INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2002; **39**:887–918 (DOI: 10.1002/fld.350)

Data assimilation of local model error forecasts in a deterministic model

V. Babovic[*,†] and D. R. Fuhrman

DHI—Water & Environment, Agern Allé 11, DK-2970 Horsholm, Denmark

SUMMARY

One of the most popular data assimilation techniques in use today are of the Kalman filter type, which provide an improved estimate of the state of a system up to the current time level, based on actual measurements. From a forecasting viewpoint, this corresponds to an updating of the *initial conditions*. The standard forecasting procedure is to then run the model *uncorrected* into the future, driven by predicted boundary and forcing conditions. The problem with this methodology is that the updated initial conditions quickly 'wash-out', thus, after a certain forecast horizon the model predictions are no better than from an *initially uncorrected* model. This study demonstrates that through the assimilation of error forecasts (in the present case made using so-called local models) entire model domains can be corrected for extended forecast horizons (i.e. long after updated initial conditions have become washed-out), thus demonstrating significant improvements over the conventional methodology. Some alternate uses of local models are also explored for the re-distribution of error forecasts over the entire model domain, which are then compared with more conventional Kalman filter type schemes. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: error prediction; data assimilation; Kalman filter; local models; hydrodynamic modelling

1. INTRODUCTION

One of the most popular data assimilation techniques in use today are of the Kalman filter type, which provide an improved estimate of the state of a system up to the current time level, based on actual measurements. In terms of forecasting, this corresponds to an updating of the *initial conditions*. Forecasts are typically obtained by then simulating the model *uncorrected* into the future, driven by predicted boundary and forcing terms. These future states are typically not corrected, as no measurements are as yet available at these time levels. The problem with this methodology is that the updated initial conditions quickly 'wash-out', thus, after a certain

[*] Correspondence to: V. Babovic, DHI—Water & Environment, Agern Allé 11, DK-2970 Hørsholm, Denmark.
[†] E-mail: vmb@dhi.dk

Contract/grant sponsor: Danish Technical Research Council (STVF); contract/grant number: 9800463

Received July 2001
Revised November 2001

forecast horizon the model results are essentially no better than if they had been made using an *initially uncorrected* model.

It is, of course, impossible to obtain actual measurements of a model state for future time levels. It is, however, possible to *forecast* these measurements, either in terms of state variables or model errors. Such forecasts can then be assimilated into the model, thus providing improved estimates of a system state at *future* time levels. This methodology potentially allows for the correction of model states for extended forecast horizons (i.e. far beyond the time it takes for updated initial conditions to become washed-out).

This paper gives a brief overview of time series forecasting with local models, including a novel approach for their optimization using genetic algorithms (GAs). The approach is then applied in the forecasting of errors in a deterministic model in a hypothetical bay, where errors are introduced artificially in the resistance, boundary, and wind-forcing terms. Subsequent local model assimilation methods are applied to re-distribute the errors over the entire model domain, and their performance is compared with some more conventional Kalman filtering approaches.

2. TIME SERIES FORCASTING WITH LOCAL MODELS

2.1. Local modelling

A rather effective method of simulating the evolution of a dynamical system is by means of a local approximation, using only the most similar trajectories from the past to make predictions of the future [1]. Such *local models* (LMs) have already been used successfully in numerous previous studies. These include error correction in deterministic models [2, 3], the forecasting of river discharges [4, 5], as well for control purposes [6].

Local models are particularly well suited for time series forecasting since they share many fundamental ideals with Takens time-delay embedding theorem [7]. This theorem essentially states that the underlying structure of a complex, multidimensional system can be equivalently viewed using a projection from a single variable (in the form of a time series) in *phase* space (i.e. an embedded space with dimensions consisting of various time lags of the variable itself). Time series can correspondingly be forecast based on this structure in their phase space. Time series forecasting with local models consists of three main steps: (1) embedding of the time series data into phase space; (2) finding the k most similar points in phase space (i.e. a local neighbourhood) to a query point; and (3) performing a regression on the local neighbourhood (using the neighbourhood co-ordinates as inputs, and their corresponding future values as outputs) to obtain a forecast.

The local regression performed in step (3) is typically of degree zero or one (i.e. averaging or linear), although this can theoretically be of any polynomial degree. Although the local approximation may not be, the resulting overall behaviour can be highly non-linear, as a local approximation is made separately for each neighbourhood.

2.2. Selection of embedding parameters

As previously mentioned the first step in making a LM forecast is to embed the time series into a phase space. This typically involves the selection of a time lag, τ, and an embedding dimension, d_e, so that a variable, y, can be represented in phase space as

$$\mathbf{y}_n = [y_n, y_{n-\tau}, y_{n-2\tau}, \ldots, y_{n-(d_e-1)\tau}] \tag{1}$$

where n is a time level. Methods for selecting *prescription* values for the necessary parameters using average mutual information (AMI) and false nearest neighbour (FNN) analyses are typically recommended in the literature (see e.g. Reference [8]), however, these have been shown to generally be sub-optimal selections (see e.g. References [1, 9]). Therefore an alternate strategy using genetic algorithms is employed throughout this work, which has been shown to demonstrate significant improvements over the prescription values.

2.3. Genetic algorithms

Genetic algorithms [10] are general purpose search algorithms based loosely on the principals of Darwinian evolution. By means of a *simulated evolution* GAs can be used to optimize highly non-linear, multidimensional problems (see e.g. Reference [11]). For details on the GA used in this study see Reference [12].

2.4. Evolutionary embedding

The bottom line for the performance of an embedding and subsequent local modelling is the resulting forecast skill, as good forecast skill implies good embedding properties [1]. Therefore, in an attempt to obtain a more optimal embedding a GA was implemented around the existing local modelling code. Essentially what is left is a global optimization problem with the input parameters consisting of the number of nearest neighbours, k; the selection of a weighting function, w; and the components in the embedding vector, $\boldsymbol{\tau}$. The number of k nearest neighbours used in this study was held globally constant, though more complex methods of neighbourhood selection could be used (see e.g. Reference [13]). The selection of a weighting function (for purposes of a weighted regression) consisted of nine possibilities (from Reference [14]): i.e. uniform, $1-d$, $1/d$, $1/d^2$, $1/(d+1)$, e^{-d*d}, e^{-d}, $(1-d^2)^2$, and $(1-d^3)^3$, where the distances, d, from a query point are normalized between zero and one. In the most general terms the embedding vector can be expressed as

$$\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2, \ldots, \tau_{d_e-1}] \tag{2}$$

Accordingly, (1) then becomes:

$$\mathbf{y}_n = [y_{n-\tau_0}, y_{n-\tau_1}, y_{n-\tau_2}, \ldots, y_{n-\tau_{d_e-1}}] \tag{3}$$

Typically τ_0 is set equal to zero, however the general notation will be maintained here. These parameters can then be optimized by minimizing some error statistic for the testing data.

Within the GA, rather than optimizing the components in the embedding vector, $\boldsymbol{\tau}$, directly, an approach using *changes* in time delay, $\Delta\tau$, as parameters has been adopted:

$$\boldsymbol{\Delta\tau} = [\Delta\tau_1, \Delta\tau_2, \ldots, \Delta\tau_{d_e-1}] \tag{4}$$

The resulting time delay vectors can then be constructed by progressively summing the preceding values. Hence (2) becomes

$$\boldsymbol{\tau} = [\tau_0, \tau_0 + \Delta\tau_1, \tau_0 + \Delta\tau_1 + \Delta\tau_2, \ldots, \tau_0 + \Delta\tau_1 + \Delta\tau_2 + \cdots + \Delta\tau_{d_e-1}] \tag{5}$$

Finally, only unique values in the τ vector are used. For example, if the GA produced an offspring for $\mathbf{\Delta\tau}$ containing the values [0 4 3 5 0 2], a progressive summation of the values yields the vector [0 4 7 12 12 14]. Finally, taking only the unique values yields the actual τ vector that would be used in the simulation: [0 4 7 12 14]. Constructing the time delay embedding vectors in this fashion allows for simultaneous variations in the delay values themselves, as well as in the effective embedding dimension, while keeping the number of optimization parameters constant. Furthermore, the range for the GA variables can be set much lower than if the time lags were evolved directly, thus significantly reducing the search space.

3. HYDRODYNAMIC MODEL

The deterministic model used in this study is the MIKE 21 HD (hydrodynamic) modelling system from DHI—Water & Environment, Horsholm, Denmark. Because this hydrodynamic model has been described in the literature numerous times a detailed description will not be provided here. For more specific details the reader is referred to References [15, 16].

4. DATA ASSIMILATION METHODS

The numerical model is, of course, far from perfect. A numerical model is indeed only a *model* of reality; i.e. it employs a number of simplifying assumptions, e.g. depth averaging of velocities in vertically integrated two-dimensional models, which inevitably produce inaccuracies. In a numerical model one also discretizes the domain, and is therefore not able to resolve numerous sub-grid scale phenomena. Errors in the model parameterization (mainly because most model parameters cannot be directly measured) may also greatly contribute to the overall error in a numerical model. Finally, it is impossible to precisely define initial conditions and forcing terms in the entire computational domain. All of these imprecisions and uncertainties can accumulate to produce fairly poor model results, despite our ‘perfect’ knowledge of the governing laws [17]. To combat the inevitable presence of such model errors methods for correcting the model results are often employed. *Data assimilation* is a methodology that utilizes information from observations, and combines it with (or assimilate it into) numerical models [18]. For an overview of various data assimilation strategies see References [19, 20].

If forecasting interest is for a considerably long forecast lead-time, a data assimilation scheme based on updating of output variables (i.e. *error prediction*) may be the most suitable approach. This cannot be done in conventional data assimilation methods (i.e. Kalman filtering), where the data must be introduced in the model state in order to be assimilated. The following sections briefly introduce conventional Kalman filtering, as well as some alternate approaches for error prediction and subsequent assimilation using local models.

4.1. Kalman filtering

A standard data assimilation procedure, which uses an updating of the model state variables, is the Kalman filter [21]. Applications of this procedure are abundant in the literature [18, 22–26], and the reader is referred to these references for further details on the method.

Table I. Summary of the different Kalman filter algorithms included in the MIKE 21 DA module (adapted from Reference [27])*.

	EnKF	RRSQRT KF	SSKF
Error propagation	Propagation of ensemble according to full non-linear model dynamics	Propagation of error covariance matrix using tangent linear model operator	No error propagation
Model error forcing	Part of ensemble propagation	Matrix algebra	No explicit model error forcing
Representation of error covariance matrix	Ensemble estimate	Reduced rank approximation of square-root of covariance matrix	Constant error covariance
Storage requirements	$M\times(N+q)$	$(M+M)\times(N+q)$	$(N+q)$
Computational costs	M model integrations	M model integrations and eigenvalue decompositions	Slightly more expensive than a model integration
Main disadvantage	Large sample required for sufficient error representation	Large rank of covariance matrix required to avoid filter divergence	Constant error covariance assumed

*M is the ensemble size in the EnKF and the rank of the covariance matrix (i.e. the number of leading eigenvalues) in the RRSQRT KF, N is the number of state variables, and q is the number of noise points.

