Stoch Envir Res and Risk Ass (2004) 18: 228–244
DOI 10.1007/s00477-004-0181-9

ORIGINAL ARTICLE

J. V. T. Sørensen · H. Madsen · H. Madsen

Data assimilation in hydrodynamic modelling: on the treatment of non-linearity and bias

Abstract The state estimation problem in hydrodynamic modelling is formulated. The three-dimensional hydrodynamic model MIKE 3 is extended to provide a stochastic state space description of the system and observations are related to the state through the measurement equation. Two state estimators, the maximum a posteriori (MAP) estimator and the best linear unbiased estimator (BLUE), are derived and their differences discussed. Combined with various schemes for state and error covariance propagation different sequential estimators, based on the Kalman filter, are formulated. In this paper, the ensemble Kalman filter with either an ensemble or central mean state propagation and the reduced rank square root Kalman filter are implemented for assimilation of tidal gauge data. The efficient data assimilation algorithms are based on a number of assumptions to enable practical use in regional and coastal oceanic models. Three measures of non-linearity and one bias measure have been implemented to assess the validity of these assumptions for a given model set-up. Two of these measures further express the non-Gaussianity and thus guide the proper statistical interpretation of the results. The applicability of the measures is demonstrated in two twin case experiments in an idealised set-up.

Keywords Data assimilation · Kalman filter · Non-linearity measure · Bias · Hydrodynamic modelling

Abbreviations *EnKF* Ensemble Kalman filter · *RRSQRT* Reduced rank square root Kalman filter · *CEnKF* Central forecast ensemble Kalman filter · *BLUE* Best linear unbiased estimator · *MAP* Maximum a posteriori

J. V. T. Sørensen (✉) · H. Madsen
DHI Water and Environment,
Agern Allé 11, DK-2970 Hørsholm, Denmark
E-mail: jts@dhi.dk
Tel.: +45 45-169304
Fax: +45 45-169292

H. Madsen
Informatics and Mathematical Modelling,
Technical University of Denmark, Richard Petersens Plads,
Building 321, DK-2800 Kongens Lyngby, Denmark

1 Introduction

The state of coastal seas has an impact on a number of socio-economic issues such as fisheries, tourism and flood warning. Thus, estimating this state is of great importance. One way of solving the state estimation problem is by combining the theoretical knowledge encapsulated in numerical models with available data at or around the time of interest. Such an approach is generally known as data assimilation.

One particular branch within data assimilation deals with sequential state estimation based on a Kalman filter approach. However, the optimality of the Kalman filter can not be preserved without imposing linearisations and constraints on the size of the state space, which are severe for the application in a realistic set-up of a hydrodynamic model. Thus, sub-optimal schemes have been introduced that attempt to reduce computational requirements by simplifying the model propagation operator and/or reducing the degrees of freedom in the model covariance estimation.

The use of a simplified process description was investigated in (Dee 1991). Such an approach is case dependent and relies on the validity of the rather strong dynamical approximations. Alternatively, the error covariance calculation can be performed on a coarser grid (Fukumori and Melanotte-Rizzoli 1995). This implies an assumption about the main model variability to be at larger scales than the model resolution. Finally, the model operator can be represented with a reduced rank approximation by applying e.g. a singular value decomposition (Cohn and Todling 1996). The simplified process description, the coarse grid approximation and the model reduction approach are all examples of applying a regularised model operator.

A different approach to speeding up a sub-optimal Kalman filter is to work with a simplified error covariance representation. One such approximation is to assume that the error covariance is in steady state (Heemink, 1986). This often works very well despite the strong assumption and has the advantage of being operational in many real time hydrodynamic forecast systems, (Cañizares et al. 2001; Heemink et al. 1997). The reduced rank square root Kalman filter (Verlaan and Heemink 1997) obtains a time varying approximation of the error covariance matrix. Based on the extended Kalman filter, the covariance is continuously approximated by its leading eigen sets. This leads to a rather smooth Kalman gain, but its application is limited when very strong non-linearities are present and only few measurements are available, (Verlaan and Heemink 2001). Alternatively the covariance can be calculated using a Monte Carlo technique as introduced in (Evensen 1994). This approach handles even strong non-linearities well, but at the price of rather noisy error covariance estimations. A larger ensemble size reduces this problem, but at the cost of an increased computational burden. Finally, hybrids of regularised model operators and approximate error covariance representations can be formed. As an example, (Sørensen et al. 2002) successfully combined the ensemble Kalman filter with a depth averaged model operator for generation of a steady Kalman gain to be used in a 3D hydrodynamic model.

Each of the sub-optimal schemes is based on a set of assumptions such as model linearity, a simplified description of the error covariance and an unbiased model operator. Often the assumptions are merely stated or even implicit in order to focus on other important issues. The schemes are typically validated by application in one or two test cases, where performance is rather good. Means of assessing the general validity of the underlying assumptions often lack and the filter performance when they are violated are generally not discussed for the different schemes. We attempt to contribute to this matter. The main aim of this paper is to highlight the assumptions of different schemes and analyse the validity of these assumptions under various conditions. In order to perform this analysis, different performance measures are introduced.

In Verlaan and Heemink (2001) a non-linearity measure is introduced, which can be used to assess the validity of the assumption of the model operating in a regime, which is weakly non-linear at worst. In this paper a simplified version of the measure is implemented in a 3D hydrodynamic model and the performance of two estimation schemes based on a central forecast is examined with respect to variation of this non-linearity measure and compared to an ensemble forecast. The Gaussianity of a solution affects the valid interpretation of the results and thus two non-Gaussianity measures are introduced. When the model noise is Gaussian, these simultaneously provide alternative non-linearity measures. Finally, the model bias is used to characterise the filter performance under various error structure assumptions. It is very important to understand the filter performance when actual errors are not well captured by the assumed error structure. This aspect will be considered in the paper.

Section 2 introduces the considered coastal ocean system, which is described by a stochastic hydrodynamic model. Section 3 discusses state estimation with particular emphasis on issues of application to a hydrodynamic model. The propagation of model error covariance is discussed in Sect. 4 along with a presentation of the ensemble Kalman filter, the central ensemble Kalman filter and the reduced rank square root Kalman filter. This section also describes the characteristics of each filter. In Sect. 5 measures of non-linearity, non-Gaussianity and bias, which will be applied to assess the validity of filter assumptions, are introduced. The simulation study is described in Section 6 and a discussion of the results is given in Sect. 7. Finally, Sect. 8 concludes the paper.

2 The stochastic state space model

The physical system under consideration consists of hydrodynamic flow in bays, estuaries, coastal regions and shelf seas. The body of water evolves according to the laws of internal dynamics of a fluid and its interaction with the atmosphere and the solid earth through the sea floor. Among the processes encompassed by this system are tidal waves, wind induced coastal upwelling, eddy formation and turbulence.

The continuity and Navier-Stokes equations state the conservation of mass and momentum in a continuum like the considered system. By developing mathematical, physical and numerical approximations of the system dynamics, the problem of estimating and predicting the state of the coastal ocean can be solved. This theoretical approach has lead to the advance of a range of numerical models, which are now routinely applied to solve a number of scientific and engineering problems. One such numerical modelling system is MIKE 3.

The MIKE 3 hydrodynamic model is part of a general finite difference modelling system and is designed to simulate non-linear, unsteady three-dimensional flows. It is developed at DHI Water and Environment (DHI 2001) and has been successfully applied to various scientific and engineering applications in domains with scales ranging from meters to thousands of kilometres (Øresundskonsortiet 1998; Vested et al. 1998; Erichsen and Rasch 2002).

MIKE 3 utilises a finite difference technique, and thus provides the discrete time evolution of the model variables defined on a mesh in the domain under consideration. Details of the finite difference scheme can be found in the scientific reference manual (DHI 2001). For the purpose of the problem at hand it is sufficient to acknowledge that the entire state of the model is uniquely determined by the variables $\zeta(t), \zeta(t-1/2)$, $\zeta(t-1), v_x(t), v_y(t-1/2), v_y(t+1/2)$ and $v_z(t-1/4)$ when the density of the water is assumed constant. The variable ζ is the water level, (v_x, v_y, v_z) are the three velocity components and t is the time index.

With knowledge of the initial conditions, sources and sinks as well as boundary conditions represented by surface elevation at open boundaries and wind velocity and pressure at the sea surface, MIKE 3 calculates a solution to the finite difference equations. Thus, an estimate of the state of the fluid is given at discrete temporal and spatial intervals and the state at time $t+1$ is completely determined by the state at time t and the forcing terms embedded in the sources and sinks and boundary conditions. Thus, let Φ_D be the model operator representing the approximate finite difference equations, u_D the forcing defined at a snapshot in time projected onto the mesh and $x_D(t)$ the model state at time t. The discrete deterministic model can then be expressed as,

$$x_D(t+1) = \Phi_D(x_D(t), u_D(t)) \tag{1}$$

The hydrodynamic model attempts to construct the best possible estimate of the state of the system within the constraints of the model structure imposed. However this estimate is based on a model and forcing terms, which we know are uncertain, but often we will have some knowledge of the second order statistical properties of the errors, $\delta(t)$. Thus, the discrete model can be extended to a stochastic model, propagating a state that is now a stochastic variable characterised by its second order statistical properties rather than the deterministic estimate in Eq. (1).

For the hydrodynamic part of a continental shelf ocean model, a main source of error comes from inaccurate meteorological and open boundary forcing. Thus, in order to simplify the error description, it is assumed that wind forcing and water level at open boundaries are the sole sources of error. No initial errors are assumed, but the model is allowed a spin-up period to propagate the forcing induced error throughout the system.

