DATA ASSIMILATION IN HYDRODYNAMIC MODELS OF CONTINENTAL SHELF SEAS

Jacob Viborg Tornfeldt Sørensen

Informatics and Mathematical Modelling
Technical University of Denmark
Ph.D. Thesis No. 126
Kgs. Lyngby 2004

IMM

ISSN 0000-0000

This document was prepared with LaTeX and printed at IMM, DTU.

Preface

This thesis was prepared at Informatics and Mathematical Modelling, Technical University of Denmark in fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with the assimilation of data in hydrodynamic models of continental shelf seas. The main contribution of to this field is the development of cost-effective Kalman filter based data assimilation schemes applicable to operational settings. Further main contributions are the interpretaion of the schemes in terms of regularisation and a proposed framework for the combination of error correction modelling and Kalman filters.

The thesis consists of a summary report and a collection of seven research papers written during the period 2000–2003, and elsewhere published or submitted for publication.

Lyngby, 16 January 2004

Jacob Viborg Tornfeldt Sørensen

Acknowledgements

First of all, a deep thank to my wife Lotte for all your support and patience and to Clara for smiles and laughter, which have carried me to the end of this endeavour maintaining a good spirit.

In carrying out the work described in this thesis I have received important assistance from many people. First of all I want to address my gratitude to my supervisors Prof. Henrik Madsen from Informatics and Mathematical Modelling at the Technical University of Denmark and Dr. Henrik Madsen from DHI Water & Environment for their help and guidance and for entering endless discussions. Further, I would like to thank Prof. Detlef Stammer at Scripps Institution of Oceanography for broadening my perspective and hosting me in the fall of 2001.

Thanks also to my colleagues at the department and at DHI Water & Environment for their invaluable cooperation, help, and discussions.

Finally, I would like to express my sincere thanks to the Industrial Ph.D. Programme (EF-835) and DHI Water & Environment who supported the project financially.

Papers included in the thesis

[A] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Parameter sensitivity of three Kalman filter schemes for the assimilation of tide gauge data in coastal and shelf sea models. *Ocean Modelling*, 2003. Submitted.

[B] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Data assimilation in hydrodynamic modelling: On the treatment of non-linearity and bias. *Stochastic Environmental Research and Risk Assessment*, 2003. Accepted.

[C] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Towards an operational data assimilation system for a three-dimensional hydrodynamic model. In I.D. Cluckie, D. Han, J.P. Davis and S. Heslop, *Proceedings of the Fifth International Conference on Hydroinformatics*, pages 1204–1209, Cardiff, Wales, July 2002.

[D] Jacob V. Tornfeldt Sørensen, Henrik Madsen, Henrik Madsen, Henrik René Jensen, Peter Skovgaard Rasch, Anders C. Erichsen and Karl Iver Dahl-Madsen. Data assimilation in an operational forecast system of the North Sea - Baltic Sea system. In H. Dahlin and N.C. Flemming and K. Nittis and S. E. Petersson *Building the European Capacity in Operational Oceanography, Proceedings of the Third EuroGOOS Conference*, Athens, Greece, December 2002.

[E] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Efficient Kalman Filter Techniques for the Assimilation of Tide

Gauge Data in Three-Dimensional Modelling of the North Sea and Baltic Sea System. *Journal of Geophysical Research*, 2003. Accepted.

[F] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Water level forecast skill of a hybrid steady Kalman filter - error correction scheme. *Ocean Dynamics*, 2003. Submitted. Revised.

[G] Jacob V. Tornfeldt Sørensen, Henrik Madsen and Henrik Madsen. Parameter estimation in a hydrodynamic model of the North Sea and Baltic Sea. Technical Report, DHI Water & Environment, Hørsholm, Denmark, 2004.

Summary

Data assimilation in hydrodynamic models of continental shelf seas

This thesis consists of seven research papers published or submitted for publication in the period 2002-2004 together with a summary report. The thesis mainly deals with data assimilation of tide gauge data in two- and three-dimensional hydrodynamic models of the continental shelf seas. Assimilation of sea surface temperature and parameter estimation in hydrodynamic models are also considered. The main focus has been on the development of robust and efficient techniques applicable in real operational settings.

The applied assimilation techniques all use a Kalman filter approach. They consist of a stochastic state propagation step using a numerical hydrodynamic model and an update step based on a best linear unbiased estimator when new measurements are available. The main challenge is to construct a stochastic model of the high dimensional ocean state that provides sufficient skill for a proper update to be calculated. Such a stochastic model requires model and measurement errors to be described, which is a difficult task independent of the computational resources at hand. Further, the need for efficient solutions necessitates further assumptions to be imposed that maintain a skillful and robust state estimate.

The assimilation schemes used in this work are primarily based on two ensemble based schemes, the Ensemble Kalman Filter and the Reduced

Rank Square Root Kalman Filter. In order to investigate the applicability of these and derived schemes, the sensitivity to filter parameters, non-linearity and bias is examined in artificial tests. Approximate schemes, which are theoretically presented as using regularised Kalman gains, are introduced and successfully applied in artificial as well real case scenarios. Particularly, distant dependent and slowly time varying or constant Kalman gains are shown to possess good hindcast and forecast skill in the Inner Danish Waters.

The framework for combining data assimilation and off-line error correction techniques is discussed and presented. Early results show a potential for such an approach, but a more elaborate investigation is needed to further develop the idea. Finally, work has been initiated on parameter estimation in two-dimensional hydrodynamic models with an approach that avoids the development of an adjoint code by using an algorithmic structure that favours application of office-grids as they are envisaged to look in the near future.

The main contribution is the development of a number of regularisation techniques for tide gauge assimilation. Further, the techniques used to assess the validity of underlying assumptions (weak non-linearity, unbiasedness or error model skill) provide a valuable tool-box for investigating a dynamical system prior to potentially selecting an assimilation approach. The combined data assimilation error correction framework may be an important contribution to future improvements of forecast skill for a number of systems. The work done on parameter estimation is expected to mature into a future standard procedure for model calibration for models with rapidly evolving complex codes.

Resumé

Data assimilering i hydrodynamiske modeller af farvande på kontinentalsoklen

Nærværende afhandling består af syv forskningsartikler, der er publiceret eller indgivet til publicering i perioden 2002-2004, og en sammenfatning. Afhandlingen beskæftiger sig hovedsageligt med data assimilering af data fra vandstandsmålere i to- og tredimensionale hydrodynamiske modeller af farvande på kontinentalsoklen. Endvidere behandles assimilering af havets overfladetemperatur og parameter estimation i hydrodynamiske modeller. Fokus har været på udviklingen af robuste og tids-effektive metoder, der kan anvendes i virkelige, operationelle problemstillinger.

De anvendte assimileringsteknikker er alle Kalman Filter baseret. De består af et skridt som propagerer den stokastiske tilstand ved brug af en numerisk hydrodynamisk model og et opdateringsskridt der baserer sig på den bedste lineære biasfrie estimator når nye målinger er tilgængelige. Hovedudfordringen er at konstruere en stokastisk model af havets tilstand, der er god nok til at en ordentlig opdatering kan udregnes. For at lave en sådan stokastisk model skal man have en god beskrivelse af model og målefejl, hvilket er svært uanset hvor store computerressourcer, der er til rådighed. For at leve op til kravet om operationel anvendelighed, er det endvidere nødvendigt at lave yderligere antagelser, der samtidig er underlagt krav om robusthed.

Assimileringsskemaerne, der bruges i nærværende afhandling, er hovedsageligt baseret på de to ensemble-baserede teknikker, ensemble Kalman

filtret og reduced rank square root Kalman filtret. For at undersøge anvendeligheden af disse og afledte skemaer undersøges deres følsomhed over for filter parametre, ikke-linearitet og bias i en række kunstige test-opsætninger. Tilnærmede skemaer præsenteres som regulariseringer af Kalman gain matricen, og demonstreres succesfuldt i kunstige såvel som virkelige scenarier. En afstandsafhængig Kalman gain med langsom eller ingen tidsvariation vises at have gode hindcast og forecast evner i de indre danske farvande.

Et framework, der kombinerer data assimilering og off-line fejlkorrektionsteknikker, præsenteres og diskuteres. Foreløbige resultater viser et potentiale for en sådan angrebsvinkel, men en mere fyldestgørende undersøgelse mangler for at kunne færdigudvikle idéen. Desuden er arbejdet med parameter estimation i todimensionale hydrodynamiske modeller påbegyndt. Der anvendes hér en teknik, som undgår den tidskrævende udvikling af en adjoint kode ved at bruge en algoritmisk struktur, som tilgodeser anvendelse af morgendagens office-grid løsninger.

Hovedresultatet er udviklingen af en række regulariseringsteknikker til assimilering af vandstandsmålere. Teknikkerne, som er brugt til at teste de underliggende antagelser (svag ikke-linearitet, biasfrihed og korrekt fejlmodel), giver en værdifuld værktøjskasse til at undersøge dynamiske systemer før der potentielt skal vælges en assimileringsmetode. Det kombinerede data assimilering og fejlkorrektion framework vil bidrage til fremtidige forbedringer af forudsigelsesevnen for et antal dynamiske systemer. Arbejdet med paramter estimation forventes i fremtiden at modne til en standard procedure for model kalibrering i modeller med hurtigt udviklende komplekse koder.

Contents

Chapter 1

Introduction

This thesis deals with data assimilation in hydrodynamic models of continental shelfs and coastal seas. Ocean scientists and coastal engineers are continuously faced with the problem of knowing what the state of the ocean was in the past, is now and will be tomorrow. Simultaneously, there is a need for better understanding why the ocean behaves in certain ways, i.e. what processes are dominating at various locations, times and spatial scales. The search for answers to these questions has been the foundation of most great scientific findings in the past centuries. However, with the advance of hydroinformatics and with the vast computational resources available today, the scene is set for pursuing new techniques for filling out the rather large number of remaining gaps in our capability of describing and understanding the seas. Data assimilation is a rather general term for incorporating observations in a physical and theoretical description of a system. Pending challenges to be solved are related security, industrial and environmental issues such as climate monitoring and prediction, risk assessment and design.

1.1 Coastal seas

The physical system under consideration consists of hydrodynamic flow and a range of other processes acting within bays, estuaries, coastal re-

gions or shelf seas. The body of water evolves according to the laws of internal dynamics and its interaction with the atmosphere and the solid earth. The system is very complex, accommodating nonlinear, turbulent mass and momentum fluxes and further a rich density structure, sediment transports as well as chemical and biological processes. Thus, a great number of interactions and physical properties describe and determine the state of the system. The important spatial scales range from micrometers for molecular dissipation to a basin scale seasonal cycle with practically every intermediate scale playing a role for one process or another. Likewise the temporal scales vary from seconds to millennia and above.

Many physical phenomena are described by the hydrodynamic and thermodynamic equations alone. Among these are tidal waves, wind induced coastal upwelling, frontal dynamics and eddy formation. Thus, as a simplest approach the treatment can be restricted to the hydrodynamic and thermodynamic variables. Hence, no chemical processes, biological processes or sediment transports are described and the thermodynamic, momentum and mass distributions alone constitute the system.

1.2 Numerical Modelling

With the advance of the computer technology and discrete mathematics, mechanistic numerical modelling became a more and more attractive approach to solving hydrodynamic problems in the marine environment. The derived techniques build on known first principles for fluid dynamics, which provide the basic mathematical formulation of a boundary value problem. A tractable solution to the problem is typically found by applying discretisation techniques. This has lead to the generation of a large number of numerical models distributed throughout the world with each their set of approximations. Any such approach requires the user to specify initial and boundary conditions along with calibration parameters. The two models applied in the present study are MIKE 21 and MIKE 3 developed at DHI Water & Environment, (DHI 2002) and (DHI 2001). MIKE 21 solves the depth integrated mass and momentum conservation equations while MIKE 3 provides a solution to the full 3-dimensional problem.

A numerical modelling approach thus has its starting point in well established theoretical knowledge. This allows for a physically consistent analysis of the results. However, a great number of approximations must be introduced in order to obtain a tractable solution. Observations are generally needed for model initialisation, specification of boundary conditions as well as model calibration and performance assessment.

1.3 Observations

A large number of measurements with a quite diverse nature exist. These range from spatial images of sea surface temperature (SST) with a temporal sparsity to tide gauge station, which possess a high temporal resolution, but are sparsely distributed in space. Other examples of observations are salt and temperature profiles from cruises and HF radar observation of surface velocities. In this study comparison with and assimilation of tide gauge water level observations are primarily reported.

Both in the satellite earth observation community and among in situ measurement providers, in increasing efford is being directed towards real time delivery. The integrated service chain from sensor to assimilation and customer service in terms of a forecast is being adressed, which directs attention to the real-time aspects of data assimilation techniques.

1.4 Real-time operations

One final aim of the work undertaken in this thesis is to provide data assimilation solutions, which can be applied in operational models used to provide value adding forecasts. First of all this requires robustness. The solution can not be marginally stable and it must handle missing data properly. For research purposes you need one good model run. In an operational setting you can not have one failed model run. Real-time operations further impose increased constraints on execution times. Often existing systems are already optimised to fill out these constraints and hence very efficient assimilation schemes are called for, if the resolution is to be maintained. Finally, the constraints on the physical consistency

of the state estimates are increased. Failure to provide a balanced estimate will result in generation of waves, which may deteriorate a forecast, where no measurements are available to correct the errors introduced.