Various approximations for the error covariance propagation inherent within a KF exist that significantly reduce the computational burden. The three that are included in the MIKE 21 DA (data assimilation) module are: (1) the ensemble Kalman filter (EnKF); (2) the reduced-rank-square-root Kalman filter (RRSQRT KF); and (3) the steady state Kalman filter (SSKF). A summary of these three variations is provided in Table I. As the emphasis of this paper is not specifically on Kalman filtering methods more detailed descriptions are not provided here. For more exact details on their implementation the interested reader is referred to References [15, 27]. In the present work the EnKF and a corresponding SSKF will be used for comparison with some alternate LM approaches described in the following sub-sections.

4.2. Local weighted spatial regression

The first LM technique is simply to redistribute measured and forecasted errors at the measurement locations to the rest of the domain based on a weighted spatial regression performed on a local neighbourhood of measurement points. For practical reasons (i.e. measurement locations are usually sparsely distributed) the regression should probably be of degree zero or one. This method can be used in essentially identical fashions to redistribute both error measurements and forecasts, thus allowing corrections to be made at both current and future time levels. The steps for each non-measurement grid point are simply to: (1) select a local neighbourhood of measurement point locations, and (2) distribute the error measurements and forecasts (for each successive time level) using a weighted spatial regression on the local neighbourhood.

The inputs for the local model forecasts with this method could potentially include the embedded values of the errors themselves, as well as the respective model outputs (i.e. water levels and fluxes). In theory, these multiple variables could be taken not only from the actual measurement location, but also from surrounding grid points. It has, however, been shown that the addition of input data from neighbouring grid points effectively increases the dimensionality of the phase space without providing a corresponding increase in data point coverage [9] (since the outputs from neighbouring grid points do not tend to vary significantly). Therefore, the inclusion of LM input data from surrounding grid points is not investigated in this work.

Such a local weighted spatial regression (LWSR) provides a very simple and efficient method for the distribution of errors throughout a model domain. To save computation time the nearest neighbours in space and regression parameters need only be determined once and then stored for reuse at all additional time levels. The computational burden is very similar to that of the SSKF for each time level.

4.3. Weighted local model ensemble

The second methodology introduced in this work is to use a weighted local model ensemble (WLME) to forecast errors at every non-measurement grid point. For this method local models are again created using data from the measurement points. These local models are then re-applied at the non-measurement grid points. Because a time series of errors is not available outside of the measurement points, these local models must only use actual model results as inputs. A weighted ensemble of the local model predictions (based on distance in space from the query grid point) is then used to obtain error forecasts for non-measurement grid points. In the manner applied throughout this work a forecast horizon of one time step corresponds to an approximation of errors at the current time level (i.e. previous model outputs are used to predict the current error). It then becomes trivial to extend the method to any forecast horizon, thus also allowing for corrections at future time levels. The steps for each non-measurement grid point, again, are to: (1) construct a local model using data from each measurement point; (2) select a neighbourhood of nearest measurement point locations; (3) make predictions for the current and extended time levels using each of the selected local models; and (4) combine the predictions from the selected local models in a weighted ensemble fashion (i.e. predictions made using local models based on data nearest to the non-measurement grid point are weighted most heavily). The idea of applying local models constructed from data at measurement points at the other grid points was taken from Reference [28], where the data from all measurement points was combined into a single database. Based on a comparison in Reference [9], however, the weighted ensemble approach presented here appears to be the better strategy.

This correction technique relies on the assumption that the dynamics throughout a model domain are related, hence allowing a LM based on data from one location to be used at another. A weighted ensemble of the forecasts is used since the underlying dynamics, though related, would also be expected to contain some spatial diversity. This method is more computationally expensive than a LWSR, largely because searching for nearest neighbours in phase space is necessary at every grid point. Such searching typically requires $O(k \log N)$ comparisons, where N is the number of training data [4].

It should also be noted that because the error measures themselves are not embedded as local model inputs, the error predictions (even for the current time step) are made with no assumed knowledge of the most recent measurements. This could be interpreted either as an advantage or a disadvantage, depending on the viewpoint. If successful the error prediction scheme can be applied without using any complicated real-time sensing (requiring only a database of previous model results and measured errors). On the other hand, not directly incorporating the most recent measurements seems to put the method at an inherent disadvantage.

4.4. Hybrid method: KF assimilation of LM forecasts

The final methodology presented in this work is a sort of hybrid approach, combining both local models and Kalman filtering. It involves making a LM forecast of the errors at the measurement points, and then assimilating these forecasts into the model using a Kalman filter. In this study the SSKF is employed for this purpose (which, again, uses a constant gain matrix), although any KF variety could be used. In order to apply this technique the uncorrected model must first be simulated into the future to gain a time series of model outputs for each measurement location. The forecasted errors at these points can then be added to these model outputs to obtain forecasted state variables (since KF is a state variable updating procedure). The Kalman filter can then be applied in the standard fashion to assimilate these forecasted state variables. This extra step could be avoided by directly forecasting the state variables, however this was not done in this paper. This hybrid technique is perhaps the most fundamentally sound of the methods presented here, as it combines the well-documented time series forecasting skill of local models with the assimilation capabilities of Kalman filtering.

5. HYPOTHETICAL BAY DESCRIPTION

In order to test the methods described in the previous section a model of a hypothetical bay was constructed using the MIKE 21 software (a similar case study is used in References [15, 28]). A description of the 'true' (i.e. before errors are introduced) hypothetical bay is provided in this section. The bay consists of a rectangular 21×20 grid using a spacing of 10 km in both the x- and y-directions. The model bathymetry is shown in Figure 1, having depths ranging from zero along the shore to -100 m at its deepest locations in the middle. The Chezy bed resistance coefficient, C, varies with the depth, having values in the range 30–45 $\mathrm{m}^{1/2}\,\mathrm{s}^{-1}$, with the largest values in the deepest areas. The bay has an open northern boundary and closed eastern, western, and southern boundaries. In previous studies (i.e. References [15, 28]) the model was driven by a simple sine wave having a period of 12 h and a range of 2 m. The flow in the present study, however, is forced by a multiperiodic sinusoidal water variation at the open boundary with periods of 12 h (representing tides) and 72 h (loosely representing some varying tidal cycle). Meteorological forcing is included using wind and pressure fields from an artificially generated moving cyclone (see Figure 2 in Reference [15]) that moves in a west–east direction with a speed of 8.33 $\mathrm{km\,h}^{-1}$. The main flow describes a Kelvin wave moving counter-clockwise in the bay region. The entire model simulation is for 288 h (12 d) and uses a time step, Δt, of 15 min, for a total of 1152 time steps. Throughout the tests described in this paper time steps 101 through 800 were used for training (where applicable), and 801 through 1152 for testing, with only the testing results reported throughout. The first 100 time steps were not used to ensure that the initial conditions were properly washed-out.

In Reference [28] nine measurement points were selected along the diagonals of the bay. Typically, however, water surface elevation measurements are available only close to the shore. Therefore, a more realistic configuration of measurement points has been adapted as in Reference [15] for the present study. This configuration consists of three measurement locations at grid points (1,16), (8,1), and (20,12), which are also shown in Figure 1.

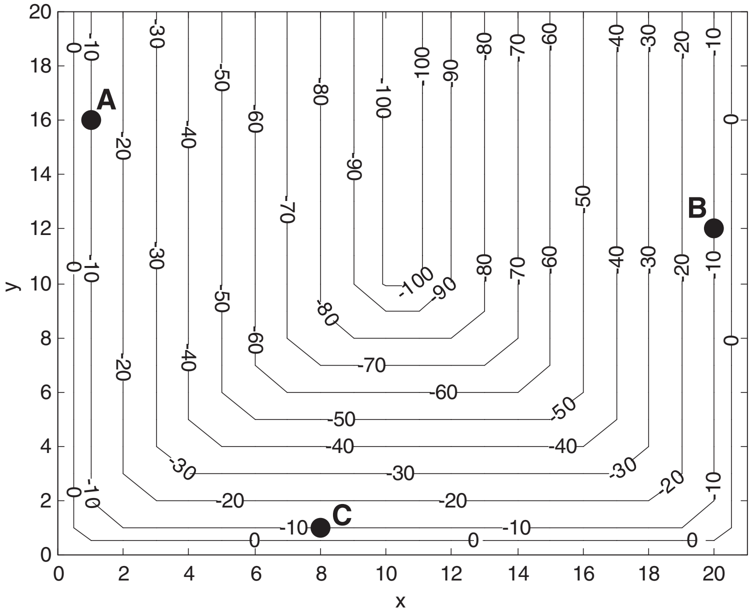

Figure 1. Bathymetry of the hypothetical bay (depths in m). The grid spacing in the x- and y-directions is 10 km.

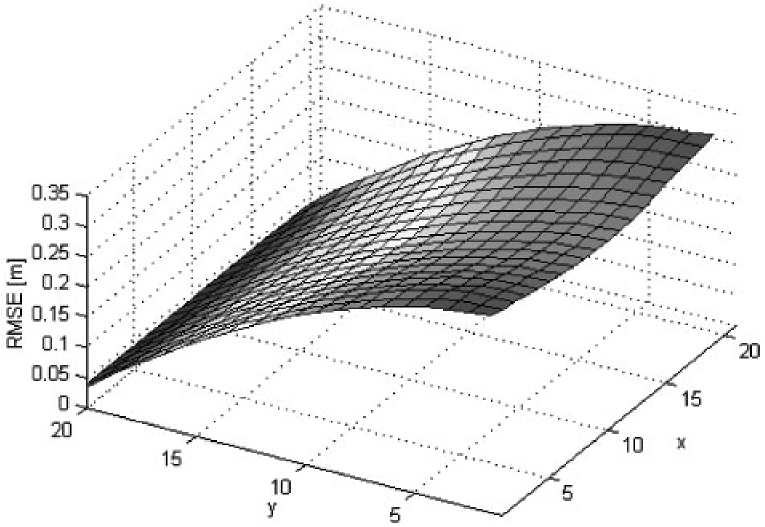

Figure 2. Spatial distribution of RMS error for the uncorrected MIKE 21 model with resistance error. The average value of this surface is 0.1788 m.

6. INTRODUCTION OF ERRORS

To test the performance of the correction schemes developed in Section 4 resistance, boundary, and wind-forcing errors were introduced into the 'true' model. The following sub-sections describe the effects of these errors when introduced individually, as well as in combination.

6.1. Resistance error

The first errors introduced to the hypothetical bay were through the resistance coefficients. These are a common source of error in hydrodynamic models, as the actual roughness coefficients cannot directly be measured from a system and are instead generally used as calibration parameters. Furthermore, the resistance formulae common in hydraulics (i.e. the Manning or Chezy equations) are empirical in nature, and therefore do not describe the effects of friction correctly in any theoretical sense. To investigate the effects and predictability of resistance error the Chezy resistance coefficients in the 'true' model were changed to a global value of $32\,\mathrm{m}^{1/2}\,\mathrm{s}^{-1}$. This new model was then re-simulated holding all other parameters constant as in the 'true' model. Throughout this work the root-mean-squared error (RMSE) statistic will be used as a fitness measure, defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6}$$

where N is the length of the time series; y_i are the observed values; and $\hat{y}_i$ are the estimates. The resulting spatial distribution of RMS error for the uncorrected MIKE 21 simulation is shown in Figure 2, having a spatially averaged value of 0.1788 m.