To reduce the computational requirements, errors can be defined on a coarser grid, G2, than the forcing grid, Gl, and thus an interpolation operator, Λ, is introduced. In general any linear reduced rank representation can be expressed by Λ, e.g. refer to Heemink (1986) and Cañizares et al. (2001) for this approach. If the errors in the forcing terms can be assumed to be uncorrelated in time, then MIKE 3 can be generalised to a stochastic model operator, Φ_{M3}:

$$x_{M3}(t+1) = \Phi_{M3}(x_{M3}(t), u_D(t) + \Lambda\delta(t)) \tag{2}$$

where

$$x_{M3}(t) = \begin{pmatrix} \zeta(t) \\ \zeta(t-1/2) \\ \zeta(t-1) \\ v_x(t) \\ v_y(t-1/2) \\ v_y(t+1/2) \\ v_z(t-1/4) \end{pmatrix} \tag{3}$$

The only difference between x_{M3} and x_D is that the elements in x_{M3} are stochastic.

However, the errors in the forcing terms are usually correlated in time. Thus it makes sense as a first approximation to construct an augmented state vector, by including the error as modelled by a first order autoregressive model (AR(1)),

$$\delta(t) = \alpha\delta(t-1) + \varepsilon_u(t) \tag{4}$$

where $\varepsilon_u(t)$ is an N_B-dimensional i.i.d. variable with zero mean and known covariance, $Q_u(t)$. In the physical system under consideration a first order autoregressive process typically explains 80–90% of the variance. If necessary it is straightforward to formulate more general correlation models still adhering to the state space description.

Finally, by assuming the error to originate from the forcing, using the augmented state vector with coloured error description as expressed by Eq. (4) and allowing for a noise to be defined on a reduced space (e.g. a coarse grid), the following stochastic finite difference model is obtained:

$$\begin{aligned} x(t+1) &= \begin{pmatrix} x_{M3}(t+1) \\ \delta(t+1) \end{pmatrix} \\ &= \begin{pmatrix} \Phi_{M3}(x_{M3}(t), u_D(t) + \Lambda\delta(t)) \\ \alpha\delta(t) + \varepsilon_u(t) \end{pmatrix} \end{aligned} \tag{5}$$

or

$$x(t+1) = \Phi(x(t), u_D(t)) + \varepsilon(t) \tag{6}$$

where Φ is the augmented model operator and

$$\varepsilon(t) = \begin{pmatrix} 0 & \varepsilon_u(t) \end{pmatrix}^{\mathrm{T}} \tag{7}$$

is a N-dimensional i.i.d. variable with zero mean and covariance $Q(t)$. Equation (6) is called the system equation and is the actual stochastic representation of MIKE 3 that will be used subsequently. The dimension of x is designated N. Note that the equation actually has additive noise even though the error is defined to enter through the forcing terms, i.e. ε enters linearly, but has a non-linear effect on x_{M3}.

Tidal gauge measurements provide an additional source of information about the state of the system. They are characterised by a high temporal resolution, but the gauges are very sparsely distributed in space. Tidal gauge sensors typically have a random instrumentation error with a standard deviation less than 1 centimetre. Let the number of measurements be designated N_m.

For the purpose of data assimilation, we need to relate the measurements to the state vector x. By doing this, a model representation error is introduced. As an example, if the model resolution is 9 nautical miles, then the model variable that would typically represent the observation is the water level averaged over the grid box at the position of the gauge, which clearly may deviate from the point measurement. Representation error is typically the main error source that needs to be considered when using tidal gauge data for modelling purposes, (Fukumori 1999). Let x be the stochastic model state defined by Eq. (6). It is assumed that the observation can be expressed as a linear combination represented by the $N_m \times N$ matrix $C(t)$ of the state variables, and an additive zero mean Gaussian distributed observational error, $\xi(t)$, with covariance $R(t)$.

$$z(t) = C(t)x(t) + \xi(t) \tag{8}$$

This is called the measurement equation. The rows of $C(t)$ will in many cases consist of zeroes and a single one. It is assumed that the separate tidal gauge stations have both spatially and temporally uncorrelated errors. Thus the measurement error covariance can be expressed as

$$R(t) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{N_m}^2 \end{pmatrix} \tag{9}$$

Tide gauge observations are performed by independent instruments and hence the instrumental errors are independent. However, their main error source is probably representation error, (Fukumori 1995 and Søerensen et al. 2002) and may depend on the system state and thus be correlated. This effect is not present in ideal scenarios as considered herein.

3 State estimation

In the previous section, stochastic descriptions of both model and measurements have been presented. The description of the system is provided by the system Eq. (6), while the measurements are described by the measurement Eq. (8). Now we will pay attention to how the best estimate of the true oceanic state can be obtained based on the available information from these two sources of information. The model gives a state estimate with high temporal and spatial resolution, but the values are hampered by the accumulation of errors. Measurements give an alternative estimate that is usually more certain when and where an observation is made, but they are sparsely distributed in space and time. The two sources of information are complimentary and both ought to be included in the state estimate.

At a given point in time consider the stochastic state, x, derived from the model, and an observation z. Note that by restricting ourselves to a single time step, the propagation and the estimation problems are separated. First, we deal with estimation. One approach is to use the information about the oceanic state provided by the model probability density function (pdf) to give e.g. a maximum likelihood (ML) estimate. However, including the information provided by available measurements will improve the estimate. By using a Bayesian approach the resulting pdf of the estimate can be calculated as the conditional probability of x given the data, z,

$$f(x|z) = \frac{f(z|x)f(x)}{f(z)} = \frac{f(z|x)f(x)}{\int f(z|x)f(x)\mathrm{d}x} \tag{10}$$

The value x^a of x that maximises $f(x|z)$ is the maximum a posteriori (MAP) estimate of x. It is common to work with the logarithmic transformation of Eq. (10) in order to ease the arithmetic expressions.

$$\log f(x|z) = \log f(z|x) + \log f(x) - \log B \tag{11}$$

In Eq. (11) B is an abbreviation for the denominator of Eq. (10). This last term does not affect the behaviour of extreme values because it only depends on the data. Thus, if the distributions $f(x)$ and $f(z|x)$ are known, then the optimum can be found. However, these distributions are in general unknown and further assumptions must be imposed in order to progress.

It will now be assumed that the distributions $f(x)$ and $f(z|x)$ are Gaussian with means x^f and Cx and known covariance matrices P^f and R respectively. Thus,

$$f(x) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(P^f)}} \times \exp\left(-1/2\left[\left(x - x^f\right)^T \left(P^f\right)^{-1}\left(x - x^f\right)\right]\right) \tag{12}$$

$$f(z|x) = \frac{1}{(2\pi)^{N_m/2}\sqrt{\det(R)}} \times \exp\left(-1/2\left[(z - Cx)^T(R)^{-1}(z - Cx)\right]\right) \tag{13}$$

where N and N_m are the sizes of P^f and R respectively. By substituting Eqs. (12) and (13) into Eq. (11) and differentiating with respect to x the maximum can be found. This provides the same solution as the minimisation in a least squares approach. For further elaboration on the least square solution refer to Wunsch (1996) and Jazwinski (1970).

The MAP estimator now reduces to,

$$x^a = x^f + K(z - Cx^f), \quad P^a = P^f - KCP^f \tag{14}$$

$$K = P^f C^{\mathrm{T}}\left[CP^f C^{\mathrm{T}} + R\right]^{-1} \tag{15}$$

The matrix P^a is the error covariance of the estimated state, x^a. Since the estimator can alternatively be derived from the least square approach as the best linear unbiased estimator (BLUE) it will always supply the minimum variance estimate under the assumption of a linear and unbiased estimate for any distribution. In the remainder of this paper we will refer to Eqs. (14) and (15) as the BLUE estimator. Note that the problem of finding the probability density of the state variables has been reduced to estimating its a posteriori mean and covariance.

The assumption about Gaussianity is certainly more an operational assumption than a justified one. Particularly, model error sources are generally far from being Gaussian. However, assimilation schemes are traditionally based on the BLUE, which is only a minimal variance estimator under the Gaussian assumption. We believe an improved estimation technique is essential in the further development of assimilation techniques. An operational Bayesian approach as discussed in Christakos (2002) could provide an interesting alternative.

The discussion above has focused on the estimate when a prior model estimate is available at the time step of a new measurement. We will now extend the discussion to encompass available information up until the time of the latest observation. In general the approach can be extended to a sequential estimator with each estimate having a similar MAP or BLUE interpretation based on all past and present measurements (Maybeck 1979). If the BLUE estimator is used with a linear model for propagation of the mean and the error covariance matrix in between updates and all variables are Gaussian distributed, then the classical Kalman filter is obtained. The Kalman filter has in many cases been the starting point in the literature and necessary generalisations have subsequently been imposed, (e.g. Verlaan and Heemink 1997). Here, we present the general problem and impose certain simplifications that allow a solution to be found on available computational resources.

