

Improving Operational Water Quality Forecasting with Ensemble Data Assimilation

Hamideh Riazi,¹ Sunghee Kim,¹ Dong-Jun Seo,¹ Changmin Shin² and Kyunghyun Kim²

¹The University of Texas at Arlington, Arlington, Texas; ²National Institute of Environmental Research (NIER), Republic of Korea.

Abstract

Being able to predict water quality in river systems accurately is critical to protecting public health from harmful water quality conditions such as algal blooms or bacterial pollution, and to allowing the decision makers to respond more quickly to emergencies such as oil spills. Water quality forecasting is subject to a number of sources of uncertainty: uncertain observations, model states, model parameters, model structures, and future input forcings. Because many of the water quality model states are never observed and the models are never perfect, the initial conditions (IC) of the model may be highly uncertain. Updating the ICs of the model based on real time observations is hence potentially a cost effective way to improve the accuracy of water quality forecasts. Data assimilation (DA) is a technique that optimally combines model-simulated observations and actual observations to provide more accurate estimates of the model ICs. In this work we describe the DA procedure for the Hydrologic Simulation Program–Fortran (HSPF) based on the maximum likelihood ensemble filter (MLEF). The resulting application, MLEF–HSPF, serves as a plugin module for the Water Quality Forecast System at the National Institute of Environmental Research (WQFS–NIER) in support of operational water quality forecasting. Also presented are the evaluation results for four catchments in three different river basins in the Republic of Korea.

1 Introduction

Harmful algal blooms (HABs) in major rivers and lakes are a large environmental issue in Korea. Controlled release of impounded water from reservoirs is one of the most active measures that can be used in responding to major HABs. For such measures, accurate short range forecasts of both water quantity and quality are necessary to minimize the release while meeting the water quality requirements. To provide reservoir managers with accurate predictive water quality information, the Water Quality Control Center of the National Institute of Environmental Research (NIER) in the Republic of Korea produces real time water quality forecasts for the four major rivers in Korea. For watershed water quality forecasting, the Hydrological Simulation Program–Fortran (HSPF, Bicknell et al. 2001) is used. For river water quality forecasting, the Environmental Fluid Dynamics Code (EFDC, Hamrick 2007) is used (see Figure 1). The HSPF model is one way coupled with EFDC such that the former provides the latter with the boundary conditions (BC). Because the HSPF results are used as the BCs of the EFDC model, one may expect that improving the accuracy of the ICs of the watershed water quality model will not only improve the accuracy of the river water quality forecasts by providing more accurate BCs for the hydrodynamical model but also increase the lead time by leveraging the hydrologic memory in the upstream catchments, which is significantly longer than the hydraulic mem-

ory in the river systems. For real time operation of these models, NIER uses the Water Quality Forecast System (WQFS) WQFS–NIER, developed by NIER and Deltares based on Delft–FEWS (Werner et al. 2004). Figure 1 shows the schematic of the daily water quality forecast process. The last step is post-processing in which a report is prepared to inform the water resources management agencies of the water quality forecast results twice a week through a dedicated website (<http://wqcast.nier.go.kr:8080/>).

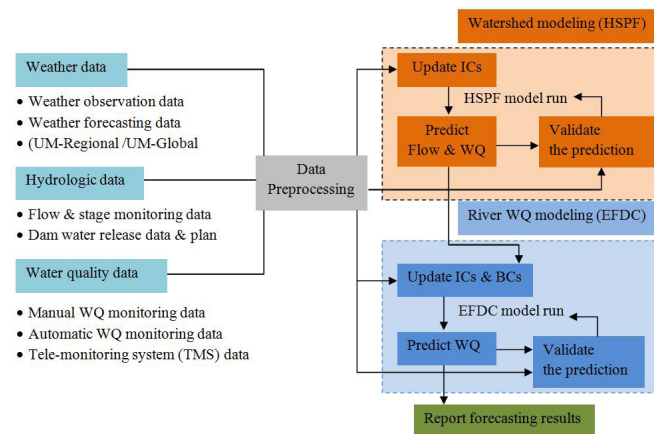


Figure 1 Schematic of the water quality forecast process using WQFS–NIER.

Water quality forecasting is subject to a number of sources of uncertainty: uncertain observations, model states, model parameters, model structures, and future input forcings (Beck 1987). A number of studies have shown that the largest errors in water quality prediction occur when the state variables change rapidly due to increased biochemical activities (see, e.g., Beck 1987 and references therein). Because most model states are never observed and the models are never perfect, the initial conditions (IC) of the model may be highly uncertain. To initialize HSPF, NIER developed the user control input (UCI) file based on available observations and laboratory experiments for each catchment of the four major rivers in the Republic of Korea. In their previous practices, NIER relied solely on long *cold-start* runs to warm up the model states for operational forecasting (Kim et al. 2015). The purpose of this work is to reduce the uncertainty in the initial conditions (IC) for real time runs by utilizing real time observations via data assimilation (DA). To keep the model states in line with the unfolding reality as reflected in the real time observations of hydrologic and water quality variables, it is necessary in operational forecasting to employ some form of state updating. DA is an objective way to optimally estimate the model states by jointly utilizing the actual observations available in real time and the model-simulated observations.