Simultaneously, real-time operations provide strong constraints on computational efficiency. For dedicated cost-efficient commercial solutions, high performance computational facilities are typically not affordable and medium size computational resources must be employed, which sets even higher demands for computational efficiency.

1.5 Outline of thesis

Chapter 2 will provide an introduction to the methodology. The system description and state estimation are treated rather cursory, but present the very basic elements. For a more elaborate discussion, the included papers must be consulted. Section 2.3 states the main challenges in ocean state estimation and reviews techniques developed to address each problem. Chapter 3 gives a condensed overview of the papers included. These should be regarded as summaries of the undertaken approaches and results. Chapter 4 discusses the results in the context of the ocean state estimation challenges of Section 2.3 and draws conclusions on the work.

Chapter 2

Methodology

A fundamental formulation of the ocean state and parameter estimation problem is to cast it in an optimisation framework. This amounts to defining a function, J, which somehow expresses a fit or misfit between a modelled state estimate of physical properties and observations thereof. E.g. J could express a mean square error or a log-likelihood function. Traditionally, there are two different approaches to solving this problem. One is based on the variational principle and has its roots in control theory. This approach is followed in Section 2.4, dealing with parameter estimation. An alternative method with its roots in estimation theory provides a sequential solution to the problem. In a linear Gaussian framework this approach reduces to the Kalman filter, (Kalman 1960). Generalisations of the Kalman filter for solving the state estimation problem are introduced and discussed in Sections 2.2 and 2.3.

2.1 System description

The first step when building a mathematical framework for estimating the state of the ocean system, is to adapt the representation in which the ocean is described and observed. This is discussed further in Section 2.3. Having decided on a state representation, the state at time t_i can be written as a vector $\mathbf{x}^t(t_i)$ and the time propagation is expressed by

the system equation:

$$\mathbf{x}^t(t_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_{i-1})) + \boldsymbol{\eta}_i \tag{2.1}$$

where $\mathbf{u}(t_i)$ is the external forcing and the system error is denoted $\boldsymbol{\eta}_i$. The state vector, $\mathbf{x}^t(t_i)$, and model operator M may in the general setting be augmented.

The observations $\mathbf{y}_i^o$ may be expressed in terms of the selected state representation in the measurement equation:

$$\mathbf{y}_i^o = h_i(\mathbf{x}^t(t_i)) + \boldsymbol{\epsilon}_i \tag{2.2}$$

where $\boldsymbol{\epsilon}_i$ is the measurement error and h_i is the measurement operator.

2.2 State estimation

Equations 2.1 and 2.2 provide a common reference frame for the two independent sources of information, model and measurements. If it is assumed that the statistical properties of the two errors, $\boldsymbol{\eta}_i$ and $\boldsymbol{\epsilon}_i$, are known, a number of estimation techniques can theoretically be employed to estimate the state of the ocean. In the present work the Best Linear Unbiased Estimator (BLUE) is adapted. This estimator only requires knowledge of the first and second order moments of the stochastic variables $\mathbf{x}^t(t_i)$ and $\mathbf{y}_i^o$. Let the mean of these be $\mathbf{x}^f(t_i)$ and $\mathbf{H}_i\mathbf{x}^f(t_i)$ and their error covariances $\mathbf{P}_i^f$ and $\mathbf{R}_i$ respectively. $\mathbf{H}_i$ is a linearised operator of h_i. The BLUE estimate of the state $\mathbf{x}^a(t_i)$ can then be written,

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}^f(t_i)) \tag{2.3}$$

The Kalman gain matrix, $\mathbf{K}_i$, is given by,

$$\mathbf{K}_i = \mathbf{P}_i^f\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \tag{2.4}$$

The error covariance, $\mathbf{P}_i^a$, of $\mathbf{x}^a(t_i)$ will always be less than or equal to $\mathbf{P}_i^f$ and can be calculated as,

$$\mathbf{P}_i^a = \mathbf{P}_i^f - \mathbf{K}_i\mathbf{H}_i\mathbf{P}_i^f \tag{2.5}$$

The BLUE estimate constitute the Kalman filter, (Kalman 1960), in combination with a linear model operator for propagating the first and second moments of state in between measurement updates.

2.3 Challenges in ocean state estimation

The main challenges in state estimation of the marine environment are embedded in the following characteristics of the problem:

- Great dimensionality of the typical setting
- Nonlinearity of the system
- Non-Gaussianity of the state
- Different representations in which the continuous reality is observed and modelled
- Complexity of errors in numerical models of the ocean
- Heterogeneity of data sets

2.3.1 Great dimensionality

The size of the state space can easily reach $n = 10^7$ in a numerical model. If for no other reason, this renders the classical Kalman filter approach intractable because of the costly error covariance propagation ($2n$ times a normal model propagation) and storage (n^2 as compared to n). Luckily the effective degrees of freedom in an ocean model error covariance, n_f, is much smaller than n and hence it can be described efficiently in a much smaller subspace of size $n \times n_f$ with a corresponding reduction in propagation time to n_f times a normal model propagation.

Much of the work on data assimilation has been centered around finding the best approximations that yields a tractable solution to the estimation problem. Early attempts assumed stationarity of the model error covariance and solved the resulting equations off-line to provide a steady Kalman gain matrix, (Heemink 1986). Subsequently, methods explicitly exploiting the low degrees of freedom in a time varying setting was introduced. The Ensemble Kalman Filter (EnKF), (Evensen 1994), uses a Markov Chain Monte Carlo technique, while the Reduced Rank SQuare Root Kalman filter (RRSQRT), (Verlaan & Heemink 1997), uses a singular value decomposition to determine the directions in state space with

the largest components of uncertainty. Both these techniques are based on defining explicit error sources in the model. The Singular Evolutive Extended Kalman filter (SEEK), (Pham, Verron & Roubaud 1997), similarly provides a low order model error covariance representation, but derives the model error space from model dynamics space.

A number of extensions and refinements to these original formulations have been developed, but the foundation for reducing the great dimensionality is well established by them. For water level forecasting the Steady approximation used in (Cañizares, Madsen, Jensen & Vested 2001) is particularly important, because is brings down the computational demands to a level, where assimilation can be applied operationally. All schemes described this far are based on a reduced rank approximation of the covariance matrix. Other approaches approximate the model operator. (Dee 1991) used a simplified dynamical model imposing geostrophical balance in the atmosphere, while (Cohn & Todling 1996) employed a singular value decomposition of the model operator for the error covariance propagation and (Fukumori & Malanotte-Rizzoli 1995) used a coarse grid for the purpose.

2.3.2 Nonlinearity

The original Kalman filter is derived for a linear model operator. The ocean contains many nonlinear processes and thus violates this premise of the filter. The Extended Kalman filter, (Kalman & Bucy 1961), was introduced as a generalisation to weakly non-linear systems. Among other, the RRSQRT filter relies on this extended formalism. The approximation has been shown to be valid for coastal areas by (Madsen & Cañizares 1999) as well as (Cañizares 1999). (Verlaan & Heemink 2001) provides a more general test of the validity of the scheme. The EnKF handles even strong nonlinearities and thus non-Gaussianity in its state propagation. However, neither of the schemes, which all employ the BLUE estimator, handles the derived non-Gaussianity of nonlinear model propagation in the estimation part of the filter.

2.3.3 Non-Gaussianity

The optimality of the BLUE estimator relies on Gaussianity and unbiasedness of model variables as well as measurements, which is generally violated. All though the EnKF approximately propagates the non-Gaussian model error distribution, even this filter assumes Gaussianity in the BLUE estimator. In (Reichle, Entekhabi & McLaughlin 2002) a general mismatch between actual model errors and the standard deviation predicted by an EnKF is accredited to the non-Gaussianity of the state, which leads to an under estimation of the uncertainty. In order to handle non-Gaussianity we must look further into the application of higher order approximations of Bayesian state updating. In (Anderson & Anderson 1999) a fully nonlinear filter was used, but the approach is not feasible for large scale application.

2.3.4 Different representations in which the continuous reality is observed and modelled

Mostly, the model spatial and temporal discretisation defines the projection of the state representation. A projection on to this particular subspace is implicit in a numerical model anyhow. Observations represent different projections of reality. E.g. a tide gauge observation may be a 10 minute temporal average of the water level in an isolated $100cm^2$ position, while the model projection provides the average over a $2km \times 2km$ square with two minute time intervals. This mismatch poses the question: Is it a model error that it does not resolve $100cm^2$ area of the tide gauge or is it a measurement representation error that it does not provide a measurement of the $2km \times 2km$ box? The answer is that it depends on the projection selected for the state representation. Hence, this choice is crucial for any model and measurement error description. A parallel of this discussion can be drawn to the dynamical filter inherent in a numerical model.

(Fukumori & Malanotte-Rizzoli 1995) discusses the measurement representation error implicitly assuming that all estimation is done in the model space. However, they provide a simplistic representation error description by simply increasing the variance of the white measurement

noise.

2.3.5 Complexity of errors

Representation error is by no means the only error source, which is given a simplistic description. Model formulation, discretisation and parameterisation as well as parameter misspecification, round-off errors and uncertain boundary conditions all contribute to model errors. Measurements errors spans a wide range of characteristics depending on the variable measured and sensor type used. Hence, both model and measurement errors may typically be biased and non-Gaussian, while they are described by unbiased and Gaussian processes in the filters. However, the generally successful application of Kalman filter based algorithms shows that they have some skill in assessing the first order characteristics of errors, but this must not elude the fact that the error descriptions still are erroneous.

A general model and measurement noise model can be formulated as an augmented state description and their parameters estimated either in a variational setting or by the filter directly. (Dee 1995) devised a technique for estimating error model parameters, but it requires large amounts of simultaneous data for estimating only a few parameters.

2.3.6 Heterogeneous data sets

In many demonstrations of data assimilation, a fairly good data coverage is used or focus is put on the area where measurements are available. A data assimilation scheme generally corrects results close to measurements, since it always basically drags the model solution towards the measurement. However, if erroneous error descriptions and hence error correlations are used, then the information from the measurement may easily be used to provide erroneous updates in areas where no other measurements constrain the solution. A derived effect of this is the observed deterioration of water level predictions on intermediate time prediction horizons as reported by (Gerritsen, de Vries & Philippart 1995) and (Vested, Nielsen, Jensen & Kristensen 1995). Thus, sparse data sets

increases the demand for good error modelling.

Another kind of data heterogeneity is their multivariate nature and different error characteristics. Large data sets require extensive computational resources if treated classically and correlated errors requires the inversion of the innovation error covariance for calculating the Kalman gain. In (Haugen & Evensen 2002) a singular value decomposition (SVD) of the model error covariance is used to limit the assimilation to a subspace of the measurement space spanned by the largest model uncertainty.

2.4 Parameter estimation

Parameter estimation can in principle be solved by the sequential state estimation techniques discussed in Section 2.2 by augmenting the state vector with the parameters and the system equation with a consistency model for the parameters. However, the use of adjoint techniques in a variational setting has been shown to provide a successful and efficient solution to the problem, (Heemink, Mouthaan, Roest, Vollebregt, Robaczewska & Verlaan 2002). In any case attention needs to be paid to the cost function. (Evensen, Dee & Schröter 1998) show the need for including prior knowledge about parameter values along with the uncertainty of initial conditions, boundary conditions and model propagation in the cost function, for a well-posed problem to be formulated.

Variational approaches discussed above apply a gradient based optimisation to find the parameters that minimize the cost function, J. Solving the adjoint equations of the numerical model is a very efficient technique for finding the gradient of J with respect to the parameters. However, the main drawback of this approach is the demand for an adjoint code. Compilers for automatically generating adjoint codes have been developed, but have not yet been applied in any coastal ocean model and thus adjoint code generation remains costly in terms of man-power. This can be circumvented by calculating gradients of the cost function by finite differencing. This is a much more computationally demanding algorithm but it is easy to implement. Further, it is highly parallisable and hence with the advance of grid computing may become an attractive alternative to algorithms based on solving the adjoint equations for medium size

model applications.

Chapter 3

Overview of included papers

The papers included in this thesis are concerned with data assimilation of tide gauge and Sea Surface Temperature (SST) measurements in numerical models of the marine system. They cover aspects ranging from water level hindcasting in 2D and 3D hydrodynamic models to water level and SST forecasting and parameter estimation in a 2D hydrodynamic model. Throughout the papers, proposed techniques are either tested in simple idealistic settings or in the North Sea and Baltic Sea system.

Paper A deals with the sensitivity to filter parameters of the three data assimilation schemes: The EnKF, the RRSQRT filter and the Steady Kalman filter. The test bed is an idealised bay with a combined tidal and wind driven circulation. The general filter performance is good when matching the filter error description to the actual errors introduced. The sensitivity to the the filter parameters is investigated. The filter performance is demonstrated to be robust with respect to low to moderate parameter variations. For more typical non-Gaussian errors such as phase errors in the open boundary water level variation or misspecified wind field, the fairly high temporal and spatial correlations characterizing these errors must be assumed in order to obtain good performance. The uncertainty estimate of the filter is quite sensitive to misspecified parameters. Hence, more care should be taken, when interpreting uncer-

tainty estimates than the actual mean state estimates.

