River Flow and Stage Estimation with Missing Observation Data using Multi Imputation Particle Filter (MIPF) Method

Z. H. Ismail[1], N. A. Jalaludin[2]
[1]*Centre for Artificial Intelligence and Robotics, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia.*
[2]*Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.*
zool@utm.my

***Abstract*—An advanced knowledge of the river condition helps for better source management. This information can be gathered via estimation using DA methods. The DA methods blend the system model with the observation data to obtain the estimated river flow and stage. However, the observation data may contain some missing data due to the hardware power limitations, unreliable channel, sensor failure and etc. This problem limits the ability of the standard method such as EKF, EnKF and PF. The Multi Imputation Particle Filter (MIPF) able to deal with this problem since it allows for new input data to replace the missing data. The result shows that the performance of the river flow and stage estimation is depending on the number of particles and imputation used. The performance is evaluated by comparing the estimated velocity obtained using the estimated flow and stage, with the measured velocity. The result shows that higher number of particles and imputation ensure better estimation result.**

***Index Terms*—Missing Data; State Estimation; Multi Imputation Particle Filter.**

I. INTRODUCTION

In hydrology, the estimation of the river flow, stage and state can contribute to better management of the water resource for human usage [1]. Whereby the estimated values give an advanced knowledge of the related parameters and helps for better improvements of water use efficiency and also balance its supply and demand [2]. By considering the energy harvesting concept, these parameters can also be used to predict the electricity that can be generated from the river system [3]. The estimation can be conducted using the Data Assimilation (DA) method. The DA method is any techniques that integrate observation data with the system model to produce an updated model state that most accurately approximates the true system state whilst keeping the model parameter fixed [4]. This method describes the flow of information from observations of the real system to the numerical model of the system, in the form of probability density function (pdf). The concern of the DA is to obtain the new posterior probability density of the system model when the new observations are involved [5]. The updated posterior pdf is then used to initiate the next model forecast. This method is desired to perform estimation in an optimal and consistent fashion even if the noisy measurements is arrived sequentially in time[6]. The general formulation of the pdf is represented by the Bayes theorem that is based on conditional probability densities [7].

The DA method is available in two class namely variational method and sequential method. The variational methods are based on the optimal control theory. Optimization is performed on the related parameters by minimizing the cost function that measures the model to data misfit [8]. The examples of this method are Variational data assimilation method (VAR), Evolutionary data assimilation method (EDA) and Maximum Likelihood Ensemble Filter (MLEF). Besides that, the sequential methods use a probabilistic framework and estimate the whole system state sequentially by propagating information only forward in time. This method does not require an adjoint model and makes it easy to adapt with the model [9]. Compared to the variational method, the sequential based method is frequently used in estimation. The examples of this method are the Extended Kalman Filter (EKF), Ensemble Kalman Filter (EnKF), Unscented Kalman Filter (UKF), Particle Filter (PF) and etc.

During estimation process, the river model and the observation by the sensors are combined together to obtain the predicted river flow, stage and cross section. The river system and observation are nonlinear since their condition may changes over time and the system is most probably disturbed by the external factors such as the evaporation, rainfall, precipitation and etc. These factors are some of the uncertainties that must be considered during estimation process[8]. Besides that, the system and observation also have their own uncertainties and errors that influence the estimation result [10]. The selection of the DA methods for estimation considers the characteristics of the system, observation and the external factors, since the ability of the DA method is very much dependent on their characteristics [11] [12].

There are two types of sensor for measurement namely Eulerian sensor and Lagrangian sensor[13]. The Eulerian sensors perform measurement as the water flow past the sensor that was placed at fixed location. While the Lagrangian sensor is more flexible since it observes the medium as it moves together with the water flow along a trajectory [2]. So, better measurement can be achieved by applying the Lagrangian sensors that provide more accurate measurement than the Eulerian [14]. However, the measurement by these sensors may

be disturbed by the obstacles and not all measurement locations are suitable for the sensors [2]. Besides that, the measurement may suffer from missing data due to the hardware power limitations, unreliable channel, sensor failure and etc. [15]. This problem may limit the ability of the standard DA method to perform prediction. Therefore, the Multi Imputation Particle Filter (MIPF) is proposed in this paper to deal with this problem by introducing new data input to replace the missing data.