So far the origins of the mean and error covariance estimates of model (system error) and measurement variables (measurement error) were avoided in order to pay attention to the estimator. However, their construction is one of the major difficulties in sequential data assimilation. For tidal gauge data the measurements at separate stations can be assumed to have no error correlation and $R(t)$ becomes diagonal as expressed by Eq. (9). This allows for an efficient sequential updating of data from different tidal gauge stations within the same time step (Madsen and Cañizares 1999). The values of the diagonal elements are set, based on reflections on the error sources discussed in Sect. 2. Estimates of $x^f(t)$ are typically based on the composite hydrodynamic and the AR(1) model in Eq. (6). The error covariance matrix, $P^f(t)$, on the other hand, has been estimated by a number of different approaches in the literature. These range from solving the Riccati difference equation (Fukumori et al. 1993) to geometric or physical assumptions (Fox et al. 2000), and transient propagation of $P^f(t)$ by the hydrodynamic equations (Verlaan and Heemink 1997; Evensen. 1994). The latter approach is pursued in this work and is treated further in the next section. Among its strengths it accommodates the calculation of non-linearity measures.

Anyone of the approaches above requires a proper definition of system noise, $Q(t)$. The error in open boundary water level or wind velocity will typically be correlated in space. The spatial error correlation patterns are here assumed to be isotropic for each error source and can thus be described by a standard deviation and a spatial correlation scale corresponding to the distance at which the correlation is 0.5. Further, because of the noise definition in Eq. (7) only the lower right $N_B \times N_B$ portion of $Q(t)$ is non-zero. The specification of $Q(t)$ poses quite a problem in real applications. Dee (1995) suggested a maximum likelihood approach for estimating the system noise from measurements. However, this is quite costly and requires 2–3 orders of magnitude of data more than the number of error parameters to be estimated. An alternative solution to the problem should be adaptive in nature, because of the generally time-varying and state dependent errors. This could be very interesting to test in ideal scenarios like the one discussed in the present paper, but they probably would be too computationally demanding for real applications.

It makes filtering seem less complex if we remind our selves that no matter what approach is taken, the procedure basically consists of two elements: Updating and propagation of model state estimates and its error covariance. We can pick and choose among various estimators for the updating and various propagation schemes, but in all cases we propagate model information in between measurement times and update the state instantaneously whenever a new measurement becomes available. The resulting updated state estimate can then be propagated onwards.

4 Error covariance propagation

This section will describe various ways of propagating the model mean and error covariance in time. The general approach for time evolution in stochastic differential equations is based on dynamic stochastic prediction (Evensen, 1994). The starting point there is a stochastic differential equation with additive noise generated by a Wiener process. The general solution is given by the Fokker-Planck equation and consists of the full probability density function of the state. In our approach, the stochastic extension was introduced in Eq. (6) at the level of the actual numerical implementation in order to make clear the physical, mathematical and numerical assumptions that we ideally attempt to capture. For both approaches, the final aim is to provide accurate estimates of the state by propagating information about the probability density in time when called for by the estimator. In both ensemble based filters presented in Sect. 4.1 and 4.2 the pdf is approximated by a finite ensemble. However, for the reduced rank square root kalman filter presented in Sect. 4.3, the propagation is restricted to first and second order statistics.

The treatment will be restricted to expressing the various moments of the state vector.

Assuming the noise sequence, $\varepsilon(t)$, to be a zero mean i.i.d. random variable, cf. Eq. (6), then the expectation of $x(t+1)$ is:

$$E\{x(t+1)\} = E\{\Phi(x(t), u_D(t))\} \qquad (16)$$

Even this first order moment is impossible to evaluate exactly for a non-linear forecast model, such as MIKE 3. Calculation of the second order moment demands even more resources for a good approximation and so forth. However, various approximate methods can be imposed, which makes the error covariance propagation manageable. In the following two different ways of approximation, which are both implemented in MIKE 3 are presented. The Ensemble Kalman Filter (Evensen 1994) is based on Monte Carlo theory, while the reduced rank square root Kalman filter (Verlaan and Heemink 1997) uses a truncated Taylor series and a square root error covariance representation.

4.1 Ensemble Kalman filter

In the Ensemble Kalman filter, (EnKF), an ensemble of possible states represents the statistical properties of the state vector. Each of these vectors is propagated according to the dynamical system subjected to model errors, and the resulting ensemble then provides estimates of the forecast state vector and the error covariance matrix. In the measurement update, the Kalman gain matrix obtained from Eq. (15) is applied for each of the forecast state vectors. To account for measurement errors, the measurements are represented by an ensemble of possible measurements (Burgers et al. 1998). The resulting updated sample provides estimates of the updated state vector and the associated error covariance matrix. The following subsections provide the mathematical detail of the scheme.

4.1.1 Forecast

Each member of the ensemble of M state vectors is propagated forward in time according to the dynamics of the augmented system in Eq. (6) and the specified model error, i.e.

$$x^f_{i,t} = \Phi(x^a_{i,t-1}, u_{D,t}) + \varepsilon_{i,t}, \quad i = 1,2,\ldots,M \tag{17}$$

where the model error $\varepsilon_{i,t}$ is randomly drawn from a Gaussian distribution with zero mean and $N \times N$ covariance matrix Q_t which represents the system noise. The time step index t has now become a subscript to shorten notation. An estimate of the state vector (forecast) is calculated as the average of the ensemble members, i.e.

$$x^f_t = \bar{x}^f_t = \frac{1}{M}\sum_{i=1}^{M} x^f_{i,t} \tag{18}$$

The error covariance matrix of the forecast is estimated from the ensemble as

$$P^f_t = S^f_t (S^f_t)^{\mathrm{T}}, \quad s^f_{i,t} = \frac{1}{\sqrt{M-1}}(x^f_{i,t} - \bar{x}^f_t) \tag{19}$$

where $s^f_{i,t}$ is the ith column in S^f_t.

4.1.2 Update

An ensemble of size M of possible measurements is generated

$$z_{i,t} = z_t + \xi_{i,t}, \quad i = 1,2,\ldots,M \tag{20}$$

where z_t is the actual measurement vector, and $\xi_{i,t}$ is the measurement error that is randomly generated from a Gaussian distribution with zero mean and covariance matrix R_t.

Each ensemble member is updated according to the updating scheme in Eq. (14). The updated state vector and error covariance matrix are derived from Eq. (18) and (19). When the data assimilation is based on in-situ measurements that are sparsely represented in space, the full error covariance matrix in Eq. (19) does not need to be calculated. In this case, the measurement matrix C_t only has a few non-zero elements and only the columns in P^f_t that correspond to these non-zero elements in C_t have to be calculated. Furthermore, since it is assumed that measurement errors are uncorrelated, a sequential updating algorithm that processes one measurement at a time can be implemented and the matrix inversion in Eq. (15) can be avoided.

The sequential updating algorithm reads (Chui and Chen 1991),

$$x^a_{t,j} = x^a_{t,j-1} + k_{t,j}\left(z_{t,j} - c_{t,j}x^a_{t,j-1}\right), \qquad j = 1,\ldots,N_m, \quad x^a_{t,0} = x^f_t \tag{21}$$

where N_m is the number of measurements, $c_{t,j}$ is the jth row in the measurement matrix C_t, $c_{t,j}x^a_{t,j-1}$ is the element in the state vector that corresponds to the measurement $z_{t,j}$, (i.e. $z_{t,j} - c_{t,j}x^a_{t,j-1}$) is the model deviation from measurement j), and $k_{t,j}$ is a Kalman gain vector corresponding to measurement j. The Kalman gain vector is given by

$$k_{t,j} = \frac{S^a_{t,j-1}h_{t,j}}{h^{\mathrm{T}}_{t,j}h_{t,j} + \sigma^2_j}, \quad h_{t,j} = (S^a_{t,j-1})^{\mathrm{T}}c^{\mathrm{T}}_{t,j}, \quad S^a_{t,0} = S^f_t \tag{22}$$

where the numerator is the covariance between the measurement j and the state vector and the denominator is the sum of the variance of measurement j and the predictive variance of the measurement. In the EnKF the sequential updating scheme is applied for each ensemble member, and after each measurement update $S^a_{t,j}$ is calculated from the ensemble cf. Equation (19). Remember that the scheme encompasses both the MIKE 3 part and the auto regressive augmented part of the state vector. For an infinite number of ensembles (∞-EnKF) and correct error description this scheme will provide an optimal estimate and is in this sense asymptotically optimal.

4.2 Central ensemble Kalman filter

A second version of the EnKF that uses a central forecast instead of the ensemble average forecast for x^f_t has also been implemented for the purpose of calculation of the non-linearity measures discussed in Sect. 5. This filter is referred to as the Central Ensemble Kalman Filter (CEnKF). A new central state vector, x^c_t, is introduced.

At initial time t_0 it is set equal to the mean estimate of x_t^a and subsequently it is propagated and updated like any other of the ensemble members, i.e:

$$x_0^{c,a} = \overline{x_0^a} \tag{23}$$

$$x_t^{c,f} = \Phi(x_{t-1}^{c,a}, u_{D,t}) \tag{24}$$

The error covariance propagation is still centred at the ensemble forecast and hence the Kalman gain is exactly the same as in the EnKF - only the state estimate is different. The computational requirements are similar to those of the EnKF, requiring only one more model execution.

4.3 Reduced rank square root Kalman filter

The reduced rank square root Kalman filter (RRSQRT) is based on the extended Kalman filter formulation in which the error propagation is calculated using a statistical linearisation of the model equation based on a first order Taylor series expansion.