Various DA techniques have been used in water quality forecasting, such as the Kalman filter (Guo et al. 2003), the ensemble Kalman filter (Chang and Latif 2011; Jin and Chang 2008), and the extended Kalman filter (Mao et al. 2009; Pastres et al. 2003), to name just a few. The above DA techniques, however, are not very effective when both the model dynamics and observation equations are highly nonlinear. In this work, we develop, implement and evaluate a DA module for the watershed water quality model, HSPF, for the open architecture real time forecast system, WQFS–NIER, in support of short range (a few to several days ahead) water quality forecasting for the four major rivers in Korea. The DA technique used is the maximum likelihood ensemble filter (MLEF, Zupanski 2005), which is capable of handling nonlinearities in both model dynamics and observation equations.

2 Methodology

The DA technique used in this work for HSPF is MLEF (Zupanski 2005). The resulting procedure and the plugin module for WQFS–NIER is referred to as MLEF–HSPF. Using the real time water quality and streamflow observations, MLEF–HSPF updates the hydrologic and water quality states of HSPF (see Table 1 below) which provides the future BCs for EFDC (see Figure 1 above).

Based on the experience of Seo et al. (2003), Seo et al. (2009) and Lee et al. (2011, 2012), we use a fixed lag smoother formulation (Schweppe 1973; Li and Navon 2001) for DA as illustrated in Figure 2.

Table 1 HSPF state variables considered by DA.

HSPF Module	Variable name	Observed?	Definition
PERLND (for pervious land)	CEPS	No	Interception Storage
	SURS	No	Surface (Overland Flow) Storage
	UZS	No	Upper Zone Storage
	IFWS	No	Interflow Storage
	LZS	No	Lower Zone Storage
	AGWS	No	Active Groundwater Storage
	GWVS	No	Index to Groundwater Slope
	SQO-NH ₄	No	Storage of NH ₄ on the surface
IMPLND (for impervious land)	SQO-NO ₃	No	Storage of NO ₃ on the surface
	SQO-PO ₄	No	Storage of PO ₄ on the surface
	SQO-BOD	No	Storage of BOD on the surface
	RETS	No	Retention storage
	SURS	No	Surface (overland flow) storage
	SQO-NH ₄	No	Storage of NH ₄ on the impervious surface
	SQO-NO ₃	No	Storage of NO ₃ on the impervious surface
	SQO-PO ₄	No	Storage of PO ₄ on the impervious surface
RCHRES (for in-stream process)	SQO-BOD	No	Storage of BOD on the impervious surface
	VOL	No	Volume of water in the RCHRES at end of interval
	TW	Yes	Water temperature
	DOX	Yes	Dissolved Oxygen Concentration
	BOD	Yes	Biochemical Oxygen Demand concentration
	NO ₃	Yes	Dissolved concentration of NO ₃
	TAM	Yes	Dissolved concentration of TAM (incl. NH ₃ , NH ₂)
	PO ₄	Yes	Dissolved concentration of PO ₄
	PHYTO	No	Phytoplankton concentration
	ORP	No	Organic Refractory Phosphorus
ORN	No	Organic Refractory Nitrogen	
ORC	No	Organic Refractory Carbon	

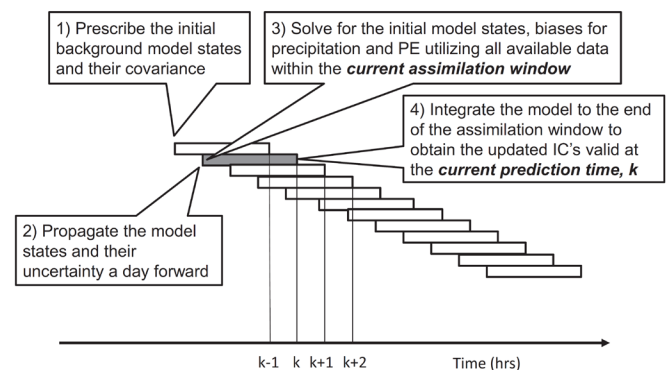


Figure 2 Schematic of the DA cycle based on the fixed lag smoother formulation.

In each assimilation cycle, all observations within the assimilation window are used to update the model ICs valid at the beginning of the assimilation window and the multiplicative adjustment factors to mean areal precipitation (MAP) and mean areal potential evapotranspiration (MAPE) valid over the assimilation window. It is well known that forcing errors greatly

impact the accuracy of ICs of both hydrologic and water quality variables. Because large errors may exist in the estimates of MAP and MAPE, two control variables representing the multiplicative adjustment factors, or biases, for MAP and MAPE were added to the control vector. Following Seo et al. (2003, 2009) and Lee et al. (2011, 2012), the biases are assumed to be spatially uniform over the subcatchment of interest and temporally uniform within the assimilation window. To implement the multiplicative adjustment factors to MAP and MAPE, segment-specific precipitation and PE within the subcatchment were weight-averaged according to the size of the area to derive MAP over the entire subcatchment. Model segments are subareas of a watershed that are connected by a river network with largely uniform parameters and meteorological input. The adjustment factors inflate or deflate MAP and MAPE while keeping the spatial pattern of MAP and MAPE among the segments the same as originally prescribed. While the choice of spatiotemporally uniform bias may seem overly simplistic, experience with hydrologic models indicates that the gain from a more complex approach such as spatiotemporally varying bias is rather small.

The ICs and the adjustment factors form the control vector for the DA algorithm. The assimilation window, or the time scale of the fixed lag, is set at 7 d in reflection of the response time of the basins in the study area (Seo et al. 2013) and the sampling frequency of about once a week for water quality observations. For prediction, HSPF is run over the assimilation window using the updated ICs valid at the beginning of the assimilation window and the adjustment factors valid over the assimilation window to produce the updated model states valid at the prediction time, which are then used to forward-integrate the model 7 d into the future. For the next assimilation cycle, HSPF is forward-integrated from the beginning of the current assimilation window to that of the next window to produce the updated model ICs valid at the beginning of the next assimilation cycle (see Figure 2 above).