The paper is structured as follows. The system model, observation model and the state space model for estimation process is described in Section II. Then, in Section III, the effect of the missing observation during estimation is explained. Next, the algorithm of the MIPF method for estimation with missing data is described in Section IV. Finally, in Section V, the detail on estimation process and numerical simulations are discussed.

II. The Model

The river flow model can be represented by one or two-dimensional Saint-Venant equations depending on the characteristic of the water flow. If the flow is in one-dimensional, the 1D Saint-Venant equations is considered. However, if the flow is not one-dimensional which may happen in flood plains or in large rivers, the 2D Saint-Venants equation is more suitable to be applied [16]. Besides that, the representation of the observation is referring to the movement of the sensor since the Lagrangian sensor is use in this research[2]. The combination of the system model and the observation is represented by the state space model and use in the DA method.

A. System Model

Consider one-dimensional flow without any uncontrolled release of water flow, the 1D Saint-Venant equations is suitable for river flow. This equation is among the most common models used for modelling the flow in open channels and irrigation systems [17]. The 1D Saint-Venant equations are two coupled first order hyperbolic partial differential equations (pde) derived from the conservation of mass and momentum. By considering a prismatic channel that have same cross-section throughout the length of channel with no lateral inflow, the equation is represented as [2].

$$T\frac{\partial H}{\partial t}+\frac{\partial Q}{\partial x}=0 \tag{1}$$

$$\frac{\partial Q}{\partial t}+\frac{\partial}{\partial x}\left(\frac{Q^2}{A}\right)+\frac{\partial}{\partial x}\left(gh_cA\right)=gA\left(S_0-S_f\right) \tag{2}$$

$$S_f=\frac{m^2Q^2P^{4/3}}{A^{4/3}} \tag{3}$$

where A is the cross section (m^2), Q is the discharge or flow(m3/s), L is the river reach(m), T is the free surface width, D is the hydraulic depth (m), S_f is the friction slope, S_o is the bed slope, g is the gravitational acceleration(m/s2), h_c is the distance of the centroid of the cross section from the free surface (m), P is the wetted perimeter, m is the Manning roughness coefficient.

B. Observation Model

The observation is represented by the velocity of the flow. Since the velocity throughout the system is change with time, the measurement can be performed using the Lagrangian sensor or drifter. The relation between the drifter velocity and the flow velocity at the corresponding cross-section relies on assumptions made about the profile of the water velocity. The profile is the combination of the average velocity in the transverse and vertical direction. In transverse direction, the surface velocity profile is assumed to be quartic, and the Von Karman logarithmic profile is assumed in the vertical direction. By considering a particle moving at a distance y from the center line and z from the surface, the relation between the particle's velocity and the water flow is represented by the following equations [2]:

$$v_p(y,z)=F_T(y)F_v(z)\frac{Q}{A} \tag{4}$$

with

$$F_T(y)=A_q+B_q\left(\frac{2y}{w}\right)^2+C_q\left(\frac{2y}{w}\right)^4 \tag{5}$$

$$A_q+B_q+C_q=0 \tag{6}$$

$$A_q+\frac{B_q}{3}+\frac{C_q}{5}=1 \tag{7}$$

$$F_v(z)=1+\left(\frac{0.1}{K_v}\right)\left(1+\log\left(\frac{z}{d}\right)\right) \tag{8}$$

where w is the channel width, d is the water depth, A_q, B_q and C_q are constants, K_v is the Von Karman log constant.

C. The state space model

During estimation process, the system and observation is represented by the state space model that consists of the parameters of the model, observation, system noise and measurement noise. The development of the model involved the discretization of the system into n cells with each cell have same length. The initial conditions and the boundary conditions of the system is included in this model as the inputs. Further, the uncertainties of the model and also the inaccuracies of the inputs measurements are considered as the system noise, ς_t. While the measurement noise, ε_t represent the errors and the uncertainties of the measurements. Both noises are represented by the zero mean Gaussian error. Thus, the state space model for the estimation is described as follow:

$$x_{t+1}=f(x_t,u_t,\varsigma_t) \tag{9}$$

$$\varphi_t=g(x_t,\varepsilon_t,t) \tag{10}$$

where x_t is the state vector at time t.

$$x_t = \left(Q_2^t, \ldots, Q_n^t, H_1^t, \ldots, H_{n-1}^t\right)^T \quad (11)$$

and the input u_t contains the boundary conditions, i.e. the upstream flow and downstream stage.

$$u_t = \left(Q_1^t, H_n^t\right)^T \quad (12)$$

where Q_i^t and H_i^t are the flow and stage at cell i at time t, respectively, and n is number of cells used for the discretization of the channel.

