INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2008; **56**:587–605
Published online 27 June 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/fld.1533

Data assimilation of forecasted errors in hydrodynamic models using inter-model correlations

D. Mancarella[1,*,†], V. Babovic[3], M. Keijzer[2] and V. Simeone[1]

[1]*Department of Environmental Engineering and Sustainable Development, Politecnico di Bari, Faculty of Engineering of Taranto, Viale del Turismo 8, Taranto 74100, Italy*
[2]*Strategic Research and Development Department, WL|Delft Hydraulics, Rotterdamseweg 185, P.O. Box 177, 2600 MH, Delft, The Netherlands*
[3]*Department of Civil Engineering, National University of Singapore 1, Engineering Drive 2, Singapore 117576, Singapore*

SUMMARY

Data-assimilation techniques of the Kalman filter type are considered to be the state-of-the-art approach for combining data information and deterministic numerical models with the objective of operational forecasting. This paper introduces, as an alternative, a faster and simpler data-assimilation technique that exploits inter-model correlations to distribute predicted errors. This scheme is performed in two steps: (i) prediction of the deterministic model errors at observation points using so-called local linear models and (ii) distribution of the forecasted errors over the computational domain employing a scheme based on deterministic inter-model correlations which describe the spatial nature of error structure. The method's advantage is that systematic error can be predicted by the error correction scheme, while the dynamics remain described by the deterministic model, which also establishes a basis for the spatial error distribution scheme. This relatively simple approach is inspired by original Kalman filter techniques but distinguishes error prediction and distribution in two different stages, hence allowing for data-driven error forecasting and off-line correction. In order to test the scheme's performance, a deterministic model of an artificial bay was constructed and run. The system was driven by specific forcing conditions and characterized by physical parameters that, in subsequent simulations, were deliberately manipulated to introduce errors into the model and test the scheme's capability. Copyright © 2007 John Wiley & Sons, Ltd.

Received 2 August 2006; Revised 29 March 2007; Accepted 20 April 2007

KEY WORDS: data assimilation; error forecast; local linear models; deterministic model; evolutionary embedding

*Correspondence to: D. Mancarella, Department of Environmental Engineering and Sustainable Development, Politecnico di Bari, Faculty of Engineering of Taranto, Viale del Turismo 8, Taranto 74100, Italy.
†E-mail: d.mancarella@poliba.it

1. INTRODUCTION

Several techniques of the Kalman filter type [1–3] have been applied where data assimilation into numerical deterministic models for operational purposes is required. Several examples pertaining to oceanographic and meteorological applications [4, 5], as well as hydrological modelling [6, 7], can be found. These traditional data-assimilation techniques offer an improved estimate of system state up to the present time based on available measurements. The procedure actually refines initial conditions by correcting model state through its error covariance. A forecast can then be obtained by simply running into the future the uncorrected deterministic model with its initial conditions updated by the data-assimilation technique.

Although these techniques demand significant computational resources, their use is rapidly becoming widespread due to the growing availability of real-time data and satellite imagery. Often, available data are in form of time series recorded at a few locations (observation points). Recently, methods different from Kalman filter schemes have been successfully applied to data assimilation into numerical models [8, 9]. In these approaches, error correction at observation points can be undertaken by predicting errors using records of past residuals between model results and actual measurements. Error forecasting can be carried out using, for example, local linear models (LLMs) in real-time applications [10, 11]. However, as it is not always sufficient to correct numerical models only at the points where observations are available, distribution of the predicted error from observation points over the remainder of the computational domain is often necessary. In this paper, a fast and simple data-assimilation scheme is proposed which uses the inter-model correlations to distribute predicted errors. This data-assimilation scheme is performed in two steps:

(i) predicting the deterministic model errors on observation points using LLMs trained by means of evolutionary embedding techniques;
(ii) distributing the forecasted errors to other grid locations *via* a straightforward plan that uses a linear model of the deterministic inter-model correlations to describe the error structure.

A similar approach was followed in [8] where weighted local spatial regression and weighted local ensemble were used to distribute errors.

The theoretical advantage of the proposed method is that systematic error can be predicted by the correction scheme while system dynamics remains modelled by deterministic model, furnishing data suitable for building an error distribution scheme based on inter-model correlations. This scheme draws inspiration from Kalman filter techniques but performs error prediction and distribution in two different stages, thus permitting integration of data-driven error forecasting and subsequent off-line model correction.

In order to test the performance of the proposed method, a set of experiments using a deterministic model of an artificial bay was carried out using Delft3D [12], a numerical hydrodynamic simulation program developed at Delft Hydraulics. The simulated system characterized by physical parameters (roughness distribution) was, in the first instance, driven by specific forcing conditions (tide and wind). Following an initial 'base' simulation, errors were deliberately introduced to the original model. These experiments yielded two different data sets: the first, being produced by the correct simulation, was regarded as field observations, while the second, obtained by introducing distortions, played the role of erroneous simulations which required correction by the data-assimilation technique. The two sets provide an environment in which one can test the performance of the proposed data-assimilation technique.

DOI: 10.1002/fld

2. ERROR CORRECTION USING LLMs

In the study of dynamical systems, a powerful abstraction is to represent the configuration of the observed process as a point (a *state*) in an appropriate space. The *dimension* of this state space is the number of coordinates required to uniquely identify the system's configuration at each instant. In such a space, the evolution of the process becomes the motion from one state to another, along an *orbit* or *trajectory*. Observations of a physical system may result in a time series $x(t)$, whose past values are sampled measurements at intervals τ_s and initiated at t_0:

$$x(t) = x(t_0 + n\tau_s) \tag{1}$$

Such a time series $x(t)$ can be seen as a projection of the system state onto the axis of observed variables from a higher dimensional space, known as the *phase space*, where system dynamics takes place. In this space, often called an *embedded space*, the actual structure of the dynamic origin of the signal $x(t)$ is represented by state vectors $X(t)$. Given a collection of sampled observations from the system, it is necessary to translate the analysis into the phase space in order to capture its dynamic evolution. The resulting problem is referred to as *phase-space reconstruction* and can be solved by means of the so-called *time-delay embedding theorem* [13, 14]. The consequence of the theorem is that delay vectors of a sufficient length d and appropriate time lag τ can recover the full geometric structure of the underlying non-linear system and represent its state $X(t)$:

$$X(t) \equiv \{x_t, x_{t-\tau}, x_{t-2\tau}, x_{t-3\tau}, \ldots, x_{t-(d-1)\tau}\} \tag{2}$$

Time series can then be forecasted on the basis of the structure in the embedded space. This process is carried out by means of LLMs and requires three main operations: (i) identification of the appropriate embedding space and the placement of time series data by means of their state vectors in this space; (ii) selection of an appropriate number of the closest neighbouring points to a forecast point in this representation and (iii) execution of a low-order regression over these neighbourhood coordinates to obtain the forecast. LLMs employ a linear approximation for each separate prediction, but this is done locally in the phase space. Consequently, the resulting overall model can be highly nonlinear since each of these linear approximations is made within separate neighbourhoods [15].