The MLEF–HSPF algorithm consists of five steps. In the first step, the analysis error covariance is obtained from the previous run. For the very first run, a lognormal distribution of control variables is assumed (Kim et al. 2014). Then ensemble runs are generated in a new analysis cycle using the analysis error covariance. The square root forecast error covariance, $P_{f,k}^{1/2}$, is calculated using Equations 1 and 2 based on the control and perturbation runs:

$$P_{f,k}^{1/2} = (b_1 b_2 \dots b_s) \quad (1)$$

$$b_i = \tilde{M} \tilde{p}_i \approx M(x_{k-1} + \tilde{p}_i) - M(x_{k-1}) \quad (2)$$

where:

- i = i th ensemble member,
- b_i = i th ensemble member of $P_{f,k}^{1/2}$,
- $\tilde{M}_{k-1,k}$ = Jacobian of the dynamical model with respect to the control vector transitioning from time step $k-1$ to time step k , and
- \tilde{p}_i = i th ensemble member of analysis error covariance.

The square root of the analysis error covariance, $P_{a,k}^{1/2}$ and \tilde{p}_i are defined as follows:

$$P_{a,k}^{1/2} = (\tilde{p}_1 \tilde{p}_2 \dots \tilde{p}_s) \quad (3)$$

$$\tilde{p}_i = (p_{1,i} p_{2,i} \dots p_{N,i}); i = 1, 2, \dots, s \quad (4)$$

where:

- N = number of control variables, and
- s = number of ensemble size.

The entry p_{ji} in Equation 4 denotes the ensemble perturbation for the j th control variable in the i th ensemble member.

In step 2, the following weakly constrained minimization problem is solved to update the control variables. The solution represents the maximum likelihood estimate of the control variables.

Minimize:

$$J(x_k) = \frac{1}{2} (x_k - x_{b,k})^T P_f^{-1} (x_k - x_{b,k}) + \frac{1}{2} (y_k - H(x_k))^T R^{-1} (y_k - H(x_k)) \quad (5)$$

subject to:

$$x_k = M(x_{k-1}) + w_{k-1} \quad (6a)$$

$$x_{lower} \leq x_k \leq x_{upper} \quad (6b)$$

where:

- x_k = control vector,
- x_{lower} = lower bound of the control vector,
- x_{upper} = upper bound of the control vector,
- $x_{b,k}$ = a priori, or background, states of the control vector,
- y_k = observation vector,
- $H()$ = nonlinear observation operator,
- R = observation error covariance matrix,
- $P_f()$ = forecast covariance matrix of the model states,
- $M()$ = dynamical model, and
- w_{k-1} = dynamical model error vector at time step $k-1$.

The observation equation associated with the above cost function is given by:

$$y_k = H(x_k) + v_k \quad (7)$$

where:

- v_k = observation error vector at time step k .

MLEF solves the nonlinear constrained minimization problem of Equations 5 and 6 in ensemble subspace with Hessian preconditioning via the variable transformation in Equation 8 (Zupanski 2005).

$$x - x_b = P_f^{1/2} (I + C)^{-T/2} \zeta \quad (8)$$

where:

- I = identity matrix,
- ζ = control vector in ensemble subspace, and

C = information matrix (see Zupanski 2005 for definition) $Z^T Z$.

The column vectors Z_i , where i denotes the i th ensemble member of Z in the information matrix, are defined and approximated via finite differencing as follows:

$$\begin{aligned} Z_i &= (R^{-1/2} \tilde{H} P_f^{1/2})_i \\ &= R^{-1/2} \tilde{H} b_i \\ &\approx R^{-1/2} H(x+b_i) - R^{-1/2} H(x) \end{aligned} \quad (9)$$

where:

\tilde{H} = Jacobian of the observation function, $H()$, with respect to the control variables.

The gradient of the cost function with respect to the control vector in ensemble subspace is given by Zupanski (2005):

$$g_\zeta = (I+C)^{-1} \zeta - (I+C)^{-1} (R^{-1/2} \tilde{H} P_f^{1/2})^T R^{-1/2} \left\{ y - H[x_b + P_f^{1/2} (I+C)^{-1/2} \zeta] \right\} \quad (10)$$

For minimization, MLEF–HSPF uses the Fletcher–Reeves–Polak–Ribière algorithm `fprmn` (Press et al. 1986).

In step 3, the control solution in ensemble subspace, ζ , is back-transformed to that in the physical space, x_{opt} , via Equation 8. In step 4, the square root analysis covariance matrix of the model state, $P_a^{1/2}$, is obtained from Equation 11 and then used as ensemble perturbations for the next analysis cycle according to Equations 3 and 4:

$$P_a^{1/2} = P_f^{1/2} [I + C(x_{opt})]^{-T/2} \quad (11)$$

In the last step, the model is integrated forward using the updated state variables to predict the state variables over the forecast horizon.

In the original formulation of MLEF, no model errors were assumed. While MLEF has since been used for estimation of systematic model errors (Zupanski and Zupanski 2006), it does not allow accounting for random errors (Zupanski 1997; Jazwinski 1970). In this work, we model the latter via state augmentation. For details, the reader is referred to Kim et al. (2014).