4.3.1 Forecast

In the case of a coloured system noise process as assumed in Eq. (6), the forecast step is given by

$$x_t^f = \Phi(x_{t-1}^a, u_{D,t}) \tag{25}$$

$$P_t^f = F_t P_{t-1}^a F_t^{\mathrm{T}} + Q_t \tag{26}$$

$$F_t = \left.\frac{\partial\Phi}{\partial x}\right|_{x=x_t^f} \tag{27}$$

The RRSQRT approximation of the extended Kalman filter uses a square root algorithm as well as a lower rank approximation of the error covariance matrix. Denote by S_{t-1}^a the approximation of rank M of the square root of the error covariance matrix P_{t-1}^a. The propagation of the error covariance matrix is then given by

$$S_t^f = \left[F_t S_{t-1}^a \middle| Q_t^{1/2}\right] \tag{28}$$

where $Q_t^{1/2}$ is the $N \times N_B$-dimensional square root of Q_t. The matrix S_{t-1}^a has M columns where M is chosen much smaller than the dimension of the state vector. To calculate the derivatives in F_t a finite difference approximation of $\Phi(\cdot)$ is adopted as follows,

$$\begin{aligned}(F_t S_{t-1}^a)_i = {} & [\Phi(x_{t-1}^a + S_{i,t-1}^a, u_{D,t}) \\ & - \Phi(x_{t-1}^a, u_{D,t})], \quad i = 1, \ldots, M\end{aligned} \tag{29}$$

where $s_{i,t-1}^a$, is the ith column of S_{t-1}^a. Thus, the propagation of the error covariance matrix requires M model integrations.

The propagation step in Eq. (28) increases the number of columns in the error covariance matrix from M to $M + N_B$. In order to reduce the number of columns and hence keep the rank of the error covariance matrix constant throughout the simulation, a lower rank approximation of S_t^f is applied by keeping only the M leading eigenvectors of the error covariance matrix. The reduction is achieved by an eigenvalue decomposition of the matrix $(S_t^f)^{\mathrm{T}} S_t^f$. For full details refer to (Cañizares 1999). For a proper reduction S_t^f must be normalised prior to the eigenvalue decomposition. Basically the normalisation is chosen to ensure that the potential energy expressed by the surface elevation and the kinetic energy expressed by the velocity get similar total weight in $(S_t^f)^{\mathrm{T}} S_t^f$ before the leading eigenvalues are found. The augmented forcing correction part of the state vector is similarly given an equal total weight.

4.3.2 Update

Based on the square root approximation of rank M, S_t^f, the error covariance matrix can be calculated as $P_t^f = S_t^f (S_t^f)^{\mathrm{T}}$, and subsequently used for the Kalman filter update. However, by using the sequential updating algorithm described for the EnKF it is not necessary to calculate the forecast error covariance matrix and the sequential updating can be performed using S_t^f directly. In this case the state vector is updated using Eq. (21), and the updated square root covariance matrix is given by (Cañizares 1999),

$$S_{t,j}^a = S_{t,j-1}^a - \frac{k_{t,j} h_{t,j}^{\mathrm{T}}}{1 + \sqrt{\frac{\sigma_j^2}{h_{t,j}^T h_{t,j} + \sigma_j^2}}}, \quad S_{t,0}^a = S_t^f \tag{30}$$

where $k_{t,j}$ and $h_{t,j}$ are defined in Eq. (22).

4.4 Filter characteristics

In the previous subsections three specific Kalman filter schemes have been presented. In the present subsection we will discuss some of their properties in greater detail. All the schemes attempt to provide time-efficient estimates of the predicted first and second order moments of the state vector. They differ primarily in the way they approximate these moments. The ensemble approach tries to make an exact propagation at the cost of an estimate that may be significantly influenced by stochastic errors due to slow convergence of the ensemble estimate (proportional to $1/\sqrt{M}$). On the other hand, the RRSQRT KF deliberately introduces a bias in both the first and second order moments, but eliminates the stochastic error.

In the EnKF stochastic errors are introduced in both first and second order moments, but when all assumptions are valid it provides an unbiased and asymptotically efficient estimate. The CEnKF maintains the stochastic error in the error covariance propagation. The state estimate inherits this stochastic error component through the update, but on top it has a bias from its first order approximation of the dynamics provided by the central forecast.

The state and its associated covariance estimate are biased in the RRSQRT KF because of the first order Taylor series truncation. A second order truncation would introduce an additional term in the estimate of the mean state, but not otherwise affect the error covariance estimate, Verlaan and Heemink (2001). Furthermore, the error covariance has a truncation error originating from the eigenvalue decomposition and reduction. Generally, the RRSQRT will underestimate the model error covariance for correctly specified $Q(t)$ and thus provide a state estimate closer to the model solution than the optimal estimate. However, there is no stochastic error in this scheme.

The various Kalman filter algorithms generally attempts to minimise the variance assuming no bias, Dee (1998). However, a bias, b, can enter the state estimate either through a bias in the system error or through non-linearities in the model operator in schemes using central forecasts such as RRSQRT KF and CEnKF. In this case the optimal estimator in a minimal prediction error sense must be calculated by using $P^f + bb^{\mathrm{T}}$ instead of P^f in the BLUE estimator, Eqs. (14) and (15), and thus the error covariance estimate provided by the ∞-EnKF is no longer optimal. Alternatively, the filter can estimate the bias by augmenting the state with the bias terms. The bias is propagated by a persistence model or a long memory auto regressive model. For a proper value of α in Eq. (4) this is exactly what the AR(1) noise description does under the assumption of all bias coming from the forcing term (Ignagni, 1990). Thus all filters accommodate bias correction in the forcing.

The reaction time of the bias correction is determined by the relative sizes of the elements in R_t and Q_t. If Q_t is comparatively large, the state will be updated to fit the measurements rather closely where available and simultaneously update all other state variables according to the assumed correlation structure of the model error and its subsequent propagation throughout the model domain. Thus the imposed error structure in Q_t is of prime importance. If the correlation between data rich and data sparse regions are poorly estimated, significant errors can be introduced into data sparse regions. For a comparatively small Q_t there will be more trust in the model and the state estimate will move slowly towards the measurements.

However, a potential structural error will still be introduced into data sparse regions albeit at a slower speed.

5 Measures of non-linearity, Gaussianity and bias

It is important to note that all schemes are imposing a number of approximations in order to make the data assimilation problem manageable. The validity of these assumptions will be case dependent for a set-up of a model like MIKE 3. Thus, before blindly relying on the schemes, the correctness of the underlying assumptions ought to be tested. In the following we will discuss a number of ways to estimate the non-linearity, Gaussianity and bias of a data assimilation algorithm.

According to Verlaan and Heemink (2001), the general aim of a non-linearity measure of a data assimilation system is, without the artificial twin experiment, to assess the accuracy of the data assimilation algorithm associated with the non-linearity of a particular application. In pursuing this goal, they developed a measure that is based on the Taylor series second order contribution to the propagation of the state estimate.

Here we would like to add that the accuracy of a filter is associated with other aspects than the expected bias accumulation induced by non-linearity, all though this is an important factor in highly non-linear applications. The applicability of the BLUE estimator as being optimal in a prediction error sense and the MAP interpretation builds on the assumption of an unbiased and Gaussian distributed state. A non-linear model propagator inherently violates the latter of these assumptions and bias is only avoided in the EnKF and when using unbiased forcing.

Verlaan and Heemink (2001) demonstrate the performance of their measure in the Burgers equation and in the Lorenz-system. Depending on the set-up, MIKE 3 possesses dynamics that can stretch over both these domains of non-linearity. Thus, it is of great interest to examine the non-linearity of a given model application in order to provide guidance in selecting the correct filter and to obtain an indication of filter performance and the accuracy of the provided error estimates. Along with validating the underlying assumptions, non-linearity measures also help the modeller configuring a data assimilation approach and obtaining a better understanding of the dynamics in the particular model domain under consideration.

Three non-linearity measures are used in the present investigation. Verlaan and Heemink's NL-measure V_2, and two measures based on skewness and kurtosis respectively, s_2 and k_2. The first of these gives information about the accumulation of bias introduced by the non-linearity, while the latter two measure the instantaneous deviation from Gaussianity. Gaussianity and linearity are closely related. In general Gaussianity implies linearity whereas the opposite is only true in the case of Gaussian distributions of the error sources and the initial field. All three measures are time varying spatial L_2-norms. Based on the derivation in Verlaan and Heemink (2001), the V_2 measure can be written as:

$$V_2(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{b_i(t)}{\gamma_i(t)}\right)^2} \tag{31}$$

$$b_i(t) = x_i^c(t) - x_i(t) \tag{32}$$

Here, N is the number of elements in the state vector and $\gamma_i(t)$ is the standard deviation of the state estimate derived as the square root of the diagonal elements of

$P^a(t)$. The bias, $b_i(t)$ is simply estimated as the difference between the central ensemble estimate and the average ensemble estimate. In the update step the EnKF scheme is used to estimate the error covariance for both state estimates. Thus, the measure includes effects from the stochastic estimate of the error covariance and average state estimate as well as the error introduced by the non-linear dynamics. For a proper assessment of non-linearity, it must be assumed that the latter is dominating, i.e. that the ensemble size is sufficiently large. The V_2-measure differs from the V-measure suggested in (Verlaan and Heemink 2001),

$$V(t) = \sqrt{b^T P^{-1} b}, \quad b = [b_1, \ldots, b_N]^{\mathrm{T}} \tag{33}$$

While V_2 measures the bias compared to the variance, i.e. the trace of the error covariance matrix, the V-measure compares the bias to the full matrix taking correlations into account.