Since the system is observed by K sensors, (10) can be reformulated into:

$$\varphi_t = \begin{pmatrix} g_1\left(x_t, \varepsilon_{t,1}, t\right) \\ \vdots \\ g_K\left(x_t, \varepsilon_{t,K}, t\right) \end{pmatrix} = \begin{pmatrix} \varphi_{t,1} \\ \vdots \\ \varphi_{t,K} \end{pmatrix} \quad (13)$$

where φ_t represent the noisy observation of the state x_t such that the $\varepsilon_{t,k}$ is an independent and identically distributed (i.i.d) measurement noise and g_k is the measurement transformation for sensor k.

III. Problem Formulation

The estimation of system states by using standard DA method apply Bayes' theorem that denoted as [18]:

$$p\left(x_t \mid \varphi_{1:t}\right) = \frac{p(\varphi_t \mid x_t)\, p(x_t \mid \varphi_{1:t-1})}{p(\varphi_t \mid \varphi_{1:t-1})} \quad (14)$$

where x_t is the system state at time t, φ_t is the observation at time t, $p\left(x_t \mid \varphi_{1:t}\right)$is the posteriori probability of state x at time t given observation φ from time 1 to time t, $p(\varphi_t \mid x_t)$ is the likelihood function of state x at time t given observation φ at time t, $p(x_t \mid \varphi_{1:t-1})$ is the prior probability of state x at time t given observation φ from time 1 to time t-1, $p(\varphi_t \mid \varphi_{1:t-1})$ is the normalizing constant. The normalizing constant is represented as[19].

$$p\left(\varphi_t \mid \varphi_{1:t-1}\right) = \int p\left(\varphi_t \mid x_t\right) p\left(x_t \mid \varphi_{1:t-1}\right) dx_t \quad (15)$$

Based on (14) and (15), the posteriori probability is very much depending on the likelihood function $p(\varphi_t \mid x_t)$. This function use the observation φ_t to modify the prior probability to obtain the desired posteriori probability that represent the estimated state.

In this research, the observation is related to y and z position of the sensors, and also the velocity of the sensors. The missing of the observation data will eventually affect the estimation process since the likelihood function could not be obtained and limit the ability of the standard DA method. Therefore, the MIPF method is introduces to perform estimation with new input data.

The availability of the observations is checked at each time instance. The missing data are handled by introducing a random indicator variable, $R_{t,k}$[15]

$$R_{t,k} = \begin{cases} 0 & \text{Observation is missing from sensor } k \text{ at time } t \\ 1 & \text{Observation is available from sensor } k \text{ at time } t \end{cases}$$

The collection of observations $\varphi_{t,k}$ at time instance t for all sensors k = 1,...,K with $R_{t,k}$ = 0 is defined as missing information set Ξ_t. While the available information set Ψ_t is the collection of $\varphi_{t,k}$ for all k = 1,...,K such that $R_{t,k}$ = 1.

IV. Multi Imputation Particle Filter

The Multi Imputation Particle Filter (MIPF) uses randomly drawn values called imputations to provide a replacement for the missing data and then uses the particle filter to perform estimation with the data. The imputations are draw from the proposal function, ϕ [20].

$$\Xi_t^j \sim \phi(\Xi_t \mid \Psi_{0:t}) = \sum_{i=1}^{N} \tilde{\omega}_t^i\, p\left(\Xi_t \mid \tilde{X}_t^i\right) \text{ for } j = 1, \ldots, M \quad (16)$$

where Ξ_t represent all missing observations at time t, $\Psi_{0:t}$ represent all available observation from time 0 to time t, $\left\{\tilde{\omega}_t^i, \tilde{X}_t^i\right\}_{i=1}^{N}$ is the particle set with no regard of missing data, N is the total number of particles, M is the total number of imputation, i is i^{th} particles and j is j^{th} imputation.