The basic framework underlying assimilation schemes based on the BLUE is discussed in Paper B, showing the equivalence between the Maximum a Posteriori (MAP) estimator and the BLUE for Gaussian distributions. Different formulations of the state space reduction allowing an error covariance propagation are then used to derive the the EnKF, the RRSQRT filter and the central EnKF combining a first order approximation of the mean state propagation with an ensemble estimate of the error covariance. These formulations are all based on assumptions of Gaussianity and unbiasedness. Further, the RRSQRT and the central EnKF assumes weak non-linearity at worst. Even the EnKF optimally assumes non-linearity, since non-linearity creates non-Gaussianity, which violates the BLUE assumption. In order to validate the underlying assumptions, measures of non-linearity, non-Gaussianity and bias are formulated based on the EnKF and the central EnKF. The measures are demonstrated in an idealised set-up in a semi-enclosed bay with a strong wind driven flow. All measures are shown to provide a realistic picture of their respective properties. Finally, sparse data coverage and approximate model error description is shown to deteriorate results far from measurements.

In Paper C a dynamical regularisation is suggested for the assimilation of tide gauge data in a three-dimensional model. It is based on the assumption that the error covariance structure is predominantly barotropic. Time averaged gains are derived from a barotropic model with an EnKF using 100 ensemble members. These are subsequently used in the three-dimensional model with a Steady Kalman filter. The filter modifications of the state are distributed to the three-dimensional velocity profile by assuming a vertically homogeneous shift of the velocity profile. The scheme is tested in the idealised bay also used in Paper A. This allows a comparison to a full three dimensional EnKF. The good performance of the elaborate EnKF in three dimensions is matched by the dynamically regularised scheme.

The regularisation technique thus demonstrated is applied in a model of the North Sea and Baltic Sea system in Paper D. This paper presents the operational Water Forecast modelling system considered and the water level and SST data chosen for assimilation in a pre-operational test. The SST assimilation builds on the work of (Annan & Hargreaves 1999).

The dynamically regularised assimilation technique shows good skill in the quite densely observed Inner Danish waters. The SST results shows a fair nowcast improvement in the mixed layer and in a 10-days forecast of the surface temperature.

The successful application of regularisation is followed up upon in Paper E. Here, the scheme introduced in Paper C is cast in a more general regularisation framework including also a smoothed Kalman gain evolution, the Steady Kalman filter and distance regularisation, where prior physically based assumptions about model error covariances can be accounted for. Only tide gauge data is considered and the proposed regularisations techniques are demonstrated in a pre-operational set-up of the Water Forecast model. Throughout all tests the dynamic regularisation is applied. The Steady Kalman filter is shown to perform as good as a low order EnKF using a smoothed Kalman gain evolution. The introduction of distance regularisation significantly increases the performance in data sparse regions which once again points to the importance of proper error covariance description when data sparsity is part of the setting.

In Paper F the water level forecast skill of the Steady Kalman filter with and without the distance regularisation introduced in Paper E and a newly introduced hybrid error correction Kalman filtering approach is investigated. The theoretical discussion focuses on the different representations of the real ocean in the model and measurements. The colored error that almost inevitably results leads to the formulation of a general system equation with augmented model and measurement error models. The properties of the innovation series is examined and it is shown that it will be colored when model and measurement errors are not well known. The information thus present in the innovation series is used to train an error correction model and hence the innovation can be forecast even after the time of forecast and assimilated by the Steady Kalman filter. The forecast skill of a barotropic model of the Water Forecast region is assessed using both the Steady filter for initialisation of the model state and using the hybrid error correction Kalman filter approach. The hybrid method was demonstrated to relatively improve results when the Steady filter forecast skill is only moderate. Distance regularisation was successfully included to vastly improve the forecast skill of the Steady initialisation. This however, left a smaller error to correct by the hybrid

scheme and hence no significant improvement was observed in this case.

Paper G reviews the work done on parameter estimation in hydrodynamic models and concludes in this respect that variational optimisation using adjoints provides the most efficient solution to the problem at present. It does however require an adjoint code and this is costly to develop despite improving automatic adjoint compilers. A more costly finite difference technique is used instead of the adjoint as part of of the optimisation problem. The approach may become a realistic future alternative to using the adjoint in models of moderate size, because of the advance of grid computing and the highly parallisable structure of the algorithm. Using this technique, wind and bottom drag friction parameters are estimated in a barotropic model of the Water Forecast region. Further, a weak constraint optimisation is approximated by employing the Steady Kalman filter in the model, thus accounting for model errors. This increases the parameter estimation skill.

Chapter 4

Conclusion and Discussion

The main issue in this thesis has been state estimation in continental shelf and coastal seas and parameter estimation in the numerical models thereof. The background and a brief methodology pointing out the main challenges of the scientific discipline have been provided in this summary report. The research consists of seven papers, which present a detailed methodology, discuss the nature of the state and parameter estimation problem and suggest operational solutions to some of the challenges posed.

The assimilation schemes used throughout this thesis build on the EnKF and the RRSQRT schemes, which have solved the challenge of the great dimensionality to a level, where data assimilation in large modelling system now has become feasible. The steady approximation provides an efficient algorithm, but its applicability can not be expected to be general and it still requires computational resources capable of generating the time-invariant gains by employing a more elaborate assimilation scheme such as EnKF or RRSQRT.

In situations with moderate variability of the Kalman gain, the smoothing factor introduced in Paper A can be used together with the EnKF to apply the right level of time variability and thus keep the ensemble size significantly lower than required by the original EnKF. The paper demonstrates good assimilation performance by the steady filter using

Kalman gain derived from an EnKF with ensemble size ten. This is to be compared to an ensemble size of 100 for the classic EnKF and 50 for the RRSQRT filter (with similar execution times as the EnKF with rank 100). This means that data assimilation can be used in a new class of applications, that previously had too high computational demands.

The dynamic regularisation introduced in Paper C and tested in the North Sea and Baltic Sea in Papers D and E provides an alternative way of making the assimilation schemes more efficient. A Kalman gain calculated by a barotropic model combined with a homogeneous vertical profile for the extrapolation to the three-dimensional velocity field is demonstrated to be sufficient for obtaining good performance matching that of applying the EnKF in the three-dimensional model directly. On existing computational resources the execution of MIKE 3 using EnKF with an ensemble size of 100 in the North Sea and Baltic Sea set-up considered was no where near feasible, but the dynamical regularisation approach made assimilation a realistic option nevertheless.

The treatment of the nonlinearity of the model operator has been a major issue in deriving the EnKF and the RRSQRT and their subsequent comparison. Hence, the schemes used in the thesis have the lessons learned last decade embedded. Paper B provides a discussion of nonlinearity and measures of the degree of non-linearity are suggested. These can be used to validate the underlying assumptions of a particular scheme in given settings and for available observations. This can guide the selection of the assimilation scheme in a subsequent application. Nonlinearity has important implications for the distribution in the stochastic state vector. This is usually assumed to be Gaussian, but with a nonlinear model operator, the distributions will inevitably be non-Gaussian. Paper B also formulates two measures of non-Gaussianity, which can be used to assess the proper statistical interpretation of the state estimates obtained.

A rather detailed discussion of the different filters through which model and observations see reality is provided in Paper F. The issue is most often not considered in data assimilation applications apart from inflating the measurement error by assuming representation error to be white and Gaussian. This simple approach is also followed in the applied Papers D, E and F. However, the implications of taking this issue properly into account is that measurement errors are most likely not white. They depend

on each other contrary to what is assumed for tide gauge measurements, and even on the system state. The importance of these dependencies and hence the error introduced by not taking them into account must be assessed in the future.

The simple description of representation error might be important, but is easily hidden behind the general problem of describing model errors. Paper F presents a general framework for describing model and measurement errors in a setting where numerical model and measurement errors are non-Gaussian. Presently, we are still some way from having developed techniques to estimate model error, and hence it makes sense to investigate the filter performance with misspecified model and measurement error descriptions. Paper A takes on such a sensitivity study and concludes that filter performance actually is pretty robust with respect to filter parameter variations in the given ideal test considered. This is encouraging for the application of the proposed tide gauge assimilation techniques in real cases. However, this does not ensure low sensitivity in other dynamical regimes and for all data types and variables.

Another important conclusion of Paper A is that the filter predicted standard deviation is sensitive to parameter variability. In any case, any filter application should accompanied by a test for whiteness of the innovation sequence or an analysis thereof. Paper F derives an expression for the autocorrelation of the innovation time series for misspecified measurement and model error covariances. The innovation sequence will only be white for correctly estimated error covariances. Paper F further suggests to use the information about the actual error covariances contained in the innovation to improve the error modelling and hence the forecast skill. Much work is still required to draw firm conclusions on the validity of such an approach, but initial results are encouraging.

Paper B introduces a bias measure for indicating erroneous error modelling and provides a simple example where a false error structure assumption gives a significant bias in data sparse regions. In the real application of Paper E, this problem is evident in the runs without distance regularisation. The hindcast results are severely deteriorated due to an inadequate model error description. In data sparse areas the model uncertainty is big and hence even a very small correlation with model estimates of a distant measurement can give a significant Kalman gain

in data sparse regions. The approximate model error description is unfortunately too poor for these correlations to be trusted and no local measurements are available to constrain the solution.

This ideally calls for improved error modelling, but the alternative of using a regularisation approach is taken in Paper E. The distance regularisation is introduced to remedy for the erroneous behaviour described above, and does so very effectively. The forecast skill when employing the distance regularisation is also significantly improved in Paper F. The regularisation approach to the filtering is general and must be expected to have a large potential in sequential filtering.

A variational parameter estimation framework was demonstrated in Paper G with the perspective of ease of implementation and efficiency in a grid computing environment. The test of the approach in the North Sea and Baltic Sea system showed the need for including the bathymetry as a control parameter, use a longer time period, to decouple the optimisation for tidal and wind driven circulation and to employ a more efficient optimisation algorithm. The Steady Kalman filter was used in one optimisation approach to approximate a weak constraint formulation for the model state. Despite the flaws of the test case, this weak constraint approach showed a more robust optimisation than the strong constraint with no data assimilation. The work done is somewhat preliminary, but now the stage is set for exploring the technique in parallel with the emergence of grid computing facilities.

Future research will extend the ideas presented to other data types such as salinity and temperature profiles, SST data, ecosystem parameters and HF radar velocity measurements. This will restate the challenges presented and the ideas on dimensionality reduction, error description, regularisation and forecasting skill improvement in a nonlinear, non-Gaussian setting presented in this thesis will be further pursued. Techniques for adaptive model error estimation should be developed and further exploration of the full potential of regularisation techniques undertaken. A parallel implementation of the EnKF will also be an objective. Finally, application of regularisation techniques in parameter estimation is a topic of interest for making optimisation techniques that do not require an adjoint code more feasible through integration with the advance of grid computing facilities.

Bibliography

Anderson, J. L. & Anderson, S. L. (1999), 'A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts', *Monthly Weather Review* **127**, 2741–2758.

Annan, J. D. & Hargreaves, J. C. (1999), 'Sea surface temperature assimilation for a three-dimensional baroclinic model of shelf seas', *Continental Shelf Research* **19**, 1507–1520.

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

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Cohn, S. E. & Todling, R. (1996), 'Approximate data assimilation schemes for stable and unstable dynamics', *Journal of Meteorological Society of Japan* **74**, 63–75.

Dee, D. P. (1991), 'Simplification of the Kalman filter for meteorological data assimilation', *Q.J.R. Meteorological Society* **117**, 365–384.

Dee, D. P. (1995), 'On-line estimation of error covariance parameters for atmospheric data assimilation', *Monthly Weather Review* **123**, 1128–1145.

DHI (2001), *MIKE 3 estuarine and coastal hydrodynamics and oceanography*, DHI Water & Environment.

DHI (2002), *MIKE 21 coastal hydraulics and oceanography*, DHI Water & Environment.

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

Evensen, G., Dee, D. & Schröter, J. (1998), Parameter estimation in dynamical models, *in* E. P. Chassignet & J. Verron, eds, 'Ocean Modeling and Parameterizations', NATO ASI, Kluwer Acad. Pub.

Fukumori, I. & Malanotte-Rizzoli, P. (1995), 'An approximate Kalman filter for ocean data assimilation; an example with an idealised Gulf Stream model', *J. Geoph. Res.* **100**(C4), 6777–6793.

Gerritsen, H., de Vries, H. & Philippart, M. (1995), The Dutch continental shelf model, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 19, pp. 425–467.

Haugen, V. E. J. & Evensen, G. (2002), 'Assimilation of sst and sla data into an ogcm for the indian ocean', *Ocean Dynamics* **52**, 133–151.

Heemink, A. W. (1986), Storm surge prediction using Kalman filtering, PhD thesis, Twente University of Technology.

Heemink, A. W., Mouthaan, E. E. A., Roest, M. R. T., Vollebregt, E. A. H., Robaczewska, K. B. & Verlaan, M. (2002), 'Inverse 3d shallow water flow modelling of the continental shelf', *Continental Shelf Research* **22**, 465–484.