Here, because the open boundary condition is assumed to be a perfect representation of the observed phenomenon, the error is essentially zero at the northern boundary. As the distance from this boundary grows, however, the error becomes more pronounced (due to shoaling of the tidal wave), with the highest errors occurring along the southern coast.

6.2. Boundary error

Next, a 1-h phase error was introduced separately into the boundary condition of the 'true' model (the resistance values were reset to their 'true' values). The resulting spatially distributed MIKE 21 RMS error without correction is shown in Figure 3, having a relatively large spatially averaged value of 0.5254 m.

From the introduction of this boundary error there is now significant error at the northern end. As in the previous section, the introduced error grows as the distance increases from the open boundary.

6.3. Wind-forcing error

The final error introduced into the model was in the wind-forcing. The moving cyclone was replaced with a constant, spatially uniform wind blowing due east at $20\,\mathrm{m\,s}^{-1}$. This is a fairly drastic alteration, but it was thought that it would certainly provide ample errors for the testing purposes here. The resulting spatial distribution of RMS error is shown in Figure 4 for the uncorrected MIKE 21 model, having an average value of 0.2359 m.

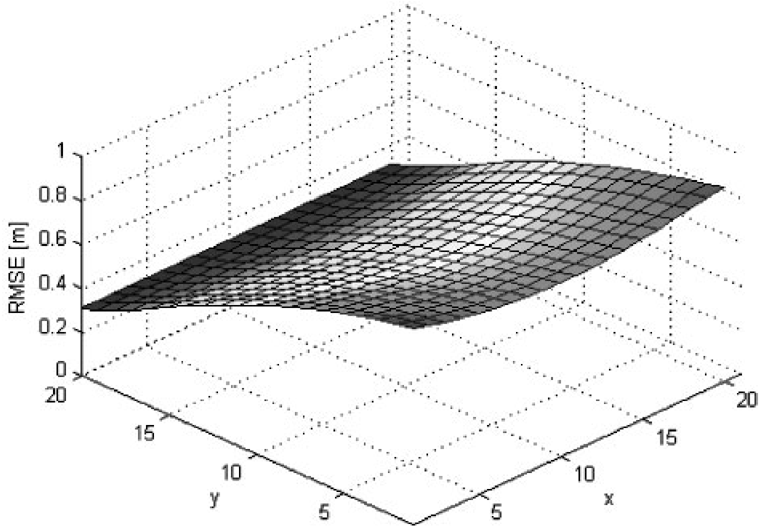

Figure 3. Spatial distribution of RMS error for the uncorrected MIKE 21 model with boundary error. The average value of this surface is 0.5254 m.

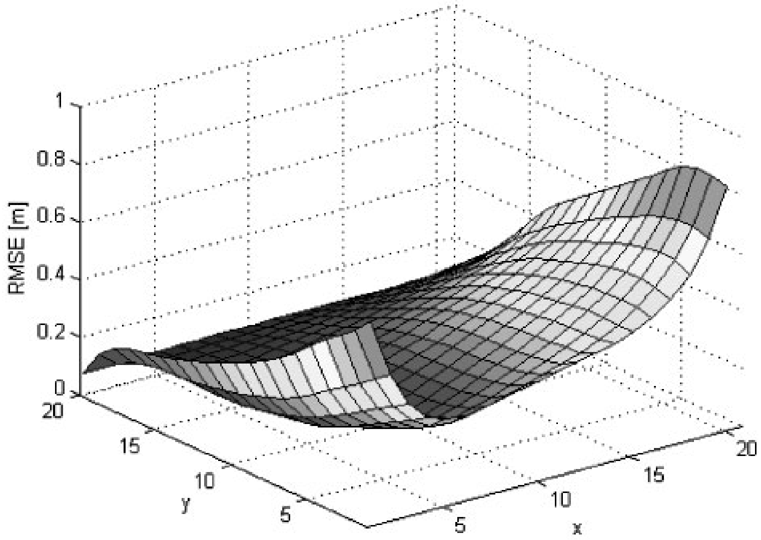

Figure 4. Spatial distribution of RMS error for the uncorrected MIKE 21 model with wind-induced error. The average value of this surface is 0.2359 m.

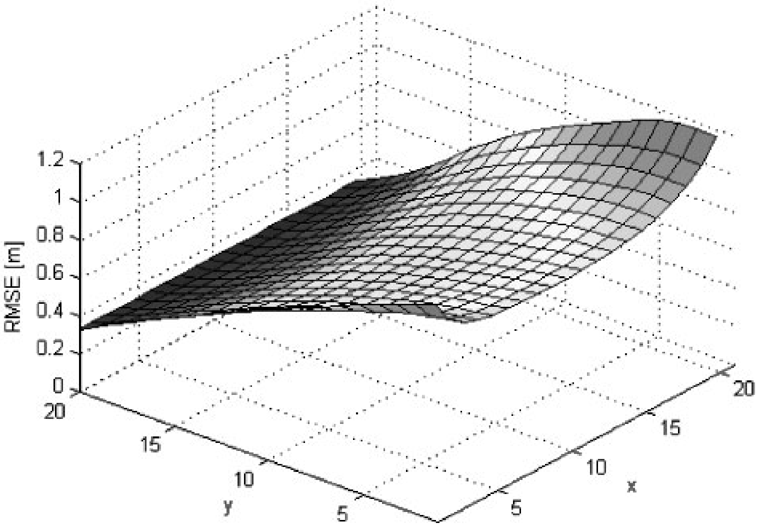

Figure 5. Spatial distribution of RMS error for the uncorrected MIKE 21 model with combined error. The average value of this surface is 0.6256 m.

Here it is seen that the errors are again somewhat amplified as the distance increases from the open boundary. More significant, however, is the concentration of errors at both the eastern and western boundaries, which lie directly perpendicular to the introduced wind.

6.4. *Combined error*

Finally, the three previously described error sources were combined into a single model. The resulting spatial distribution of RMS error for the uncorrected MIKE 21 model is shown in Figure 5, having an average value of 0.6256 m. Here, the individual components from the various error sources are very apparent, as the combined error is clearly a composite of these individual surfaces. This model provides the most realistic testing environment, as errors in deterministic models are generally due to multiple factors. Therefore, although corrections were made on all of the described models (see Reference [9] or alternatively Reference [29] for complete details), the results correcting this combined error model will be given the most attention in this paper.

7. ASSIMILATION AT CURRENT TIME LEVEL

As an initial testing atmosphere the error correction schemes described in Section 4 were used to correct the errors at the current time levels. For the local weighted spatial regression and Kalman filtering techniques this corresponds to a re-distribution of actual measurements. For the weighted local model ensemble scheme this corresponds, again, to a forecast horizon of one time step. As previously noted, the results described here will concentrate on the

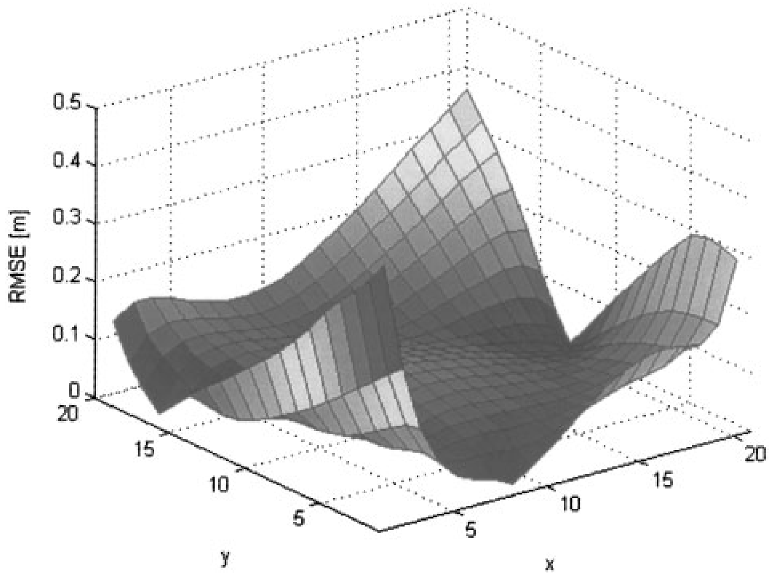

Figure 6. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using a local weighted spatial regression. The average value of this surface is 0.1356 m.

correction of the combined errors, however, a summary of all corrections (i.e. including the individually introduced errors) is presented later in Table III.

7.1. Local weighted spatial regression of measurements

Firstly, results using a local weighted spatial regression (see Section 4.2) of the current measurements to correct the combined error model are presented. An optimal configuration was found using a GA, which minimized the average RMS error at four grid points in the model domain having locations (7,7), (7,14), (14,7), and (14,14). In reality this means that measurements at these locations were assumed to be available, which effectively increases the assumed number of measurement points from three to seven. However, it was decided to take advantage of the current 'data-rich' situation, in the hope that the findings might be generally applied to real modelling situations where enough data were not available for such optimization. The optimal configuration found consisted of a weighted spatial linear regression, $w = 1/d^2$, of all three measurement points. The resulting spatial distribution of RMS error after correction is shown in Figure 6, having an average error of 0.1356 m, which is substantially lower than the uncorrected MIKE 21 error of 0.6256 m.

The time series of errors and corresponding water surface elevations for point (11,11) are shown in Figure 7 for the testing duration. The uncorrected MIKE 21 simulation contains both phase and amplitude errors, which are nearly eliminated using this approach, reducing the uncorrected error at this point from 0.6277 to 0.1015 m.

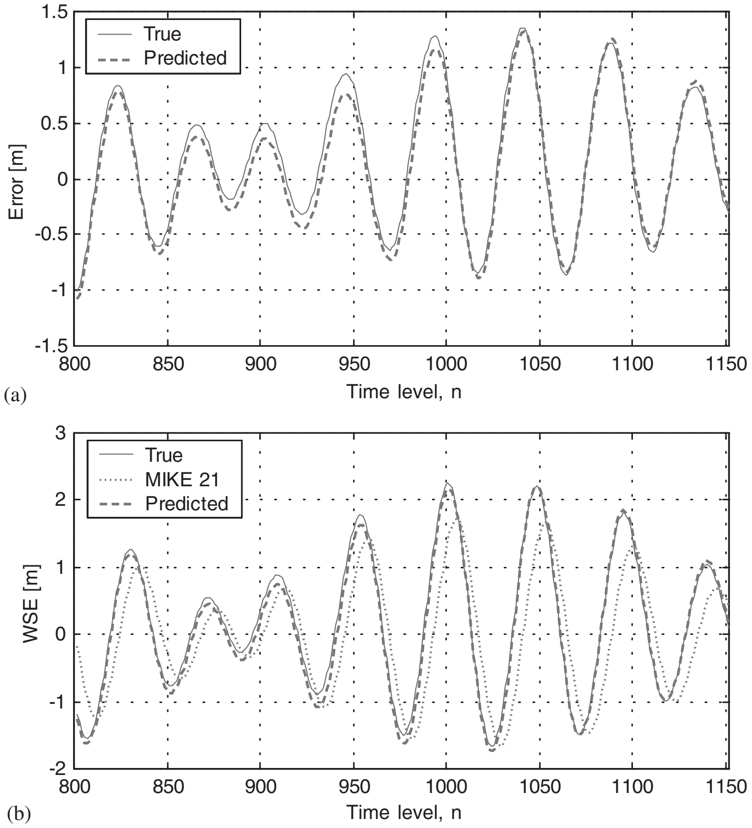

Figure 7. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) in the combined error model for the correction at current time levels using a local weighted spatial regression. The RMSE for this point after correction is 0.1015 m compared with 0.6277 m in the uncorrected MIKE 21 model.