Next, the imputations are reformulated into imputed data sets

$$U_t^j = \left\{\Xi_t^j, \Psi_t\right\} \quad (17)$$

where Ξ_t^j represent all missing observation during jth imputation and time t, and Ψ_t represent all available observation at time t.

The posterior probability density with missing observation is represented by

$$p(X_t \mid \Psi_{0:t}) = \int p(X_t \mid \Gamma_{0:t-1}, \Psi_t)\, p(\Xi_t \mid \Psi_{0:t}) d\Xi_t \quad (18)$$

where X_t is the system state at time t, $\Gamma_{0:t-1}$ is the complete observation by the sensors that include available and missing observation, Ξ_t and $\Psi_{0:t}$ are defined in (16), and Ψ_t is defined (17).

By considering Monte Carlo approximation and imputations, (18) can be written as follows:

$$p(X_t \mid \Psi_{0:t}) \approx \frac{1}{M} \sum_{j=1}^{M} p(X_t \mid \Gamma_{0:t-1}, U_t^j) \quad (19)$$

where M is defined in (16), U_t^j is defined in (17), and X_t, $\Psi_{0:t}$, $\Gamma_{0:t-1}$ are defined in (18).

For each data set U_t^j, the probability density from particle filtering is written as follows:

$$p(X_t \mid U_{0:t-1}, U_t^j) \approx \sum_{i=1}^{N} \omega_t^{j,i} \delta\left(X_t - X_t^{j,i}\right) \quad (20)$$

where $X_t^{j,i}$ is the system state at *i*th particle and *j*th imputation at time instance t, and $\omega_t^{j,i}$ is the related weight.

The overall representation of the posterior probability density with missing data is determined by substituting (20) into (19) and form

$$p\left(X_t \mid \Psi_{0:t}\right) \approx \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} \omega_t^{j,i} \delta\left(X_t - X_t^{j,i}\right) \quad (21)$$

where $\Psi_{0:t}$, *M*, *N* are defined in (16), X_t is defined in (18), $X_t^{j,i}$, $\omega_t^{j,i}$ are defined in (20).

V. Results and Discussion

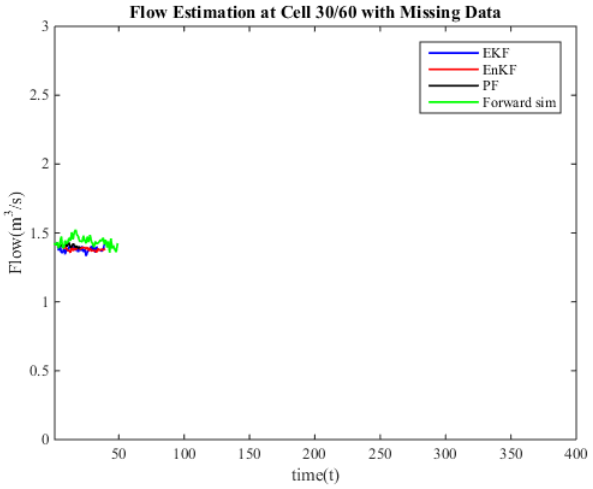
The estimation the system state is carried out by blending the system model with the observation from the sensors, via the DA method. The availability of the observation data is demanded by the likelihood function and will certainly influence the estimation process as explained in section III. So, for missing observation problem, the MIPF can be applied since it has the ability to replace the missing data with the new data during estimation process.

A. Description of the estimation process

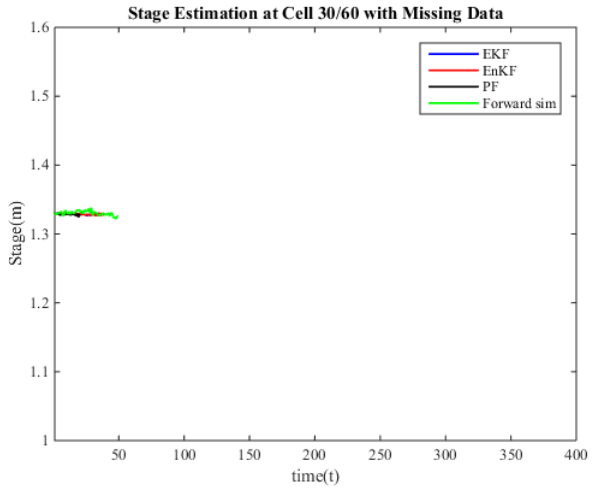
In this research, five sensors are used to measure the velocity of the flow for approximately 400 second. The sensors are released one by one with 30 second of interval. The river system has gate at the end and the gate was opened as soon as the final drifter was released. The estimation is conducted to predict river flow, stage and cross section by integrating the system model with the observation using the DA method. Next, the estimated states are used to predict the velocity of the river flow and compared with actual velocity by the sixth drifter to evaluate the performance of this method.