This forecasting procedure has been successfully applied directly to measured time series [11] as well as to time series of deterministic model errors [10] computed as residuals between simulation outputs and measurements at different sampling times. The second approach is desirable when an important contribution to describe the system dynamics can come from physically based numerical models.

It is important to note that, in order to produce a good forecast, the embedding space must be identified through an appropriate choice of time delay τ and phase-space dimension d (i.e. the so-called embedding parameters). Some systems can be more generally described by state vectors characterized by variable rather than constant time delays since average mutual information and false nearest neighbour analyses [16] have been shown to be generally sub-optimal selections [17]. Identifying the appropriate phase space becomes therefore a global optimization problem that is solved here using genetic algorithms [18], in an overall process referred to as evolutionary embedding [19]. From an error-forecasting viewpoint, the problem consists in finding the optimal representation E_t of the deterministic model residual $\{\varepsilon_t\}$ in the embedding space:

$$E = (\varepsilon_t, \varepsilon_{t-\tau 1}, \varepsilon_{t-\tau 2}, \ldots, \varepsilon_{t-\tau d}) \tag{3}$$

DOI: 10.1002/fld

Therefore, the embedding parameters $\tau_1, \tau_2, \ldots, \tau_d$, (i.e. the variable time delays in the lag vectors), lag vector length and the number k of nearest neighbours upon which linear regression is performed are the decision variables to be optimized. Once the error has been forecasted at an observation point, it can be superimposed on the simulation output in order to improve the accuracy of the deterministic model state.

3. ERROR DISTRIBUTION

Having accurate predictions at the observation stations is important when the stations themselves are the only points of interest. Usually, however, interest is focused on points at other locations or even on the entire grid. Since it is not always possible to collect measurements at a point of interest, distributing the predicted error from available observation stations to other points in the grid is required. Here a simple distribution scheme is proposed that uses the inter-model correlations to distribute the predicted error at measurement locations. To distribute the error that is predicted at local points, a distribution scheme must be defined that is capable of either:

(i) predicting values at the points of interest;
(ii) predicting model error at these points.

Prediction in the sense of (i), if possible, dispenses with the deterministic model altogether, as none of the known physical interactions are used. This is often undesirable since physical insight is wasted. Conversely, prediction in the sense of (ii) offers some advantages. By predicting model error at observation stations and trying to apply this information to the other points using an appropriate distribution, systematic error can be modelled by the error correction scheme and the physical dynamics can still be described by the deterministic model.

In order to distribute the error, a very simple scheme was investigated in which the inter-model correlations are used to describe the error structure. A linear model, based *only* on deterministic model results, was created and subsequently used to distribute the predicted residuals. The latter can be added to the simulation output in order to provide a better estimate of the system state. Given certain observation points, for each grid point a vector is created that models the linear (first-order) interaction between the observation point and the grid point of interest. The weight vector W_{MU} is found using least-squares optimization in such a way that

$$M^{\mathrm{d}} W_{\mathrm{MU}} = U^{\mathrm{d}} \tag{4}$$

where M^{d} is a data matrix of the deterministic model output time series at locations where measurements are available and U^{d} represents the deterministic model output at grid points where measures are unavailable. For brevity, the points without measurements are referred to here as 'unobserved'. The linear model W_{MU} is created to describe relationships between grid point outputs in the deterministic model. Thus, given the errors forecasted at measurement points, $\hat{\varepsilon}_M$, the corrected output in a set of unobserved grid points in the deterministic model is given by

$$\hat{U} = U^{\mathrm{d}} + \hat{\varepsilon}_M * W_{\mathrm{MU}} \tag{5}$$

As the linear model is a first-order approximation of the model itself, the distribution scheme works under the assumption that when a model produces errors with respect to observations in measurement grid locations, these errors will be distributed similar to the distribution of model

DOI: 10.1002/fld

values. Especially, in order to resolve bias error in the model, such a scheme is likely to be effective. In such case, systematic under- or over-predictions by the model are corrected locally and the distribution scheme will distribute the error to regions where equivalent over- and under-predictions are present.

The method thus critically relies on correlations within the model itself. If the results at the point(s) of interest are uncorrelated with the measurement points, the distribution scheme is not expected to perform well. Indeed, the underlying assumption is that the higher the correlation among different grid locations in the deterministic model output, the higher the correlation among errors as well. However, this situation can be assessed by examining the deterministic model and does not require observations.

The distribution scheme entails several desirable properties. At the forefront, it is simple, fast, and the weight vectors can be found using only model results; that is, no observations are necessary to create the scheme. As the expected error of this distribution scheme is related to the correlations between observed and unobserved points in the model, it gives ample opportunity to use the distribution scheme to determine good, or even optimal, observation points in order to guide planning of the measurement campaign. Suitable locations for deploying measurement instruments will be the highly correlated observation points.

Moreover, this scheme permits generalization over the time frames used in the prediction. Given a realistic model set-up, it is likely that the instantaneous correlations between observed and unobserved points are not the most important correlations in the model. Depending on the model's overall hydrodynamics, it can very well be noted that previous output values at the observed points are better correlated with the points of interest. This information can easily be incorporated in the distribution scheme proposed here. Correlation in time among water level variations at different locations in the deterministic model can be explored and exploited in this scheme by simply shifting the time series that constitute U^{d} and M^{d} at different lags. This scheme could also be performed in an ensemble fashion. Another key advantage is the possibility to use a large amount of data to build the error distribution model since only simulation data are needed.

4. THE HYDRODYNAMIC MODEL

The hydrodynamic model used in the present work is Delft3D-FLOW, distributed by WL|Delft Hydraulics, Delft, The Netherlands. In this paragraph we report only a brief description of the model, since this is not the main focus of this paper.