Kalman, R. E. (1960), 'A new approach to linear filter and prediction theory', *Journal of Basic Engineering* **82**(D), 35–45.

Kalman, R. E. & Bucy, R. S. (1961), 'New results in linear filter and prediction theory', *Journal of Basic Engineering* **83**(D), 95–108.

Madsen, H. & Cañizares, R. (1999), 'Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling', *International Journal of Numerical Methods in Fluids* **31**(6), 961–981.

Pham, D. T., Verron, J. & Roubaud, M. C. (1997), 'Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography', *Journal of Marine Systems* **16**, 323–340.

Reichle, R. H., Entekhabi, D. & McLaughlin, D. B. (2002), 'Downscaling of radiobrightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach', *Water Resources Research* **37**, 2353–2364.

Verlaan, M. & Heemink, A. W. (1997), 'Tidal flow forecasting using reduced rank square root filters', *Stochastic Hydrology and Hydraulics* **11**, 349–368.

Verlaan, M. & Heemink, A. W. (2001), 'Nonlinearity in data assimilation applications: A practical method for analysis', *Monthly Weather Review* **129**, 1578–1589.

Vested, H. J., Nielsen, J. W., Jensen, H. R. & Kristensen, K. B. (1995), Skill assessment of an operational hydrodynamic forecast system for the North Sea and Danish belts, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 17, pp. 373–396.

Papers

A

Paper A

Parameter sensitivity of three Kalman filter schemes for the assimilation of tide gauge data in coastal and shelf sea models

Submitted to *Ocean Modelling*.

Parameter sensitivity of three Kalman filter schemes for the assimilation of tide gauge data in coastal and shelf sea models

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

Abstract

In applications of data assimilation algorithms, a number of poorly known parameters usually needs to be specified. Hence, the documented success of data assimilation methodologies must rely on a moderate sensitivity to these parameters. This study presents three well known Kalman filter approaches for the assimilation of tidal gauge data in a three dimensional hydrodynamic modelling system. It undertakes a sensitivity analysis of key parameters in the schemes for a setup in an idealised bay. The sensitivity of the resulting RMS error is shown to be low to moderate. Hence the schemes are robust within an acceptable range and their application even with misspecified parameters is to be encouraged in this perspective. However, the predicted uncertainty of the assimilation results are sensitive to the parameters and hence must be applied with care.

1 Introduction

Data assimilation methodologies are becoming increasingly applied in the ocean modelling community. The methods employed can be categorised according to two basic approaches: Sequential estimation and variational optimisation. In this paper only the former approach is considered although most of the conclusions drawn on the error structure formulation carries over to the latter.

The standard approach and hence terminology of sequential estimation techniques is that of the Kalman filter, (Kalman 1960). The original Kalman filter was derived for a linear system with Gaussian error sources.

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

When applied to non-linear and high dimensional systems, the formulation demands vast computational resources and its limitations in terms of Gaussian error assumptions and linearity become clear. Several extensions have been made in an attempt to accommodate for such deficiencies.

Primarily, the problem needs to be solvable on available computational resources. The most widespread techniques for making the problem tractable are ensemble based. Basically these schemes represent the information contained in the error covariance matrix in a reduced space spanned by a small number of ensembles. The Ensemble Kalman Filter (EnKF), (Evensen 1994) and the Reduced Rank SQuare RooT Kalman filter (RRSQRT), (Verlaan & Heemink 1997), are examples presented in this paper. Two alternative popular ensemble based approaches are the SEEK filter, (Pham et al. 1997), and the SEIK filter, (Pham, Verron & Gourdeau 1998). A recent review of ensemble based Kalman filters is provided in (Evensen 2003). Another approach reducing the computational cost uses a simpler description of model dynamics. This can either be done by using a coarser grid for the error covariance modelling in the numerical model, (Cohn & Todling 1996) and (Fukumori & Malanotte-Rizzoli 1995), or by approximating time consuming elements of the numerical model, such as employing cheaper numerical schemes, simpler turbulence closure schemes or assuming geostrophic balance for the error covariance propagation, (Dee 1991).

A significant reduction in computational time can be obtained with the Steady Kalman filter, where the model error covariance or the Kalman gain is assumed to be the same at each update time. (Fukumori & Malanotte-Rizzoli 1995) derives such a steady gain from limiting theory solving the time invariant Riccati equation. (Cañizares et al. 2001) also uses a steady approach, but here the steady gain is calculated as a time average of the EnKF. The steady approach generally reduces computational times with two orders of magnitude compared to the EnKF and is only slightly more computationally demanding than a single execution of a numerical model.

Extensions to the Kalman filter need to accommodate for non-linearities in the model propagation and the measurement equation. Also, bias or coloured noise in the numerical model and the measurements requires attention. Most schemes use a non-linear numerical model for the state

propagation, while the foreward operator employed for the error covariance propagation ranges from a steady linear operator, (Fukumori & Malanotte-Rizzoli 1995), to a linear expansion in extended Kalman filter applications such as the RRSQRT filter and a full non-linear error propagation in the EnKF.

While the handling and nature of non-linearities in a data assimilating system thus have been widely examined, the importance of using a proper error structure and robustness to error misspecification has gained only sporadic attention. The optimality of the Kalman filter assumes known and unbiased model and measurement errors. However, the estimation of these errors is to some extent subjective and can typically never be estimated from the limited data sets available. Further, structural model errors often lead to biased model states. (Dee & da Silva 1998) present a scheme for the simultaneous estimation of the unbiased state and the model bias. (Cañizares 1999) and (Verlaan 1998) both uses a coloured noise implementation. (Sørensen, Madsen & Madsen 2004a) investigates the behaviour under misspecification of the model error in the case of a biased forcing. In all cases a clear improvement of the estimate results from correct error structure specification.

In a general data assimilation application the error sources are typically only known to a first or second order approximation and hence misspecification is part of the working conditions. However for storm surge models, good performance is nevertheless demonstrated in schemes, which do not explicitly account for the actual error structure, e.g. (Madsen & Cañizares 1999). This must be accredited to a sufficient information content of the measurements and subsequent distribution. Bias is also corrected by a Kalman filter approach assuming no bias, albeit in a suboptimal way, (Dee & da Silva 1998). The specification of error structure and its subsequent propagation only need to provide a good interpolation of the innovation in space and time. Hence, when many data are available, the importance of a proper error model is reduced. In the case of assimilation of tidal gauge data, as considered herein, the measurements are usually sparsely distributed in space. Thus, the error structure provides the mean for updating state elements situated far from points of observation and hence its description becomes more important.

Focusing at the three state-of-the-art assimilation schemes, the EnKF,

the RRSQRT filter and the Steady filter, with a coloured noise assumption implemented in a 3D hydrodynamic model, this papers sets out to perform a sensitivity study of the schemes for various parameter settings. Acknowledging that misspecifications are often part of the working conditions such a study provides insight to the effect on performance of uncertain parameters. Hence calibration can be focused at key parameters and in case of low sensitivity, confidence can be build in the schemes even for moderately misspecified parameters.

Section 2 will introduce the building blocks of the assimilation approach, which provides the Kalman filter as a special case. The three schemes, which constitute the basis of this study will be described briefly - namely the EnKF, the RRSQRT and the Steady Kalman filter. In Section 3 the filter parameters in the schemes are presented and discussed. In Section 4 results are presented for a range of sensitivity twin experiments using an idealised bay test case. Finally, Section 5 summarises and concludes the paper. The notation suggested by (Ide, Courtier, Ghil & Lorenc 1997) is used throughout.

2 Assimilation approach

The foundation of sequential estimation schemes is a linear model for combining the information contained in a model with measurements in an estimate of state variables. Hence, let $\mathbf{x}^t(t_i) \in \mathbb{R}^n$ be a representation of the true state at time t_i. This could be an array of grid averaged water levels and velocities at all model grid points in the area of interest. It can also contain additional augmented elements from an error model. Let $\mathbf{x}^f(t_i) \in \mathbb{R}^n$ be the model estimate of $\mathbf{x}^t(t_i)$ and $\mathbf{y^o}_i \in \mathbb{R}^p$ be a vector of observations at time t_i, which is assumed related to the state vector through the measurement equation,

$$\mathbf{y^o}_i = \mathbf{H}_i \mathbf{x}^t(t_i) + \boldsymbol{\epsilon}_i \tag{1}$$

The operator $\mathbf{H}_i \in \mathbb{R}^{p \times n}$ projects the state space onto the measurement space. The measurement noise is assumed additive and represented by the random variable, $\boldsymbol{\epsilon}_i \in \mathbb{R}^p$. The relation in (1) is assumed linear.

With the definitions given above and assuming both $\mathbf{x}^f(t_i)$ and $\mathbf{y^o}_i$ to be

unbiased, a linear unbiased estimate, $\mathbf{x}^a(t_i)$, of $\mathbf{x}^t(t_i)$ can be obtained as

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i(\mathbf{y}_i^o - H_i\mathbf{x}^f(t_i)) \tag{2}$$

A good sequential assimilation scheme is characterised by a proper estimation of the elements of the linear operator, $\mathbf{K}_i \in \mathbb{R}^{n\times p}$, which is denoted the Kalman gain. What is meant by proper depends on the application at hand and the properties of the estimate, $\mathbf{x}^a(t_i)$, sought for. Usually assumptions about linearity and unbiasedness are imposed and a least squares approach is taken. This is followed in the next section.

2.1 The BLUE estimator

The linear projection of $\mathbf{x}^t(t_i)$ on $\mathbf{y^o}_i$ provides the best (minimum variance) linear unbiased estimate (BLUE) of $\mathbf{x}^t(t_i)$, (Jazwinski 1970). The first moment of $\begin{bmatrix} \mathbf{x}^t(t_i) \\ \mathbf{y^o}_i \end{bmatrix}$ is given as $\begin{bmatrix} \mathbf{x}^f(t_i) \\ \mathbf{H_i x}^f(t_i) \end{bmatrix}$ and its covariance matrix is $\begin{bmatrix} \mathbf{P}_i^f & \mathbf{P}_i^f\mathbf{H}_i^T \\ \mathbf{H}_i\mathbf{P}_i^f & \mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^T+\mathbf{R}_i \end{bmatrix}$. Here, $\mathbf{R}_i$ is the covariance of the noise process $\boldsymbol{\epsilon}_i$ in (1). The linear projection $\mathbb{E}(\mathbf{x}^t(t_i)|\mathbf{y^o}_i)$ is given by (2) with $\mathbf{K}_i$ given by

$$\mathbf{K}_i = \mathbf{P}_i^f\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \tag{3}$$

The error covariance of the estimated state, $\mathbf{x}^a(t_i)$, is given by

$$\mathbf{P}_i^a = \mathbf{P}_i^f - \mathbf{K}_i\mathbf{H}_i\mathbf{P}_i^f \tag{4}$$

The quest now becomes the estimation of $\mathbf{P}_i^f$ and $\mathbf{R}_i$. The measurement error is usually assumed constant in time and is prescribed according to measurement uncertainty and its representation in the model state. For sequential Kalman filter based algorithms, the discrepancies lies in the approximations made in the estimation of $\mathbf{P}_i^f$ and $\mathbf{x}^f(t_i)$. Three approaches are described in Section 2.2.

It should noted that the BLUE estimator assumes that $\mathbb{E}(\mathbf{x}^t(t_i)|\mathbf{y^o}_i)$ is linear in $\mathbf{y^o}_i$. This is true in the measurement point if the linear relation in (1) is valid. However, the state variables are generally not linearly related and hence this assumption is not valid. The minimal variance property of the BLUE estimator and hence the Kalman filter

only applies to class of linear functions. Further, the numerical model does in most cases not provide an unbiased estimate of the true state. Since most sequential assimilation schemes employ this estimator, they are also subject to these sources of sub-optimality.

2.2 State and error propagation

The basis of the predictions is a numerical hydrodynamic model. In this study the model adapted is MIKE 3, which is developed at DHI Water & Environment, (DHI 2001). The code that constitutes a one-time-step-ahead prediction can be regarded as a model propagation operator, M_{M3}. With knowledge of the state at time t_{i-1} and the forcing, $\mathbf{u}(t_i)$, it provides the state at time t_i. The state considered consists of velocities and water levels on a specified grid. The forcing is open boundary water levels, sources and sinks, wind velocities and atmospheric pressure. Hence in a standard non-assimilating application of the numerical model, the one-time-step-ahead prediction can be written as,

$$\mathbf{x}^f(t_i) = M_{M3}(\mathbf{x}^f(t_{i-1}), \mathbf{u}(t_i)) \tag{5}$$

In this case the state description and propagation are deterministic.

