Application of hybrid Kalman filter for improving water level forecast
Xuan Wang and Vladan Babovic

ABSTRACT
Numerical modeling is one of the popular means to simulate and forecast the state of oceanographic systems. However, it still suffers from some limitations, e.g., parameter uncertainties, simplification of model assumptions, and absence of data for proper boundary and initial conditions. This paper proposes a hybrid data assimilation scheme, which combines the Kalman filter (KF) with a data-driven model (local linear model (LM)), to directly correct numerical model outputs at locations without measurements. Two different types of KF (unscented Kalman filter and two-sample Kalman filter) are tested and compared. A local LM is utilized to describe the evolution of the model state and then assimilated into the KF. This in turns simplifies the application of KF for highly complex nonlinear systems such as the dynamic motion of Singapore regional water. The proposed scheme is first examined using a simple hypothetical bay experiment followed by an operational model of the Singapore Regional Model (SRM), in which both are set up in the Delft3D modeling environment. This combination of KF and data-driven model provides insights into the influence of different error covariance estimations on the model updating accuracy. This research also provides guidance to offline utilization of KF in updating of numerical model output.

Key words | data assimilation, model correction, spatial correlation, two-sample Kalman filter, unscented Kalman filter

INTRODUCTION
Understanding oceanographic systems is important for safe navigation and offshore operations. In principle, deterministic equations that describe physical phenomena can be numerically solved to forecast a future state based on initial condition and evolution of forcing terms. However, these types of numerical models tend to produce imperfect results for several reasons (e.g., model resolutions, parameter uncertainties, simplification of model assumptions, and absence of data for proper boundary and initial conditions). By combining the results of numerical models with the measurements, data assimilation can serve as a useful tool to combat the inevitable presence of model error. This has enhanced the model accuracy of sea approximation (Babovic et al. 2001; Vojinovic & Kecman 2003; van den Boogaard & Mynett 2004; Sun et al. 2009; Sun 2010).

The Kalman filter (KF) (Kalman 1960; Chui & Chen 1999), a widely used data assimilation approach, has been applied in several oceanographic and meteorological studies (Evensen & Van Leeuwen 1996; Madsen & Canizares 1999). Sun et al. (2009) have used conventional KF to describe the spatial distribution of the measurements in the case of the Singapore Regional Model (SRM). Although the simple assumption of linear propagation of the model state has facilitated the KF application in the SRM case study, this relative simplification of state updating equation may not completely describe highly complex nonlinear systems (e.g., the dynamic motion of Singapore regional water), and the estimation of the predicted covariance matrix is also lacking. In addition, real error covariance was shown to be poorly represented, and hence limited the KF
performance. In contrast, the extended Kalman filter (EKF) (Jazwinski 1970) is a natural choice for nonlinear systems. The nonlinear transition and observation function are simply linearized and transformed into a matrix of partial derivatives (Jacobians). It is a computationally efficient and recursive update form of KF. However, the linearized transformations of the nonlinear system are only reliable when the error propagations are well approximated by a linear function. Furthermore, such linearization in EKF can only be applied if the Jacobian matrix exists. Besides, the calculation of the Jacobian matrix is an error-prone process (Julier & Uhlmann 2004; Aguirre et al. 2005). Therefore, it may lead to sub-optimal performance and sometimes divergence of the filter (Aguirre et al. 2005). On the other hand, the ensemble Kalman filter (EnKF), one of the most advanced sequential assimilation methods (Evensen 1994), extends the conventional KF and replaces the error covariance matrix by the sample covariance computed from ensembles. It is an advantageous approach for highly dimensional application, and has been applied in different complex models (Hamill 2006; Sakov & Sandery 2015). However, the optimal ensemble size is uncertain and is chosen based on a heuristic evaluation (Medina et al. 2014). Its optimality in terms of reducing error variance depends on the linearity assumption of the model and observation operators (Luo & Moroz 2009; Lei & Baehr 2013). The estimation of the EnKF based on small ensemble sizes can also be affected by large biases, even if the ensemble mean and covariance are correct (Simon et al. 2015). Another problem with the above KFs is that, in the case of a forced stable system, the updated initial conditions quickly ‘wash-out’ over a certain forecast horizon (Babovic & Fuhrman 2002). For example, the study of Singapore regional water by Karri et al. (2015) shows that EnKF is only suitable for short-time forecasting (<6 hours) in predicting water levels and currents. As the lead time increases, the forecasting accuracy deteriorates and tends to be consistent with the original numerical model result. This is actually a common issue for conventional utilization of KF. Moreover, the KF implementation requires the numerical model to be run online at each iteration, and thus is highly time-consuming.

To overcome the limitations in the above-mentioned KF approaches, two different KF methods are applied in this study, namely, two-sample Kalman filter (two-SKF) (Sumihar et al. 2008) and unscented Kalman filter (UKF) (Uhlmann 1994; Julier & Uhlmann 1997). These two approaches estimate the error covariance and thereby Kalman gain without the limitations described above (e.g., Jacobian matrix or amounts of ensembles). The two-SKF derives steady-state Kalman gain in the updating procedure, while the UKF derives dynamic Kalman gain based on the unscented transform (UT). In addition, a local linear model (LM) is introduced to first simulate the dynamic of the original numerical model offline. The forecasting results from the LM are then assimilated into the KF algorithm to predict the state variable. Based on the updated results using the LM method, updating of the initial conditions at each time step during the forecasting period is undertaken offline without re-running a time-consuming original numerical model. This combination not only enables both KF methods to be updated offline without the initial conditions being washed out, but also reduces the computational cost. Thus it would be more applicable for a real-time updating system.

