varying multiplier. The derived TSM measurements from satellite images also contain errors because of the transformation by the Eq. (12). It would be interesting to examine the evolution of turbidity including simulating the effect of organic substances, but this is not supported in the current version of COHERENS.

The use of the automatic station data with the satellite data might improve the situation further, as the model will learn from this other dataset, too.

The use of high-resolution TSM data should be considered in future studies. With the current resources, this was not feasible because all simulations have been run with a desktop personal computer, with normal resources. The use of a supercomputer is desirable in that case and in any real and difficult simulations.

APPENDIX

Broyden–Fletcher–Goldfarb–Shanno minimization of the cost function (1) yields model state analysis \( \hat{\xi}_k \) and low-storage estimates for the covariance matrix \( \hat{C}_k \). New ensemble \( \psi_k \) is sampled as follows

\[
\psi_{k,i} \sim \mathcal{N} \left( \hat{\xi}_k, \hat{C}_k \right), \quad i = 1, 2, \ldots, n
\]

The inverse Hessian of (1) can be estimated by using either full rank low-memory representation obtained by the LBFGS unconstrained optimizer [12] or by the reduced rank representation in the Krylov space obtained by conjugate gradient minimization of (1) [32]. Consider a problem \( \min_x f(x) \), the inverse Hessian at time point \( k \) \( (H_k) \) update rule can be referred in [12] is given by

\[
H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k s_k y_k^T) + \rho_k s_k s_k^T
\]

where

\[
s_k = x_{k+1} - x_k, \quad y_k = \nabla f_{k+1} - \nabla f_k, \quad \rho_k = \frac{1}{y_k^T s_k}
\]

For efficiency, LBFGS uses the most recent \( m \) vectors to approximate the Hessian matrix. At iteration \( k \), the current iterate is \( x_k \) and the pairs is \( \{s_i, y_i\} \) for \( i = k - m, k - m + 1, \ldots, k - 1 \). If the initial approximation \( H_0 \). The \( H_k \) approximation in (A.1) is found to satisfy Eq. (A.2) [12], where \( V_k = I - \rho_k s_k y_k^T \)

\[
H_k = (V_{k-1}^TV_{k-m}) H_0 (V_{k-m} \ldots V_{k-1})
+ \rho_{k-m} (V_{k-1}^T \ldots V_{k-m+1}) s_{k-m} s_{k-m}^T (V_{k-m+1} \ldots V_{k-1})
+ \rho_{k-m+1} (V_{k-1}^T \ldots V_{k-m+2}) s_{k-m+1} s_{k-m+1}^T (V_{k-m+2} \ldots V_{k-1})
+ \ldots
+ \rho_{k-1} s_{k-1} s_{k-1}^T
\]

© 2016 The Authors. *International Journal for Numerical Methods in Fluids* Published by John Wiley & Sons Ltd.

*Int. J. Numer. Meth. Fluids* 2017; 83:544–558

DOI: 10.1002/fld
The inverse Hessian matrix $H_k$ can be decomposed into the form $H_k^0 = L_0 L_0^T$, for example, by Cholesky decomposition, then as in [13], an efficient sampling is achieved by giving the $H_k$ update rule (A.1) into

$$H_k = Q_0 Q_0^T + \sum_{i=1}^{m} q_i q_i^T$$

where

$$Q_0 = (V_{k-1}^T \cdots V_{k-m}^T) L_0$$

$$q_1 = \sqrt{p_{k-1}} s_{k-1}$$

$$q_i = \sqrt{p_{k-i}} (V_{k-1}^T \cdots V_{k-i+1}^T) s_{k-1} \cdots i = 2, \ldots m \text{ and } p_i > 0 \forall i$$

$Q_0$ is a $N \times N$ matrix, $N$ is the dimension of the model state vector, and $q_i$ are vectors of size $N \times 1$. With the aforementioned representation, a random vector $r$ with zero mean and covariance $H_k (r \sim N(0, H_k))$ can be sampled with the help of (A.3)

$$r = Q_0 z + \sum_{i=1}^{m} \omega_i q_i$$

(A.3)

where $z \sim N(0, I)$ and $w \sim N(0, 1)$.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the following people: to Professor Heikki Haario of Lappeenranta University of Technology (LUT) for his contribution to developing the VEnKF and sound advice; to Olli Malve and Janne Ropponen of Finnish Environmental Institute (SYKE), for access to lake Pyhäjärvi materials and valuable advice regarding COHERENS; to Dr. Patrick Luyten of RBINS-MUMM for valuable advice and for providing a very well-structured and thoroughly documented model code as the chief architect of COHERENS; and, finally, but not least, to Aleksandr Bibov of LUT for providing the VEnKF library as its chief designer and to Juhani Saastamoinen of Arbonaut Ltd. for providing measurement data and illustrations on Lake Säkylän Pyhäjärvi. And our special thanks go to Prof. Ionel Michael Navon of Florida State University (FSU) for an inspiring and unrelenting deluge of ideas and references; without which, we would often have been totally lost!

REFERENCES

1. Madsen H, Skotnert C. Adaptive state updating in real-time river flow forecasting – a combined filtering and error forecasting procedure. *Journal of Hydrology* 2005; **308**:302–312.

2. Reichle RH, Crow WT, Keppenne CL. An adaptive ensemble Kalman filter for soil moisture data assimilation. *Water Resources Research* 2008; **44**(3):W03423.