Delft3D-FLOW is a numerical model based on a finite difference discretization of the three-dimensional Navier–Stokes equations averaged over turbulence time-scales (Reynolds-averaged Navier–Stokes equations). For further details, the reader is referred to [12, 20, 21]. In the present case study (see Section 5.1), the depth is much smaller than the horizontal length scales of flow and bathymetry and the vertical accelerations are assumed small in comparison to the horizontal scales (hydrostatic flow). Under these conditions the shallow water assumption is valid and the general governing volume, mass and momentum conservation equations for this free surface flow can be written as follows:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{\omega}{d+\zeta}\frac{\partial u}{\partial \sigma} - fv = -\frac{1}{\rho}P_x + F_x + \frac{1}{(d+\zeta)^2}\frac{\partial}{\partial \sigma}\left(v_V\frac{\partial u}{\partial \sigma}\right) \tag{6}$$

DOI: 10.1002/fld

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+\frac{\omega}{d+\zeta}\frac{\partial v}{\partial \sigma}+fu=-\frac{1}{\rho}P_x+F_y+\frac{1}{(d+\zeta)^2}\frac{\partial}{\partial \sigma}\left(\nu_V\frac{\partial v}{\partial \sigma}\right) \tag{7}$$

$$\frac{\partial \omega}{\partial \sigma}=-\frac{\partial \zeta}{\partial t}-\frac{\partial[(d+\zeta)u]}{\partial x}-\frac{\partial[(d+\zeta)v]}{\partial y}+H(q_{\text{in}}-q_{\text{out}})+P-E \tag{8}$$

$$\frac{\partial \zeta}{\partial t}+\frac{\partial[(d+\zeta)U]}{\partial x}+\frac{\partial[(d+\zeta)V]}{\partial y}=Q \tag{9}$$

where d (m) is the water depth below the horizontal plane of reference (datum); F_x, F_y (m/s^2) the radiation stress gradient in the x- and y-direction; f (1/s) the Coriolis coefficient (inertial frequency); H (m) the total water depth; P, E (m/s) the precipitation and evaporation; P_x, P_y (kg/m^2 s^2) the gradient hydrostatic pressure in x- and y-direction; q_{in}, q_{out} (1/s) the local source and local sink per unit volume; Q (m/s) the global source or sink per unit area including the contributions of discharge or withdrawal, evaporation and precipitation; t (s) the time; u, v (m/s) the flow velocity in the x- and y-direction; U, V (m/s) the depth-averaged velocity in the x- and y-direction; x, y, z (m) the Cartesian coordinates; σ the scaled vertical coordinate $\sigma=(z-\zeta)/(d+\zeta)$; ζ (m) the water level above some horizontal plane of reference (datum); ω (m/s) the velocity in the σ-direction in the σ-coordinate system.

Here the equations are expressed in Cartesian rectangular coordinates in the horizontal direction and, as introduced by Philips [22], σ-coordinates in the vertical. In hydrostatic flow, the vertical momentum equation reduces to the hydrostatic pressure relation.

The vertical velocities ω in the σ-coordinate system can be computed from the continuity equation by integrating vertically from the bottom ($\sigma=-1$) to a level σ ($-1\leqslant\sigma\leqslant 0$).

The grid used to discretize the equations is a three-dimensional orthogonal staggered grid with a particular arrangement of variables known as Arakawa C-grid. The numerical solution is achieved through an alternating direction implicit time-stepping scheme incorporating a combination of second-order central and third-order upwind spatial discretization. For a detailed description of the method and its implementation within Delft3D the reader is referred to [12, 20].

5. DESCRIPTION OF THE HYPOTHETICAL BAY EXPERIMENT

In order to test the performance of the proposed data-assimilation scheme, a deterministic model of an artificial bay was first established using Delft3D. The simulated system was driven by specific forcing conditions and characterized by physical parameters that were, in a later simulation, altered to induce model error. This experiment returned two different data sets: the first, being produced by the correct simulation, was treated as field observations, while the second, obtained by introducing distortions, played the role of actual deterministic model predictions, necessarily erroneous due to incomplete knowledge of the system.

The alterations introduced to the uncorrected models affected the type and intensity of wind forcing, the phase of the boundary condition and the spatial distribution of bed roughness coefficients. Similar experiments have been conducted in [5, 8, 23]. In this work, the efficiency of the scheme has been tested for each type of error introduced and for their overall combination.

DOI: 10.1002/fld

5.1. The source model

The artificial bay has a rectangular shape with only one open boundary to the north and its bathymetry is characterized by increasing depths from the closed boundaries on the coast to the central-northern part of the bay. The rectangular grid used for the simulation was comprised of 21×20 square cells having a constant spacing of 10 km. The bathymetry is shown in Figure 1.

Also, bed resistance is spatially variable in the model domain with the value of the Chezy coefficient ranging between 30 and 45 $m^{0.5}/s^{-1}$; the largest values associated with the deepest areas. Figure 2 portrays the distribution of Chezy roughness coefficient values in the model domain.

The open boundary condition for this model is represented by a multi-sinusoidal water level variation whose two components are characterized by a period of 12 h (representing tides) and 72 h (representing a varying tidal cycle). The forcing wave at the open boundary has maximum amplitude of 2 m.

Wind forcing was introduced into the simulation using pressure fields from a moving cyclone, artificially generated employing the WES model [24–26]. The hurricane is characterized by a radius of 600 km, maximum wind speed of 75 knots, and its eye moves from west to east at a speed of 7.4 km/h along the central line of the bay on a wider grid than that for simulation (800 km), hence forcing the model for 4 days.

The entire simulation covers a period of 12 days and uses a time step of 15 min, providing output time series of 1152 data points in each location considered. From these locations, the first 100 time steps were excluded from the used data sets and discarded in order to ensure that the effects of initial conditions were negligible. The time steps from 101 up to 800 were used for training the LLMs, and the remaining were applied to testing the capability of the proposed data-assimilation scheme. This deterministic model, as previously described, driven by these forcing conditions and characterized by these parameters, was assumed to be perfectly representative of the actual phenomenon and hence its output was treated as measurements. Time series of water levels at seven grid points were stored, three of them to be used as 'true' measurement stations

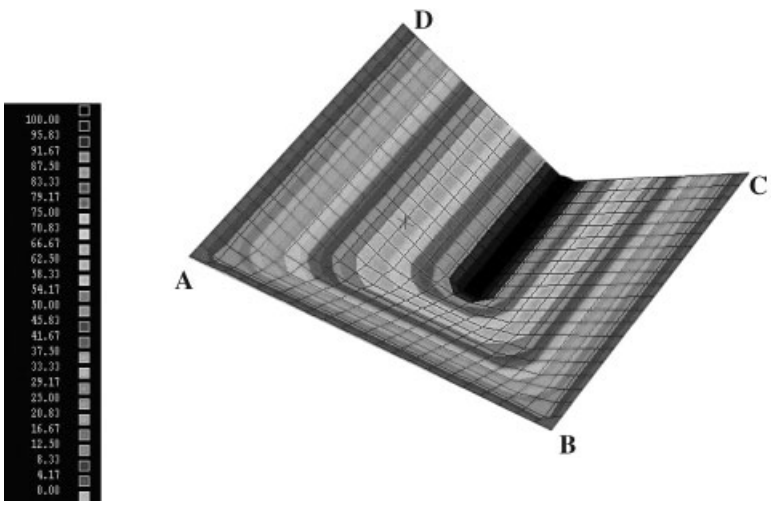

Figure 1. Perspective view of the artificial bay domain and simulation grid. The bay is characterized by a space varying bathymetry, deeper in the north-central area.