The main purpose of this paper is to introduce these two hybrid KF methods and examine their performance based on case studies. The proposed scheme is tested in a hypothetical scenario of an artificial bay and then applied in a real case of SRM to correct the water level outputs at non-measured locations. Based on the comparison of predicted results between the two KF methods, their advantages and corresponding potential applications to complex nonlinear systems are discussed.

**ALGORITHM**

For KF and its derivative, the internal state of process is described by the state transition function. The LM introduced by Babovic & Fuhrman (2002), as one kind of data-driven model, can produce accurate forecasting for a short forecasting horizon (Wang & Babovic 2014). Therefore, the LM is used to simulate the state process based on numerical model output, to simplify the application of KF (e.g., two-SKF and UKF in this study) in complex nonlinear systems (e.g., SRM). This hybrid scheme is illustrated in Figure 1. The LM is applied here to forecast the original numerical model output directly at all the locations of interest. It
aims to simulate the system dynamic of outputs of the original model. These forecasting results and measurements are then assimilated into the KF as inputs. The measurements provide background information to tune the Kalman gain and update the state at the current time step. The updated state is then forecasted using the LM. In this way, the state variable at all non-measured stations can be updated online with actual measurements and predicted results based on LM simulation for the forecast period.

**Local linear model**

For a time series in a chaotic system, it is known that LM can utilize the inner nonlinear deterministic rule based on Taken’s embedding theorem (Takens 1981) to reconstruct a phase (or embedded) space. This embedded space is equivalent to the original state from a scalar time series.

Given time series $x_n$ in reference period $t_n$, the phase space can be constructed in terms of embedded vector $X_n(t)$, through the parameters of time lag $\tau$ and embedding dimension $d_e$:

$$X_n(t) = \{x_{tn}, x_{tn-\tau}, \ldots, x_{tn-(d_e-1)\tau}\}$$

Then, a Euclidean metric is imposed onto the phase space to find the $k$ nearest neighbors $X_i(t)$, $i = 1, 2, \ldots, k$ of the current state. For each neighborhood point $X_i(t)$, there is a projected value in time series $\hat{x}_i(T_f)$, which is an element in the expected value vector corresponding to the forecast time horizon. These neighborhood points and corresponding projected value can be used to derive coefficient matrix $\beta_T$, in which the value of the forecast period of $T_f$ can be calculated through:

$$\hat{x}_{n+\Delta t} = \beta_T \cdot X_n(t)$$

It should be noted that this study utilizes the results from LM directly. Details of LM implementation are elaborated by Babovic et al. (2005) and Wang & Babovic (2014).

**Two-sample Kalman filter**

For most KF, the system state is estimated based on the information available in agreement with system statistical uncertainty, where the error covariance matrix and Kalman gain are updated at each time step. One of the less computationally demanding assimilation algorithms is based on the steady-state KF. Two-SKF, proposed by Sumihar et al. (2008), computes the steady Kalman gain based on two forecast realizations. It assumes that the error process of the system of interest is weakly stationary. It may also be applicable for a system where error statistics vary slowly in time (Sumihar et al. 2008). In this study, two-SKF is used to correct the water level model output at non-measured stations.

Given a system:

$$x_k = F(x_{k-1}, u_{k-1}) + q_{k-1}$$

$$y_k = Hx_k + r_k$$

where, $x_k \in R^n$ is the state vector; $y_k \in R^m$ the measurement; $F$ the dynamic model governing the process,
representing the function to compute the predicted state from the previous estimate (in this algorithm, it can be a linear (i.e., \( x_k = A x_{k-1} + B u_{k-1} + q_{k-1} \)) or nonlinear function \( x_k = f(x_{k-1}, u_{k-1}) + q_{k-1} \); \( u_{k-1} \) the forcing item; \( q_{k-1} \sim N(0, Q) \) the Gaussian process noise; \( r_k \sim N(0, R) \) the Gaussian measurement noise with covariance of \( R \); \( H \in \mathbb{R}^{m \times n} \) the linear observation function.

The main steps include an open-loop and closed-loop step, described as follows.

**Open-loop step**

In the open-loop estimation, estimation of the covariance matrix of the random process is done without making use of any observations. Two realizations of open-loop process \( x_{f(1)} \in \mathbb{R}^n \) and \( x_{f(2)} \in \mathbb{R}^n \) are generated to estimate covariance matrix \( P^f \in \mathbb{R}^{m \times n} \). These two forecast realizations can be generated by running the dynamical model twice with statistically independent perturbations.

\[
\begin{align*}
    x^f_k & = F(x^f_{k-1}, u_{k-1}) + q_{k-1} \\
    P^f & = \frac{1}{2} \frac{1}{N-1} \sum_{k=1}^{N} e_{k(1,2)} e_{k(1,2)}^T \\
    e_{k(1,2)} & = x^f_{k(1)} - x^f_{k(2)}
\end{align*}
\]

where, \( e_{k(1,2)} \) is the difference between two series \( x^f_{k(1)}, x^f_{k(2)} \), \( k = 1, \ldots, N \), two independent samples from the process in Equation (5).