3. Moradkhani H, Sorooshian S, Gupta HV, Houser PR. Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources* 2005; **28**:135–147.

4. Amour I, Mussa Z, Bibov A, Kauranne T. Using ensemble data assimilation to forecast hydrological flumes. *Nonlinear Processes in Geophysics* 2013; **20**:955–964.

5. Heemink AW, Metelkaar IDM. Data assimilation into a numerical shallow water flow model: a stochastic optimal control approach. *Journal of Marine Systems* 1995; **6**:145–158.

6. Luyten PJ, Jones JE, Proctor R. A numerical study of the long- and short-term temperature variability and thermal circulation in the North Sea. *Journal of Physical Oceanography* 2003; **33**:37–56.

7. Ponsar S, Luyten P. Data assimilation with the EnKF in a 1-D numerical model of a North Sear station. *Ocean Dynamics* 2009; **59**:983–996.

8. Auvinen H, Bardsley JM, Haario H, Kauranne T. The variational Kalman filter and an efficient implementation using limited memory BFGS. *International Journal for Numerical Methods in Fluids* 2009; **64**:314–335.

9. Evensen G. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Geophysical Research* 1994; **99**:10143–10162.

10. Li H, Kalnay E, Miyoshi T. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quarterly Journal of Royal Meteorological Society* 2009; **135**:523–533.

11. Courtier P, Andersson E, Heckley W, Vasiljevic D, Hamrud M, Hollingsworth A, Rabier F, Fisher M, Pailleux J. The ECMWF implementation of three-dimensional variational assimilation (3D-VAR). I: formulation. *Quarterly Journal of Royal Meteorological Society* 1998; **124**(550):1783–1807.

© 2016 The Authors. *International Journal for Numerical Methods in Fluids* Published by John Wiley & Sons Ltd. DOI: 10.1002/fld
Nocedal J, Wright SJ. *Numerical Optimization* (2nd edn). Springer: New York, NY 10013, USA, 2006.

Solonen A, Haario H, Hakkarainen J, Auvinen H, Amour I, Kauranne T. Variational ensemble Kalman filtering using limited memory BFGS. *Electronic Transaction on Numerical Analysis* 2012; 39:271–285.

Nigham N. *Accuracy and Stability of Numerical Algorithms* (2nd edn). SIAM: Philadelphia, 2002.

Zapanski M. Maximum likelihood ensemble filter: theoretical aspects. *Monthly Weather Review* 2005; 133(6):1710–1726.

Clark MP, Rupp DE, Woods RA, Zheng X, Ibbitt RP, Slater AG, Schmidt J, Uddstrom MJ. Hydrological data assimilation with the ensemble Kalman filter: use of streamflow observations to update states in a distributed hydrological model. *Electronics Transactions on Numerical Analysis* 2008; 31:1309–1324.

Xie X, Zhang D. Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. *Advances in Water Resources* 2010; 33:678–690.

Kim K, Park M, Min J, Ryu I, Kang M, Park LJ. Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. *Journal of Hydrology* 2014; 519:2810–2821.

Komma J, Bloshch G, Reszler C. Soil moisture update by ensemble Kalman filtering in real-time flood forecasting. *Journal of Hydrology* 2008; 357:228–242.

Kim S, Seo D, Riaz H, Shin C. Improving water quality forecasting via data assimilation – application of maximum likelihood ensemble filter to HSPF. *Journal of Hydrology* 2014; 519:2797–2809.

Reichle RH, Mclaughlin DB, Entekhabi D. Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review* 2002; 130:103–130.

Martin N, Gorelick SM. MODFreeSurf2D: A MATLAB surface fluid flow model for rivers and streams. *Computers and Geosciences* 2005; 31:929–946.

Bellus C, Soulis J, Sakkas J. Computation of two-dimensional dam-break induced flows. *Advances in Water Resources* 1991; 14:31–41.

Luyten P (ed.) *COHERENS - A Coupled Hydrodynamical-Ecological Model for Regional and Shelf Seas: User Documentation*. Version 2.6. RBINS Report, Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, 2014.

Huttula T. Suspended sediment transport in lake Säkylän Pyhäjärvi. *Aqua Fennica* 1994; 24(2):171–185.

Rasanen M, Salonen VP, Salo J, Walls M, Sarvala J. Recent history of sedimentation and biotic communities in lake Pyhajarvi, SW Finland. *Journal of Paleolimnology* 1992; 7(2):107–126.

Mano A, Malve O, Koponen S, Kallio K, Taskinen A, Koponen J, Juntunen J, Liukko N. Assimilation of satellite data to 3D hydrodynamic model of lake Säkylän Pyhäjärvi. *Water Science and Technology* 2015; 71(7):1033–1039.

Ştefănescu R, Sandu A, Navon I. POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation. *Journal of Computational Physics* 2015; 295:569–595.

Attia A, Ştefănescu R, Sandu A. The reduced-order hybrid Monte Carlo sampling smoother. *Computing Research Repository* 2016; abs/1601.00129.

Houtekamer P, Mitchell H. A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review* 2001; 129(1):123–137.

Hamill T, Whitaker J, Snyder C. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review* 2001; 129(11):2776–2790.

Bardsley J, Solonen A, Parker A, Haario H, Howard M. An ensemble Kalman filter using the conjugate gradient sampler. *International Journal of Uncertainty Quantification* 2013; 3(4):357–370.