DOI: 10.1002/fld

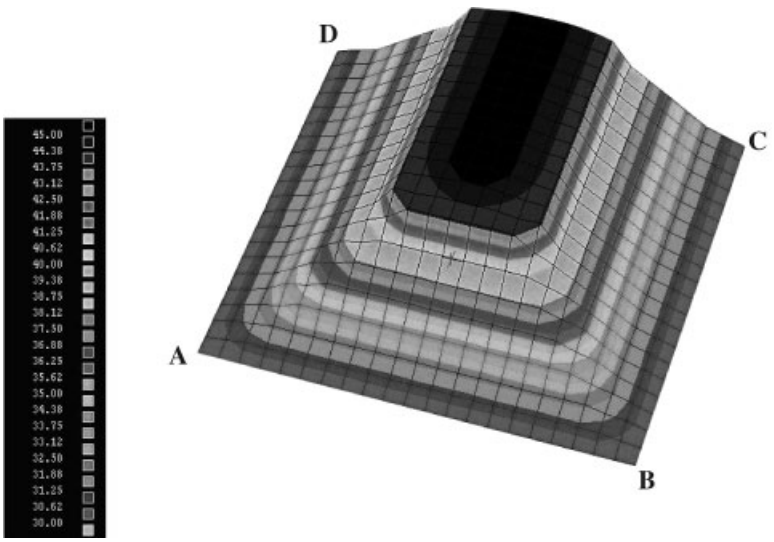

Figure 2. Spatial distribution of bed resistance coefficient (Chezy) in the artificial baysource model.

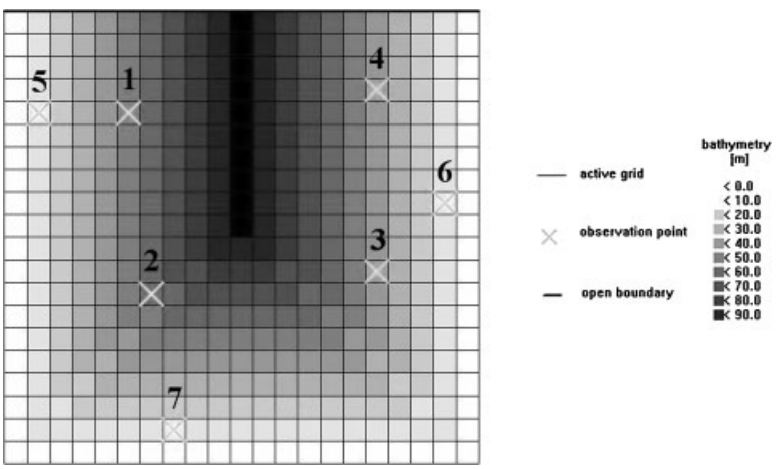

Figure 3. Top view of the model domain and location of grid points considered in the present work.

while the remaining four were used to test the efficiency of the error prediction and distribution scheme (only after the LLMs were trained and the error distribution had been determined). The locations of these grid points are depicted in Figure 3.

5.2. The uncorrected deterministic models

Starting from the original 'source model', four different simulations were set up by introducing distortions, initially applied separately and then subsequently combined. At first, a phase error of 1 h was introduced to the model's boundary condition, while the other physical parameters and

DOI: 10.1002/fld

Table I. Overview of artificial bay experiments and corresponding errors introduced.

Experiment	Abbreviation	Distortion introduced	Max RMSE (m)	Mean RMSE (m)
Boundary error	BND	One hour phase error in the boundary condition	0.870	0.639
Roughness error	RGH	Change the values of Chezy coefficient	0.334	0.222
Wind error	WND	Cyclone wind field replaced by constant wind	0.202	0.124
Combined error	CMB	Combination of all errors	0.704	0.568

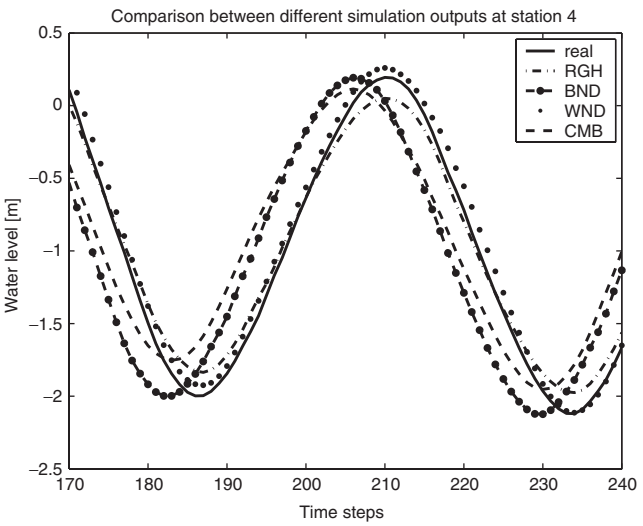

Figure 4. Effect of distortions introduced in the model. This figure reports an example of different simulation outputs in station 4 concerning the source model and the four different uncorrected models.

driving forces were kept unmodified (BND experiment). The time series of deterministic model errors were obtained from the comparison between this simulation output and the source model as reported:

$$\varepsilon = \text{source model output} - \text{deterministic model output} \tag{10}$$

The second experiment (RGH experiment) consisted of introducing only an alteration in the Chezy resistance coefficients, whose space varying distribution was replaced by a global constant of $32\,\mathrm{m}^{0.5}/\mathrm{s}^{-1}$ for the entire model domain.

The last separately introduced error was a distortion in the wind field: the moving cyclone was replaced by a spatially uniform, and temporally constant, wind blowing towards east at a speed of 20 m/s (WND experiment). In the final scenario, all the distortions were applied simultaneously since the model was conceived to be driven by a combination of the previously described erroneous parameters and forcing conditions (CMB experiment).

The whole set of experiments was carried out to provide synthetic but reasonable data series of errors of the types commonly present in hydrodynamic models, with the aim of testing the efficiency

DOI: 10.1002/fld

of the scheme in predicting, distributing and correcting the error of deterministic numerical models. In Table I, an overview of the error introduced in each experiment is provided. For brevity, the distorted models and their associated experiments are denoted by the abbreviations reported in this chart. Figure 4 shows an example of the effect due to the distortions introduced into the simulation.

In the remaining article, the time series produced by the source model are referred to as the *observations* or *real water levels* and the series referred to as *deterministic model* or *simulation* are those generated by the distorted models. Finally, the model obtained by applying the error correction scheme to the simulations is termed the *corrected model*.