Then, the Kalman gain \( K \) can be calculated by:

\[
    K = P^f H^T (H P^f H^T + R)^{-1}
\]  

**Closed-loop step**

In the closed-loop estimation, observation of the dynamic system is available and related to unknown state \( x_k \) in Equation (4). The Kalman gain \( K \) is inserted into Equation (8) to update the state at each time step, and generate two realizations of the closed-loop process \( x^f_{k+1(1)}, x^f_{k+1(2)} \) through Equation (9):

\[
\begin{align*}
    x^f_k & = x^f_k + K (y_k - H x^f_k) \\
    x^f_{k+1} & = F(x^f_k, u_k) + q_k
\end{align*}
\]

Then, \( P^f \) and \( K \) can be estimated using Equations (6) and (7), respectively. The closed-loop step is repeated until \( P^f \) and \( K \) converge. Convergence of the algorithm is indicated by insignificant difference between the gain matrices estimated by two consecutive iterations.

**Unscented Kalman filter**

The UKF has been developed to overcome the deficiencies of the linearization in the EKF (Uhlmann 1994; Julier & Uhlmann 1997; Linares-Perez & Hermoso-Carazo 2011). It provides a direct and explicit mechanism to transform mean and covariance information, and has been previously shown to be a superior tool to EKF in various aspects, especially in strongly nonlinear systems (Aguirre et al. 2005).

Given a nonlinear system:

\[
\begin{align*}
    x_k & = f(x_{k-1}, u_{k-1}) + q_{k-1} \\
    y_k & = h(x_k) + r_k
\end{align*}
\]

where, \( h \) is the observation function describing the mapping from observation to state. In this paper, the nonlinear observation functions \( h \) are estimated using genetic programming (GP). Unlike the general genetic algorithm, GP induces equations or models that describe the problems without requiring the user to know or specify the form or structure of the solution in advance (Babovic & Keijzer 2000; Poli et al. 2008; Babovic 2009). It has been used to generate mathematically meaningful models, which are composed of functions and terminals of the problem, without needing any prior knowledge of the governing mechanism and process. For a detailed description of GP, readers are referred to Babovic & Abbott (1997) and Khu et al. (2001).

The UKF makes use of the UT to reduce the potentially large number of state vectors to a small representative group
(i.e., sigma points) and gives an approximation to the filtering solutions of the nonlinear optimal filtering problems mentioned earlier. The UKF steps are described as follows.

1. Compute the set of sigma points $X$:

$$
X^{(0)}_{k-1} = x^{0}_{k-1} \\
X^{(i)}_{k-1} = x^{n}_{k-1} + [\sqrt{(n+\lambda)P_{k-1}}]_{i} \\
X^{(i+1)}_{k-1} = x^{n}_{k-1} - [\sqrt{(n+\lambda)P_{k-1}}]_{i} \quad i = 1, \ldots, n
$$

(12)

The sigma point is derived in this way to capture the mean and covariance information, at the same time permitting direct propagation of this information through an arbitrary set of nonlinear equations. The associated weights $w$ are defined as:

$$
w_{m}^{(0)} = \frac{\lambda}{n+\lambda} \\
w_{c}^{(0)} = \frac{\lambda}{n+\lambda} + (1-\alpha^{2}+\beta) \\
w_{m}^{(i)} = w_{c}^{(i)} = \frac{\lambda}{2(n+\lambda)}, \quad i = 1, \ldots, 2n \\
\lambda = \alpha^{2}(n+\kappa) - n
$$

(13)

where, $P$ is the covariance estimated; $\lambda$ a scaling parameter; $\alpha$, $\beta$ and $\kappa$ the positive constants used in the method. $\alpha$ controls the spread of the sigma points; $\beta$ is used to incorporate prior knowledge of the distribution of $x$; and $\kappa$ is a secondary scaling parameter. For a Gaussian distribution of $x$, the optimal values for $\alpha$, $\beta$ and $\kappa$ are $10^{-3}$, 2 and 0 respectively.

1. Compute predicted state $x^{p}_{k}$ and predicted covariance $P^{p}_{k}$:

$$
\begin{align*}
\hat{x}^{(i)}_{k} &= f(\hat{x}^{(i)}_{k-1}, u_{k}) \\
x^{p}_{k} &= \sum_{i=0}^{2n} w_{m}^{(i)} x^{(i)}_{k} \\
P^{p}_{k} &= \sum_{i=0}^{2n} w_{m}^{(i)} (x^{(i)}_{k} - x^{p}_{k})(x^{(i)}_{k} - x^{p}_{k})^{T}
\end{align*}
$$

(14)

2. Compute predicted $\hat{y}_{k}$ and covariance of measurement $S_{k}$, and cross-covariance of the state and measurement $C_{k}$:

$$
\left\{ \begin{array}{l}
\hat{y}^{(i)}_{k} = h(\hat{x}^{(i)}_{k}) \\
\hat{y}_{k} = \sum_{i=0}^{2n} w_{m}^{(i)} \hat{y}^{(i)}_{k}
\end{array} \right. \\
S_{k} = \sum_{i=0}^{2n} w_{m}^{(i)} (\hat{y}^{(i)}_{k} - \hat{y}_{k})(\hat{y}^{(i)}_{k} - \hat{y}_{k})^{T}
$$

(16)

$$
C_{k} = \sum_{i=0}^{2n} w_{m}^{(i)} (x^{(i)}_{k} - x^{p}_{k})(\hat{y}^{(i)}_{k} - \hat{y}_{k})^{T}
$$