DA as formulated above addresses uncertainties in the ICs and observed BCs of MAP an MAPE only. If significant parametric or structural errors exist in the model, DA is likely to adjust the model ICs for the wrong reasons to compensate for any systematic biases that may exist from other sources of error. Ideally, issues such as model biases and limited dynamic range should be addressed by improving model physics and calibration before DA. Such an effort, however, is expected to occur incrementally over time. As an alternative, MLEF–HSPF employs a statistical bias correction procedure to account for systematic errors so that the DA solution may be found within the dynamic range of the model. To correct conditional biases in model simulations, particularly in the right tail end of their distributions, we employ conditional

bias-penalized optimal linear estimation (Seo 2013; Seo et al. 2014). This linear regression model is incorporated into the observation equation of the DA formulation. In this procedure, the HSPF simulation, X , is related to the truth, Y , via the following linear relationship:

$$Y = aX + b + \varepsilon \quad (12)$$

where:

- a = slope parameter,
- b = intercept parameter, and
- ε = zero-mean random error.

The best estimator of the truth, Y^* , is assumed to take the form:

$$Y^* = aX + b \quad (13)$$

In conventional linear regression, the parameters a and b are solved for by minimizing the error variance, $E_{Y^*,X}[(Y^* - X)^2]$. In conditional bias-penalized optimal linear estimation, we add a type-II conditional bias penalty term, following Seo (2013), to minimize:

$$J = E_{Y^*,X} \left[(Y^* - X)^2 \right] + \alpha E_X \left[\left\{ E_{Y^*} [Y^* | X] - X \right\}^2 \right] \quad (14)$$

where:

- α = weight given to the type-II conditional bias penalty term.

It can be easily shown (Seo 2013) that the solution to Equation 14 is given by:

$$a = \frac{1 + \alpha}{1 + \alpha \rho_{XY}^2} \rho_{XY} \frac{\sigma_Y}{\sigma_X} \quad (15)$$

$$b = m_Y + a m_X \quad (16)$$

where:

- ρ_{XY} = correlation between X and Y ,
- σ_X = standard deviation of X ,
- σ_Y = standard deviation of Y ,
- m_X = mean of X , and
- m_Y = mean of Y .

3 Evaluation

The MLEF–HSPF module has been extensively tested and evaluated for multiple catchments in the Republic of Korea. In this paper, we present the results for four catchments. The Kumho catchment in the Nakdong River basin, the Miho catchment in the Geum River basin and the Jojong and Yanghwa catchments in the Han River basin (see Figure 3) have respectively drainage areas of 2 170.65 km², 1 854.49 km², 262.49 km² and 352.65 km² and respectively comprise 31, 39, 10 and 10 model segments resulting in totals of 331, 412, 103 and 103 control variables (i.e. x in Equation 8) that are associated with the 28 HSPF state variables (see Table 1 above) and the two multiplicative adjustment factors.

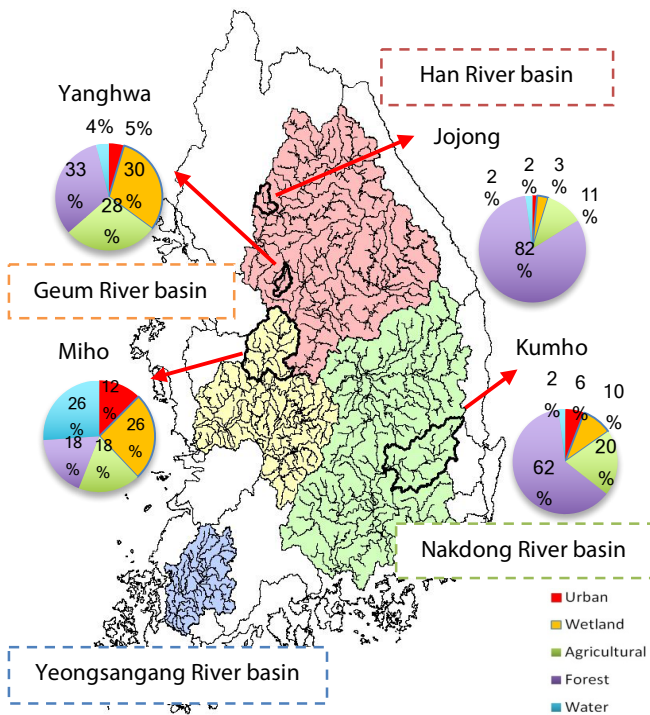


Figure 3 The four catchments used in the evaluation.

The criteria used for selecting the catchments included the quality of HSPF simulation, number of water quality variables observed, period of record of observations, and observation frequency. The streamflow and water quality data used in the evaluation are the instantaneous observations of streamflow, biochemical oxygen demand (BOD), chlorophyll a (CHL-a), dissolved oxygen (DO), nitrate (NO_3), phosphate (PO_4), ammonium (NH_4), water temperature (TW), total nitrate (TN), total phosphate (TP) and total organic carbon (TOC), taken approximately once a week at the water quality and point source monitoring stations. The observation time, typically between 10:00 and 17:00, varies from station to station and according to the sampling schedule.

For DA to be effective, the model has to be reasonably skillful. If the model cannot simulate the biophysiochemical processes that occur in the catchment, one may not expect DA to add skill. As such, understanding the performance of the model is important in assessing the performance of DA. In this work, skill refers to the ability of the DA-aided forecast to outperform DA-less forecast. There are multiple aspects to consider in evaluating the quality of base model simulation for application of DA. The model may be better at simulating certain processes and variables than others. The model may be able to simulate many variables well individually, but may not be able to simulate co-variability among them. The model may capture co-variability well, but may have phase, or timing, errors in some variables. To illustrate how the quality of base model simulations was assessed in this work, we show in Figure 4 the correlation between the observed and simulated variables. We also examined the inter-variable correlation which also showed significant variability among the catchments

and paired variables. They suggest that one may expect significant variations in the performance of DA among different catchments and variables.