With M still being the ensemble size, the s_2-measure is simply the spatial L_2-norm of the skewness, $s_i(t)$:

$$s_i(t) = \frac{M\sum_{j=1}^{M}\left(x_j(t) - \bar{x}(t)\right)^3}{(M-1)(M-2)\gamma_i^3(t)} \tag{34}$$

$$s_2(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i(t))^2} \tag{35}$$

A positive skewness expresses that the distribution has a longer tail towards larger values and vice versa for a negative value. Likewise the k_2-measure is the spatial L_2-norm of the kurtosis, $k_i(t)$:

$$k_i(t) = \frac{M(M+1)\sum_{j=1}^{M}(x_j(t) - \bar{x}(t))^4}{(M-1)(M-2)(M-3)\gamma_i^4(t)} - \frac{3(M-1)^2}{(M-2)(M-3)} \tag{36}$$

$$k_2(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(k_i(t))^2} \tag{37}$$

A positive or negative kurtosis respectively expresses that the distribution is peaked or flat relative to the Gaussian distribution.

The two latter measures are introduced in order to measure the point by point non-Gaussianity of the ensemble distribution. Having a Gaussian initial distribution and Gaussian error sources, the non-Gaussianity is an expression of the effect of accumulated non-linearity in the modelled state. However, the measures have both a bias and a variance due to a limited ensemble size. Keep in mind that the forcing function is part of the model operator when employing the augmented state description. Thus, the squared dependence between wind velocity and surface momentum transfer will introduce a skewness into the velocity components. An important operational issue is the robustness of these measures to the ensemble size. All three measures have an off-set that vary with ensemble size. Further, the larger the ensemble size the smaller variance of the measures.

The V_2 measure corresponds to a 2nd order Taylor series expansion in the error covariance propagation. Thus, it can provide information about the validity of the extended Kalman filter (EKF) and its size can be used to measure the linear deviation from this EKF validity regime as long as third and higher order moments can be neglected. The s_2 and the k_2 measures will provide measures of non-linearity that exceeds the point at which the V_2 measure levels out. However, their interpretation as measures of non-linearity depends on having Gaussian system errors. Further, measures based on higher order moments could be introduced to measure higher order non-linearity. E.g. the deviation between the EKF and the EnKF error covariance estimates could be applied in an appropriate way.

Finally, a bias measure is introduced, which compares the updated model to observations where available. The measurements should include validation stations not assimilated, since assimilation might actually increase bias in validation stations. For every measurement, j, the bias measure, β_j, is defined as,

$$\beta_j = \frac{1}{T}\sum_{t=1}^{T}\frac{z_{t,j} - c_{t,j}x^a}{\sqrt{c_{t,j}P_t^a c_{t.j}^T + \sigma_j^2}} \tag{38}$$

T is the number of time steps. The β-measure is applicable to any run in which a model standard deviation is estimated. Taking the L_2-norm over all available measurements, possibly divided into assimilated and non-assimilated stations can aggregate the information of the measure further.

6 Simulation study

A twin test in an idealised set-up is used to demonstrate the application of the non-linearity measures in MIKE 3. The study also investigates the model performance in a set-up with biased forcing using different error correlation structures to estimate the state both with and without a long memory AR(1) error assumption. Both investigations have been designed in order to assess the validity of filter assumptions and the performance when they are violated. However, first attention must be paid to the performance measures used.

Only water level is used in the performance measures, which are as such different from the cost function that the scheme attempts to minimise. However water level is considered the most important forecast variable, and it is the variable that has the largest correlations with the tidal gauge measurements and therefore most clearly shows the strengths and weaknesses of the various approaches. Practically all results transfer to the velocity part of the state vector, albeit with a smaller amplitude. Similarly, the non-linearity, non-Gaussianity and bias measures defined in Section 5 are restricted to include only water levels as well.

A standard performance measure of data assimilation schemes is the root mean square error (RMSE) between the true (*true*) and assimilating or perturbed solutions (*pert*) in a twin experiment (Verlaan and Heemink 2001; Madsen and Cañizares 1999). It can be expressed in a way that collapses either the temporal or the spatial dimension. In the present paper, the following definition is used,

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\zeta_t^{true}(i) - \zeta_t^{pert}(i)\right)^2} \tag{39}$$

where n is the number of water level grid points, T is the number of time steps included in the estimate and ζ the water level. Similarly bias and standard deviation can be defined as,

$$\mathrm{Bias} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\zeta_t^{true}(i) - \zeta_t^{pert}(i)\right)\right]^2} \tag{40}$$

$$\mathrm{St.dev.} = \frac{1}{n} \times \sum_{i=1}^{n}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left(\zeta_t^{true}(i) - \bar{\zeta}_t^{true}(i)\right) - \left(\zeta_t^{pert}(i) - \bar{\zeta}_t^{pert}(i)\right)\right)^2} \tag{41}$$

The filter theory is based on ensemble statistics, but in order to estimate the filter performance time sampled statistics must be used. This requires ergodicity and a sufficiently long time period for the statistics to have acceptable accuracy.

For ergodicity to apply a basin with constant wind forcing and constant open boundary elevation provides the basis of the test case. The basin contains a simple horse shoe island and the initial state has a constant surface elevation at 0 m and is at rest. The spatial resolution is 10 km and the time step is 15 min. The northern open boundary has surface elevation 1.0 m and the eastern has surface elevation 0.0 m. The bathymetry, which is shown in Fig. 1, was chosen to mimic a typical application of MIKE 3 in shelf seas, while remaining simple enough for fairly fast execution and ease of interpretation. In the non-linearity twin test, NL, the false run uses a steady 20 m/s westerly wind, while the true run is forced by the same wind field with a realisation of two similar AR(1) processes added to the x- and y-components of the wind velocity, respectively. Each AR(1) process has a time constant of one hour and 25 min and is forced with a Gaussian distributed white noise with a standard deviation of 5 m/s. In the error structure twin test case, ES, the false run is similar, but the true run uses a steady 19.8 m/s south-westerly wind corresponding to x and y wind velocity components equal to 14 m/s.

The model was run for 16 days and statistics were calculated during the last 15 days. The realised wind errors (the two AR(1) processes) added to the 20 m/s westerly wind in the true NL run had spatially averaged standard deviations of 9.0 m/s and 9.3 m/s in the x and y directions respectively, and maximum norms of the mean of 0.5 m/s and 0.3 m/s with spatial averages of minus 0.01 and 0.03. This is taken to provide a sufficiently good representation of the assumed error statistics of zero mean and standard deviation of 9.2 m/s. Thus, any bias introduced in the system in the NL false run must be due to non-linearity.

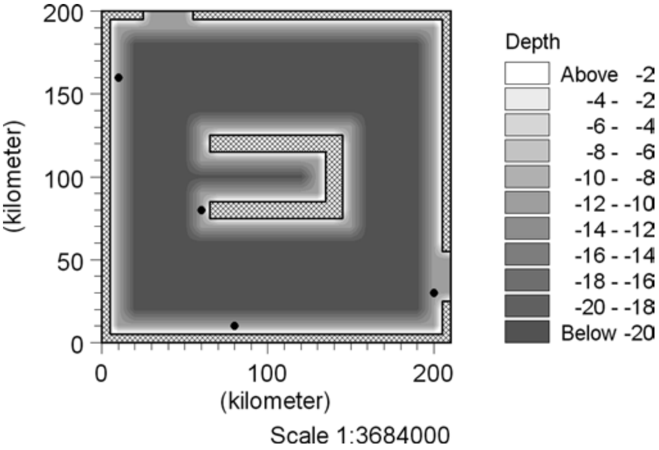

Fig. 1 Test case bathymetry [m]. The black dots indicate measurement positions (10, 160 km), (60, 80 km), (80, 10 km) and (200, 30 km)

Note here that it is not sufficient to work with a period much longer than the time constant of the noise itself, since the model operator potentially filters the input and thus transforms the characteristic time scales. This is clearly seen when an auto-regressive noise is used, but even in the case of direct Gaussian wind stress perturbation, the model operator performs a filtering. In order to make sure that the time statistics are reliable, the time average of the model output from an execution with the assumed true run should compare well with the result of a ∞-EnKF of the false run. For the NL true run this was successfully validated against a 1000 EnKF run without assimilation.

Measurements were extracted from four points in each of the true runs to be assimilated into the false runs. The positions shown in Fig. 1 were chosen at boundaries, as is typically the case for tidal gauge stations. The asymmetry of the positions suggests a similar asymmetry in the standard deviation of the state estimate to be provided by the assimilation schemes. The measurement positions are also chosen to investigate the filter performance in data sparse regions as compared to data rich regions for various error structure assumptions.

The NL-experiments were designed to provide a comparison between the various non-linearity measures and relate these to the filter performance of the three filters presented in Section 4. The design enables the ∞-EnKF to provide the optimal estimate since care has been taken not to have significant reminiscent bias in the system apart from that introduced by the non-linearity in the schemes based on a central forecast. Thus the relation between bias and non-linearity should stand clear. The non-linearity is expected to increase with increasing update time intervals, (Verlaan & Heemink 2001). Therefore, update interval (ui) is chosen as a control parameter of non-linearity. The update interval is given in time steps and thus the ui12 run updates the state every 12th time step or equivalently every 3 h. Update intervals equal to 1,4, 8, 12, 24 and 48 are chosen and compared to a simulation without updating.