7.2. Weighted local model ensemble

Secondly, the combined error model was corrected using a weighted local model ensemble as described in Section 4.3. The parameters for the weighted local model ensemble were optimized over the same four grid points, with the optimal configurations shown in Table II. Throughout this work only zero-degree local approximations (i.e. averaging models) were considered due to the relatively small number of training data. Also, the model results (which, again, serve as LM inputs) were modified slightly i.e. the water depths were converted to water surface elevations, and fluxes were divided by the water depths to obtain a flux per unit depth. This was to ensure that the input values remain more or less constant for the re-application of the LMs throughout the grid (for simplicity, however, their original notation will be maintained). This configuration was then applied over the entire system. Because the measured errors were assumed to be known at the measurement locations these were artificially set to zero for this simulation, however (as discussed previously) no knowledge

Table II. Optimal results found with a genetic algorithm for the correction of the combined error model (at current time levels) using a weighted local model ensemble.

Parameter	Value	Description
Include h	1	Do include
Include q_x	0	Do not include
Include q_y	0	Do not include
Surrounding cells	[0 0 1 0 0]	Include only centre-point (j,k)
Spatial weight dist.	5	$w = e^{-d*d}$
Spatial k	3	Use weighted ensemble prediction of all meas. points
τ_A	[0 10 11 12 14]	Time delay embedding for local model A
τ_B	[0 13 28 39 54]	Time delay embedding for local model B
τ_C	[0 13 26 33 39 40]	Time delay embedding for local model C
k_A	19	No. of nearest neighbours in phase space for A
k_B	11	No. of nearest neighbours in phase space for B
k_C	23	No. of nearest neighbours in phase space for C
Phase space weight dist.	8	$w = (1-d^3)^3$

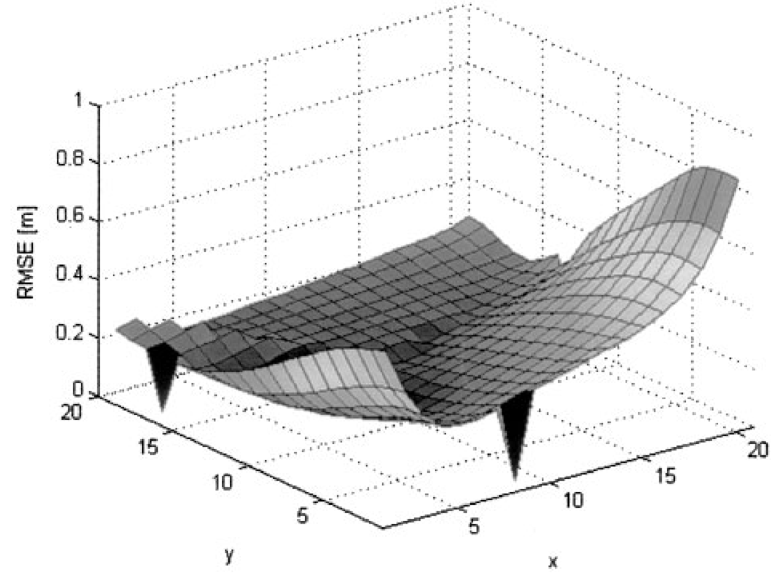

Figure 8. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using a weighted local model ensemble. The average value of this surface is 0.2487 m.

of these most recent measurements is incorporated into the rest of the model. The resulting spatial distribution of RMS error is shown in Figure 8, having an average error of 0.2487 m. The wind error, when introduced alone, produced a spatially averaged error of 0.2063 m after correction, thus by far being the largest contributor (see Table III). This dominance is even more apparent as significant errors still exist along the western and eastern shores after correction (compare with Figure 4). For the combined errors this approach performs

Table III. Summary of spatially averaged errors. The lowest values obtained for each error source are highlighted in bold.

Error	RMSE (m)				
	MIKE 21	EnKF	SSKF	LWSR	WLME
Resistance	0.1788	0.09030	**0.03785**	0.04044	0.05508
Boundary	0.5254	0.09992	**0.06700**	0.07540	0.1458
Wind	0.2359	**0.09955**	0.1289	0.1044	0.2063
Combined	0.6256	0.1616	0.1774	**0.1356**	0.2487

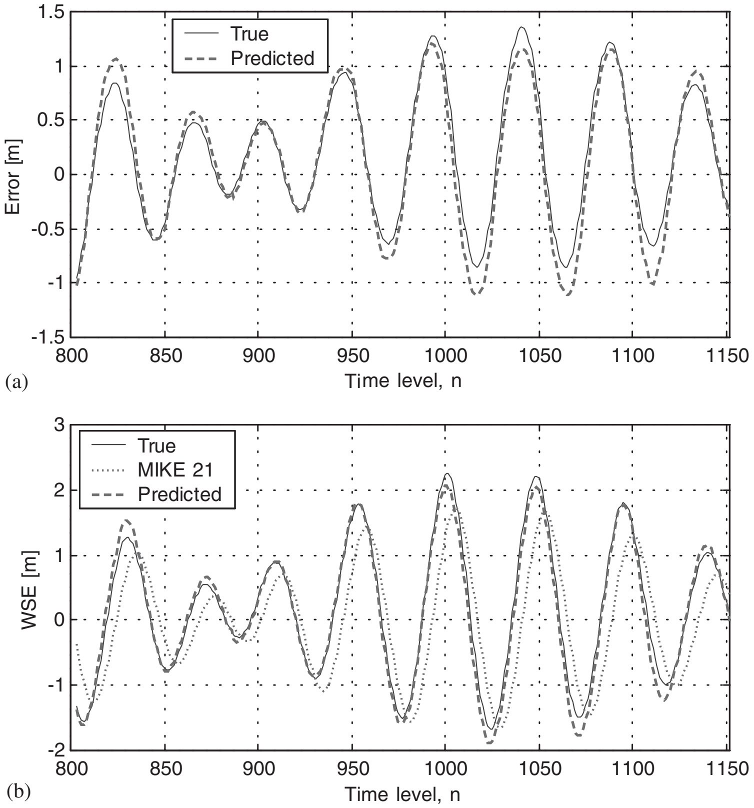

Figure 9. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) in the combined error model for the correction at current time levels using a weighted local model ensemble. The RMSE for this point after correction is 0.1486 m compared with 0.6277 m in the uncorrected MIKE 21 model.

significantly worse than the local weighted spatial regression of measurements seen in the previous sub-section.

The time series of errors and corresponding water surface elevations are plotted for point (11,11) in Figure 9. Once again there is a substantial improvement in both the phase and

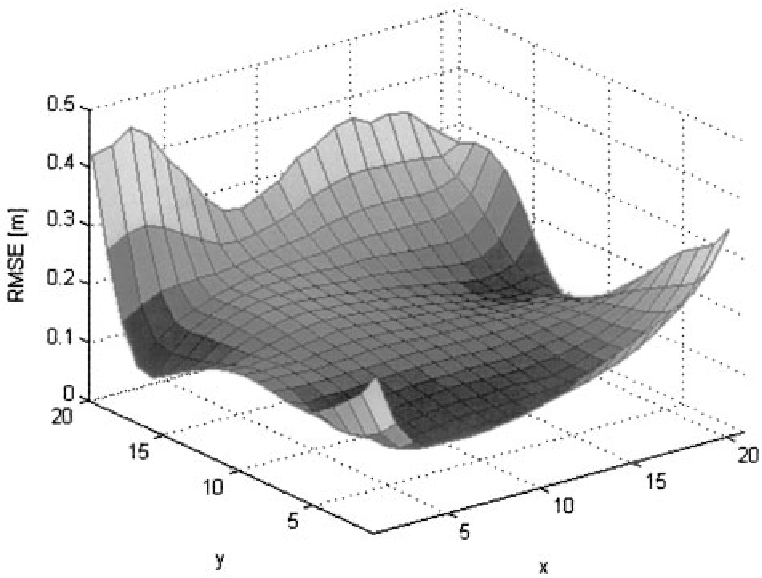

Figure 10. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using an ensemble Kalman filter. The average value of this surface is 0.1616 m.

amplitude errors evident in the uncorrected MIKE 21 simulation, though not as good as was seen in Figure 7.

7.3. Kalman filtering

For comparison the Kalman filtering schemes were also employed. The following Kalman filtering parameters for the EnKF were calibrated manually for the correction of this combined error model. The introduced boundary error was assumed to have a standard deviation of 0.1 m, with both spatial and temporal (lag-one) autocorrelation coefficients of 0.9. The introduced wind errors were assumed to have a standard deviation of 1.0 m s^{-1} with a spatial autocorrelation coefficient of 0.98 and a temporal autocorrelation coefficient of 0.97. The measurement errors were assumed to have a standard deviation of 0.10 m. The correction scheme did not seem to be very sensitive to any of the parameters, and they were therefore held constant throughout this work. The EnKF simulations also used an ensemble size, M, of 100, which was verified to be enough for convergence. The SSKF used the calculated average gain matrix between time levels 101 through 800 of the EnKF simulation. Both KF varieties used a noise grid that was four times courser than the actual model grid. These correction schemes resulted in spatially averaged RMS errors of 0.1616 and 0.1752 m for the EnKF and SSKF, respectively. These are also substantial improvements over the original MIKE 21 simulation, but not quite as low as the 0.1356 m obtained using a local weighted spatial regression of the measured errors. The spatial distribution of RMS error for the EnKF simulation is shown in Figure 10. The time series of errors and corresponding water levels for point (11,11) are also shown for the EnKF in Figure 11.