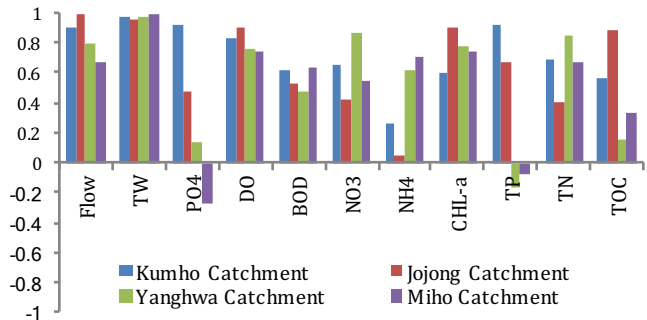


Figure 4 Correlation between observed and simulated variables at the outlets of the Kumho, Miho, Jojong and Yanghwa catchments for 2012.

For evaluation, a set of hindcasting experiments was carried out in which HSPF was run (1) without bias correction or DA; (2) with bias correction only; and (3) with bias correction and DA under the assumption of perfectly known future MAP and MAPE. They are denoted as Base, BC-Base and BC-DA, respectively. Below we present selected results for the four catchments. In the hindcasting experiment, the observed variables were predicted out to 7 d into the future in each assimilation cycle for 2012. Based on the sensitivity analysis carried out for the Kumho catchment (Kim et al. 2014), 9 ensemble traces, which represent plausible, equally likely realizations of what may occur, were used for all four catchments. In all four catchments, the sample size varies significantly between analysis and prediction due to missing observations, and the fact that the observations are made only once a week at the most. For the Kumho and Miho catchments, all observations are available weekly for the entire period. For the Jojong and Yanghwa catchments, however, only monthly observations are available in the first six months of 2012. Because the number of observations used in DA is different among different catchments, one may expect the effectiveness of DA to vary among different catchments due not only to different levels of predictability and predictive skill but also to the varying amount of data assimilated. As such, caution is necessary in comparing the results among different catchments. Two types of flow observation are available for most of the catchments in Korea: (1) mean daily flow based on measurement made every 10 min over a 24 h period; and (2) weekly observations made at the total maximum daily load (TMDL) stations where water quality samples are taken. Using only the weekly streamflow observations significantly reduces the hydrologic information content available for DA due to the sparse sampling frequency. Our experience, however, is that assimilating simultaneously the two types of streamflow observations generally has negative impact on DA performance due to the fact that, at the mean daily flow stations, there are no water quality observations available. Also, the locations of mean daily and weekly flow observations are not the same which poses inconsistent spatial sampling for DA. The

use of TMDL streamflow observations, on the other hand, ensures spatiotemporal consistency in information content over all assimilation cycles. To overcome the above limitation, we are currently investigating *decomposed DA* in which all available streamflow data are assimilated first before the water quality observations are assimilated in a second DA operation, and the results will be reported in the near future.

With the parameter settings used for the Kumho catchment, hindcasting of the DA procedure for a single assimilation cycle takes ~4 min on a desktop with 3.3 GHz Intel Core i5 CPU. Because the DA procedure is run only once a day for each catchment, the computational cost is not a large issue in the real time mode. Figure 5 shows the time series of the BC-DA (○) vs. Base (○) and BC-Base (○) results and the verifying observation (○) for CHL-a at the outlet of the Kumho catchment. Note that there are multiple BC-DA results for each verifying observation. This is because DA is performed every day over an assimilation window of 7 d, resulting in as many as seven analysis results. It is readily seen in the figure that BC-DA tracks the verifying observations much more closely than Base or BC-Base.

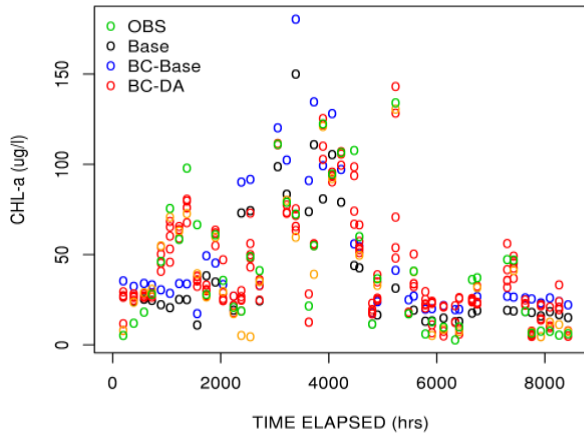


Figure 5 Base, BC-Base and BC-DA vs the verifying observations of CHL-a at the outlet of the Kumho catchment.

To assess the quality of the analysis and prediction results of water quality variables and streamflow, a set of statistical measures, including root mean square error (RMSE) and mean square error (MSE) decomposition, were used. Figure 6 shows the analysis RMSE of BASE, BC-Base and BC-DA based on assimilating 1 d old to 5 d old observations of CHL-a at the outlet locations of the four catchments. Note that the DA analysis results differ depending on the age of the observations. It was found that the DA analysis results based on 1 d old to 5 d old observations have significantly larger predictive skill than those based on 6 d old observations. Not surprisingly, 7 d old observations provided little predictive skill and hence were not considered for assimilation.