The ES-experiments are meant to expose the importance of using the augmented AR(1) error description in the presence of bias. More generally they highlight the performance of the Kalman filter using true and false descriptions of the error structure and how insight can be obtained from the bias measure. The ES false run has a clearly biased wind forcing having a direction, which is turned 45 degrees. The EnKF is used to assimilate the true results under the assumptions of biased and unbiased wind, i.e. time constants of 0 and 10^6 seconds equivalent to an AR(1) parameter α equal to 0.0 and 0.9994 respectively. This is done in combination with four different spatial correlation scales of 0, 100, 495 and 10,000 km for the wind error.

7 Results and discussion

7.1 Non-linearity (NL) experiments

7.1.1 Solution without data assimilation

In order to give an impression of the general solution of the NL true and false run and central and ensemble forecast without assimilation, Fig. 2 shows a time series of water level at the measurement point (60, 80 km) for each case. The ensemble run is based on 1000 ensembles. All variability in the true run is due to a changing wind field. A rather large variation has been imposed and the shortcomings of the false runs are obvious. Further, the bias introduced by the central forecast stands out clearly.

An alternative view of the false run is provided by Fig. 3, which shows the bias and standard deviation over the last 15 days for the central forecast false run. The spatial distribution of the bias reflects the non-linearity from the squared dependence of wind speed in the momentum transfer. The distribution of the standard deviation arises from the coloured wind error showing its peak values close to the closed boundaries. Similar statistics are shown for a 1000 ensemble forecast in Fig. 4. Note the reduction in bias, while the standard deviation remains literally unaltered.

7.1.2 General filter performance

The assimilation schemes all improve the rather poor false solution significantly. Figures 5 and 6 show the bias and standard deviation for the RRSQRT with 40 leading eigenvalues and EnKF based on 1000 ensembles, respectively, and should be compared to Figs. 3 and 4. In both cases the state was updated at every time step, i.e. $ui = 1$. An obvious error reduction is seen in either case, which proves the efficiency of the assimilation schemes. Figure 7 shows the standard deviation estimated by the EnKF averaged over the last 15 days. Ensuringly, the structure of this estimate is seen to correspond closely to the actual standard deviation in Fig. 6. Neither of the schemes have a significant bias. Note that $ui = 1$ is the most linear of the NL model runs. The good estimation of standard deviation generalises to all assimilation runs.

Table 1 sums up the performance of the schemes as estimated by the RMSE between the true run and each run with false forcing with assimilation and varying update interval and a run without assimilation. The good filter performance already demonstrated in the figures generalises to all cases. The larger the update interval the worse performance, as expected when longer periods of time with possible

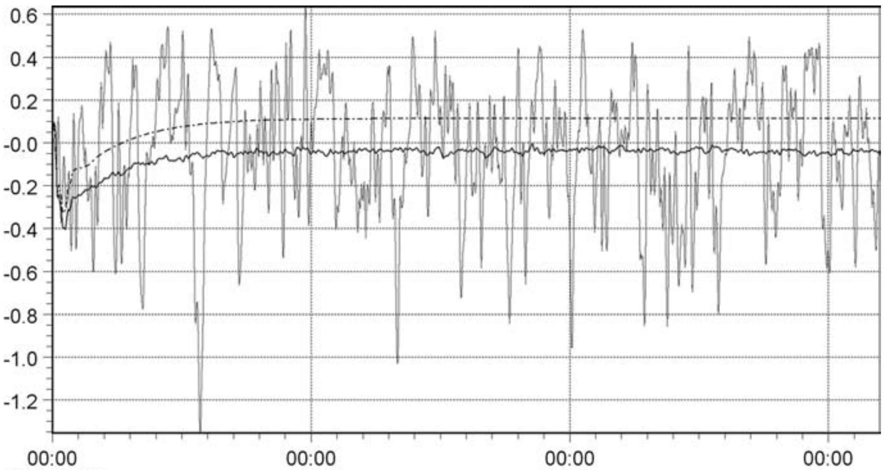

Fig. 2 Water levels extracted at (60, 80 km). Grey: True run. Black dot-dashed: Central forecast false run. Black: Ensemble forecast false run

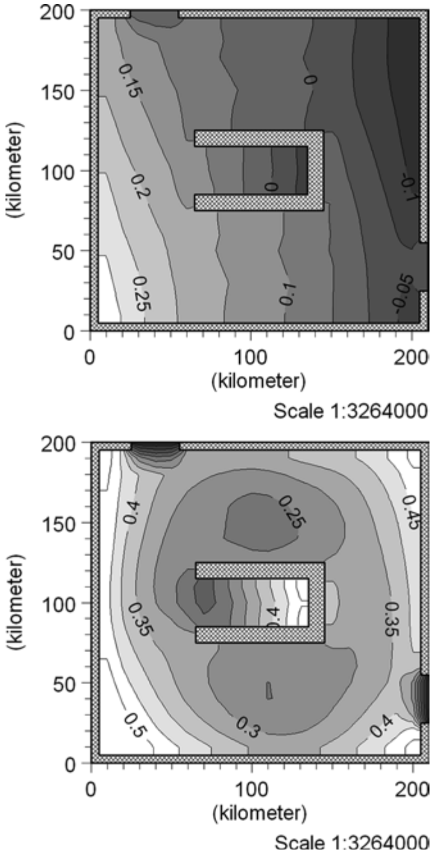

Fig. 3 Top: Central forecast NL false run water level bias [m]. Bottom: Central forecast NL false run water level standard deviation [m]

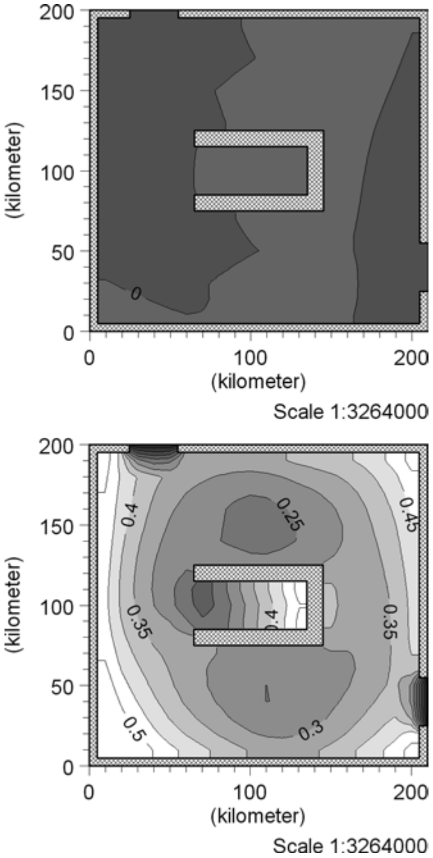

Fig. 4 Top: 1000 ensemble forecast NL false run water level bias [m]. Bottom: 1000 ensemble forecast NL false run water level standard deviation [m]

drift away from the true state is allowed. An EnKF with 100 ensembles is also included in the present study and has approximately the same execution time as the RRSQRT KF with 40 leading eigenvalues. These numbers have been shown by Madsen and Cañizares (1999) to be sufficient in the kind of system under consideration. In order to assess the stochastic variation of the EnKF, five realisations of the 100 EnKF have been calculated. Only one of these is included in Table 1, but the variability of the RMSE is generally less than 0.005.

7.1.3 Assessment of non-linearity and non-Gaussianity

Consider the bias in the central forecast provided by the CEnKF versus the EnKF based forecast with no assimilation. Figures 3 and 4 show maps of their time averaged bias for 1000 ensembles in the extreme case of no updates. Table 2 shows the spatial L_2-norm of the time averaged bias for the range of different runs with false forcing and using the various schemes and update intervals. It is clear how the non-linear model equation introduces a model bias as the update interval increases in the schemes relying on central forecasts, RRSQRT and CEnKF, whereas the ensemble forecast has a negligible bias.

This behaviour is well captured by the non-linearity measure, V_2, defined in Section 5. As can be seen in Table 3 the effect of changing the update interval is consistently to increase V_2. This is a consequence of the bias demonstrated in Table 2. As assumed, the non-linearities in the model operator introduces progressively more bias in the system as the update interval (ui) is increased. However even when no assimilation is used at all, the V_2 non-linearity measures remain small. The main source of non-linearity in the model is the conversion of wind velocity to wind stress in the interplay between the augmented and the model part of the state vector. Thus, the present set-up is not highly non-linear, but on the other hand non-linearities are not negligible either.