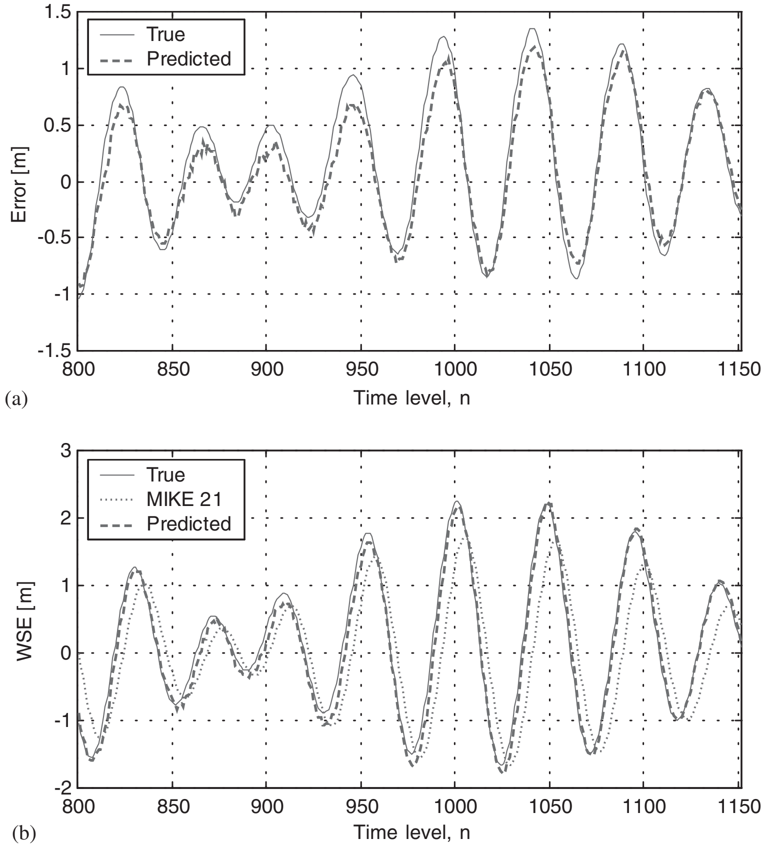

Figure 11. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) in the combined error model for the correction at current time levels using an ensemble Kalman filter. The RMSE for this point after correction is 0.1464 m compared with 0.6277 m in the uncorrected MIKE 21 model.

8. TOTAL WATER VOLUME

In order to gain insight into how the various error correction schemes also affected the mass balance of the system the total water volume (TWV) was calculated for the combined error models. The volume of water, V, represented by each grid point can be approximated by

$$V_{j,k} = \Delta x \cdot \Delta y \cdot h_{j,k} \tag{7}$$

The sum of these values throughout the grid can then be used to approximate the TWV in the system at any time level. This approximation was carried out for the 'true' model, as well as for all of the correction schemes for the testing time series. A portion of these results is shown in Figure 12, as are the respective differences from the 'true' model.

All of the correction schemes demonstrate a significant improvement over the uncorrected MIKE 21 simulation, with the weighted local spatial regression method, again, giving the

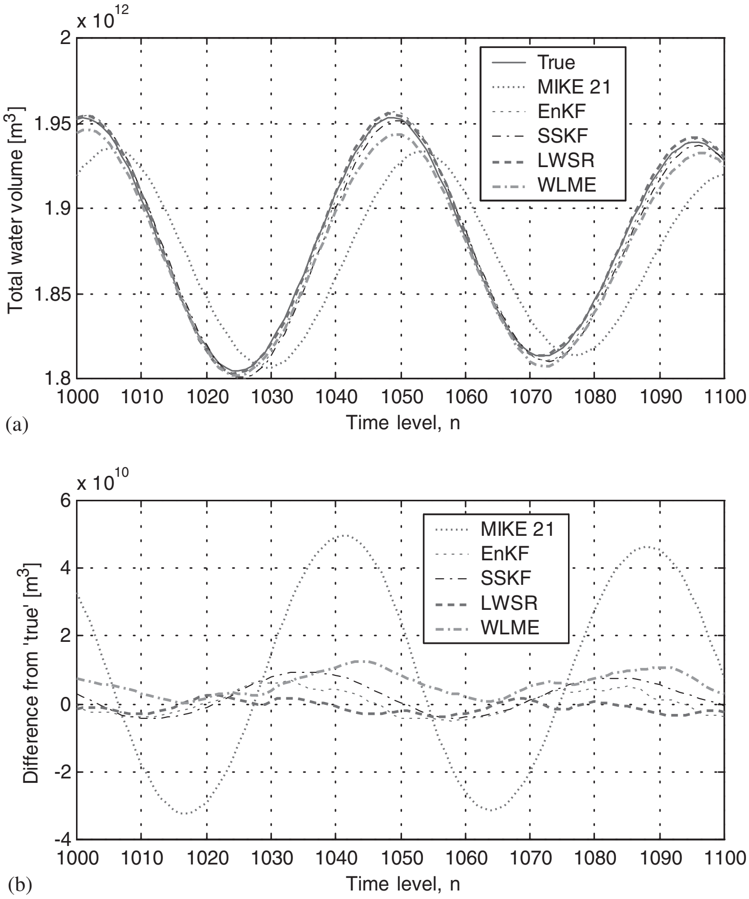

Figure 12. A portion of the testing time series of (a) total water volume, and (b) difference from the 'true' volume. The RMS errors for the entire testing duration were found to be $2.288\times10^{10}\,\mathrm{m}^3$, $3.067\times10^9\,\mathrm{m}^3$, $4.156\times10^9\,\mathrm{m}^3$, $1.467\times10^9\,\mathrm{m}^3$, and $5.225\times10^9\,\mathrm{m}^3$ for the MIKE 21, EnKF, SSKF, WLSR, and WLME model results, respectively.

best results. These are consistent with the overall correction errors (see Table III) in that the schemes with the best error correction in water levels also produced the best correction in total water volume. These results give evidence, if only empirically, that the corrected models are indeed also roughly conserving mass, since the TWV values after correction are quite close to the original 'true' values (which are based in part on solving the continuity equation).

9. COMPARISON UNDER SPARSE MEASUREMENT CONDITIONS

To test the performance of the various correction schemes under more sparse measurement conditions two of the three measurement points were removed (leaving only point C, see

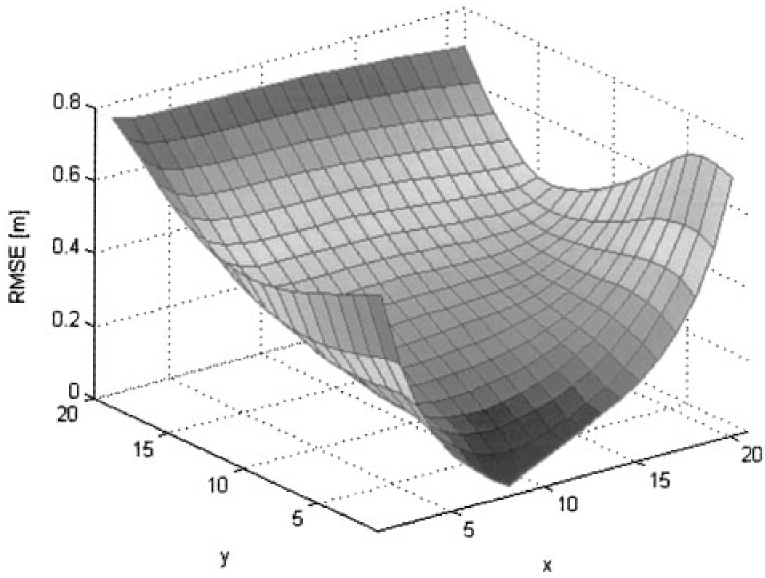

Figure 13. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using a local weighted spatial regression with a single measurement. The average value of this surface is 0.4017 m.

Figure 1). Simulations using the combined error model were then carried out under this much thinner coverage of measurement points. Figure 5, again, shows the spatially distributed error for the uncorrected MIKE 21 simulation.

9.1. Local weighted spatial regression

For this simple example, a local weighted spatial regression of measurements actually defaults to applying the lone measured error at all other grid points. This seems illogical in that it neglects obvious spatial variations, though it does clearly demonstrate limitations associated with this method. Larger model domains that are sparsely populated with measurement points would essentially result in the same type of behaviour throughout the domain. The spatial distribution of RMS error after correction is shown in Figure 13, having an average value of 0.4017 m. The error at and surrounding the measurement point is still quite low, however, the error grows rapidly with distance from the measurement point, actually increasing the error at the northern end of the bay. This is obviously due to the fact that errors from the measurement point continue to be applied even at points that are too far away to be considered relevant. Although this technique seems to work well in a model domain with a more dense population of measurement locations, it breaks down for obvious reasons as the measurements become more sparsely populated in space. Similar results might also be expected with more complex geometric configurations.

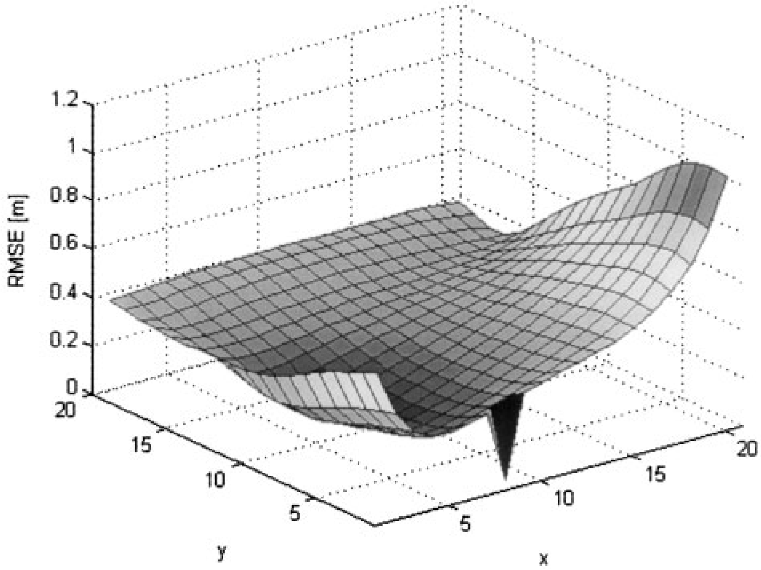

Figure 14. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using a weighted local model ensemble from a single measurement location. The average value of this surface is 0.3681 m.

9.2. *Weighted local model ensemble*

Next, a weighted local model ensemble was applied to correct errors at the current time levels. For this simple example with only one measurement location this technique defaults to the application of the lone local model at all other grid points. The configuration used for this local model is the same as in Table II (for measurement point C). The spatial distribution of RMS error for this simulation is shown in Figure 14, having an average value of 0.3681 m. The corrections throughout the domain are spatially more consistent than in the previous sub-section, as they are predicted based on system dynamics alone. Because the dynamics are generally related throughout the domain, the predicted error levels are allowed to vary at each grid point depending upon where it happens to be in its respective dynamic cycle.

9.3. *Kalman filtering*

For comparison the EnKF and SSKF were also employed to correct the combined errors using only measurement point C, resulting in spatially averaged RMS errors of 0.2614 and 0.2385m, respectively. These are significantly lower than the values of 0.4017 and 0.3681 m obtained in the previous two sub-sections. The spatial distribution for the SSKF simulation is shown in Figure 15. This demonstrates a clear superiority of the KF methods when faced with sparse measurements.