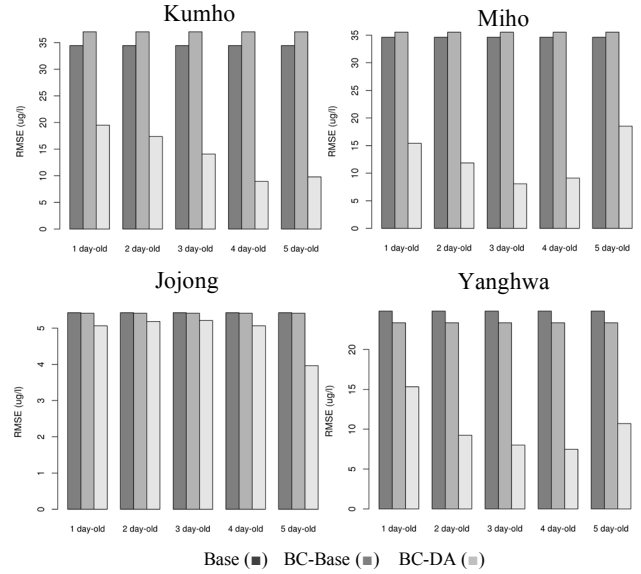


Figure 6 RMSE Base (left bars), BC-Base (middle bars) and BC-DA (right bars) for analysis using 1 d old to 5 d old observations for CHL-a for the Kumho, Miho, Jojong and Yanghwa catchments.

Jojong is a rather small catchment (262.49 km²). As such, one may expect smaller hydrologic memory and hence smaller potential for DA. Note also that the absolute magnitude of CHL-a concentration is much lower for Jojong than for the others. It is possible that there may not be such large uncertainty in the ICs for this basin for DA to readily realize significant improvement. The most likely explanation for the increase in RMSE for DA using 5 d old observations for Miho and Yanghwa is that the biophysiochemical processes in these catchments may operate at time scales <5 d, which the DA is not able to capture using older observations. Unfortunately, the above postulation could not be verified due to the lack of daily observations of water quality. All results shown in this paper are based on true validation; that is, validation was carried out using the data that were not used in calibration or parameter estimation. Because BC is purely a statistical correction, its performance is necessarily susceptible to inter-annual variability of the hydrologic and water quality processes. For example, if the observed flow and water quality vary greatly from one period to another, one may expect the performance of BC to deteriorate. Also, being more urbanized catchments, one may expect Miho and Kumho to be more susceptible to anthropogenic nonstationary effects which would also reduce the performance of BC.

To ascertain where the improvement in the DA analysis may be coming from, we carried out MSE decomposition. The MSE can be decomposed into three terms (Murphy and Winkler 1987; Nelson et al. 2010):

$$MSE = \sum_{i=1}^N (f_i - o_i)^2 = (m_f - m_o)^2 + (\sigma_f - \sigma_o)^2 + 2\sigma_f\sigma_o(1-\rho) \quad (17)$$

where:

- f_i = i th forecast,
- o_i = verifying observation associated with f_i ,
- N = number of pairs of forecast and verifying observation,
- m_f = mean of forecast,
- m_o = mean of verifying observation,
- σ_f = standard deviation of forecast,
- σ_o = standard deviation of the verifying observation,
- and
- ρ = correlation between the forecast and the verifying observation.

In the right hand side of Equation 17, the first and second terms measure the bias in the mean and standard deviation, respectively, and the third term measures the strength of covariation (the smaller, the stronger) between the prediction and the verifying observation (Nelson et al. 2010). Figure 7 shows the MSE decomposition of the analysis results of BASE, BC-Base and BC-DA for CHL-a at the outlet locations in the Kumho, Miho, Jojong and Yanghwa catchments. It is seen that BC alone does not consistently reduce errors, due presumably to sampling uncertainties from interannual variability. However, BC-DA reduces biases in the mean and standard deviation, and significantly improves the strength of covariation between observed and simulated values. It is interesting to note that, for the heavily forested Jojong catchment for which the DA analysis results are relatively poor (see Figure 6 above), DA improves covariation but increases biases. Additional investigation is needed to explain the above.

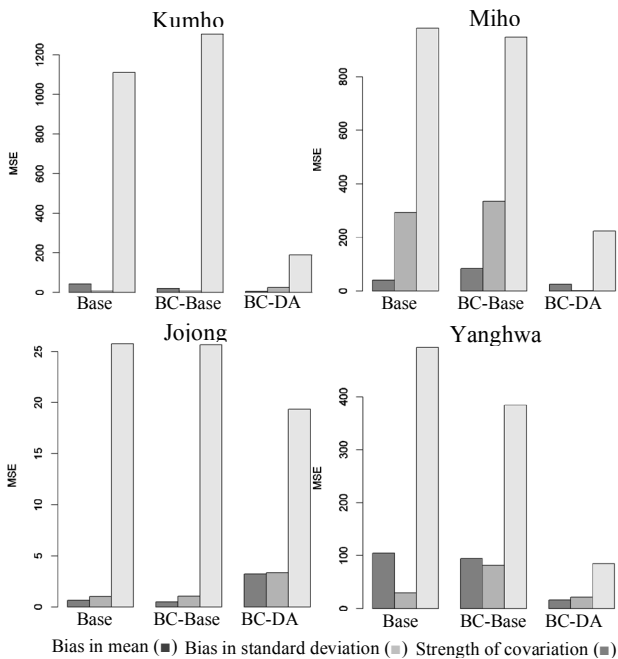


Figure 7 MSE decomposition for Base, BC-Base and BC-DA analysis for CHL-a for the Kumho, Miho, Jojong and Yanghwa catchments.