(17)

3. Compute gain $K_{k}$, updated state $x^{a}_{k}$ and covariance $P_{k}$:

$$
\left\{ \begin{array}{l}
K_{k} = C_{k}S_{k}^{-1} \\
x^{a}_{k} = x^{p}_{k} + K_{k}(y_{k} - h(x^{p}_{k})) \\
P_{k} = P^{p}_{k} - K_{k}S_{k}K_{k}^{T}
\end{array} \right. \\
\left\{ \begin{array}{l}
K_{k} = C_{k}S_{k}^{-1} \\
x^{a}_{k} = x^{p}_{k} + K_{k}(y_{k} - h(x^{p}_{k})) \\
P_{k} = P^{p}_{k} - K_{k}S_{k}K_{k}^{T}
\end{array} \right.
$$

(19)

The numerical models in which the above combination of the data-driven model and the KF method are examined are presented in the next section.

One difference between the above two KFs scheme is that two-SKF estimates steady covariance based on two forecast samples, while the UKF estimates the covariance based on a minimal set of chosen sample points through unscented transformation at each time step. Furthermore, instead of linear transformation of observation, UKF allows for a nonlinear observation function, which will be implemented and compared in the next section.

### DESCRIPTION OF NUMERICAL MODEL

The numerical model used in this study is set up in the Delft3D modeling environment. The modeling system consists of several modules, including flows, waves, water quality, particle tracking, ecology, sediment transports, and morphological development. Delft3D-FLOW is the core of Delft3D, providing hydrodynamic information for the other modules. This numerical model solves the Navier–Stokes equations for an incompressible fluid,
under shallow water and Boussinesq assumptions (Stelling 1984; WL DelftHydraulics 2003a, 2003b). This paper employs the Delft3D modeling system to test the performance of the proposed data assimilation scheme based on two case studies. One is a simple hypothetical bay experiment and the other is a real case of SRM. The water level outputs from the numerical model at non-measured points are corrected based on information from monitoring points.

**Description of hypothetical bay experiment**

The hypothetical bay experiment is first simulated by the specific-driven system, and the corresponding data set is considered as the field measurement. Subsequent simulations are carried out by introducing distortions to generate model error. This data set is treated as a deterministic model prediction, which is inevitably less accurate due to incomplete and inadequate knowledge of the system.

The artificial bay experiment conducted by Mancarella et al. (2008) is of a rectangular shape with only one open boundary in the north. The bay stretches up to 200 km from north to south and 210 km from east to west. The model grid consists of $20 \times 21$ cells with a constant spacing of 10 km. The bathymetry increases from the closed boundaries on the coast to the central-northern part of the bay. Bed roughness based on the Chezy coefficient varies spatially, ranging from 30 to $45 \text{ m}^{0.5}/\text{s}^{-1}$, and corresponds to the bathymetry. The grids and bathymetry are shown in Figure 2. The flow is driven by a multi-sinusoidal water level variation at open boundaries with a maximum amplitude of 2 m and for a period between 12 and 72 hours. Wind and pressure fields of a moving cyclone generated by a wind enhanced scheme are applied on the surface of the model.

In this study, the simulation was carried out for 12 days with a time step of 15 minutes. The first 100 time steps were excluded to eliminate the influence of the initial condition. The deterministic model driven by the forcing, as described above, is treated as a source model assumed to be perfect.

![Figure 2](http://iwaponline.com/jh/article-pdf/18/5/773/390436/jh0180773.pdf)
and representing the actual phenomenon. Therefore, its output can be treated as the measurements.

Distortions are introduced to generate a set of model outputs, which are considered as uncorrected model outputs. The distortion covers boundary error, roughness error and wind error, namely, CMB here. A phase error of 1 hour is applied on the open boundary, and the spatially varied Chezy coefficient is replaced by a constant value of $32 \text{m}^{0.5}/\text{s}^{-1}$ over the entire domain. Also, the moving cyclone was replaced by a uniform wind of 20 m/s.

Figure 2 shows the seven selected stations exemplified in the experiment. Stations 5, 6, and 7 are assumed to be measured points, and the water level outputs generated by the source model are used as measurements. The other four stations (Stations 1–4) are assumed to be non-measured stations, and water levels generated by distorted forcing at these four stations will be corrected by the real data at the other three measured points.

**Description of SRM**

Following application of the proposed hybrid scheme in the hypothetical bay experiment, it is reasonable to conduct a further test in the real world situation – the SRM case study.

The SRM is developed in the Delft3D-Flow module and aims to provide hydrodynamic information, in particular sea level anomalies (residual water levels) in the Singapore Straits (Kurniawan et al. 2011).

The model is set up with three open boundaries, which are the South China Sea in the east, Andaman Sea in the west, and a small part of the Java Sea in the south. The grid and bathymetry of SRM are shown in Figure 3. The model covers a large area of water body, and the three open boundaries are set far away from Singapore Island, which reduces the error from the open boundaries. Tidal forcing is prescribed as water level variations by means of amplitudes and phases (relative to time zone GMT + 8) of eight tidal constituents: Q1, O1, P1, K1, N2, M2, S2, and K2. The bathymetry in SRM is based on Admiralty charts. The maximum depth of the model is about 2,000 m in the Andaman Sea and >160 m in the Singapore Straits.