6. EXPERIMENTAL RESULTS

6.1. Introduction

As shown in Figure 3, seven stations were considered during the present work. In this case study, stations 5, 6 and 7, served as measurement stations and, thus, past errors were assumed to be perfectly known at these locations. Errors were forecasted at these locations using LLMs and then spatially distributed to the non-measurement stations, represented by grid points numbered from 1 to 4. At the non-measurement locations, the time series obtained from the source model were used as real water levels and the corrected model performance was tested against them. Therefore, the simulation was corrected in stations 5, 6 and 7 by directly adding the forecasted error to the deterministic model outputs, while in grid points 1–4 it was rectified by spatially distributing the errors previously forecasted in measurement stations.

6.2. Global performance of the data-assimilation scheme

Within each experiment, the proposed data-assimilation scheme was applied to correct the distorted deterministic model outputs in stations 1, 2, 3 and 4, on the basis of error records available in stations 5, 6 and 7. The corrective effect of the applied scheme on the deterministic model is then tested against 'unseen' error records also available for stations 1, 2, 3, but previously unused.

A summary of results is presented in Tables II and III. The first chart reports values of mean and maximum RMSE between corrected and source model outputs after spatial distribution of forecasted errors in each of the four different experiments. Table III expresses the same statistics but as percentages. It is evident from Table III that the decrease in error is usually about one order of magnitude, with lower reduction rates for the WND experiment due to the drastic distortion introduced in this case. An example of error correction in a non-measurement station is given in Figure 5 where the CMB model output at grid point 2 was corrected using the proposed error correction scheme. For a forecast horizon of 12 h, the phase error was greatly removed and only small deviations from the observed water levels can be noticed in correspondence to the peak values. For reasons of brevity, only figures associated with the most relevant cases are presented here, such as those pertaining to the CMB experiment, which deals with the most common situation in practice, or the WND model, whose error appears more difficult to be resolved. Other results are presented only in the tables.

In Table IV, the value of mean square correlation coefficients between the deterministic model error at non-measurement grid points and the error predicted at the same locations using the proposed data-assimilation scheme are reported for various forecast horizons. The table illustrates

DOI: 10.1002/fld

Table II. Mean and max RMSE of corrected model output in non-measurement locations.

Hours	Time steps	BND	RGH	WND	CMB
Max RMSE (m) for different forecast horizons					
1	4	0.068	0.050	0.032	0.065
12	48	0.084	0.051	0.037	0.074
18	72	0.095	0.051	0.036	0.087
24	96	0.072	0.049	0.037	0.078
Mean RMSE (m) for different forecast horizons					
1	4	0.050	0.032	0.021	0.052
12	48	0.060	0.034	0.022	0.062
18	72	0.066	0.035	0.023	0.072
24	96	0.052	0.033	0.022	0.063

For each performed experiment, these values are averaged over the four non-measurement stations.

Table III. Percent reduction in mean and max RMSE of corrected model output in non-measurement locations.

Hours	Time steps	BND	RGH	WND	CMB
Percent reduction in max RMSE for different forecast horizons					
1	4	92.1	85.1	83.9	90.8
12	48	90.3	84.8	81.5	89.5
18	72	89.1	84.9	82.0	87.7
24	96	91.8	85.4	81.5	89.0
Percent reduction in mean RMSE for different forecast horizons					
1	4	92.2	85.4	83.2	90.9
12	48	90.7	84.8	82.4	89.1
18	72	89.7	84.1	81.7	87.4
24	96	91.8	85.3	82.6	88.8

For each performed experiment, these values are averaged over the four non-measurement stations.

for each experiment which amount of deterministic error at non-measurement points is actually resolved through the data-assimilation scheme. Once again, the lowest values are those corresponding to the WND case, where almost half of the deterministic model inadequacy can be still resolved, in terms of correlation through the spatial error distribution scheme.

6.3. Performance of the LLM error forecast

Errors in measurement stations have been forecasted with univariate LLMs, trained using only measurement point error dynamics. The embedding parameters of LLMs were optimized through a genetic algorithm search in a space of solutions made up of the possible time delays vectors for each measurement point time series, the weighting distribution and the number of nearest neighbours in phase space.

DOI: 10.1002/fld

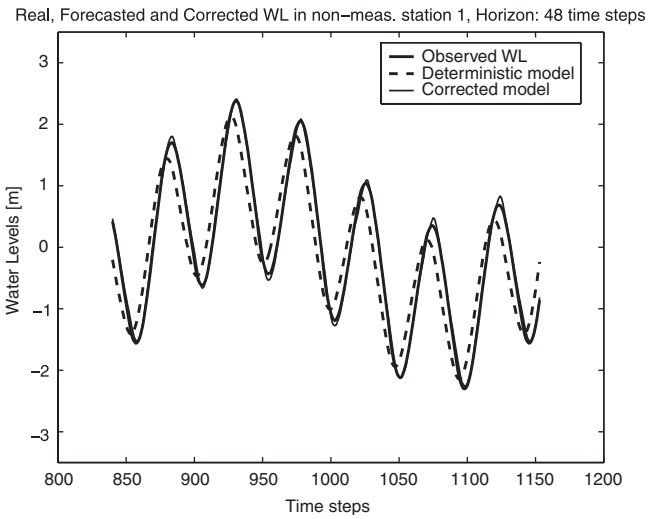

Figure 5. Observed, simulated and corrected water level in station 2, in the CMB case, for a forecast horizon of 12 h concerning the source model and the four different uncorrected models.

Table IV. Square correlation coefficient between deterministic model error and error forecasted through the data-assimilation scheme in non-measurement grid points, for the different experiments and for various forecast horizons.

Square correlation coefficient					
Forecast Horizon		Experiments			
Hours	Time steps	BND	RGH	WND	CMB
1	4	0.991	0.974	0.439	0.990
12	48	0.986	0.972	0.474	0.987
18	72	0.982	0.968	0.386	0.983
24	96	0.990	0.973	0.478	0.986

All values are averaged over the four non-measurement grid points.

In Figure 6, an example of error forecasting is given for the WND experiment and corresponding residual errors that remain after error correction at measurement locations are shown. In this case, the error is consistently negative for the test set owing to the water level established by the uncorrected hydrodynamic model which results from the continuous wind blowing towards the east.

Table V summarizes the results of the error forecasting capability shown by the LLMs through evaluating the respective residual errors in terms of RMSE.

The increment in RMSE of the Corrected model at station 6 for various forecast horizons is plotted in Figure 7 for the combined error experiment. The deterioration in the performance with increasing forecast lead times is even lower for other experiments and at other stations.