Figure 8 shows the RMSE of Base (left bars), BC-Base (middle bars) and BC-DA (right bars) results for Day-1 through Day-3 predictions for CHL-a for the outlets of the four catchments. Note that, except for Miho, BC-DA reduces RMSE over Base or BC-Base for Day-1 through Day-3 predictions. The reductions in RMSE in the BC-DA results of Yanghwa and Jojong are probably contributed by small sample size and may not hold over a long run. The lack of improvement by BC-DA for Miho probably reflects lack of predictability of hydrologic and water quality processes in that catchment; note in Figure 3 that Miho has by far the largest fraction of urban areas, and in Figure 4 that the quality of the base model simulation tends to be lower, particularly for PO_4 and TP.

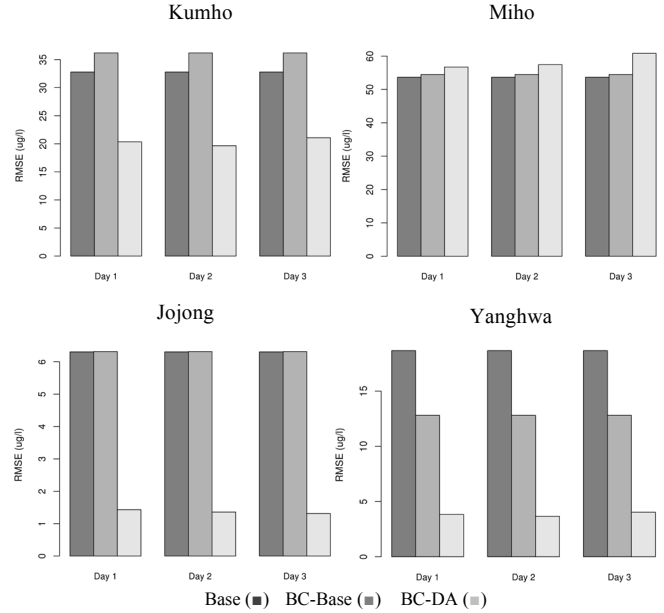


Figure 8 RMSE of Day-1 through 3 predictions by Base (left bars), BC-Base (middle bars) and BC-DA (right bars) for CHL-a for the Kumho, Miho, Jojong and Yanghwa catchments.

4 Conclusions and Research Recommendations

In watershed water quality modeling, only a very small subset of the model state variables is actually observed. As such, they are subject to large errors, reduction of which via data assimilation (DA) may significantly improve short range water quality forecasting. In this work, we apply maximum likelihood ensemble filter (MLEF) to the Hydrologic Simulation Program–Fortran (HSPF) to improve the accuracy of real time water quality prediction by updating the initial conditions (IC) of the model. MLEF combines strengths of variational assimilation (VAR) and ensemble Kalman filter (EnKF). Unlike VAR, however, it does not need adjoint code and unlike EnKF it does not assume a linear observation equation. The resulting plugin module for the Water Quality Forecast System at the National Institute of Environmental Research (WQFS–

NIER), which is based on the Flood Early Warning System (FEWS), is referred to as MLEF–HSPF and provides boundary conditions for river water quality forecasting using the environmental fluid dynamics code (EFDC) within WQFS–NIER.

For evaluation, we applied MLEF–HSPF to the Kumho catchment in the Nakdong River basin, Miho catchment in the Geum River basin, and Jojong and Yanghwa catchments in the Han River basin in the Republic of Korea. The results indicate that MLEF–HSPF significantly improves short range prediction of key water quality variables for Kumho, Jojong and Yanghwa. The margin of improvements, however, varies significantly from catchment to catchment and from variable to variable. Improvement was larger for more natural catchments compared to more urbanized catchments. For Miho, which has a large fraction of urban areas and lower skill in the base HSPF simulation, however, DA is seen to add little predictive skill even though significant improvement is seen in analysis. It suggests that HSPF is not able to simulate the anthropogenic effects in the Miho catchment and hence state updating via DA is not able to add skill. Evaluation of the ensemble members from DA analysis indicates that the HSPF model is not able to propagate uncertainty in the ICs realistically, resulting in underspread analysis ensembles. Work is under way to employ a statistical post-processor that operates on the HSPF output to produce reliable ensembles.

Acknowledgments

This work is supported by the Water Quality Control Center of the National Institute of Environmental Research (NIER), Republic of Korea, under the Agreement of the Cooperative Study between Geosystem Research Corporation, Korea and The University of Texas at Arlington. The second and third authors were supported also in part by the National Science Foundation under Grant No. CyberSEES-1442735 (Dong-Jun Seo, University of Texas at Arlington, PI). These supports are gratefully acknowledged.