The s_2 and k_2 measures in Table 3 show a similar dependence on update interval and thus provide interesting complimentary measures. While describing the non-linearity they simultaneously provide an indicator of non-Gaussianity and thus the reliability of interpreting the results as MAP-estimates. For s_2 and k_2 the variability with update interval is somewhat different from V_2. They increase rather steadily with update

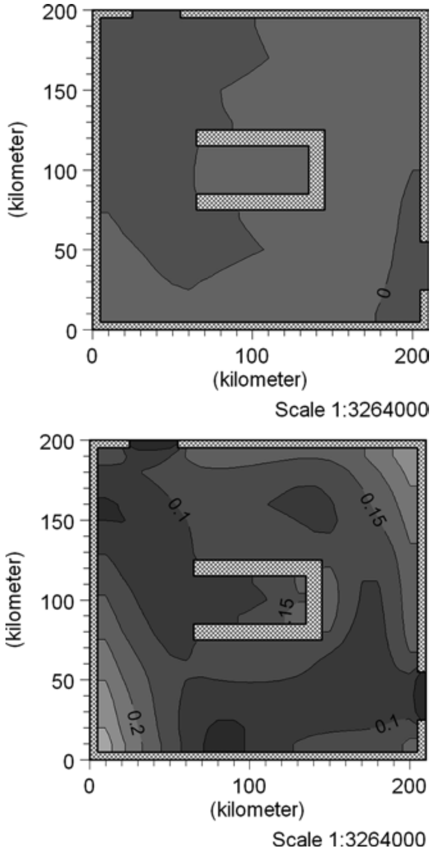

Fig. 5 Top: Forecast NL false run water level bias [m] using the RRSQRT. Bottom: Forecast NL false run water level standard deviation [m] using RRSQRT

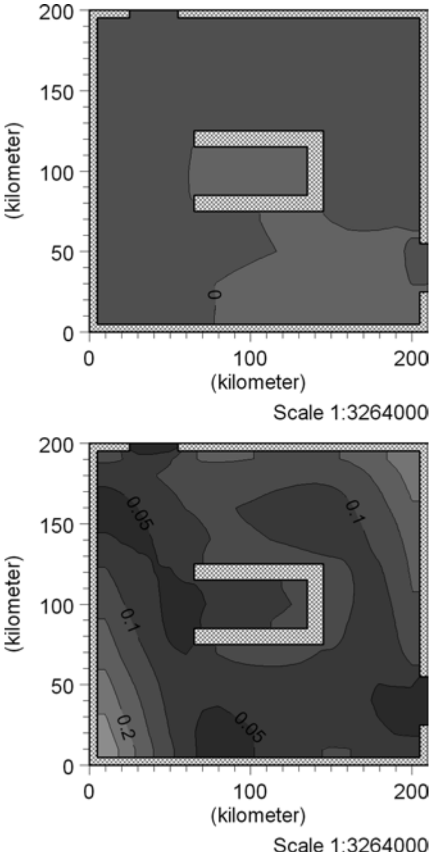

Fig. 6 Top: Forecast NL false run water level bias [m] using the 1000 EnKF. Bottom: Forecast NL false run water level standard deviation [m] using 1000 EnKF

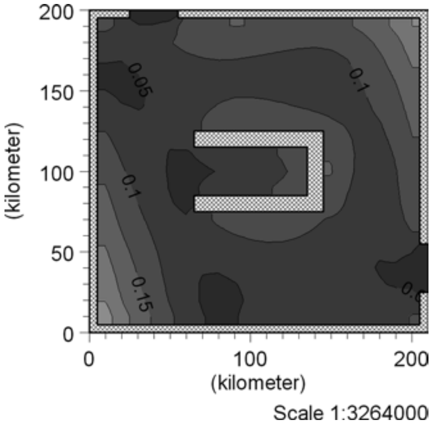

Fig. 7 Estimated water level standard deviation [m] by the 1000 EnKF

interval for the chosen intervals and even in the most linear case ($ui = 1$) the solution seems to be non-Gaussian. Therefore, even the 1000 ensemble estimate does not give the state with the maximum a-posteriori probability, but rather the state estimate with the lowest mean square error using linear and unbiased estimators.

All three measures are merely stochastic realisation and their variability should be assessed. First of all, the measures obviously vary with ensemble size. This is to be expected since they rely on sample estimates of second and higher order moments. However, for a given ensemble size there might still be a stochastic variability, due to limited ensemble size. Five realisations of 100 EnKF have been used to assess this variability. In all cases, the maximum difference is less than 0.02 in the RMSE estimate, 0.02 in V_2, 0.03 in s_2 and 0.06 in k_2. Thus the single run estimates can be considered sufficiently accurate to indicate the relative non-linearity and Gaussianity of various data assimilating set-ups.

Bias has been introduced as a product and measure of non-linearity, but simultaneously it is the source of trouble for schemes based on the extended Kalman filter, such as the RRSQRT, in strongly non-linear applications. In Segers et al. (2000) a second order RRSQRT filter was introduced, which handles significantly more non-linear situations. However, the only enhancement as compared to the regular RRSQRT filter is to estimate and correct the bias introduced in the state estimate by non-linearities. The forcing induced bias, which can have a similar impact on the filter performance, is most often not considered in literature, but

Table 1 Root mean square error (RMSE) in the NL assimilation runs for varying update interval (ui) and assimilation scheme. Runs with no update are denoted ui-∞

RMSE	ui1	ui4	ui8	ui12	ui24	ui48	ui-∞
1000 EnKF	0.10	0.12	0.17	0.22	0.27	0.31	0.34
100 EnKF	0.10	0.12	0.17	0.22	0.27	0.32	0.34
40 RRSQRT	0.12	0.13	0.19	0.24	0.30	0.34	0.36
1000 CEnKF	0.10	0.12	0.17	0.22	0.28	0.33	0.36
100 CEnKF	0.10	0.13	0.18	0.23	0.29	0.33	0.36
No assim.	0.36	0.36	0.36	0.36	0.36	0.36	0.36

Table 2 Spatial L_2-norm of the bias in the NL assimilation runs for varying update interval (ui) and assimilation scheme. Runs with no update are denoted ui-∞

Bias	ui1	ui4	ui8	ui12	ui24	ui48	ui-∞
1000 EnKF	0.01	0.01	0.01	0.01	0.01	0.02	0.01
100 EnKF	0.02	0.01	0.01	0.02	0.02	0.02	0.01
40 RRSQRT	0.01	0.01	0.02	0.04	0.07	0.11	0.13
1000 CEnKF	0.01	0.01	0.03	0.04	0.07	0.11	0.13
100 CEnKF	0.01	0.02	0.02	0.04	0.07	0.11	0.13
No assim.	0.13	0.13	0.13	0.13	0.13	0.13	0.13

Table 3 Non-linearity measures for the NL assimilation runs for varying update interval (ui). Runs with no update are denoted ui-∞

NL-measure	ui1	ui4	ui8	ui12	ui24	ui48	ui-∞
V_2	0.10	0.12	0.14	0.17	0.22	0.29	0.33
s_2	0.21	0.25	0.35	0.40	0.49	0.54	0.60
k_2	0.37	0.43	0.52	0.56	0.61	0.64	0.69

much more attention needs to be paid to this aspect for operational use of Kalman filtering techniques. The next part of the discussion attempts to examine this bias source and how the implemented schemes can handle it in the case of true as well as false error structure assumptions.

7.2 Error structure (ES) experiments

7.2.1 Solution without data assimilation

Both the true and the false ES runs reach a steady state rather fast and thus the false run error is essentially determined by the bias, which is shown in Fig. 8. The bias is created by a constant difference in wind direction throughout the domain. Thus, the error source is known to be a bias in the wind velocity with infinite spatial correlation. The bias is evident and has an L_2-norm of 0.27 m. However, the bias varies throughout the domain. In real applications the bias can only be estimated in measurement points. Thus, sufficient data coverage is required for a proper assessment of bias. The bias in Fig. 8 does not necessarily suggest a spatially constant bias to the untrained eye. Only with the proper physical insight and sufficient sampling, this can be anticipated.

By running one of the data assimilation schemes with no updates, the model standard deviation and thus the β-measure can be estimated. Assuming the entire field to be known, the L_2-norm of β is 1.8 and if we restrict ourselves to the measurement points the corresponding value is also 1.8, but obviously a different set of points could yield a substantially different value. Four validation points were selected: (10, 80 km), (160, 10 km), (130, 90 km) and (190, 190 km). Based on these the L_2-norm of β is 2.1. In all cases the measure shows that the model-

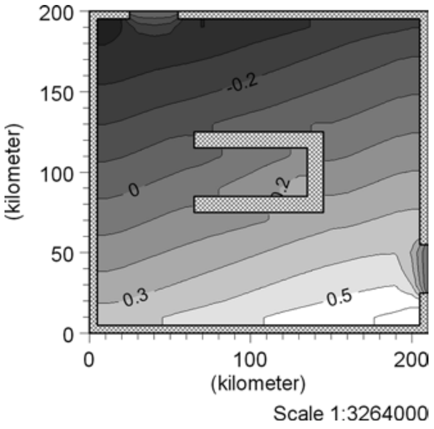

Fig. 8 Forecast ES false run water level bias [m]

measurement difference is significantly larger than its standard deviation. Knowing that the measurements are unbiased in this idealised test case, we can conclude that the model has a bias.

7.2.2 Solution with data assimilation

The biased error structure can be cast within the assimilation schemes presented in Section 4 and thus these ought to give a very good estimation of the bias. This is demonstrated in Fig. 9 showing the bias from the 100 EnKF scheme correctly assuming a biased error with a very high spatial correlation of 10.000 km. Alternatively, if the error is assumed to be white, a bias will always remain as shown in Fig. 10, still assuming a spatial correlation of 10.000 km. The results are summarised in Table 4, showing the L_2-norm of bias and β for varying spatial correlation lengths with a white noise or bias assumption corresponding to a time constant of zero and 10^6 s, respectively. The effectiveness in bias correction is seen to clearly depend on the validity of the imposed error assumptions. The assimilation runs assuming coloured and spatially correlated noise leave a bias, which is smaller than the estimated standard deviation of the model-measurement difference. Since the model believes it is correcting an error in all assimilation runs, this standard deviation is rather quickly dominated by the measurement standard deviation of 0.05 m. However, for the assimilation runs assuming white noise the resulting bias is only barely within the bounds of the uncertainty even for the correct spatial correlation.