DOI: 10.1002/fld

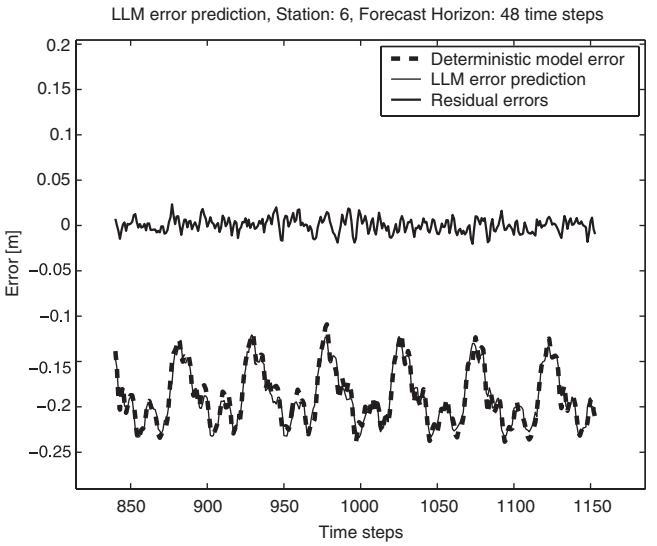

Figure 6. Example of error forecast in station 6 using LLMs in the WND experiment. Residual errors after error correction in the same station are also reported. Forecast horizon is 48 time steps ahead, equivalent to 12 h.

Table V. Summary of residual errors in measurement stations.

Forecast Horizon		Stations		
Hours	Time steps	5	6	7
RMSE (m)—BND experiment				
1	4	0.036	0.032	0.007
12	48	0.049	0.047	0.006
18	72	0.061	0.071	0.012
24	96	0.037	0.041	0.010
RMSE (m)—RGH experiment				
1	4	0.013	0.016	0.014
12	48	0.015	0.019	0.019
18	72	0.019	0.028	0.019
24	96	0.016	0.022	0.015
RMSE (m)—WND experiment				
1	4	0.020	0.009	0.009
12	48	0.009	0.008	0.007
18	72	0.022	0.011	0.009
24	96	0.013	0.010	0.009
RMSE (m)—CMB experiment				
1	4	0.032	0.022	0.032
12	48	0.045	0.036	0.045
18	72	0.062	0.045	0.052
24	96	0.054	0.036	0.047

In these locations the distorted deterministic models are corrected using LLM error forecast.

DOI: 10.1002/fld

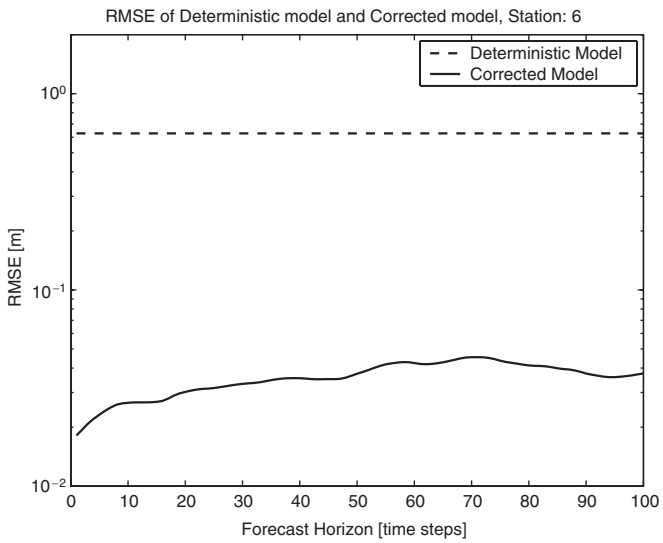

Figure 7. Comparison of RMSE of deterministic and corrected model in measurement station 6 for the CMB experiment. LLM performance deteriorates very slowly with increasing forecast lead time.

Table VI. Summary of residual errors in non-measurement stations.

Forecast Horizon		Stations			
Hours	Time steps	1	2	3	4
RMSE (m)—BND experiment					
1	4	0.050	0.039	0.041	0.068
12	48	0.066	0.042	0.046	0.084
18	72	0.072	0.042	0.053	0.095
24	96	0.055	0.040	0.043	0.072
RMSE (m)—RGH experiment					
1	4	0.025	0.050	0.032	0.023
12	48	0.026	0.051	0.035	0.024
18	72	0.026	0.051	0.038	0.026
24	96	0.024	0.049	0.033	0.024
RMSE (m)—WND experiment					
1	4	0.024	0.012	0.015	0.032
12	48	0.022	0.012	0.016	0.037
18	72	0.027	0.012	0.016	0.036
24	96	0.021	0.012	0.016	0.037
RMSE (m)—CMB experiment					
1	4	0.056	0.049	0.038	0.065
12	48	0.067	0.058	0.048	0.074
18	72	0.082	0.064	0.055	0.087
24	96	0.072	0.057	0.047	0.078

In these locations the distorted deterministic models are corrected by distributing error forecasted in measurement stations.

DOI: 10.1002/fld

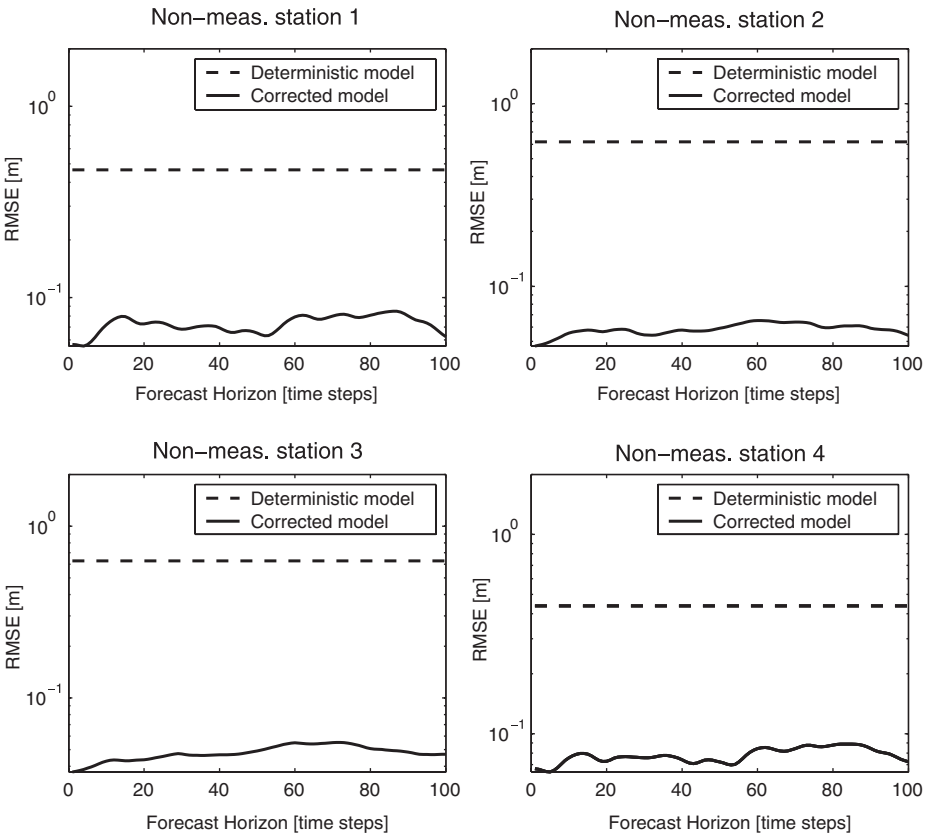

Figure 8. RMSE after error distribution in non-measurement stations for various forecast lead times, CMB experiment.