B. Result and Discussion

The observation data is considered to be suffered from 10% and 20% of missing data. Three types of the basic DA method namely the EKF, EnKF and PF are applied but only able to perform prediction before the observation data become unavailable for the first time as shown in Figure 1. Since the velocity estimation is very much related to the prediction of the flow and stage, this problem affects the obtained estimated velocity as in Figure 2. The result shows that the missing data affect the blending process whereby the probability calculation for prediction could not be carried out without the reference data. In order to solve this problem, external data input is required as a replacement to the missing data and can be implemented using the MIPF.



(a)



(b)

Figure 1: The (a) flow and (b) stage estimation with missing observation data by using forward simulation, EKF, EnKF and PF at 30th cell

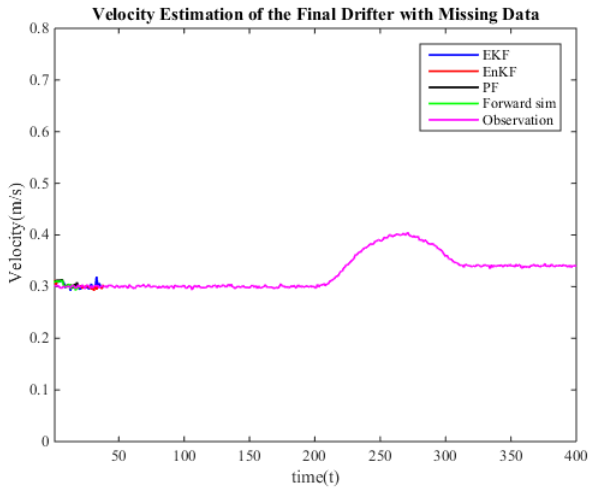


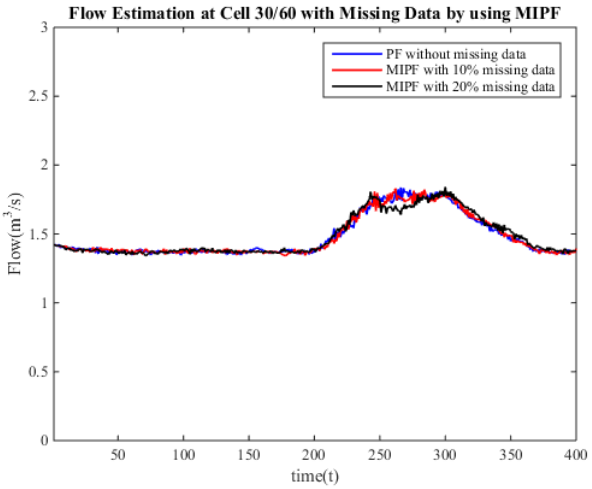
Figure 2: The velocity estimation with missing observation data by using forward simulation, EKF, EnKF and PF.

The MIPF allows for any number of input data that are generated based on the previously available data and weight. The new generated data and weight are used in the prediction