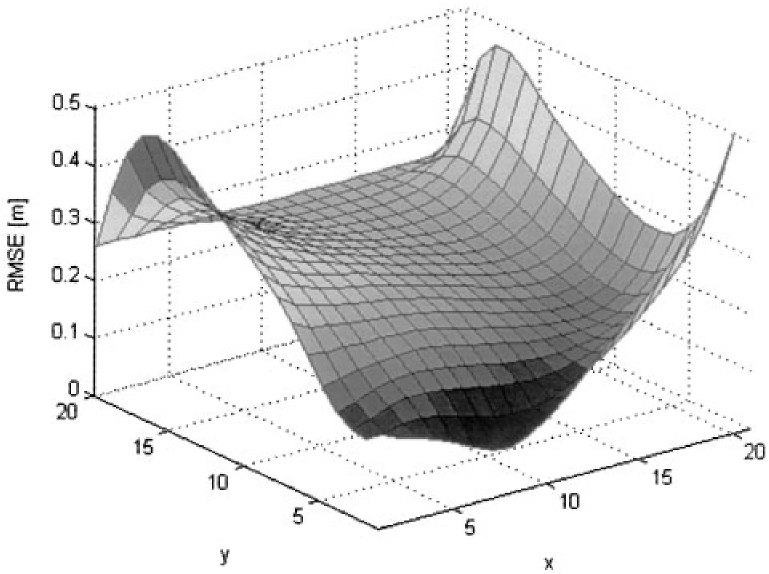

Figure 15. Spatial distribution of RMS error after the correction of the combined error model (at current time levels) using a steady state Kalman filter with a single measurement. The average value of this surface is 0.2385 m.

10. INTERMEDIATE SUMMARY AND STEPS TO FOLLOW

A summary of the spatially averaged RMS errors for the tests performed for the correction of errors at current time levels (using all three measurement locations) is provided in Table III. Clearly, all of the correction schemes were able to significantly improve the uncorrected MIKE 21 results.

Of the applied correction methods, the weighted local model ensemble consistently produced inferior results when compared to the other techniques. One reason is its inability to significantly correct wind-induced errors. This is due to the fact that these wind-induced errors are largely independent of the underlying tidal dynamics, thus leaving no clear input–output relationship for the LM forecasts. It therefore becomes difficult to obtain the same level of correction as observed with the other techniques, since the actual measurements are not in any way incorporated into the updating.

The Kalman filter type schemes provided the best results (though usually marginally) for all three tests where the errors were introduced individually. The local weighted spatial regression of measurements provided the best results for the correction of the combined error, which is probably the most realistic case. In general, however, the results between these methods are quite similar, and their differences would likely be considered largely insignificant. Under sparse measurement conditions (see Section 9), however, a significant shortcoming in the LWSR methodology becomes very apparent in that it is unable to recognize proper spatial regions for updating based on a given measurement.

Table IV. Summary of embedding parameters used for forecasting errors at the measurement locations. The forecasts were made using a uniformly weighted local averaging (degree zero) model with both water surface elevations and errors as inputs.

Measurement point	k	τ
A	7	[0 1 6 8 18 19]
B	8	[0 1 4 17 21 22]
C	7	[0 1 2 10 19 24]

Therefore, based on these results it does not appear that either of the introduced methodologies for re-distributing error measurements throughout a model domain are capable of replacing more standard Kalman filtering. Results using these alternate methods will continue to be displayed for comparison, however. Also noteworthy is the relatively good performance of the SSKF when compared to the much more computationally expensive EnKF. This would not be expected to hold in models having strongly non-linear dynamics, but in this simple example the results are quite impressive given the relative difference in computational demand (see Table I).

11. ASSIMILATION AT FUTURE TIME LEVELS

As previously mentioned, the problem with conventional data assimilation from a forecasting perspective is that the updated initial conditions quickly become washed-out. Thus, after a certain forecast horizon the predictions are no better than if made using an initially uncorrected model. This section provides results obtained at *future* time levels through the assimilating of error forecasts (made using local models, see Section 2). The subsequent assimilation methods used are, again, described in Section 4. For the KF and LWSR schemes this corresponds to a re-distribution of error forecasts, while for the WLME a forecast horizon greater than unity is used. All forecasts are made with the assumption of perfect knowledge of the incorrect boundary and wind-forcing conditions.

11.1. Local model configurations

To assimilate model error forecasts it is once again necessary to optimize the local model configurations. This was done previously for a forecast horizon, T, of one time step for the combined errors (see Table II). For the purposes in this section, however, a forecast much further into the future is desired. Therefore, the local models were again optimized with a GA as before, but this time for a forecast horizon equal to 16 time steps (i.e. 4 h). The resulting optimal configurations for each LM are shown in Table IV. Because the time series of errors are actually available at the measurement points these were also included as potential model inputs as it is generally an advantage to forecast using previously measured values from the output time series. The resulting optimal local models used both the time series of errors as well as water levels for inputs.

Actually, an optimal embedding could be constructed for each desired forecast horizon. This would, however, require a nearest neighbour search at *each* measurement location for *each*

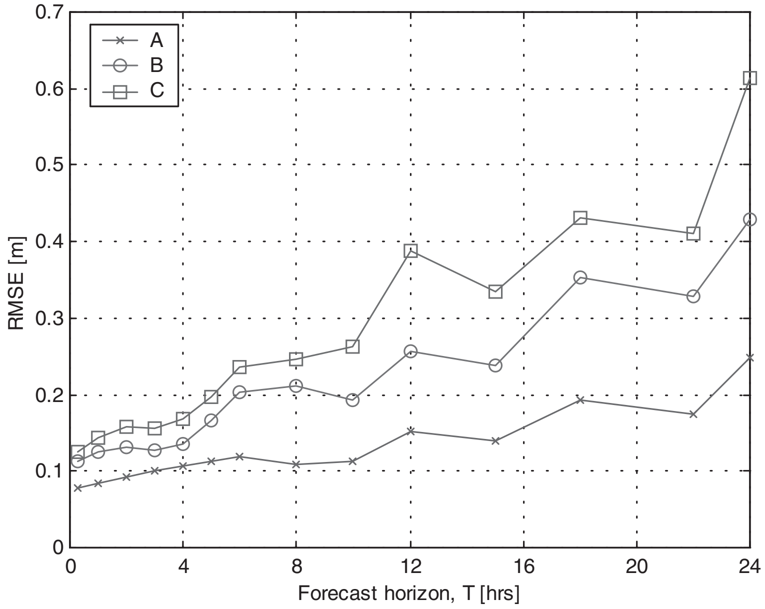

Figure 16. Error forecasting skill at measurement points A, B, and C for the testing duration.

lead-time. In the present example a single embedding was used for all forecast horizons, thus requiring nearest neighbours to be found only once per location for each series of forecasts.

The forecast skill for the errors at all three measurement points (based on their own data) can be seen in Figure 16 for a range of forecast horizons up to 24 h. A relative increase in error can be seen at 6-h intervals, which correspond to the highs and lows of the tidal cycle. The increase in error for all three local models is roughly linear with the forecast horizon, which is not surprising as these are not chaotic time series.

11.2. Local weighted spatial regression of forecasts

After determining the local model configurations, the error forecasts were assimilated as before using a LWSR. The spatial distribution of RMS error for a forecast horizon of 12 h (i.e. 48 time steps) is shown in Figure 17. The spatially averaged RMSE of 0.2973 m is still less than half that of the uncorrected MIKE 21 simulation, which, again, had an RMS error of 0.6256 m. This is quite remarkable since the updated initial conditions from a conventional Kalman filtering type scheme would likely be completely washed-out by this time. There are still noticeable areas of low error surrounding the measurement locations, but these are much less apparent than before (compare with Figure 6).

The time series of errors and corresponding water surface elevations are shown for point (11,11) in Figure 18, again for a forecast horizon of 12h. In general, the corrected time series are much better than the uncorrected MIKE 21 model, though there are portions where the error is actually increased. The phase error is essentially removed, however, there seems to be more difficulty in accurately correcting amplitude error this far in advance. This difficulty

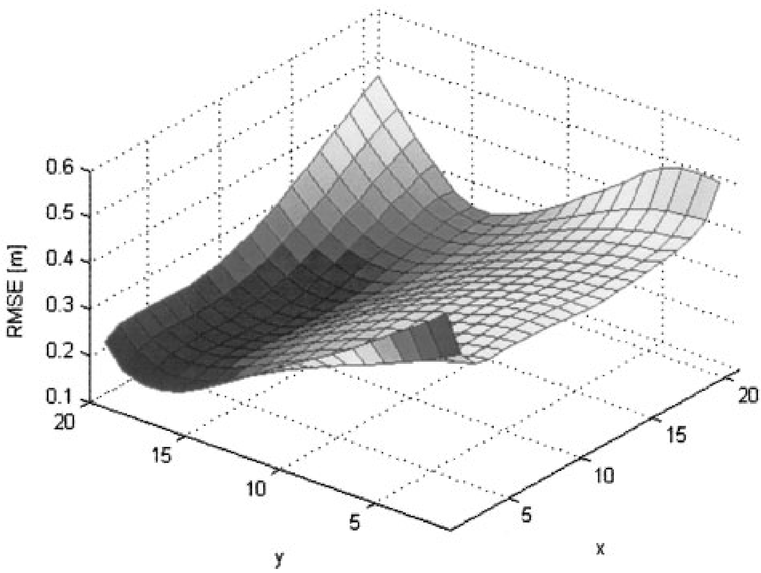

Figure 17. Spatial distribution of RMS error after the correction of the combined error model (at a forecast horizon of 12 h) using a local weighted spatial regression. The average value of this surface is 0.2973 m.

would almost certainly lessen if more training data were made available to create a more densely populated phase space. The remaining error at this particular location is still less than half that of the uncorrected time series, which is quite impressive given the extended forecast horizon.

11.3. Weighted local model ensemble

A WLME was also used to forecast the errors at every grid point. For these tests the local model configurations from Table IV were used (using only water surface elevations as model inputs). The spatial distribution of errors for a forecast horizon of 12 h is shown in Figure 19, having a spatially averaged value of 0.3291 m. The time series of errors and corresponding water surface elevations for point (11,11) are also shown in Figure 20. The correction at this particular point is, again, a significant improvement over the uncorrected MIKE 21 simulation. In fact, it is even better than was seen in Figure 18, though this is certainly not the trend throughout the entire model domain.