6.4. Performance of the error distribution scheme

An overview or residual errors after error distribution in non-measurement stations is reported in Table VI in terms of RMSE between corrected model output and observations for various forecast horizons. In Figure 8, deterioration of RMSE with increasing forecast lead times in the different non-measurement stations is plotted: an increasing trend, though weak, is evident. Nonetheless, the scheme's performance is, however, excellent.

Figures 9 and 10 clearly show the consequence of discarding measurement stations that are dynamically well correlated to the one which is being corrected in the proposed data-assimilation scheme. In these figures, time series of the corrected model, uncorrected model output and observations are plotted together. In particular, Figure 9 shows the result of correcting water levels at station 4 without making use of forecasted error and deterministic model output from station 5 in the CMB experiment. In other words, the distribution model is built using those measurement stations that are less correlated to the one that is being corrected. In such case, as illustrated in the figure, although the phase shift has been removed, large errors still persist in the corrected model output.

DOI: 10.1002/fld

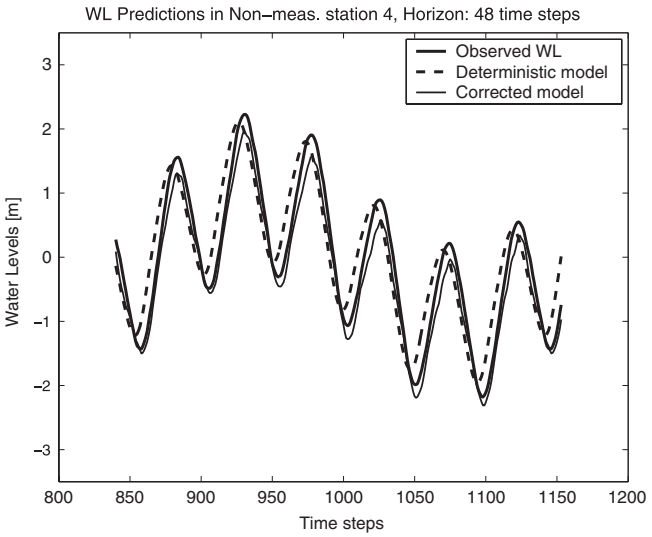

Figure 9. Real, forecasted and corrected water levels in station 4 in the CMB experiment. In this case measurements from station 5 were not used in the correction scheme.

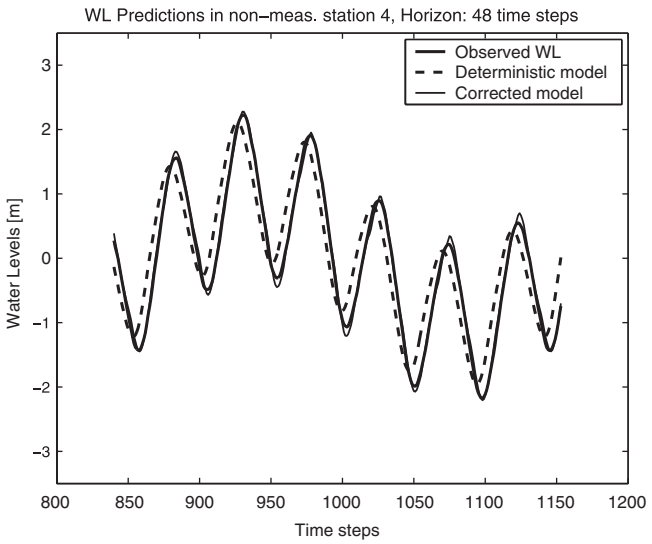

Figure 10. Real, forecasted and corrected water levels in station 4 in the CMB experiment. In this case also measurements and simulated output from station 5 were used in the error correction scheme and the improvement in error distribution is evident.

Indeed, as shown in Table VII, the dynamics of these two grid points is highly correlated and thus the information contained in the error time series of station 5 is expected to have a large impact on the corrective capability of the proposed scheme.

DOI: 10.1002/fld

Table VII. RSQ between deterministic model water level timeseries in measurement and non-measurement stations, for the CMB experiment.

RSQ CMB	Station 1	Station 2	Station 3	Station 4
Station 5	0.994	0.885	0.866	0.982
Station 6	0.842	0.996	0.997	0.809
Station 7	0.701	0.974	0.982	0.657

Notice that for station 4 the model output time series in station 5 is the most correlated.

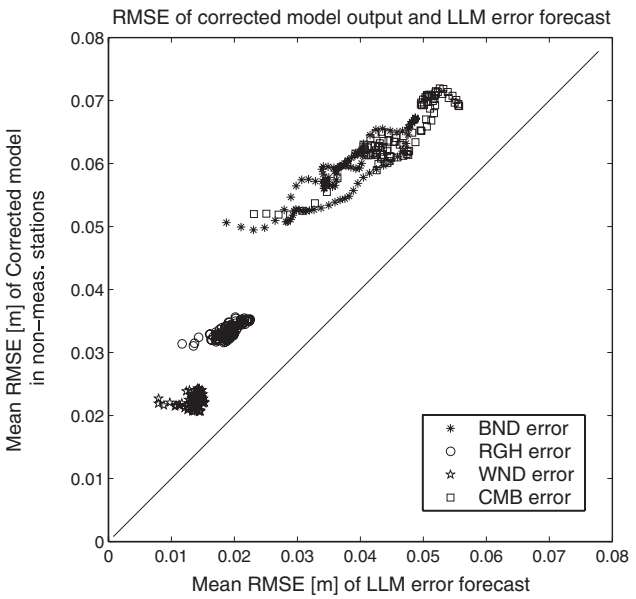

Figure 11. Scatter plot of mean RMSE of corrected model output in measurement stations and after error distribution in other grid points. RMSE is averaged over the three measurement stations for the x-axis and over the four non-measurement stations for the y-axis. Data points in the plot correspond to the various forecast horizons.