References

- Beck, M. B. 1987. "Water Quality Modeling: A Review of the Analysis of Uncertainty." *Water Resources Research* 23 (8): 1393–442.
<https://doi.org/10.1029/WR023i008p01393>.
- Bicknell, B. R., J. C. Imhoff, J. L. Kittle Jr., T. H. Jobes and A. S. Donigian Jr. 2001. *Hydrological Simulation Program–Fortran (HSPF): User's Manual for Release 12*. Athens, GA: USEPA National Exposure Research Laboratory (NERL).
- Chang, S.-Y. and S. Latif. 2011. "Use of Regional Covariance in Data Assimilation Method to Improve the Estimation Accuracy of a Three Dimensional Contaminant Transport Model." In *Proceedings, World Environmental and Water Resources Congress 2011*, 1118–26. Reston, VA: American Society of Civil Engineers.
[https://doi.org/10.1061/41173\(414\)115](https://doi.org/10.1061/41173(414)115).
- Guo, H. C., L. Liu and G. H. Huang. 2003. "A Stochastic Water Quality Forecasting System for the Yiluo River." *Journal of Environmental Informatics* 1 (2): 18–32.
- Hamrick, J. M. 2007. *The Environmental Fluid Dynamics Code: User Manual, Version 1*. Fairfax, VA: USEPA.
- Jazwinski, A. H. 1970. *Stochastic Processes and Filtering Theory*. Mathematics in Science and Engineering, vol. 64. New York: Academic Press.
- Jin, A. and S.-Y. Chang. 2008. "Kalman Filter for Subsurface Transport Models with Inaccurate Parameters and Unknown Sources." *Journal of Environmental Informatics* 12 (1): 37–43.
- Kim, S., D.-J. Seo, H. Riazi and C. Shin. 2014. "Improving Water Quality Forecasting via Data Assimilation—Application of Maximum Likelihood Ensemble Filter to HSPF." *Journal of Hydrology* 519:2797–2809.
- Kim, S., D.-J. Seo, C. Shin and H. Song. 2015. "HSPF Restart Function for Short-Range Water Quality Forecasting and Data Assimilation." In *Proceedings, World Environmental and Water Resources Congress 2015*, 2443–8. Reston, VA: American Society of Civil Engineers.
<https://doi.org/10.1061/9780784479162.239>.
- Lee, H., D.-J. Seo and V. Koren. 2011. "Assimilation of Streamflow and in Situ Soil Moisture Data into Operational Distributed Hydrologic Models: Effects of Uncertainties in the Data and Initial Model Soil Moisture States." *Advances in Water Resources* 34 (12): 1597–615.
- Lee, H., D.-J. Seo, Y. Liu, V. Koren, P. McKee and R. Corby. 2012. "Variational Assimilation of Streamflow into Operational Distributed Hydrologic Models: Effect of Spatiotemporal Scale of Adjustment." *Hydrology and Earth System Sciences* 16 (7): 2233–51.
- Li, Z. J. and I. M. Navon. 2001. "Optimality of Variational Data Assimilation and its Relationship with the Kalman Filter and Smoother." *Quarterly Journal of the Royal Meteorological Society* 127 (572): 661–83.
- Mao, J. Q., J. H. W. Lee and K. W. Choi. 2009. "The Extended Kalman Filter for Forecast of Algal Bloom Dynamics." *Water Research* 43 (17): 4214–24.
- Murphy, A. H. and R. L. Winkler. 1987. "A General Framework for Forecast Verification." *Monthly Weather Review* 115:1330–8.
- Nelson, B. R., D.-J. Seo and D. Kim. 2010. "Multisensor Precipitation Reanalysis." *Journal of Hydrometeorology* 11 (3): 666–82.
- Pastres, R., S. Ciavatta and C. Solidoro. 2003. "The Extended Kalman Filter (EKF) as a Tool for the Assimilation of High Frequency Water Quality Data." *Ecological Modelling* 170 (2): 227–35.
- Press W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling. 1986. *Numerical Recipes*. Cambridge: Cambridge University Press.
- Scheweppe, F. C. 1973. *Uncertain Dynamic Systems*. Englewood Cliffs, NJ: Prentice-Hall.

- Seo, D.-J. 2013. "Conditional Bias-Penalized Kriging (CBPK)." *Stochastic Environmental Research and Risk Assessment* 27 (1): 43–58.
- Seo, D.-J., L. Cajina, R. Corby, and T. Howieson. 2009. "Automatic State Updating for Operational Streamflow Forecasting Via Variational Data Assimilation." *Journal of Hydrology* 367 (3): 255–75.
- Seo, D.-J., V. Koren and N. Cajina. 2003. "Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational Hydrologic Forecasting." *Journal of Hydro-meteorology* 4 (3): 627–41.
- Seo, D.-J., R. Siddique, Y. Zhang and D. Kim. 2014. "Improving Real-Time Estimation of Heavy-to-Extreme Precipitation Using Rain Gauge Data via Conditional Bias-Penalized Optimal Estimation." *Journal of Hydrology* 519 (B): 1824–35. <https://doi.org/10.1016/j.jhydrol.2014.09.055>.
- Werner M. G. F., M. Van Dijk and J. Schellekens. 2004. "DELFT-FEWS: An Open Shell Flood Forecasting System." In *Hydroinformatics: Proceedings of the 6th International Conference, Singapore*, edited by S. Y. Liong, K. Phoon and V. Babovic, 1205–12. Singapore: World Scientific. https://doi.org/10.1142/9789812702838_0149.
- Zupanski, D. 1997. "A General Weak Constraint Applicable to Operational 4DVAR Data Assimilation Systems." *Monthly Weather Review* 125 (9): 2274–92.
- Zupanski, D. and M. Zupanski. 2006. "Model Error Estimation Employing an Ensemble Data Assimilation Approach." *Monthly Weather Review* 134 (5): 1337–54.
- Zupanski, M. 2005. "Maximum Likelihood Ensemble Filter: Theoretical Aspects." *Monthly Weather Review* 133 (6): 1710–26.