The one year simulation was carried out using 2004 from January 1st 00:00 to December 31st 00:00 with a time step of 4 minutes and hourly output, which produces 8,761 hourly time series data for all grids in the domain. The first 10 days’ data points are discarded in order to eliminate the influence of the initial condition. Thirteen stations are considered in the present study: West Coast, Tanjong
Changi, Tanah Merah, Sembawang, and Raffles are located around Singapore Island; Langkawi, Kelang, Lumut, and Penang are located in the Malacca Straits; and Tioman, Getting, Kuantan as well as Sedili are located along the east of the Malaysia Peninsula. Their locations are shown in Figure 3. The water level measurements are available at these stations for the year 2004. Two stations are selected from each region (West Coast, Tanjong Changi; Langkawi, Kelang; Tioman and Getting) as measured stations, while the others are assumed to be non-measured ones.

The simulated water level and the actual measurements are compared to calculate the residuals. The corresponding statistical results calculated as Equations (20) and (21) (root mean square error (RMSE) and correlation coefficient \( R \)) are summarized in Table 1. It is found that the model errors unavoidably exist at all measurement locations. In addition to general limitations of the numerical model, such residuals may also be attributed to the complex hydrodynamic behavior in this area resulting from water exchange between the Indian and Pacific oceans. Therefore, it is important to apply a data assimilation method to correct SRM and improve its accuracy:

\[
RMSE = \sqrt{\frac{1}{n} \sum (x^m - \bar{x})^2} \tag{20}
\]

\[
R = \sqrt{\frac{\sum (x^m - \bar{x})(x - \bar{x})}{\sqrt{\sum (x^m - \bar{x})^2 \sum (x - \bar{x})^2}}}
\]

where, \( x^m \) is the observed water level; \( x \) the water level from the numerical model; and \( n \) the number of records.

The first half of the year 2004 is considered as the hindcast period. The corresponding measurements in this period are available, and used in the proposed data assimilation scheme to update and estimate the state variables at non-measured stations. The second half of 2004 is treated as the forecasting period in which the measurements are used for the model validation.

**APPLICATION AND RESULTS IN THE CASE OF BAY EXPERIMENT**

**Application in hypothetical bay experiment**

The water level outputs from the uncorrected model (CMB) are used as the system state in the following KF, i.e., \( x_k \in \mathbb{R}^7 \). The dynamics of the water level are represented through LM. This LM is carried out at seven sampling stations, respectively, with a 15-minute forecast horizon, which is one time step interval consistent with the numerical model output. The optimal embedding parameters are estimated through the genetic algorithm described by Babovic et al. (2005) and summarized in Table 2.

![Table 1](http://iwaponline.com/jh/article-pdf/18/5/773/390436/jh0180773.pdf)

**Table 1** | Statistics of model residue at sample stations

<table>
<thead>
<tr>
<th>Station</th>
<th>RMSE (m)</th>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Coast</td>
<td>0.2172</td>
<td>0.966</td>
</tr>
<tr>
<td>Tanjong Changi</td>
<td>0.2670</td>
<td>0.947</td>
</tr>
<tr>
<td>Tanah Merah</td>
<td>0.2122</td>
<td>0.957</td>
</tr>
<tr>
<td>Sembawang</td>
<td>0.3196</td>
<td>0.933</td>
</tr>
<tr>
<td>Raffles</td>
<td>0.2090</td>
<td>0.963</td>
</tr>
<tr>
<td>Langkawi</td>
<td>0.1305</td>
<td>0.989</td>
</tr>
<tr>
<td>Kelang</td>
<td>0.2834</td>
<td>0.983</td>
</tr>
<tr>
<td>Lumut</td>
<td>0.1465</td>
<td>0.981</td>
</tr>
<tr>
<td>Penang</td>
<td>0.1883</td>
<td>0.948</td>
</tr>
<tr>
<td>Tioman</td>
<td>0.2323</td>
<td>0.979</td>
</tr>
<tr>
<td>Getting</td>
<td>0.2887</td>
<td>0.874</td>
</tr>
<tr>
<td>Kuantan</td>
<td>0.2439</td>
<td>0.978</td>
</tr>
<tr>
<td>Sedili</td>
<td>0.2199</td>
<td>0.961</td>
</tr>
</tbody>
</table>

**Table 2** | The embedding parameter of LM for sample stations in bay experiment

<table>
<thead>
<tr>
<th>Station</th>
<th>( d^* )</th>
<th>( k )</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Such a forecasting model is then utilized in the UKF and two-SKF procedure to correct the water level at non-measured locations 1–4 based on observed vector \( y_k \). The measured water levels at the three observed stations 5–7 are selected as variables of vector \( y_k \in \mathbb{R}^3 \). The observation function is linear as indicated by \( H \).

\[
\begin{align*}
    x &= [D_1, D_2, D_3, D_4, D_5, D_6, D_7]^T \\
    y &= [M_5, M_6, M_7]^T \\
    H &= \begin{bmatrix}
        0 & 0 & 0 & 0 & 1 & 0 & 0 \\
        0 & 0 & 0 & 0 & 0 & 1 & 0 \\
        0 & 0 & 0 & 0 & 0 & 0 & 1 
    \end{bmatrix}
\end{align*}
\]

where, \( x_k \) and \( y_k \) are the system state and observation, respectively, as indicated in Equation (3) or Equation (10) and Equation (4) or Equation (11); \( D_j \) the water level output from CMB simulation at all the stations of interest; \( M_j \) the measured water level at monitoring stations 5, 6, and 7; \( H \) the observation transformation function used in both the UKF and two-SKF.