Acknowledging the approximate nature of a numerical model, a more detailed description must incorporate the model error introduced at each time step and its propagation throughout the system. Thus the propagation operator and the the state become stochastic. The error introduced by the model propagation, $\boldsymbol{\xi}_i \in \mathbb{R}_i^r$ is evident in the system equation,

$$\mathbf{x}^t(t_i) = M_{M3}(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_i), \boldsymbol{\xi}_i) \tag{6}$$

Optimally, knowledge of the correct time varying probability density function (pdf) of the noise sequence $\boldsymbol{\xi}_i$ and an initial state could be used to provide an exact stochastic forecast of the state. However, representation and propagation of the full pdf is not a tractable approach. A common approximation is to consider only the first and second order moments of the distribution. In case of Gaussian random fields and a linear model operator this further describes the full probability distribution. In addition, only these two moments are needed for the BLUE estimator in (2) and (3). Another approach is to approximate the propagation of the

pdf to a precision, which allows a confident estimation of the first and second moments. Two schemes for forecasting the first and second order moments in the hydrodynamic model and subsequent update of the state conditioned on observations are presented in Sections 2.2.1 and 2.2.2. A steady Kalman gain reduction of these elaborate techniques is presented in Section 2.2.3.

Of prime importance to any assimilation scheme is a proper model and measurement error description. As previously stated, the measurement error is usually based on simple assumptions. The specification of model error is a more difficult task. It is generally assumed proportional to the model dynamics variability in some way or to originate solely from external forcing fields. The latter approach is applied here. Errors are introduced in the open boundary water level and in the wind velocity. Hence, the numerical model is in itself assumed to be perfect. In coastal and shelf seas this is often a good approximation to actual model inaccuracies. A further step is taken by assuming the error to be coloured as described by a first order autoregressive model $M_{AR(1)}$,

$$\boldsymbol{\xi}_i = M_{AR(1)}(\boldsymbol{\xi}_{i-1}, \boldsymbol{\eta}_i) \;=\; \mathbf{A}\,\boldsymbol{\xi}_{i-1} + \boldsymbol{\eta}_i \tag{7}$$

Cross correlations are neglected, suc that $\mathbf{A} = diag(\boldsymbol{\alpha})$, where the vector $\boldsymbol{\alpha}$ contains the coeficient of the autoregressive model. The noise process $\boldsymbol{\eta}_i$ is assumed Gaussian with zero mean and error covariance matrix, $\mathbf{Q}_i^{\eta} \in \mathbb{R}^{r\times r}$. Hence, $\mathbf{x}^t(t_i)$ is augmented with the boundary and wind error description and an extended operator, $M = (M_{M3}, M_{AR(1)})^T$, is introduced. This leads to a system equation with additive noise, which will be used in the remainder of this work,

$$\mathbf{x}^t(t_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_i), \boldsymbol{\eta}_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_i)) + \begin{bmatrix} 0 \\ \boldsymbol{\eta}_i \end{bmatrix} \tag{8}$$

The error covariance of $\begin{bmatrix} 0 \\ \boldsymbol{\eta}_i \end{bmatrix}$ is $\mathbf{Q}_i = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{Q}_i^{\eta} \end{bmatrix}$. Running the hydrodynamic model alone with error propagation according to one of the schemes described below, but with no assimilation, will yield a model mean estimate and its error covariance matrix. Hence, the accuracy of model results based on the given assumptions can be addressed and compared to observational evidence.

2.2.1 The ensemble Kalman Filter

In the Ensemble Kalman Filter (EnKF) the propagation of the full pdf is approximated by an ensemble propagation. Both the first and second moments are calculated as ensemble statistics and used for the update of each ensemble member. The strength of the approach lies in its representation of the full pdf, its handling of non-linearities and its ease of implementation for complex state and error descriptions.

An ensemble of q state realisations is defined at an initial time. In this work, a single model initial state defines all ensembles with zero spread at a pre-initial point of time as the starting conditions of a spin-up period. During this period the forcing errors are propagated throughout the system to provide the initial model error covariance matrix and mean state estimate.

All ensemble members are propagated according to model operator in (8),

$$\mathbf{x}_j^f(t_i) = M(\mathbf{x}_j^a(t_{i-1}), \mathbf{u}(t_i), \boldsymbol{\eta}_{j,i}), \quad j = 1, ..., q \tag{9}$$

The state estimate $\mathbf{x}_j^a(t_{i-1})$ is the update from the previous time step. If no new data were available for update then $\mathbf{x}_j^a(t_{i-1}) = \mathbf{x}_j^f(t_{i-1})$. The model error, $\boldsymbol{\eta}_{j,i}$ is randomly drawn from a predefined Gaussian distribution with zero mean and covariance, $\mathbf{Q}_i$. With each ensemble member propagated by (9), the mean state estimate and model error covariance estimate are provided by the following equations,

$$\hat{\mathbf{x}}^f(t_i) = \frac{1}{q}\sum_{j=1}^{q} \mathbf{x}_j^f(t_i) \tag{10}$$

$$\mathbf{P}_i^f = \mathbf{S}_i^f(\mathbf{S}_i^f)^T, \quad \mathbf{s}_{j,i}^f = \frac{1}{\sqrt{q-1}}(\mathbf{x}_j^f(t_i) - \hat{\mathbf{x}}^f(t_i)) \tag{11}$$

The vector, $\mathbf{s}_{j,i}^f \in \mathbb{R}^n$, is the $j'th$ column of $\mathbf{S}_i^f \in \mathbb{R}^{n\times q}$. The update can be performed by (2) and (3), when given the proper interpretation in an ensemble setting. For computational efficiency an algebraically equivalent set of equations are used.

Each ensemble member must be updated rather than the ensemble state estimate, in order to maintain correct statistical properties of the updated ensemble. For the same reason an ensemble of measurements

must be generated and used for each ensemble member update accordingly rather than the measurement itself, (Burgers, van Leeuwen & Evensen 1998). Hence,

$$\mathbf{y}^o_{j,i} = \mathbf{y}^o_i + \boldsymbol{\epsilon}_{j,i}, \quad j = 1, ..., q \tag{12}$$

Randomly generated realisations, $\boldsymbol{\epsilon}_{j,i}$, of $\boldsymbol{\epsilon}_i$ are drawn from a Gaussian distribution with zero mean and error covariance, $\mathbf{R}$, and added to each member. The update scheme presented here specifically considers that measurement errors are uncorrelated to assimilate simultaneous measurements sequentially. The updating algorithm for every ensemble member, j, reads, (Chui & Chen 1991),

$$\mathbf{x}^a_{j,m}(t_i) = \mathbf{x}^a_{j,m-1}(t_i) + \mathbf{k}_{i,m}(y^o_{j,i,m} - \mathbf{h}_{i,m}\mathbf{x}^a_{j,m-1}(t_i)), \quad m = 1, ..., p \tag{13}$$

and $\mathbf{x}^a_{j,0}(t_i) = \mathbf{x}^f_j(t_i)$. In (13) $y^o_{j,i,m}$ is the $m'th$ element in $\mathbf{y}^o_{j,i}$ and $\mathbf{h}_{i,m}$ is the $m'th$ row of $\mathbf{H}_i$. Treating one measurement at a time the Kalman gain is a vector, $\mathbf{k}_{i,m}$, given by,

$$\mathbf{k}_{i,m} = \frac{\mathbf{S}^a_{i,m-1}\mathbf{c}_{i,m}}{\mathbf{c}^T_{i,m}\mathbf{c}_{i,m} + \sigma^2_{i,m}}, \quad \mathbf{c}_{i,m} = (\mathbf{S}^a_{i,m-1})^T\mathbf{h}^T_{i,m} \tag{14}$$

The $m'th$ diagonal element in $\mathbf{R}_i$ is denoted $\sigma^2_{i,m}$. The matrix $\mathbf{S}^a_{i,m}$ in (14) is calculated as

$$\mathbf{S}^a_{i,m} = [\mathbf{s}^a_{1,i,m}...\mathbf{s}^a_{q,i,m}], \quad \mathbf{s}^a_{j,i,m} = \frac{1}{\sqrt{q-1}}(\mathbf{x}^a_{j,m}(t_i) - \hat{\mathbf{x}}^a_m(t_i)) \tag{15}$$

for $m = 1, ..., p$ and $\mathbf{S}^a_{i,0} = \mathbf{S}^f_i$. Now, (13), (14) and (15) provides the update equations of all ensemble members, one measurement at a time.

2.2.2 The reduced rank square root Kalman Filter

The Reduced Rank Square Root Kalman filter (RRSQRT) is based on the extended Kalman filter formalism, in which the error propagation is calculated using a statistical linearisation of the model propagation operator. It further uses a square root algorithm and a lower rank approximation of the error covariance matrix. Thus, it handles weak nonlinearities and it has a concise and smooth representation of the error covariance matrix.

The state propagation is the model forecast of the central estimate,

$$\mathbf{x}^f(t_i) = M(\mathbf{x}^a(t_{i-1}), \mathbf{u}(t_i), 0) \tag{16}$$

The error covariance propagation basically performs the following truncated Taylor Series approximation,

$$\mathbf{P}^f(t_i) = \mathbf{M}_i \mathbf{P}^a(t_{i-1}) \mathbf{M}_i^T + \mathbf{Q}_i, \mathbf{M}_i = \left. \frac{\partial M}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^f(t_i), \mathbf{u}=\mathbf{u}(t_i), \boldsymbol{\eta}=0} \tag{17}$$

A square root implementation of this propagation and subsequent update has been performed. Denote by $\mathbf{S}^a(t_{i-1})$ the approximation of rank q of the square root of the error covariance matrix $\mathbf{P}^a(t_{t-1})$. The propagation of the error covariance matrix approximated according to (17), is then given by,

$$\mathbf{S}^f(t_i) = [\mathbf{M}_i \mathbf{S}^a(t_{i-1}) | \mathbf{Q}_i^{1/2}] \tag{18}$$

To calculate the derivatives needed in $\mathbf{M}_i$ a finite difference approximation of M is column-wise adopted as follows,

$$(\mathbf{M}_i \mathbf{S}^a(t_{i-1}))_j = \frac{M(\mathbf{x}^a(t_{i-1}) + \delta s^a_{j,i-1}, u(t_i), 0) - M(\mathbf{x}^a(t_{i-1}), u(t_i), 0)}{\delta} \tag{19}$$

The value of the parameter δ has been discussed in (Segers, Heemink, Verlaan & van Loon 2000) and is set equal to one according to their recommendation. The propagation step in (18) increases the number of columns in the error covariance matrix from q to $q + r$. Thus a complimentary part of the scheme must provide a mean for reducing the rank of the space similarly. In order to do this a lower rank approximation of $\mathbf{S}^f(t_i)$ in (18) is applied through an eigenvalue decomposition of the matrix $(\mathbf{S}^f(t_i))^T \mathbf{S}^f(t_i)$, (Verlaan 1998). This approach provides efficient calculations, but introduces the need for normalisation in a multivariate setting. The optimal normalisation is application dependent which is approximated by using normalisation based on energy consideration. Basically, the constribution from potential and kinetic energy to each element of $(\mathbf{S}^f(t_i))^T \mathbf{S}^f(t_i)$ are equal.

Now, having calculated the forecast state and error covariance, the algorithms developed for the EnKF can be followed to provide the update. The state is updated using (13). However, an additional ingredient is needed, namely the update of the error covariance estimate, when the

state is updated. This is provided by, (Cañizares 1999),

$$\mathbf{S}^a_{i,m} = \mathbf{S}^a_{i,m-1} - \frac{\mathbf{k}_{i,m}\mathbf{c}^T_{i,m}}{1 + \sqrt{\frac{\sigma^2_{i,m}}{\mathbf{h}^T_{i,m}\mathbf{h}_{i,m}+\sigma^2_{i,m}}}} \tag{20}$$

The vectors, $\mathbf{k}_{i,m}$ and $\mathbf{h}_{i,m}$ are defined in (14).

2.2.3 The Steady Kalman Filter

For the Ensemble Kalman filter and the reduced rank square root Kalman filter, the error covariance propagation typically takes of the order 10^2 model executions, (Cañizares 1999), which may be too many in operational settings. A well known work around assumes time invariant model and measurement error covariance matrices, $\mathbf{P}^f_i$ and $\mathbf{R}_i$, rendering a time constant Kalman gain, $\mathbf{K}$. However, it can still be difficult to estimate $\mathbf{K}$ without the help from more elaborate methods.

When time invariance is approximately true, both the EnKF and the RRSQRT can provide robust estimates of the gain. Hence a Kalman gain that is still based on model dynamics can be obtained as a time average of the gain from one of these two elaborate methods, (Cañizares et al. 2001). The update is still done using (13), but now with a fixed, $\mathbf{K}$, and operating on just a single state forecast.

3 Filter parameters

In an actual implementation of the filters above, several parameters need to be specified. These mainly relate to the model and measurement error covariance description. This section describes each parameter, while the sensitivity to parameter variations is tested in Section 4.2.

Rank: q. The rank of the model error covariance matrix is essential to the performance of an assimilation scheme. For the ensemble Kalman filter this is equal to the ensemble size, while for the RRSQRT Kalman filter it is the number of leading eigenvalues preserved in the covariance

reduction. In either case the rank needs to be large enough to describe the error covariance field with sufficient accuracy, but with the trade-off of increased computational time.

Measurement standard deviation: σ_m. This study only considers uncorrelated water level measurements. Hence, the error specification is simply given by a value of σ_m for each measurement. Both instrumentation error and the error due to lack of representation of state variables need to be taken into account in the specification.