Figure 10 clearly reveals the improvement realized with the deterministic model and error forecast information of Point 5 when distributing errors to Point 4. Based on this, if a gauge location has to be chosen for future measurements, a former correlation analysis among possible sites in the hydrodynamic model is advisable, especially if these data are intended also for model correction purposes.

Finally, from Figure 11, it is possible to read the amount of extra error introduced by the distribution method itself with respect to the pure LLM error forecasting performance at measurement points. A RMSE scatter plot of the corrected model at measurement locations *versus* RMSE of the same model at other grid points is presented here, giving an overall picture of the entire study case since all the experimental results are presented in terms of average values at the stations for each experiment. Data points in this plot do not lie far from the bisecting line of the graph. Therefore, in the ideal experiment considered here, the distribution method itself can correct the

DOI: 10.1002/fld

model output in non-measurement locations using LLM error forecasting at observational stations with low additional error.

7. CONCLUSIONS

A computationally efficient and operationally viable data-assimilation scheme is proposed in this paper. The approach is useful when employing physically based numerical models to describe system dynamics for operational purposes. The updating procedure is carried out in two steps: the model errors are forecasted at observation points using historical records. These discrete error forecasts are then conveyed to the entire domain based on the error covariance structure.

The assimilation scheme is based on the derivation of a spatial error covariance structure. The scheme updates the entire field within a forecast horizon with the forecasted observations using a stochastic time series prediction based on the local linear modelling approach. The error distribution procedure exploits the inter-model correlations to distribute the predicted error over the rest of the computational domain and, therefore, the overall scheme allows integration of data-driven error forecasting and subsequent off-line model correction. The procedure can thus be operated in a parallel environment with dynamic updating of the error covariance structure undertaken at frequent intervals. Such an approach provides the opportunity to use the distribution scheme to locate suitable observation points in order to guide planning of the measurement campaign. If a gauge location has to be chosen for future measurements, a former correlation analysis among possible sites in the hydrodynamic model is advisable if these data are also intended for model correction purposes according to this scheme. In the ideal experiment here described, the reduction in deterministic model error is usually about one order of magnitude, with lower reductions in the more drastic WND experiment where still almost half of the error can be resolved. In the present paper it is shown that the distribution method itself can correct the model output at non-measurement locations using LLM error forecasts in observational stations with low additional error.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Dr Arthur Mynett of WL|Delft Hydraulics, Prof. Dr Orazio Giustolisi from Politecnico di Bari and Dr Andrew Colombo from University of Toronto for their valuable advices in preparation of the paper.

REFERENCES

1. Kalman RE. A new approach to linear filtering and prediction theory. *Journal of Basic Engineering* 1960; **82D**:35–45.
2. Jazwinski AH. *Stochastic Processes and Filtering Theory*. Academic Press: New York, 1970.
3. Evensen G. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics* 2003; **53**:343–367.
4. Evensen G, van Leeuwen PJ. Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasi-geostrophic model. *Monthly Weather Review* 1996; **124**:85–96.
5. Madsen H, Canizares R. Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling. *International Journal for Numerical Methods in Fluids* 1999; **31**:961–981.
6. Drécort J-P. Data assimilation in hydrological modelling. *Ph.D. Thesis*, Danish Technical University, Lyngby, 2004.

DOI: 10.1002/fld

7. Graham WD, Tankersley CD. Forecasting piezometric head levels in the floridan aquifer: a Kalman filtering approach. *Water Resources Research* 1993; **29**:3791–3800.
8. Babovic V, Fuhrman DR. Data assimilation of local model error forecasts in a deterministic model. *International Journal for Numerical Methods in Fluids* 2002; **39**:887–918.
9. Keijzer M, Babovic V. Error correction of a deterministic model in Venice lagoon by local linear models. *Proceedings of Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*, Venice, 1999.
10. Babovic V, Sannasiraj SA, Chan ES. Error correction of a predictive ocean wave model using local model approximation. *Journal of Marine Systems* 2005; **53**:1–17.
11. Sannasiraj SA, Babovic V, Chan ES. Local model approximation in the real time wave forecasting. *Coastal Engineering* 2005; **52**:221–236.
12. WL|Delft Hydraulics. *Delft3D-FLOW User Manual*, version 3.10. Delft, 2003.
13. Takens F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, Rand DA, Young L-S (eds), Lecture Notes in Mathematics, vol. 898. Springer: Berlin, 1980; 366–381.
14. Babovic V. A data mining approach to time series modelling and forecasting. In *Proceedings of the Third International Conference on Hydroinformatics*, Copenhagen, Denmark, Babovic V, Larsen LC (eds). Balkema: Rotterdam, 1998.
15. Babovic V, Keijzer M. Forecasting of river discharges in the presence of chaos and noise. In *Coping with Floods*: *Lessons Learned from Recent Experiences*, Marsalek J (ed.), NATO ARW Series. Kluwer: Dordrecht, 1999.
16. Abarbanel HDI. *Analysis of Observed Chaotic Data*. Springer: New York, 1998.
17. Babovic V, Keijzer M, Bundzelm M. From global to local modeling: a case study in error correction of deterministic models. *Proceedings of the Fourth International Conference on Hydro Informatics*, Iowa City, 2000.
18. Houck CR, Joines JA, Kay MG. A genetic algorithm for function optimization: a MATLAB implementation. *Technical Report TR/NCSU-IE/95-09*, North Carolina State University, 1995.
19. Babovic V, Keijzer M, Stefansson M. Optimal embedding using evolutionary algorithms. *Proceedings of the 4th International Conference on Hydroinformatics*, Iowa City, 2000.
20. WL|Delft Hydraulics. *Delft3D-FLOW 2003 Validation Document for the 3D Hydrodynamics Modelling Software*, Version 1.0. Delft, 2003. Available at http://www.wldelft.nl/soft/d3d/intro/validation/valdoc_flow.pdf
21. Stelling GS. On the construction of computational methods for shallow water flow problems. *Rijkswaterstaat Communications No. 35*, Rijkswaterstaat, The Hague, 1984.
22. Phillips NA. A coordinate system having some special advantages for numerical forecasting. *Journal of Meteorology* 1957; **14**:184–185. Available at http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/
23. Yu X. Time series analysis using multivariate chaotic techniques. *M.Sc. Thesis*, International Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE), Delft, 2000.
24. Holland GJ. An analytical model of the wind and pressure profiles in hurricanes. *Monthly Weather Review* 1980; **108**:1212–1218.
25. Vatvani DK, Gerritsen H, Stelling GS, Krishna Rao AVR. Cyclone-induced storm surge, flood forecasting system for India. In *Solutions to Coastal Disasters*, Ewing L, Wallendorf L (eds). ASCE: San Diego, Reston, VA, 2002; 473–487.
26. WL|Delft Hydraulics. *WES User Manual*. APCHMP Project Document, June 2001.

DOI: 10.1002/fld