For the application of the two-SKF, one sample realization used is the uncorrected model CMB, described above, which is also tested by the UKF. Another sample realization is generated with a phase error of \( \frac{\pi}{12} \) hour at the open boundary, \( 32 \text{ m}^{0.5}/\text{s}^{-1} \) for the Chezy coefficient, and a uniform wind of \( 20 \text{ m/s} \). Other settings are the same with the UKF.

**Assessment of results in hypothetical bay experiment**

The corrected water levels are compared with the measurements in terms of RMSE, and percentage of improvement (\( \text{imp}\% \)) based on Equation 20 (\( x \) here is replaced by \( x^e \), i.e., the water level after correction) and Equation 22:

\[
\text{imp}\% = \frac{\text{RMSE}_{SRM} - \text{RMSE}}{\text{RMSE}_{SRM}} \times 100\%
\]

where, \( \text{RMSE}_{SRM} \) is the RMSE of the original numerical model.

The forecasted water levels using the LM approach in the case of CMB at points 5, 6, and 7, are shown in Figure 4. Some noise is found around time step 190, and this is probably due to LM dependence on embedding parameter optimization. For this case, the embedding parameters are estimated based on the historical residuals generated through adding turbulence in the forcing. This generated error may not fully obey the characteristics of chaotic time series. As well, searching nearest neighborhoods needs large amounts of preliminary data. It will be more accurate if a larger dataset is available. Therefore, some unstable turbulence is observed during the early period. In spite of this noise, the LM approach describes the propagation of original water level output with good accuracy, especially after 200 time steps as compared with the original numerical model output, suggesting the feasibility of using such a data-driven model in KF.

The hindcast results after the correction of UKF and two-SKF at non-measured stations are compared and summarized in Table 3, where the best estimates at each station are underlined. It can be seen that the two-SKF gives more accurate predictions of water level at points 2 and 3 comparing with the results of the UKF. This may be attributed to the fact that the process covariance which affects the filter efficiency is determined by two realizations in the two-SKF. As points 2 and 3 are located further away from the open boundary than points 1 and 4, their results computed by numerical model are less affected by the distortion at the open boundary. Therefore, the covariance calculated through realizations gives more promising results at these two points.

**APPLICATION AND RESULTS IN CASE OF SRM**

**Application of combination of data-driven model and KF in SRM**

Water level outputs from SRM at 13 sample stations are used as the system state \( x_k \in \mathbb{R}^{13} \). Similar to the previous application of the hypothetical bay experiment, the two KFs are also coupled with the trained LM. These water levels are simulated through the LM at these 13 stations individually, with a 1 hour forecast horizon, which is consistent with the numerical model output interval. The optimal embedding parameters used are shown in Table 4.
The measured water levels at six observed stations (West Coast, Tanjong Changi, Langkawi, Kelang, Tioman, and Getting) are considered as six variables of the observed vector $y_k \in \mathbb{R}^6$. The linear observation function $H$ is used in both two-SKF and one scenario of UKF. As described in Figure 1, based on the system driven by LM at each time step, the information from both observations and numerical model are combined to construct the Kalman gain according to Equations (7) and (19). Based on these Kalman gains, the scheme further corrects the model state, which will be the initial state for the next time step:

$$x = [D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}]^T$$

### Table 3
<table>
<thead>
<tr>
<th>Station</th>
<th>RMSE (m)</th>
<th>UKF RMSE (m)</th>
<th>Imp%</th>
<th>RMSE (m)</th>
<th>Imp%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4722</td>
<td>0.0657</td>
<td>86.1%</td>
<td>0.0676</td>
<td>85.7%</td>
</tr>
<tr>
<td>2</td>
<td>0.6619</td>
<td>0.0588</td>
<td>91.1%</td>
<td>0.0293</td>
<td>95.6%</td>
</tr>
<tr>
<td>3</td>
<td>0.6710</td>
<td>0.0597</td>
<td>91.1%</td>
<td>0.0290</td>
<td>95.7%</td>
</tr>
<tr>
<td>4</td>
<td>0.4558</td>
<td>0.0959</td>
<td>79.0%</td>
<td>0.1029</td>
<td>77.4%</td>
</tr>
</tbody>
</table>

Underlined figures denote the best estimates for each method.

### Table 4
<table>
<thead>
<tr>
<th>Stations</th>
<th>$d_e$</th>
<th>$k$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Coast</td>
<td>3</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>Tanjong Changi</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Langkawi</td>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Kelang</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Tioman</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Getting</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Tanah Merah</td>
<td>3</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>Sembawang</td>
<td>4</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Raffles</td>
<td>4</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Lumut</td>
<td>4</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>Penang</td>
<td>4</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Kuntan</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Sedili</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 4 | Comparison of LM produced water level (solid line) and numerical model simulated water level (dot), as well as their difference (dashed line). The full colour version of this figure is available in the online version of this paper: http://dx.doi.org/10.2166/hydro.2016.085.
\[ y = [M_1, M_2, M_3, M_4, M_5, M_6]^T \]

\[ H_{6 	imes 13} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \]

where, \( x, y \) is the system state \( x_k \) and observation \( y_k \) as indicated in Equations (10) and (11); \( D_i \) the water level output from SRM simulation at all the stations of interest; \( M_j \) the measured water level at monitoring stations; \( H \) the linear observation transformation function.