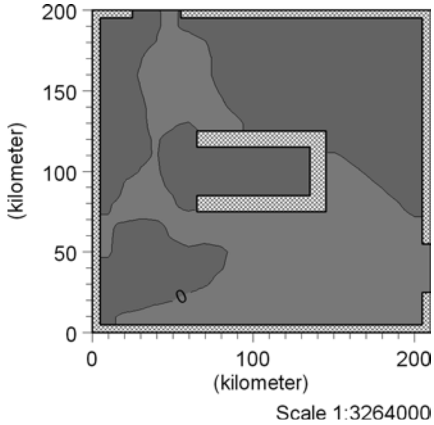

Fig. 9 Forecast ES false run water level bias [m] using 100 EnkF with a time constant of 10^6 and a spatial correlation scale of 10,000 km

Applying a wrong spatial correlation scale can potentially increase the bias in data sparse areas as demonstrated in Fig. 11, which shows the bias for a spatial correlation scale of 0 kilometres and a time constant of 10^6 seconds. Compared to Fig. 8 there is an evident bias increase in the data sparse bay of the horse shoe island.

All together, these experiments show the importance of treating the error structure correctly. Making false assumptions can severely affect the filter performance. Both in the deterministic case and when employing an assimilation scheme the bias in measurement points ought to be examined. The β-measure can be used to indicate whether the bias is within the range of uncertainty for every point of interest.

Table 4 Top: The L_2-norm of the bias. Bottom: The L_2-norm of β. The time constant and the spatial correlation scale vary along the vertical and horizontal axes respectively. All runs are based on the 100 EnKF scheme

Bias	0 km	100 km	495 km	10,000 km
0 s	0.25 m	0.13 m	0.08 m	0.05 m
10^6 s	0.17 m	0.04 m	0.01 m	0.00 m
β				
0 s	4.13	1.63	1.12	0.84
10^6 s	3.03	0.64	0.21	0.06

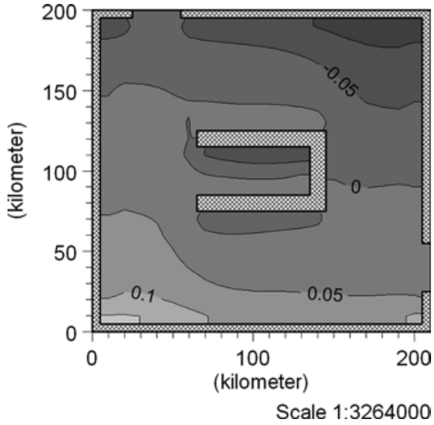

Fig. 10 Forecast ES false run water level bias [m] using 100 EnkF with a time constant of zero and a spatial correlation scale of 10,000 km

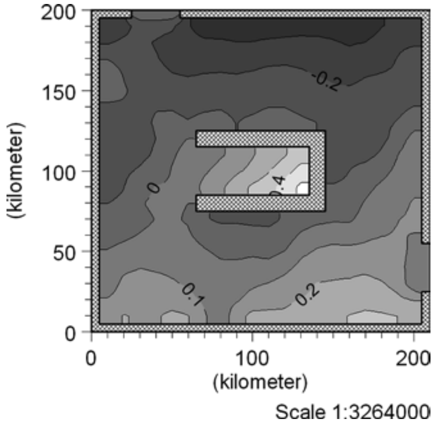

Fig. 11 Forecast ES false run water level bias [m] using 100 EnkF with a time constant of 10^6 and a spatial correlation scale of 0 km

8 Summary and conclusions

A stochastic model of the physical system consisting of hydrodynamic flow in coastal and continental shelf seas has been formulated. This stochastic model and observations are the foundation of providing statistically based estimates of the oceanic state. However, in order to obtain such estimates a number of assumptions must be imposed. A non-linearity measure, two measures for non-Gaussianity and a bias measure have been presented with the aim of providing means of assessing the validity of these assumptions.

The non-linearity measure has been demonstrated to vary consistently with the non-linearity of the set-up. The EnKF handles the non-linearity well, leaving only a minor bias, whereas procedures based on central forecast have significant biases for more non-linear set-ups. The correspondence between the non-linearity and non-Gaussianity has been verified. The MAP interpretation of the estimated state must be discredited in the case of strong non-linearities or lack of Gaussian noise input. Finally, it has been demonstrated how wrong error structure assumptions may severely hamper the results. This is particularly true for data sparse regions.

For the simple test case examined in this paper, the wind driven coastal circulation does not require data assimilation schemes, which handles strongly non-linear dynamics for assimilation of tidal gauge data. This might not be the case for all bathymetries and thus it is recommended to employ non-linearity measures to assess the applicability of the various schemes. The non-Gaussianity measures provide complimentary measures that simultaneously guides the user to a proper interpretation of the results. In many real case applications, the bias introduced by non-linearity is not the dominating source of bias. Rather the forcing induced bias will often be larger. A general bias measure, which is easy to calculate, has been formulated. This measure indicates the presence of bias, but not whether the source is model non-linearity or biased forcing. However, in combination with the non-linearity measures, the contribution from each can be approximately assessed. Hence work can proceed to take the bias properly into account in the data assimilation scheme. In any case, the presence of bias indicates that the filter is working under the wrong assumptions and therefore is not optimal in a least square sense. Another prerequisite of optimality of the estimator is a correct error structure description. It is demonstrated that the specification of a correct error structure is important in practical application and wrong assumptions can induce severe errors in data sparse regions.

Acknowledgements This work was in part funded by the Danish Academy for Technical Sciences under the industrial Ph.D. fellowship programme, EF 835, and the Danish Technical Research Council under the Talent Project programme, contract No. 9901671.

References

Burgers G, van Leeuwen PJ, Evensen G (1998) Analysis scheme in the ensemble Kalman filter. Monthly Weather Rev 126: 1719–1724

Chui CK, Chen G (1991) Kalman Filter with Real-Time Applications. Volume 17 of Springer Series in information sciences, Springer-Verlag

Cañizares R (1999) On the application of data assimilation in regional coastal models. PhD thesis Delft University of Technology

Cañizares R, Madsen H, Jensen HR, Vested HJ (2001) Developments in operational shelf sea modelling in Danish waters. Estuarine Coastal Shelf Sci 53: 595–605

Cohn SE, Todling R (1996) Approximate data assimilation schemes for stable and unstable dynamics. J Meteorol Soc Japan 74: 63–75

Christakos G (2002) On the assimilation of uncertain physical knowledge bases: Bayesian and non-Bayesian techniques. Adv Water Res 25: 1257–1274

DHI (2001) MIKE 3 estuarine and coastal hydraulics and oceanography; users guide. DHI Water & Environment

Dee DP (1991) Simplification of Kalman filter for meteorological data assimilation. QJR Meteorol Soc 117:365–384

Dee DP (1995) On-line estimation of error covariance parameters for atmospheric data assimilation. Monthly Weather Rev l23: 1128–1l45

Dee DP, da Silva AM (1998) Data assimilation in the presence of forecast bias. QJR Meteorol Soc 124: 269–296

Erichsen AC, Rasch PS (2002) Two- and three-dimensional model system predicting the water quality of tomorrow. In: Spaulding ML (ed) Proceedings of the Seventh International Conference on Estuarine and Coastal Modeling. American Society of Civil Engineers

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99(C5): 10143-10162

Fox AD, Haines K, de Cuevas BA, Webb DJ (2000) Altimeter assimilation in the OCCAM global model. Part I: A twin experiment. J Marine Sys 26: 303–322

Fukumori I, Raghunath R, Wunch C, Haidvogel DB (1993) Assimilation of sea surface topography into an ocean circulation model using a steady-state smoother. J Phys Ocean 23: 1831–1855

Fukumori I, Melanotte-Rizzoli P (1995) An approximate Kalman filter for ocean data assimilation; an example with an idealised gulf stream model. J Geophys Res 100: 6777–6793

Fukumori I, Raghunath R, Fu L-L, Chao Y (1999) Assimilation of TOPEX/Poseidon altimeter data into a global ocean circulation model: How good are the results? J Geophys Res 104(C11): 25647–25665

Heemink AW (1986) Storm surge prediction using Kalman filtering. Ph.D. Thesis Twente University of Technology

Heemink AW, Bolding K, Verlaan M (1997) Storm surge forecasting using Kalman filtering. J Meteorol Soc Japan 75(1B): 305–318

Ignagni MB (1990) Separate-bias Kalman estimator with bias state noise. IEEE Trans Automatic Control 35: 338–341

Jazwinski AH (1970) Stochastic processes and filtering theory. Mathematics in Science and Engineering, Vol 64 Academic Press

Madsen H, Cañizares R (1999) Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling. Int J Numer Meth Fluids 31(6): 961–981

Øresundskonsortiet (1998) The resund link. Assessment of the Impacts on the Marine Environment of the resund link. Update, Øresundskonsortiet

Segers AJ, Heemink AW, Verlan M, van Loon M (2000) Kalman filtering for nonlinear atmospheric chemistry models: second (order) experiences. Technical report Delft University of Technology. pp 28

Sørensen JVT, Madsen H, Madsen H (2002) Towards an operational data assimilation system for a three-dimensional hydrodynamic model. Proceedings of the fifth International Conference on hydroinformatics 1204–1209

Verlaan M, Heemink AW (1997) Tidal flow forecasting using reduced rank square root filters. Stochastic Hydrology Hydraulics ll: 349–368

Verlaan M, Heemink AW (2001) Nonlinearity in data assimilation applications: A practical method for analysis. Monthly Weather Rev 129: 1578–1589

Vested HJ, Berg P, Uhrenholdt T (1998) Dense water formation in the Northern Adriatic. J Marine Sci 18: 135–160

Wunsch C (1996) The ocean circulation inverse problem. Cambridge University Press pp 422