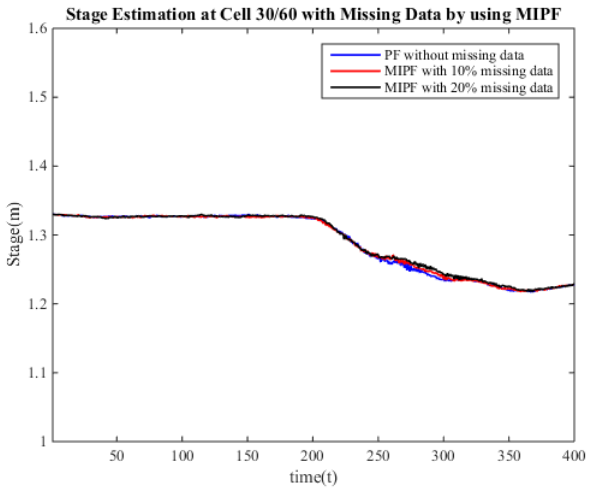
of the desired parameter. In this research, few sets of MIPF with different number of particles and imputations are applied to different level of missing data as listed in Table 1. Since the MIPF is activated at the point where the PF unable to perform prediction, the performance of this method is observed through the overall estimation using this method and PF. The performance is evaluated based on the relative error between the estimated velocity and the measurement. The result shows that the estimation during missing data is comparable with the estimation during no missing problem. The percentage of missing data influences the number of imputation to be applied. For small percentage of missing data, small number of imputation is required and vice versa. Besides that, the number of particles also affects the estimation result. Whereby, better result can be achieved with the increasing number of particles before the degeneracy problem is occurred like the standard PF method. In this research 50 particles are used by the particle filtering method and for the missing data problem, 10 imputations and 20 imputations are required by the MIPF to have good estimation result for 10% missing data and 20% missing data respectively. However, the increase of the number of particles and imputation will increase the computational time. The application of the MIPF for river flow, stage and state estimation with incomplete observation data able to produce good estimation result and ensures the chance to have good velocity estimation as shown in Figure 3 and Figure 4.

Table 1
The performance comparison between the MIPF during missing data and PF without missing data

Method	Particles	Imputation	Relative error (%)
PF (no missing data)	50	-	4.0576
MIPF (10% missing data)	50	5	4.0610
MIPF (20% missing data)	50	20	4.1641



(a)



(b)

Figure 3: The (a) flow and (b) stage estimation without missing observation data by using PF, and 10% and 20% of missing observation data by using MIPF at 30th cell.

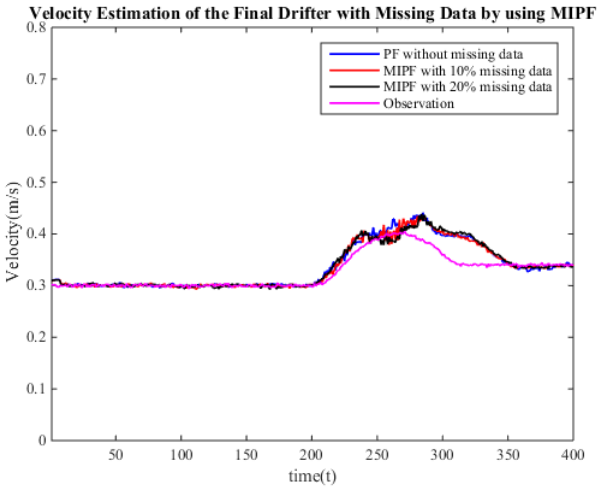


Figure 4: The velocity estimation without missing observation data by using PF, and 10% and 20% of missing observation data by using MIPF.

VI. Conclusion

Missing data affect the quality of the observation. This problem disturbs the estimation process whereby the standard DA method such as EKF, EnKF and PF are only able to perform estimation whenever the observation is available. By applying MIPF, the missing data is replaced with new set of data according to the particle set during no missing data problem. The performance of this method is depending on the number of particles and imputation used. The proper combination of the number of particles and imputation ensure good estimation result.

Acknowledgment

This work was supported in part by the Ministry of Higher Education, Malaysia and Universiti Teknologi Malaysia under Grant no. Q.K130000.2543.13H80 and Centre for Artificial Intelligence and Robotics no. U.K091303.0100.00000.

REFERENCES

[1] A. Y. Sun, D. Wang, and X. Xu, "Monthly streamflow forecasting using Gaussian Process Regression," *Journal of Hydrology.*, vol. 511, pp. 72–81, Apr. 2014.

[2] A. Tinka, M. Rafiee, and A. M. Bayen, "Floating sensor networks for river studies," *IEEE System Journal*, vol. 7, no. 1, pp. 36–49, 2013.

[3] S. Michelin and O. Doaré, "Energy harvesting efficiency of piezoelectric flags in axial flows," *Journal of Fluid Mechanics.*, vol. 714, pp. 489–504, Jan. 2013.