Model error standard deviation: σ_η. The model error specification has a more complex description. The model errors are assumed to be due to an error in the open boundary water level and/or in the wind forcing. The assumed standard deviation, σ_η, of the white noise process, η_i is naturally a key parameter. An independent set of parameters are specified for each boundary and wind velocities. The relative sizes of model and measurement uncertainty basically determines which source of information that ought to be trusted the most in the state estimation.

Temporal correlation scale: τ. The temporal correlation scale defines the coefficient, α, in the AR(1) process by giving the half time τ of the exponential process.

$$\alpha = 0.5^{(\frac{\Delta t}{\tau})} \tag{21}$$

where Δt is the model time step. Note that since the noise enters into an autoregressive process the actual standard deviation, $\sigma_{forcing}$ of the boundary or wind forcing is given by,

$$\sigma^2_{forcing} = \frac{\sigma^2_\eta}{1-\alpha^2} \tag{22}$$

A higher temporal correlation allows a more distant effect of an error in the forcing. The formulation ensures that the standard deviation can be specified from (21) and (22) to construct a forcing perturbation that is independent of time step length, Δt, and changes in τ.

If $\alpha = 1$ and σ_η is zero with a suitable initial covariance, then the Kalman filter provides a bias estimate of the forcing terms. An α close to one and a moderate σ_η approximates this bias estimation, hence enabling the filters to detect a slowly varying errors in the external forcing.

Spatial correlation length: l_c. An exponential correlation model is employed in the definition of the error covariance model, $\mathbf{Q}^\eta$. The spatial correlation length of the model errors plays a key role in defining the correlation structure in the model that ultimately determines the update. A too large spatial correlation scale assumption in the wind velocities can cause an update in sparse regions based on a measurement, which does not contain any information about this distant area in the real system. On the other hand, a too small spatial correlation scale underestimates the correlation in the model errors and thereby provides a filtered estimate, which is too close to the model solution. It also increases the effective number of degrees of freedom in the error model, which in turn makes the estimated parameters of the error model more uncertain.

Grid factor: g. The grid factor is introduced as an ad hoc approach to reducing the dimension, r, of the error space. This integer factor simply expresses the number of model grid points in between each error point. The errors are subsequently redistributed using a kriging technique. Such a space reduction is viable because the spatial correlation length often is considerably larger than the grid spacing. However, when the spatial correlation length approaches the distance between error points, this assumption is violated and the spatial correlation length looses its interpretation. Other space reduction techniques, e.g. EOF decomposition, can also be cast in the present framework.

Smoothing factor: s. The Steady Kalman filter uses a time average from one of the more elaborate schemes for the generation of the Kalman gain. A greater deal of smoothness can be obtained in the time varying Kalman filter schemes as well through the introduction of an exponential smoothing factor. This number is the proportion of weight given to the Kalman gain calculated at the present time, $\mathbf{K}_i^{KF}$. The applied gain matrix then becomes

$$\mathbf{K}_i = (1-s)\mathbf{K}_{i-1} + s\mathbf{K}_i^{KF}, s \in [0;1] \tag{23}$$

Update interval: d. In practical applications it must be considered whether all data shall be assimilated. Trying to drive the model into an observed regime with effects not represented by the model, can introduce noise and ultimately cause model instabilities. In practice, water levels may typically be provided at half hourly intervals, but the implementation of the data assimilation schemes interpolates the measurements

to every model time step and assimilates it. Hence, an update interval is introduced as the final parameter to test the effect of using various subsets of the interpolated water level measurements.

4 Idealised bay experiment

For the purpose of investigating filter performance, when basic assumptions are violated, an idealised, controllable and stable setup was chosen. This choice also facilitates a large number of sensitivity runs to be performed and a comparison to the full true field to be done.

4.1 Setup and basic results

The region under consideration is the hypothetical Ideal Bay situated at $51°N$. It is a 200 km by 200 km square bay with an open Northern boundary and simple bathymetry with a maximum depth of $100m$ as shown in Figure 1. The vertical grid spacing is 10 meters and the horizontal resolution is 10 km.

Density is constant in this study, which is conducted over a 48 hour period. The open Northern boundary is forced with a spatially constant water level signal with a sinusoidal variation in time. The period is 12 hours and the amplitude is one meter. The model is further forced by an artificially generated passing cyclone, which moves 50 km across the bay every six hours. It has a maximum wind speed of 26 m/s.

The basic solution has a main flow, which is dominated by a Kelvin wave moving cyclonically in the bay. This is superposed with a wind generated flow.

4.2 Parameter sensitivity and robustness assessment

The performance of an assimilation scheme should be examined under ideal conditions as well as under conditions, where assumptions that are

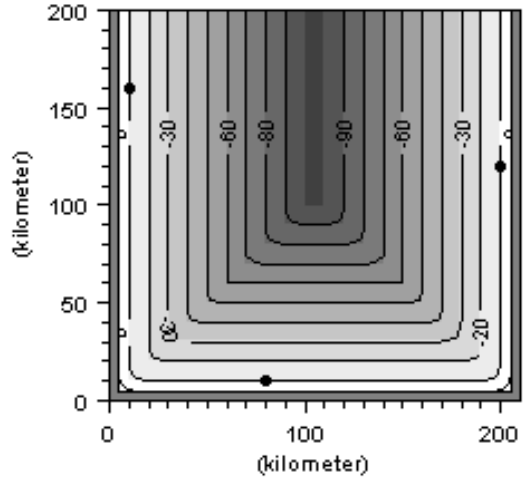

Figure 1: Bathymetry of Ideal Bay. The three dots indicate the positions, where water level time series were extracted to be used as measurements

typically violated break down. This section does both, but with the focus aimed at the nature of the solution in the latter case.

A large number of twin experiments has been performed. In each case the basic run from Section 4.1 is taken as the true state of the system. Water levels are extracted at three locations indicated in Figure 1 at every 15-minute time-step interval. Subsequently uncorrelated Gaussian white noise with a standard deviation of 5 cm is added to each time series in order to represent measurement noise. Only these three time series provide information about the true state in the assimilation procedures. In each perturbed run a different error source is introduced and the ability of the assimilation scheme to correct this error is examined.

As a measure of the filter performance, the spatially averaged root mean square error ($\overline{RMSE}$) of water levels, l, calculated over the last 24 hours

is used,

$$\overline{RMSE} = \frac{1}{J}\frac{1}{K}\sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(l_{true}(x_j, y_k, t_i) - l_{pert}(x_j, y_k, t_i))^2} \tag{24}$$

The constants $J, K = (20, 21)$ are the number of grid points in the x and y direction respectively. The constant $I = 96$ is the number of time steps in the period, in which the statistic is calculated. The indices *true* and *pert* refers to results from the true experiment and the perturbed results respectively. The error types in the perturbed runs are divided into two groups: Gaussian errors in Section 4.2.1 and typical errors in Section 4.2.2. Gaussian errors refer to error structures that basically fulfill the assumptions of the assimilation schemes if the parameters are chosen correctly. This is where the actual parameter sensitivity study is performed. The typical errors, on the other hand, refers to errors that include other distributions and sources than those assumed in the filters. This latter case is thought to closer resemble a real filter application.

4.2.1 Gaussian errors

The model error assumption lies in two forcing terms: The boundary water level and the wind velocity components. Thus, the investigation and the presentation of the results are divided according to this division. All the results presented in this section attempt to correct the same two perturbed runs, where a random coloured noise realisation has been added to the boundary forcing and wind field respectively. The parameters used to generate these Gaussian perturbations are stated in Tables 1 and 2.

Boundary spatial correlation scale	100 km
Boundary grid factor	1
Boundary temporal correlation scale	2 h
Boundary st.dev.	0.10 m
Autocorrelation coefficient, α	0.92
Boundary actual st.dev.	0.25 m

Table 1: Characteristics of the imposed errors in the open boundary perturbed runs

Wind spatial correlation scale	300 km
Wind grid factor	3
Wind temporal correlation scale	6 h
Wind st.dev.	3 m/s
Autocorrelation coefficient, α	0.97
Wind actual st.dev.	13 m/s

Table 2: Characteristics of the imposed errors in the wind velocity perturbed runs

Comparisons between the perturbed runs without data assimilation and the true runs give $\overline{RMSE}$ values of 0.33 m and 0.58 m for boundary and wind errors, respectively. The maps of the $RMSE$ values for each case is shown in Figure 2 and 3. Using the correct parameters in the assimilation schemes, but varying the rank of the covariance matrix, yields the $\overline{RMSE}$ results displayed in Figures 4 and 5. Since the the EnKF uses randomly generated noise, an average over five runs were used to smooth out the worst stochastic variations in the statistics. The assimilation schemes simultaneously provide the standard deviation they use for the update. The spatial averages of these are also included in the figures. A previous study by (Madsen & Cañizares 1999) has shown that the RRSQRT filter converges to a good performance at a lower rank than the EnKF, but similar execution times gave similar performance. This is also evident here for the wind error case, but in the boundary error, low rank EnKF outperforms the RRSQRT. Further investigation shows that this can be accredited to the normalisation procedure required in the eigenvalue decomposition of the RRSQRT scheme and hence a tuning of the normalisation makes the RRSQRT converge faster for the boundary error case as well. In general the figures show good performance with respect to the reduction of overall prediction error as well as the estimation of the prediction error.

Also included in the Figures 4 and 5 are the results of the Steady Kalman filter using a constant Kalman gain obtained as the average gain estimated by the EnKF over the last 24 hours. For a low rank of the error covariance the steady filter out-performs the EnKF. This is an example of bias-variance trade-off. The EnKF attempts to estimate the best unbiased state. However, this is done at the price of a high variance of the estimated parameters in the Kalman gain. The time averaging operation

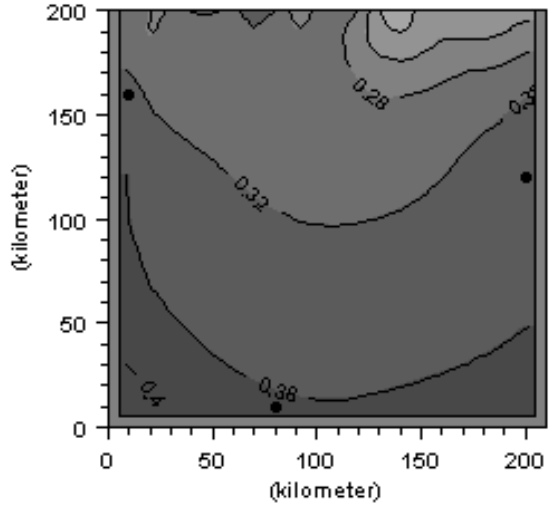

Figure 2: Spatial distribution of $RMSE$ values between the true run and the false boundary run, where a realisation of the Gaussian process described by Table 1 has been added. The positions of the measurements are indicated by dots.

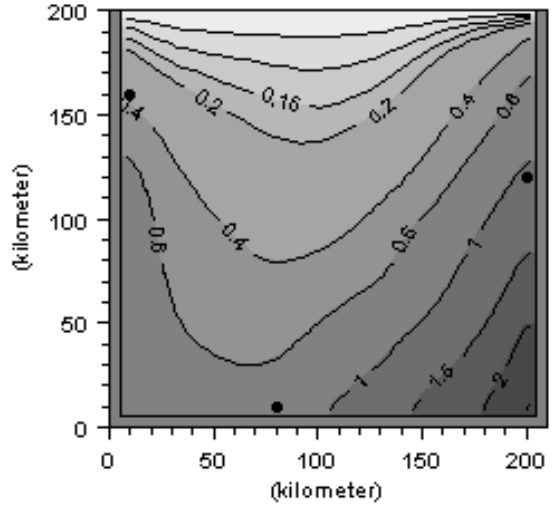

Figure 3: Spatial distribution of $RMSE$ values between the true run and the false wind run, where a realisation of the Gaussian process described by Table 2 has been added. The positions of the measurements are indicated by dots.

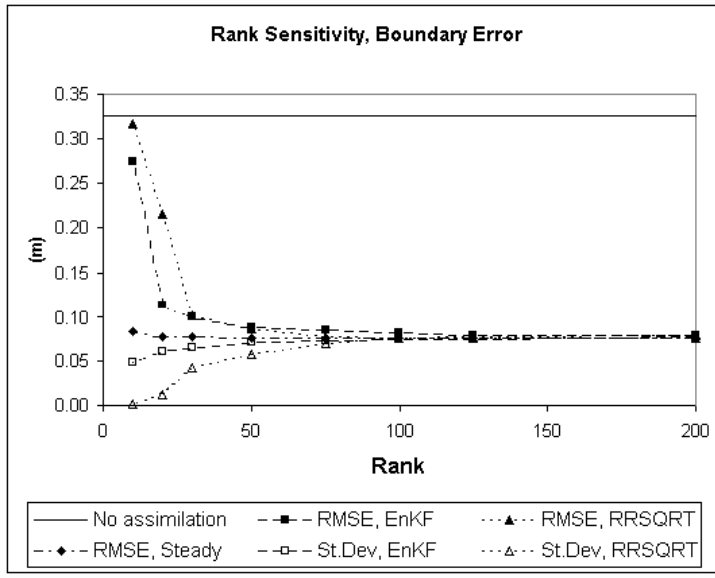

Figure 4: Sensitivity to rank of error covariance for perturbed boundary runs. All results are $\overline{RMSE}$ and given in meters (m)

employed for the generation of the time constant gain used in the Steady Kalman filter reduces this variance considerably, but deliberately introduces a bias in the estimation by assuming the gain to be time-invariant. For few ensembles, the latter trade-off between bias and variance gives the better performance. As the rank is increased the ensemble based estimate becomes more and more certain and the performance gets much better. For the boundary error case the estimated gain is actually rather time invariant and hence, even when the EnKF and RRSQRT filter are converged their performance is matched by that of the cheaper Steady filter. The converged gain of the wind error case is more time varying. Thus, in this case the Steady Kalman filter performs a little worse than the two time varying filters.