11.4. Hybrid approach: KF assimilation of LM forecasts

The final method presented in this study is the hybrid type model described in Section 4.4, which uses a Kalman filter to assimilate local model forecasts. This hybrid-type may be more theoretically sound and robust than is a simple local weighted spatial regression of measurements or forecasts. From Table III, the SSKF also provides a good approximation of the much more computationally expensive EnKF (at least in this model), and therefore was

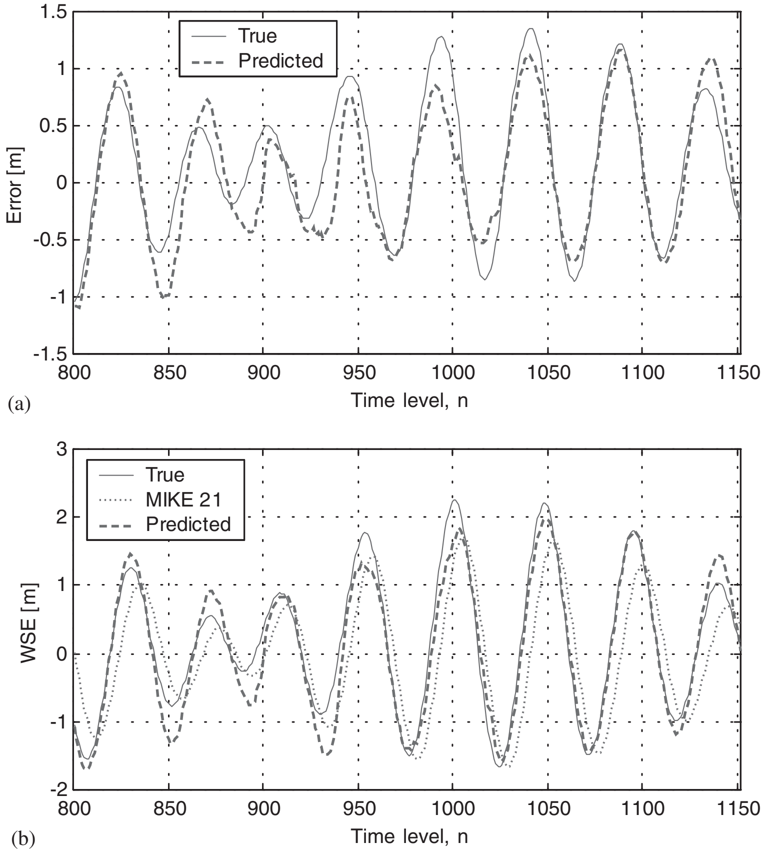

Figure 18. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) for corrections in the combined error model (for a forecast horizon of 12 h) using a local weighted spatial regression. The RMSE for this point after correction is 0.2785 m compared with 0.6277 m in the uncorrected MIKE 21 model.

used for the assimilation step in this work. In reality any KF variety could be applied in essentially the same manner.

The spatial distribution of RMS error for a forecast horizon of 12 h is shown in Figure 21, having an average value of 0.2709 m. The time series of errors and corresponding water levels for point (11,11) are likewise shown in Figure 22 (also for a forecast horizon of 12 h). For this extended forecast horizon this method produced the lowest spatially averaged error of the three correction techniques.

11.5. Summary

A summary of the resulting spatially averaged RMS errors for the various methods of error forecasting applied in this work is shown in Figure 23 for forecast horizons up to 24 h (i.e. 96 time steps). For comparison, results using the EnKF with various updating intervals were

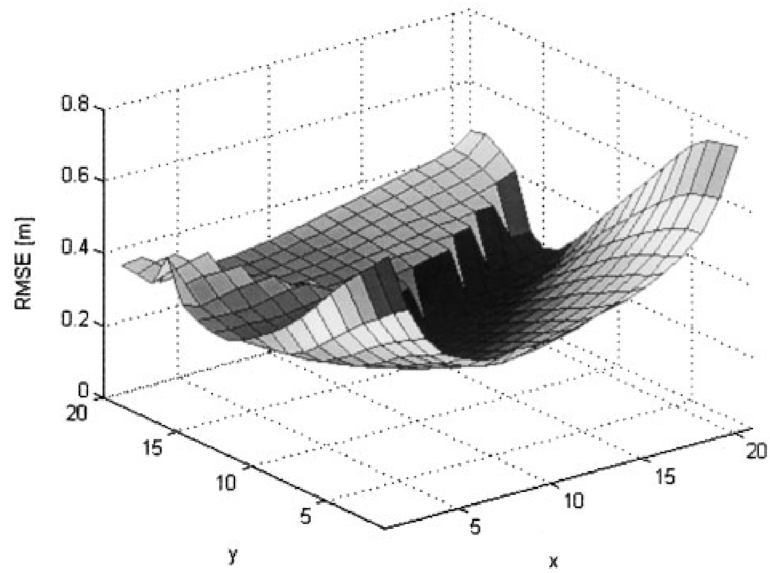

Figure 19. Spatial distribution of RMS error after the correction of the combined error model (at a forecast horizon of 12 h) using a weighted local model ensemble. The average value of this surface is 0.3291 m.

Table V. Spatially averaged errors using the EnKF with various updating intervals. These errors were then used to estimate the deterioration of updated initial conditions under the conventional methodology (sample calculations: 0.1772 m = (2 × 0.1694 m − 1 × 0.1616 m)/(2 − 1); 0.1771 m = ((95 − 1) ∗ 0.1772 m + 1 × (0.6256 m × 0.1772 m/0.6512 m))/95).

Updating interval	Time (h)	MAE (m)	RMSE (m)	Approximate EnKF forecasts			
				T	Time (h)	RMSE (m)	Adj. RMSE (m)
1	0.25	0.1325	0.1616	0	0	0.1616	0.1616
2	0.5	0.1394	0.1694	1	0.25	0.1772	0.1771
4	1	0.1527	0.1881	3	0.75	0.2068	0.2065
8	2	0.1918	0.2430	7	1.75	0.2979	0.2970
16	4	0.2706	0.3586	15	3.75	0.4742	0.4713
32	8	0.3893	0.4837	31	7.75	0.6088	0.6010
48	12	0.4543	0.5588	47	11.75	0.6339	0.6215
96	24	0.5023	0.6050	95	23.75	0.6512	0.6256

used to approximate the deterioration of the updated initial conditions (see Table V) under the conventional methodology. This is only an approximation, and the values were adjusted slightly to ensure that they did not actually rise above the level of the uncorrected MIKE 21 simulation. From Figure 23, the updated initial conditions are washed-out almost completely after 12 h (i.e. 48 time steps), after which the forecasts would be essentially the same as if an initially uncorrected model were used. Similar behaviour would be expected for any updating procedure that fails to correct at the forecasting time levels. The extended improvement with

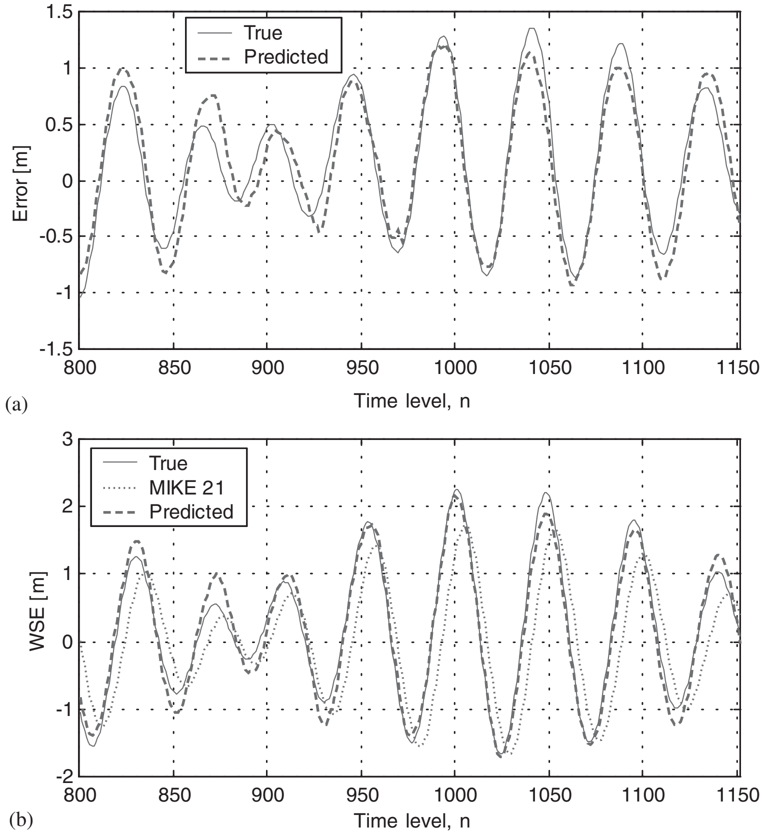

Figure 20. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) for corrections in the combined error model (for a forecast horizon of 12 h) using a weighted local model ensemble. The RMSE for this point after correction is 0.1780 m compared with 0.6277 m in the uncorrected MIKE 21 model.

the error forecasting methods is quite impressive, with the weighted local model ensemble again performing worse than the other two approaches for reasons previously explained. The weighted spatial regression performs the best of the three correction schemes up to a forecast horizon of about 6 h. Between a T of 6 and 10 h it is essentially equal to the performance of the hybrid model (i.e. local model forecasts with SSKF assimilation). After 10 h the hybrid model demonstrates the best performance. The more gentle slope associated with this hybrid model means that it is more robust against poor predictions than is the simple weighted spatial regression. Because of this and the other previously mentioned reasons (see again e.g. Section 10) it is felt that this is the most fundamentally sound of the methods presented in this paper.

To demonstrate the significance of these improvements, a few examples are discussed here. Firstly, we will consider an RMSE of 0.3 m. If only the initial conditions of a forecast are updated (as is conventionally done) this level of error will be exceeded after a forecast horizon

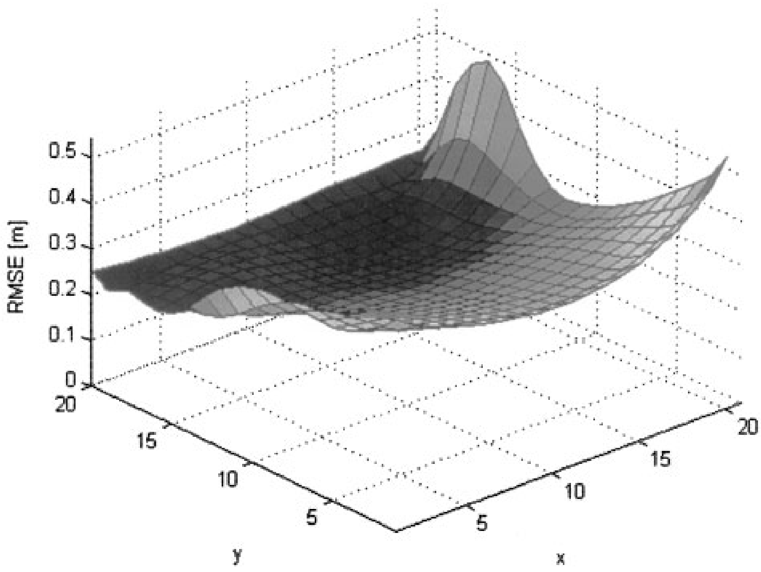

Figure 21. Spatial distribution of RMS error after the correction of the combined error model (at a forecast horizon of 12 h) using a steady state Kalman filter to assimilate local model error forecasts. The average value of this surface is 0.2709 m.

of only 2 h. By also correcting the model at future time levels through the assimilation of error forecasts this horizon was extended to 21 h (with the hybrid method) for an extension of 19 h (i.e. 76 time steps)! Similarly, in a comparison at a forecast horizon of 12 h (where the updated initial conditions are for all intents and purposes washed-out), the improvement in error over the uncorrected model goes from essentially zero to a very significant 35 cm!