One advantage of the UKF is that the observation transformation function \( h(x) \) can be more complex than the simplified linear functions. In addition to the linear observation function, which is the same as that of the two-SKF, another scenario based on the UKF approach is also tested. In order to further explore the correlation between non-measured and measured stations, model residuals at the six measured stations are added as the other six variables of \( y_k \), i.e. \( y_k \in \mathbb{R}^{12} \). The \( h(x) \) is a nonlinear observation function estimated based on the GP approach. This scheme is termed as UKF-GP. The innovation vector \( y_k - h(x_k^*) \in \mathbb{R}^{12} \) can be spread over the entire state space through Kalman gain and thereby improve the correction efficiency:

\[ x = [D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}]^T \]

\[ y = [M_1, M_2, M_3, M_4, M_5, M_6, e_1, e_2, e_3, e_4, e_5, e_6]^T \]

\[ h_1(x) = x(1); h_2(x) = x(2); h_3(x) = x(3); \]

\[ h_4(x) = x(4); h_5(x) = x(5); h_6(x) = x(6); \]

\[ h_7(x) = -0.093182 \times x(3) + 0.673811 \times x(5) + 0.193984 \times x(7) - 0.364351 \times x(8) - 0.160330 \times x(10) - 0.546005 \times x(12) - 0.146429; \]

\[ h_8(x) = 0.108322 \times x(3) - 0.021807 \times x(4) + 0.738639 \times x(5) - 0.738639 \times x(6) + 0.073050 \times x(10) + 0.234522 \times x(12) - 0.708541 \times x(13) + 0.173657; \]

\[ h_{10}(x) = 0.124845 \times x(1) x(3) - 0.124845 \times x(4) x(10) - 0.324859 \times x(4) + 0.337520 \times x(10) - 0.337520 \times x(11) + 0.117244; \]

\[ h_{11}(x) = -0.092249 \times x(3) - 0.277404 \times x(5) + 0.138702 \times x(6) + 0.184625 \times x(7) - 0.099214 \times x(11) - 0.175513 \]

\[ h_{12}(x) = -0.642559 \times x(5) + 0.042968 \times x(9) - 0.203183 \times x(11) + 0.529664 \times x(13) - 0.188984; \]

where, \( x(i) \) is the \( i^{th} \) element of vector \( x \); \( e_j = M_j - D_j \) the model residual; \( h(x) \) the nonlinear observation transformation function.

The water levels simulated in the hindcast period were tested against the actual measurements. Also, the performance of the proposed scheme is then assessed based on LM forecasting results at eight prediction horizons from 1 to 72 hours in the forecast period.

**Assessment of results in Singapore Regional Model**

As an example, the forecasted water levels through the LM approach in the case of SRM at the Tanjong Changi, Langkawi and Tioman stations are selected and shown in **Figure 5**. It can be seen that the LM method offers promising forecasting results and can reproduce the variation of the water level state.

The distribution efficiencies of the UKF and two-SKF in terms of RMSE for all forecast horizons at non-measured stations are summarized in **Figure 6**. In consideration of the temporal trend at the same stations, there is a general tendency that the correction efficiencies of these two methods are degraded with increased lead time \( T \). It suggests that the data-driven forecasting process seems to have an equivalent
effect on the results of these two methods. The difference between their efficacies is mainly determined by the estimated Kalman gain. It should also be noted that the RMSE increases considerably when the lead time $T$ varies from 1 to 4 hours, whereas there is no discernible increase (with a slight decrease observed in some situations) when the lead time $T$ increases to 12 hours and then 24 hours. Afterwards, it increases every 24 hours. This may be attributed to the period of some main tidal constituents, for example $K_1$ which is 23:53 hours, $S_2$ 12:00 hours, and $M_2$ 12:25 hours.

Considering the different stations, the SRM is found to have larger errors at Penang compared to Lumut in the Malacca Straits. This may be due to a sudden change of bathymetry from 1,600 to 400 m (see Figure 3) near the open boundary. The two-SKF, however, seems to lead to a similar RMSE for both locations after correction. This suggests that the errors caused by the open boundary could be corrected by the proposed scheme. Considering all corrected stations in the entire domain, it is interesting to note that the RMSE is reduced by the two-SKF to 15 cm in the Malacca Straits (i.e., Lumut and Penang). For the remaining area (east of the Malaysia Peninsula and Singapore Island), the correction results consistently remain at 10 cm. The UKF performs similarly but with a lower error (10 cm in the Malacca Straits and 5 cm in the remaining area). Considering the original SRM results, the SRM prediction error in this area of the Malacca Straits is found to be relatively lower than those in other areas. The phenomenon in the Malacca Straits is distinctly different from other areas which are strongly influenced by the monsoon winds over the South China Sea. It creates differences in water levels that drive residual flows through...
Singapore regional water which are not reproduced in the Malacca Straits (Kurniawan et al. 2013). Therefore, the SRM, which does not consider wind effect, leads to lower error in this area. These findings may indicate that the two-SKF could be more effective to correct the model residual caused by insufficient consideration of the forcing effect, but may not help to remove the errors from other sources. In general, the UKF yields more accurate results than the two-SKF at all the non-measured stations. It may be because the two-SKF estimates the Kalman gain based on two realizations, where only the tidal error is considered and may not reflect the realistic error. The Kalman gain in this algorithm may have difficulty capturing the realistic correlation among different stations. In contrast, in the case of the UKF, the dynamic feature of the model is less important and the filter stability is determined based on correctly mapping the sigma points. However, the advantage of the UKF is not obvious in the case of the hypothetical bay experiment, and the two-SKF corrects the numerical model with the higher improvement instead. This may be due to a faster dynamic process in SRM compared with the bay experiment. These results imply that the two-SKF may be more suitable for slow dynamic systems, while the UKF can be well adapted to nonlinear systems especially for faster dynamic systems.