In the subsequent experiments, an ensemble size of 100 is used for the EnKF and a rank of 50 is used for the RRSQRT scheme. The Steady Kalman filter will be based on the EnKF with 100 ensemble members. For these choices the EnKF gives a $\overline{RMSE}$ of 0.08 m and 0.10 m and RRSQRT 0.09 m and 0.09 m for boundary and wind error, respectively.

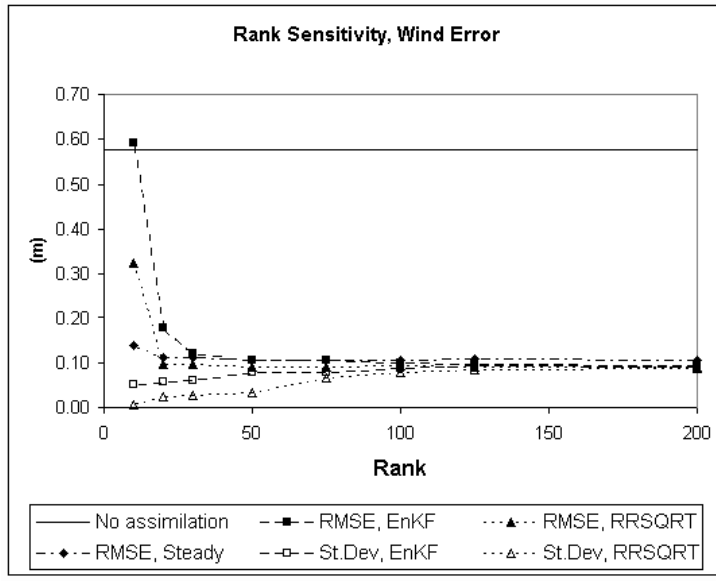

Figure 5: Sensitivity to rank of error covariance for perturbed wind velocity runs. All results are $\overline{RMSE}$ and given in meters (m)

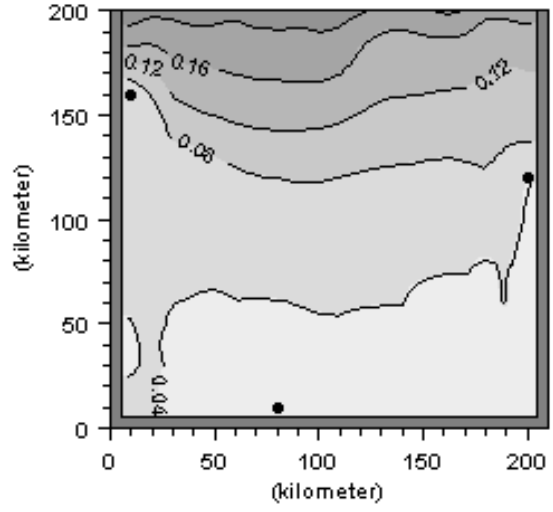

Figure 6: Spatial distribution of $RMSE$ values between the true run and the 100 EnKF run for the boundary error case. The positions of the measurements are indicated by dots.

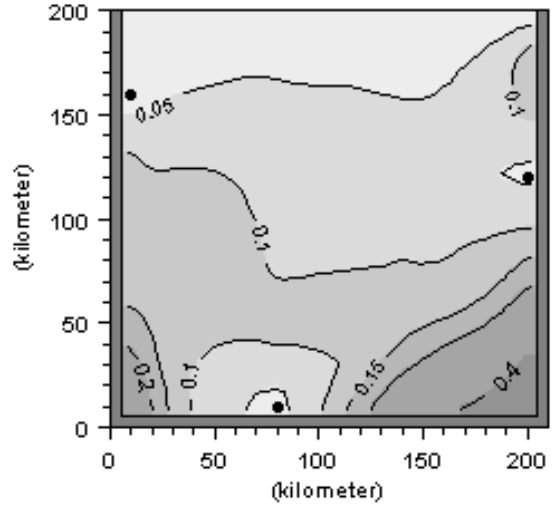

Figure 7: Spatial distribution of $RMSE$ values between the true run and the 100 EnKF run for the wind error case. The positions of the measurements are indicated by dots.

Similarly the Steady Kalman filter gives a $\overline{RMSE}$ of 0.08 m and 0.10 m. Maps of the $RMSE$ for the EnKF case are shown in Figures 6 and 7. Compared to Figures 2 and 3 these demonstrate the good performance of the filter. The results also show the importance of a good network design for assimilation purposes. For instance, in the boundary error case the Northwestern measurement corrects most of the noise in the Kelvin wave, leaving little error in its cyclonic propagation further South in the bay. Experiments show that leaving out the Southern measurement hardly alters the performance of the schemes.

Now the sensitivity to the assumed measurement standard deviation is examined. For each of the three measurements extracted from the basic solution a measurement standard deviation must be specified. The same value is used for all three stations. Table 3 summarises the $\overline{RMSE}$ results. The most notable result is the robustness of the filters for varying measurement standard deviation. The general picture is a degradation both when the solution is pulled too strongly toward the measurement, where the innovation has an excessive impact on unobserved regions, and when little trust is put in the measurements leading to only minor corrections of the perturbed solution. However, in both directions, extreme and unrealistic values must be assumed to significantly degrade the results.

σ_m	Bound			Wind		
	EnKF	RRSQRT	Steady	EnKF	RRSQRT	Steady
0.005	0.26	0.13	0.14	–	0.14	–
0.01	0.12	0.09	0.10	0.24	0.11	0.11
0.02	0.09	0.08	0.08	0.13	0.10	0.10
0.05	**0.08**	**0.09**	**0.08**	**0.09**	**0.09**	**0.10**
0.15	0.09	0.10	0.09	0.10	0.10	0.11
0.40	0.12	0.12	0.12	0.13	0.12	0.13
1.00	0.18	0.19	0.18	0.19	0.18	0.21

Table 3: Sensitivity to measurement standard deviation for perturbed boundary runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

The sensitivity to model standard deviation is summarised in Tables 4 and 5 as $\overline{RMSE}$ values. Here too, a quite robust performance is achieved. In general the behaviour degrades as the values become too large or too

small. An interesting point is that the RRSQRT scheme has its best performance for slightly overestimated model errors. This is explained by the standard deviations estimated by the RRSQRT scheme itself. As shown in Figures 4 and 5 the RRSQRT in general tends to underestimate the size of the model error and hence it provides a more correct standard deviation estimate with an excessive error assumption.

σ_η	EnKF	RRSQRT	Steady
0.001	0.31	0.31	0.31
0.005	0.20	0.19	0.19
0.01	0.13	0.13	0.13
0.05	0.09	0.09	0.08
0.10	**0.08**	**0.09**	**0.08**
0.25	0.10	0.08	0.08
1.50	0.13	0.09	0.10

Table 4: Sensitivity to model error standard deviation for perturbed boundary runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

σ_η	EnKF	RRSQRT	Steady
0.05	0.32	0.32	0.36
0.10	0.23	0.22	0.26
0.50	0.12	0.11	0.13
1.00	0.11	0.10	0.11
3.00	**0.09**	**0.09**	**0.10**
5.00	0.11	0.09	0.11
10.00	–	0.10	–

Table 5: Sensitivity to model error standard deviation for perturbed wind runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

Now, the sensitivity to the temporal correlation scale is examined. A temporal correlation scale of 0 hours gives a white noise assumption and a temporal correlation of ∞ hours gives a random walk error process. These extremes corresponds to values of $\sigma_{forcing}$ equal to σ_η and ∞ m respectively. Tables 6 and 7 summarise the results. Here too, quite robust performance can be observed. The results are most degraded when there is a small time correlation. The explanation for this is a combination of the smaller resulting standard deviation of the forcing terms and

a worse description of the spatial correlations due to the approximate white noise assumption. When coloured noise is adopted, the dynamical propagation transfers the coloured signal into spatial correlations.

τ	EnKF	RRSQRT	Steady
0.00 hours	0.15	0.15	0.14
0.25 hours	0.12	0.12	0.11
2.00 hours	**0.08**	**0.09**	**0.08**
6.00 hours	0.09	0.09	0.08
24.00 hours	0.10	0.09	0.09

Table 6: Sensitivity to model error temporal correlation scale, τ, for perturbed boundary runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

τ	EnKF	RRSQRT	Steady
0.00 hours	0.18	0.15	0.20
0.25 hours	0.16	0.13	0.16
1.00 hours	0.12	0.10	0.12
6.00 hours	**0.09**	**0.09**	**0.10**
24.00 hours	0.13	0.10	0.10

Table 7: Sensitivity to model error temporal correlation scale, τ, for perturbed wind runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

This dependence on a proper spatial correlation is also evident when considering the spatial correlation scale (Tables 8 and 9). In line with the discussion above about temporal correlations, the case of no spatial correlation has a poorer performance, which shows the importance of properly describing the correlations in the state vector. This can potentially be an important factor to control in real setups of the assimilation schemes. The results are in general insensitive to grid factor variations as long as the resulting coarse error grid resolves the assumed spatial error covariance.

The smoothing factor shows hardly any sensitivity at all. This can be accredited to the similar performance of the time varying and the steady schemes. In cases, with more time varying correlation structures, this factor must be expected to play a greater role, providing a smooth transition from the time varying to the steady performance. The smoothing

factor opens a possibility of obtaining stable time varying runs of e.g. a low rank EnKF. If the time variation of the error covariance fields are slow, but important to resolve, then this approach might provide an important operational option.

Altering the update interval degrades the result shown in Table 10. This proves a continuous transition to the runs with no assimilation, which corresponds to an update interval equal to ∞. Using less information from the true state gives a lower resemblance with the truth.

All together, in the present test case the assimilation schemes are robust to moderately misspecified parameters for Gaussian error sources that resembles the specified error models. This is encouraging, but does not guarantee good performance for any setup. In particular, care must be taken to ensure a proper model error covariance in sparsely observed systems.

l_c	EnKF	RRSQRT	Steady
0 km	0.11	0.15	0.09
25 km	0.09	0.08	0.08
100 km	**0.08**	**0.09**	**0.08**
250 km	0.08	0.09	0.08
1000 km	0.09	0.10	0.08

Table 8: Sensitivity to model error spatial correlation scale, l_c, for perturbed boundary runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

l_c	EnKF	RRSQRT	Steady
30 km	0.19	0.18	0.20
100 km	0.11	0.09	0.13
300 km	**0.09**	**0.09**	**0.10**
1000 km	0.08	0.10	0.10

Table 9: Sensitivity to model error spatial correlation scale, l_c, for perturbed wind runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

4.2.2 Typical errors

The demonstrated robustness in the Gaussian error case gives some hope that even for more typical error sources, not elaborately taken into account by the schemes, an improved performance can be obtained using good first guess estimates of parameters in the Gaussian framework assumed. This section investigates such behaviour.

Perturbed runs with ten different errors have been conducted. The results of the twin experiments both without and with the three data assimilation schemes are summarised in Table 11. The runs with 'Bound:' and 'Wind:' indicates that only open boundary or wind errors was added and subsequently assumed in the assimilation procedures. The 'Bound:' runs apply one and three hour phase errors as well as half a meter amplitude error by itself and in combination with the three hours phase error. The two systemamic wind error runs are forced by 20 m/s Westerly winds and a a stronger cyclone with a perturbed path referred to as 'False cyclone'. Further, a run was conducted with an erroneous bed friction using a Nikuradse roughness coefficient of 0.5 instead of the true 0.05. The 'Bathymetry' perturbed run refers to a run applying a modified bathymetry with a standard deviation of 1 meter compared to the truth. Also, a run, 'All errors with †'s', applying a composite of these errors is included in the study. Finally, a 'No forcing' run was conducted with no wind and open boundary forcing at all, thus giving a solution at rest. In all assimilation runs, the parameters of Tables 1 and 2 were used in the assimilation schemes.

d	Bound			Wind		
	EnKF	RRSQRT	Steady	EnKF	RRSQRT	Steady
1	**0.08**	**0.09**	**0.08**	**0.09**	**0.09**	**0.10**
2	0.09	0.10	0.09	0.10	0.09	0.11
4	0.11	0.12	0.11	0.11	0.10	0.12
8	0.16	0.16	0.15	0.17	0.15	0.16
16	0.22	0.21	0.21	0.27	0.25	0.29
32	0.26	0.25	0.26	0.42	0.33	0.38

Table 10: Sensitivity to update interval for perturbed runs. Reference run is marked in bold. All results are $\overline{RMSE}$ and given in meters (m)

A much improved performance is observed in all cases. In the worst case the residual $\overline{RMSE}$ is 0.51, which is too large for many applications, but a convincing result considering the size of the errors introduced into the model. Even in the 'No Forcing' run, the schemes are actually able to generate a large portion of the signal from the missing forcing terms including the forcing terms themselves through the augmented state vector description.