12. DISCUSSION

The potential implications of these results are very exciting. Clearly, using any of the three approaches presented in this work allows for the errors throughout a model domain to be forecast and assimilated far into the future, without requiring a costly updating sequence. In this example the errors that were forecast and assimilated as far as 24 h ahead of the current time step still showed significant improvements over the uncorrected MIKE 21 simulation. Such extended corrections are simply not possible with standard KF updating type schemes (i.e. those that only update the initial conditions).

Data assimilation is a huge field, having applications in meteorology and oceanography, as well as in engineering. Usually, the ultimate goal in correcting the model is the improvement at not only the current time step, but also at subsequent forecasting time levels. The problem with conventional techniques, as has been shown here (see Figure 23), is that the corrected initial conditions are quickly washed-out. By also assimilating error forecasts the resulting model predictions can be significantly improved throughout a model far beyond the time it

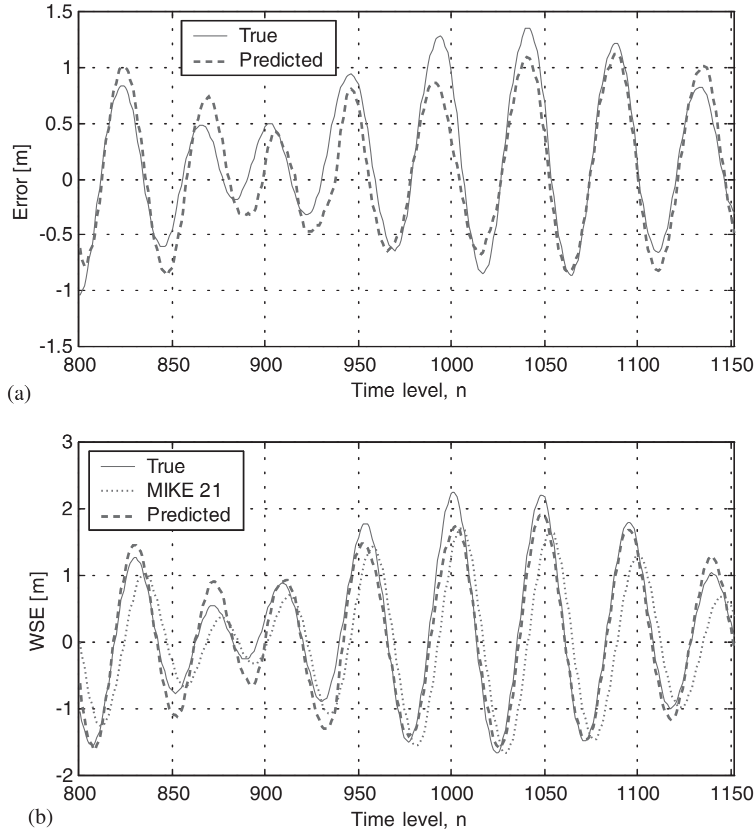

Figure 22. Time series of (a) observed and predicted errors, and (b) corresponding water levels at point (11,11) for corrections in the combined error model (for a forecast horizon of 12 h) using a steady state Kalman filter to assimilate local model forecasts. The RMSE for this point after correction is 0.2312 m compared with 0.6277 m in the uncorrected MIKE 21 model.

takes for the updated initial conditions to become washed-out, thus demonstrating a clear advantage over the traditional approach.

13. CONCLUSIONS

Clearly, the assimilation of error forecasts into deterministic models is capable of providing substantial improvements over the conventional methodology of only correcting the initial conditions. For the purposes in this paper three different methodologies have been tested: (1) a local weighted spatial regression; (2) a weighted local model ensemble; and (3) a hybrid approach using a SSKF to assimilate local model forecasts. The first two methods, though they still demonstrated impressive improvements over the conventional methodology, were found to have considerable limitations. A simple spatial regression has been shown

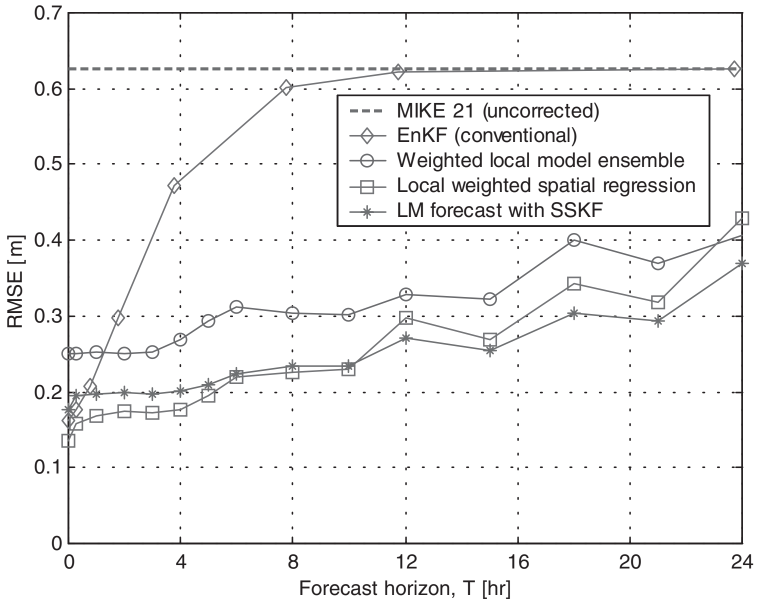

Figure 23. Summary of spatially averaged errors after correction of the combined error model for a variety of forecast horizons. The baseline uncorrected MIKE 21 error is also shown, as is the estimated deterioration of updated initial conditions under the conventional methodology.

not to be robust in sparse measurement conditions, and would also likely have difficulty in dealing with more complex geometric configurations. Likewise, the weighted local model ensemble technique was unable to significantly correct wind-induced errors, while also failing to incorporate knowledge of the most recent measurements into its predictions. It is unlikely that methods based on either of these premises would be able to replace existing Kalman filtering methods in the general case. A LWSR does, however, seem to work quite well under a dense enough coverage of measurements. Application of a WLME approach could also be used to provide significant corrections *without* requiring any real-time sensing. In general, a hybrid approach using KF techniques to assimilate local model forecasts appears to be the most robust scheme.

14. RECOMMENDATION

This work on data assimilation of error forecasts is based on synthetic data from a hydrodynamic model of a hypothetical bay. The recommendation for continued research in this area is to apply the methods presented here in a real-world case study, and compare the findings with those in this paper. This work is currently in progress.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Henrik Madsen (DHI—Water & Environment) for providing the MIKE 21 DA module used in this study. This work was in part funded by the Danish Technical Research Council (STVF) under the Talent Project No. 9800463 entitled "Data to Knowledge—D2K". Their support is greatly appreciated. More information on the project can be obtained through http://www.d2k.dk/.

REFERENCES

1. Babovic V, Keijzer M, Stefansson M. Optimal embedding using evolutionary algorithms. In *Proceedings of the 4th International Conference on Hydroinformatics*, Iowa City, 2000.
2. Keijzer M, Babovic V. Error correction of a deterministic model in Venice lagoon by local linear models. In *Proceedings of the* '*Modeli Complessi e Metodi Computazional Intensivi per la Stima e la Previsione*' *Conference*, Venice, 1999.
3. Babovic V, Keijzer M, Bundzelm M. From global to local modelling: a case study in error correction of determinstic models. In *Proceedings of the 4th International Conference on Hydroinformatics*, Iowa City, 2000.
4. 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.). Kluwer: Dordrecht, 1999.
5. Shamseldin AY, O'Conner KM. A nearest neighbour linear perturbation model for river flow forecasting. *Journal of Hydrology* 1996; **179**:353–375.
6. Bontempi G, Birattari M, Bersini H. Lazy learning for local modelling and control design. *International Journal of Control* 1999; **72**:643–658.
7. Takens F. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, vol. 898. Springer-Verlag: Berlin, 1981.
8. Abarbanel HDI. *Analysis of Observed Chaotic Data*. Springer-Verlag: New York, 1996.
9. Fuhrman DR. Data Assimilation and Error Prediction Using Local Models. *MSc Thesis*, International Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE), Delft, 2001.
10. Holland JH. *Adaptation in Natural and Artificial Systems*. University of Michigan Press: Ann Arbor, MI, 1975.
11. Babovic V. *Emergence*, *Evolution*, *Intelligence*: *Hydroinformatics*. Balkema: Rotterdam, 1996.
12. Houck CR, Joines JA, Kay MG. A genetic algorithm for function optimization: a MATLAB® Implementation. Technical Report: North Carolina State University, TR/NCSU-IE/95-09, 1995 Available: http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/.
13. Birattari M, Bontempi G, Bersini H. Lazy learning meets the recursive least-squares algorithm. In *Advances in Neural Information Processing Systems 11*, Kearns MS, Solla SA, Cohn DA (eds). MIT Press: Cambridge, MA, 1999.
14. Atkeson CG, Moore AW, Schaal S. Locally weighted learning. *Artificial Intelligence Review*, 1996, submitted.
15. 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.
16. Danish Hydraulic Institute. MIKE 21 HD (Hydrodynamic Module) Release 2.5, *User Guide and Reference Manual*. Danish Hydraulic Institute: Hørsholm, 1995.
17. Babovic V, Canizares R, Jensen HR, Klinting A. Neural networks as routine for error updating of numerical models. *Journal of Hydraulic Engineering* 2001; **127**:181–193.
18. Canizares R. On the Application of Data Assimilation in Regional Coastal Models. *PhD Thesis*. Balkema: Rotterdam, 1999.
19. WMO. Simulated real-time intercomparison of hydrological models. *Operational Hydrology Report No. 38*, Geneva, 1992.
20. Refsgård JC. Validation and intercomparison of different updating procedures for real-time forecasting. *Nordic Hydrology* 1997; **28**:65–84.
21. Kalman RE. A new approach to linear filtering and prediction theory. *Journal of Basic Engineering* 1960; **82D**:35–45.
22. Ghil M, Malonotte–Rizzoli P. Data assimilation in meteorology and oceanography. *Advances in Geophysics* 1991; **33**:141–266.
23. Miller RN, Ghil M, Gauthiez F. Advanced data assimilation in strongly non-linear dynamical systems. *Journal of the Atmospherical Sciences* 1994; **51**:1037–1056.
24. Canizares R, Heemink AW, Vested HJ. Sequential data assimilation in fully non-linear hydrodynamic model. In *Proceedings of the 2nd International Conference on Hydroinformatics*. Muller A (ed.). Balkema: Rotterdam, 1996; 463–470.
25. Canizares R, Heemink AW, Vested HJ. Application of advanced data assimilation methods for the initialisation of storm surge models. *Journal of Hydraulic Research* 1998; **36**:655–674.

26. Evensen G. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast the error statics. *Journal of Geophysical Research* 1994; **99-C5**:10 143–10 162.
27. DHI—Water & Environment. MIKE 21 DA (Data Assimilation Module), *Technical Documentation and User Guide*. DHI—Water & Environment: Hørsholm, 2001.
28. Yu X. Time series analysis using multivariate chaotic techniques. *MSc Thesis*. IHE: Delft, 2000.
29. Babovic V, Fuhrman DR. Data assimilation and error prediction using local models. *D2K Technical Report* 0401-2, 2001. Available: http://www.d2k.dk/Publications/index.htm.