[4] P. J. Smith, G. D. Thornhill, S. L. Dance, A. S. Lawless, D. C. Mason, and N. K. Nichols, "Data assimilation for state and parameter estimation: Application to morphodynamic modelling," *Quarterly Journal of Royal Meteorological Society*, vol. 139, no. 671, pp. 314–327, 2013.

[5] Y. Liu and H. V. Gupta, "Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework," *Water Resource Research*, vol. 43, pp. 1–18, 2007.

[6] S. A. Gadsden, M. Al-Shabi, I. Arasaratnam, and S. R. Habibi, "Combined cubature Kalman and smooth variable structure filtering: A robust nonlinear estimation strategy," *Signal Processing*, vol. 96, no. PART B, pp. 290–299, 2014.

[7] X. He, R. Sithiravel, R. Tharmarasa, B. Balaji, and T. Kirubarajan, "A spline filter for multidimensional nonlinear state estimation," *Signal Processing*, vol. 102, pp. 282–295, 2014.

[8] S. Kim, D.-J. Seo, H. Riazi, and C. Shin, "Improving water quality forecasting via data assimilation – Application of maximum likelihood ensemble filter to HSPF," *Journal of Hydrology*, Oct. 2014.

[9] L. Bertino, G. Evensen, and H. Wackernagel, "Sequential Data Assimilation Techniques in Oceanography," *International Statistical Review*, vol. 71, no. 2, pp. 223–241, 2003.

[10] J. Samuel, P. Coulibaly, G. Dumedah, and H. Moradkhani, "Assessing model state and forecasts variation in hydrologic data assimilation," *Journal of Hydrology*, vol. 513, pp. 127–141, May 2014.

[11] G. G. Rigatos, "A derivative-free kalman filtering approach to state estimation-based control of nonlinear systems," *IEEE Transaction on Industrial Electronics*, vol. 59, no. 10, pp. 3987–3997, 2012.

[12] H. Chen, D. Yang, Y. Hong, J. J. Gourley, and Y. Zhang, "Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting," *Advances in Water Resource*, vol. 59, pp. 209–220, Sep. 2013.

[13] T.-J. Chang, H.-M. Kao, K.-H. Chang, and M.-H. Hsu, "Numerical simulation of shallow-water dam break flows in open channels using smoothed particle hydrodynamics," *Journal of Hydrology*, vol. 408, pp. 78–90, Sep. 2011.

[14] Y. Yuan, J. W. C. Van Lint, R. E. Wilson, F. Van Wageningen-kessels, and S. P. Hoogendoorn, "Real-Time Lagrangian Traffic State Estimator for Freeways," *IEEE Transaction on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 59–70, 2012.

[15] X. Zhang, A. S. Khwaja, J. Luo, A. S. Housfater, and A. Anpalagan, "Convergence Analysis of Multiple Imputations Particle Filters for dealing with Missing Data in Nonlinear Problems," *IEEE Journal of Selected Topic in Signal Processing*, vol. 9, no. 8, pp. 2567–2570, 2014.

[16] X. Litrico and V. Fromion, "Modeling of Open Channel Flow," in *Modeling and Control of Hydrosystems*, 1st ed., Springer-Verlag London, 2009, pp. 17–41.

[17] X. Liu, A. Mohammadian, J. Angel, and I. Sedano, "Irrigation & Drainage Systems Engineering One Dimensional Numerical Simulation of Bed Changes in Irrigation Channels using Finite Volume Method," *Irrigation and Drainage Systems Engineering*, vol. 1, no. 2, pp. 1–6, 2012.

[18] D. Crisan and A. Doucet, "A survey of convergence results on particle filtering methods for practitioners," *IEEE Transaction on Signal Processing*, vol. 50, no. 3, pp. 736–746, 2002.

[19] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transaction on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[20] X. Zhang, A. S. Khwaja, J. Luo, A. S. Housfater, and A. Anpalagan, "Multiple Imputations Particle Filters: Convergence and Performance Analyses for Nonlinear State Estimation With Missing Data," *IEEE Journal of Selected Topic in Signal Processing*, vol. 9, no. 8, pp. 1536–1547, 2015.