The above results using UKF are further illustrated in Figure 7, in which the time series of corrected water levels at seven non-measured stations when the prediction horizon is 1 hour are compared with the actual measurements and the water levels before correction. It can be seen that the proposed method can effectively correct the water level simulated by SRM and capture its tendencies of rising and falling.
The efficacy for large prediction horizons after correction can be found in Figure 8. Figure 8(a) presents the scatter plot of the water level simulated from the numerical model against the measured data at station Tanah Merah. For comparison, the corrected results obtained by the two methods for lead times of 1 hour and 72 hours are also plotted against measurements in Figures 8(b)–8(e). It is found that the UKF corrected water level is much closer...
to the observation than that of the SKF even for 72 hours forecasting. Therefore, the UKF can effectively distribute a full range of time series water level forecasting.

Figure 9 shows a comparison of the percentage improvement obtained by UKF and UKF-GP for $T$ is 1 and 72 hours. It is noted that, in the case of the short forecast horizon, the two methods show equally competent performance in correcting the numerical model at all the non-measured stations. For the long forecasting horizon, however, the combination of the UKF and GP has
significantly improved accuracy. It may be because the additional consideration of the residual as the observed vector can provide more information, and the innovation vector in this approach can capture as much available information between different stations as possible. Thus, the correction efficiency can be improved, with filter stability eventually enhanced for long lead time.

In terms of the computational cost, the time consumed for a one-year period is shown in Table 5. Similar to the case study of the hypothetical bay experiment, the UKF approach is computed based on the sigma points. Although the Kalman gain is updated at each time step, such implementation is rapid because it is not necessary to evaluate the Jacobians required for an EKF. However, the implementation of GP would demand extra computational cost. Therefore, in the case of short time correction, the UKF can provide accurate results with less time, while the UKF-GP is more useful for the long time correction. In the two-SKF approach, the constant Kalman gain is based only on two forecast realizations and can be computed offline until it reaches a constant solution. Such steady gain application can greatly reduce the computational demands, resulting in lower computational cost compared to the UKF. In this case, since the state vector consists of small variables, their computational differences can be negligible. However, the computational load will be substantial when applied to a large-scale system where there exist thousands or even tens

<table>
<thead>
<tr>
<th></th>
<th>Two-SKF</th>
<th>UKF</th>
<th>UKF-GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (mins)</td>
<td>33</td>
<td>36</td>
<td>39 (UKF) + 20 (GP)</td>
</tr>
</tbody>
</table>

Figure 9 | Percentage of improvement after correction based on the UKF and UKF-GP at non-measured stations when the forecasting horizon is (a) 1 hour and (b) 72 hours.
of thousands of states. Therefore, in terms of the computational time, using the two-SKF is more efficient for large-scale problems.

CONCLUSIONS

The inherent limitations of numerical modeling necessitate the data assimilation to further correct the numerical model output. In this study, the two-SKF and UKF are combined with a data-driven model (LM) to map the spatial distribution and followed by correcting the water level output from the numerical model at non-measured points. Our results from the hypothetical bay experiment indicate that both of the proposed hybrid KF schemes can improve the original numerical model effectively. It has been demonstrated that the UKF is more flexible for correction at locations with a complex hydrodynamic situation. The application in the SRM case study further confirms its efficacy. The performance of the two-SKF depends on the selection of the two initial realizations from the numerical model simulation, hence the correlation between different locations may be affected by errors generated by the numerical model. However, the two-SKF can greatly reduce the computational demands due to the derived steady Kalman gain. On the other hand, the UKF, which calculates the Kalman gain based on sigma points, is proven to be less influenced by the source of model errors, implying its applicability in nonlinear systems. In addition, given that the UKF can utilize the nonlinear observation function, combining the UKF and GP enables the model state at non-measured points to be effectively associated with all available variables at measured points. The innovation vector in this approach can capture available information between different stations as much as possible, thus improving the correction efficiency when applied to a long forecast horizon. The proposed hybrid KF method can be successfully implemented at stations of interest without running a numerical model online for the entire domain. Therefore, this scheme is flexible and suitable for complex nonlinear systems. Furthermore, this scheme is not limited to water level, as it can also be extended to correct any numerical model outputs (e.g., currents) for both linear and nonlinear systems as long as observations are available at nearby related points.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support and contributions of the Singapore-Delft Water Alliance (SDWA). The research presented in this work was carried out as part of the SDWA’s ‘Must-Have Box’ (MHBox) research program (R-264-001-003-272). We also thank our colleagues working for the MHBox project. Finally, we thank MPA and UHSLC for providing the local maritime data for analysis. More information can be obtained through http://www.sdwa.nus.edu.sg/.

REFERENCES


Hamill, T. M. 2006 Ensemble-based atmospheric data assimilation. In: Predictability of Weather and Climate


First received 27 April 2015; accepted in revised form 1 February 2016. Available online 2 March 2016