However, the good performace is not matched by a good uncertainty estimate. For instance, both time varying filters estimate a standard deviation of 0.07 for the 3h phase lag experiment. The standard deviation estimates for the 'False cyclone' run were 0.06 and 0.07 for the EnKF and RRSQRT schemes, respectively, while the numbers for the 'All errors with †'s' are 0.13 and 0.14. This clearly shows the violation of the underlying filter assumptions and the error estimates provided by the filters must be applied with care. At least, data must be retained in an attempt to perform a subsequent validation of the standard deviation estimates.

The biased nature of the errors makes the performance even more dependent on fairly high temporal and spatial correlations than was the case in the previous section. Hence, these are key parameters to consider in the calibration of any data assimilation setup. In general the Steady

Error Type	No ass.	EnKF	RRSQRT	Steady
Bound: 1h phase lag	0.69	0.07	0.07	0.07
Bound: 3h phase lag†	1.87	0.18	0.18	0.20
Bound: 1.5 × amplitude†	0.62	0.07	0.07	0.07
Bound: 3h phase lag + 1.5 × amplitude	2.23	0.23	0.22	0.26
Wind: 20 m/s West	0.17	0.07	0.09	0.09
Wind: False cyclone†	0.25	0.12	0.12	0.12
Bed friction 0.5†	0.26	0.06	0.06	0.13
Bathymetry†	0.05	0.04	0.04	0.03
All errors with †'s	2.53	0.51	0.44	0.38
No forcing	1.33	0.19	0.18	0.20

Table 11: Sensitivity to typical errors for perturbed runs. All results are $\overline{RMSE}$ and given in meters (m)

filter performs well with the 'All errors with †'s' case being the most impressive case. Basically, when all the error assumptions are violated, the elaborate schemes can not be expected to give superior performance. Rather, a certain regularisation of the Kalman gain acknowledges the bias in the estimate hence allowing a reduced variance and in this case, a better performance. However, time dependent schemes are clearly superior for bed friction error by itself. This study demonstrates the success of assimilation schemes despite the unavoidable wrong error assumptions imposed, and also shows how different filters handle different real error sources the best.

5 Summary and Conclusions

This paper presented three known assimilation schemes and described the filter parameters that can typically be varied in an application of the schemes. In a set of experiments in an idealised bay a sensitivity study has been conducted to investigate the filter performance for misspecified error structure in the schemes. The sensitivity to key parameters are vital for the practical use of sequential data assimilation techniques in hydrodynamic modelling. It is demonstrated that the filter performance is robust with respect to low to moderate parameter perturbations in the specification of the noise statistics. For more typical errors such as phase lags, bathymetry, etc., care must be taken to ensure a specification of fairly high temporal and spatial correlations. However, thought should be put into properly setting up every application. For some, an EnKF with a low smoothing factor is the best choice, while a sum of a range of significant error sources not assumed by the filter is best handled by a steady filter.

In general the steady filter seems like a good candidate for tide gauge assimilation in coastal areas. This is particularly true if an operational setting is considered. The spatial distribution of the filter performance has further demonstrated that proximity to the stations or dynamically well chosen positions enhances the estimation skill. Hence, the denser and better designed the measurement network is, the better the overall performance. Such a design increases the representative information available. Further, dense networks diminishes the importance of the

spatial distribution of the information and thus the correct parameter settings.

References

Burgers, G., van Leeuwen, P. J. & Evensen, G. (1998), 'Analysis scheme in the ensemble Kalman filter', *Monthly Weather Review* **126**, 1719–1724.

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

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Chui, C. K. & Chen, G. (1991), *Kalman filter with real-time applications*, Vol. 17 of *Springer Series in information sciences*, Springer-Verlag.

Cohn, S. E. & Todling, R. (1996), 'Approximate data assimilation schemes for stable and unstable dynamics', *Journal of Meteorological Society of Japan* **74**, 63–75.

Dee, D. P. (1991), 'Simplification of the Kalman filter for meteorological data assimilation', *Q.J.R. Meteorological Society* **117**, 365–384.

Dee, D. P. & da Silva, A. M. (1998), 'Data assimilation in the presence of forecast bias', *Q.J.R. Meteorological Society* **124**, 269–296.

DHI (2001), *MIKE 3 estuarine and coastal hydrodynamics and oceanography*, DHI Water & Environment.

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

Evensen, G. (2003), The ensemble Kalman filter: Theoretical formulation and practical implementation, In print: Ocean Dynamics.

Fukumori, I. & Malanotte-Rizzoli, P. (1995), 'An approximate Kalman filter for ocean data assimilation; an example with an idealised Gulf Stream model', *J. Geoph. Res.* **100**(C4), 6777–6793.

Ide, K., Courtier, P., Ghil, M. & Lorenc, A. C. (1997), 'Unified notation for data assimilation: Operational, sequential and variational', *Journal of Meteorological Society of Japan* **75**(1B), 181–189.

Jazwinski, A. H. (1970), *Stochastic Processes and filtering theory*, Vol. 64 of *Mathematics in Science and Engineering*, Academic Press.

Kalman, R. E. (1960), 'A new approach to linear filter and prediction theory', *Journal of Basic Engineering* **82**(D), 35–45.

Madsen, H. & Cañizares, R. (1999), 'Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling', *International Journal of Numerical Methods in Fluids* **31**(6), 961–981.

Pham, D. T., Verron, J. & Gourdeau, L. (1998), 'A singular evolutive Kalman filter for data assimilation in oceanography', *C. R. Acad. Sci. Paris* **326**, 255–260.

Pham, D. T., Verron, J. & Roubaud, M. C. (1997), 'Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography', *Journal of Marine Systems* **16**, 323–340.

Segers, A. J., Heemink, A. W., Verlaan, M. & van Loon, M. (2000), Kalman filtering for nonlinear atmospheric chemistry models: second (order) experiences, Technical report, Delft University of Technology.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2004a), Data assimilation in hydrodynamic modelling: On the treatment of nonlinearity and bias, Accepted: Stochastic Environmental Research and Risk Assessment.

Verlaan, M. (1998), Efficient Kalman filtering algorithms for hydrodynamic models, PhD thesis, Delft University of Technology.

Verlaan, M. & Heemink, A. W. (1997), 'Tidal flow forecasting using reduced rank square root filters', *Stochastic Hydrology and Hydraulics* **11**, 349–368.

B

Paper B

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

Accepted by *Stochastic Environmental Research and Risk Assessment*.

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

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

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.

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

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

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 & Malanotte-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 & 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, Bolding & Verlaan 1997). The reduced rank square root Kalman filter (Verlaan & 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 nonlinearities are present and only few measurements are available, (Verlaan & 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, Madsen & Madsen 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 & 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 Section 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 Section 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 Section 7. Finally, Section 8 concludes the paper.

2 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, Berg & Uhrenholdt 1998), (Erichsen & Rasch 2002).

MIKE 3 utilises a finite difference technique, and thus provides the dis-

crete 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 $\zeta(t_i)$, $\zeta(t_{i-1/2})$, $\zeta(t_{i-1})$, $v_x(t_i)$, $v_y(t_{i-1/2})$, $v_y(t_{i+1/2})$ and $v_z(t_{i-1/4})$ 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_i 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_{i+1} is completely determined by the state at time t_i and the forcing terms embedded in the sources and sinks and boundary conditions. Thus, let M_D be the model operator representing the approximate finite difference equations, $\mathbf{u}_D(t_i)$ the forcing defined at a snapshot in time projected onto the mesh and $\mathbf{x}_D(t_i)$ the model state at time t_i. The discrete deterministic model can then be expressed as,

$$\mathbf{x}_D(t_{i+1}) = M_D(\mathbf{x}_D(t_i), \mathbf{u}_D(t_i)) \tag{1}$$

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, $\boldsymbol{\eta}_{x,i}$. 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, G1, and thus an interpolation operator, $\mathbf{\Lambda}$, is introduced. In general any linear reduced rank representation can be expressed by $\mathbf{\Lambda}$, e.g. refer to (Cañizares 1999) 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, M_{M3}:

$$\mathbf{x}_{M3}(t_{i+1}) = M_{M3}(\mathbf{x}_{M3}(t_i), \mathbf{u}_D(t_i) + \mathbf{\Lambda}\boldsymbol{\eta}_{x,i}) \tag{2}$$

where

$$\mathbf{x}_{M3}(t_i) = \begin{pmatrix} \zeta(t_i) \\ \zeta(t_{i-1/2}) \\ \zeta(t_{i-1}) \\ v_x(t_i) \\ v_y(t_{i-1/2}) \\ v_y(t_{i+1/2}) \\ v_z(t_{i-1/4}) \end{pmatrix} \tag{3}$$

The only difference between $\mathbf{x}_{M3}$ and $\mathbf{x}_D$ is that the elements in $\mathbf{x}_{M3}$ 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)),

$$\boldsymbol{\eta}_{x,i} = \mathbf{A}\boldsymbol{\eta}_{x,i-1} + \boldsymbol{\eta}_{\eta,i} \tag{4}$$

where $\boldsymbol{\eta}_{\eta,i}$ is an n_η-dimensional i.i.d. variable with zero mean and known covariance, $\mathbf{Q}_\eta(t_i)$. $\mathbf{A} = diag(\boldsymbol{\alpha})$ is a linear diagonal model. 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:

$$\mathbf{x}(t_{i+1}) = \begin{pmatrix} \mathbf{x}_{M3}(t_{i+1}) \\ \boldsymbol{\eta}_{x,i+1} \end{pmatrix} = \begin{pmatrix} M_{M3}(\mathbf{x}_{M3}(t_i), \mathbf{u}_D(t_i) + \mathbf{\Lambda}\boldsymbol{\eta}_{x,i}) \\ \mathbf{A}\boldsymbol{\eta}_{x,i} + \boldsymbol{\eta}_{\eta,i} \end{pmatrix} \tag{5}$$

or

$$\mathbf{x}(t_{i+1}) = M(\mathbf{x}(t_i), \mathbf{u}_D(t_i)) + \boldsymbol{\eta}_i \tag{6}$$

where M is the augmented model operator and

$$\boldsymbol{\eta}_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boldsymbol{\eta}_{\eta,i} \end{pmatrix}^T \tag{7}$$

is a n-dimensional i.i.d. variable with zero mean and covariance $\mathbf{Q}(t_i)$. Eq. (6) is called the system equation and is the actual stochastic representation of MIKE 3 that will be used subsequently. The dimension of $\mathbf{x}$ 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. $\boldsymbol{\eta}_i$ enters linearly, but has a non-linear effect on $\mathbf{x}_{M3}$.

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 p.

For the purpose of data assimilation, we need to relate the measurements to the state vector $\mathbf{x}$. 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, Raghunath, Fu & Chao 1999). Let $\mathbf{x}$ be the stochastic model state defined by Eq. (6). It is assumed that the observation, $\mathbf{y}_i^o$, can be expressed as a linear combination represented by the $p \times n$ matrix $\mathbf{H}(t_i)$ of the state variables, and an additive zero mean Gaussian distributed observational error, $\boldsymbol{\varepsilon}_i$, with covariance $\mathbf{R}(t_i)$.

$$\mathbf{y}_i^o = \mathbf{H}_i \mathbf{x}(t_i) + \boldsymbol{\varepsilon}_i \tag{8}$$

This is called the measurement equation. The rows of $\mathbf{H}_i$ 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 er-

rors. Thus the measurement error covariance can be expressed as

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

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 & Malanotte-Rizzoli 1995) and (Sørensen, Madsen & Madsen 2003d) 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 equation (6), while the measurements are described by the measurement equation (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, $\mathbf{x}$, derived from the model, and an observation $\mathbf{y}^o$. 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 $\mathbf{x}$ given the data, $\mathbf{y}^o$,

$$f(\mathbf{x}|\mathbf{y}^o) = \frac{f(\mathbf{y}^o|\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{y}^o)} = \frac{f(\mathbf{y}^o|\mathbf{x})\,f(\mathbf{x})}{\int f(\mathbf{y}^o|\mathbf{x})\,f(\mathbf{x})d\mathbf{x}} \tag{10}$$

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

$$\log f(\mathbf{x}|\mathbf{y}^o) = \log f(\mathbf{y}^o|\mathbf{x}) + \log f(\mathbf{x}) - \log B \tag{11}$$

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 $f(\mathbf{x})$ and $f(\mathbf{y}^o|\mathbf{x})$ 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 $f(\mathbf{x})$ and $f(\mathbf{y}^o|\mathbf{x})$ are Gaussian with means $\mathbf{x}^f$ and $\mathbf{Hx}$ and known covariance matrices $\mathbf{P}^f$and $\mathbf{R}$ respectively. Thus,

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

$$f(\mathbf{y}^o|\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\mathbf{R})}} \exp\left(-1/2\left[\left(\mathbf{y}^o-\mathbf{Hx}\right)^T\left(\mathbf{R}\right)^{-1}\left(\mathbf{y}^o-\mathbf{Hx}\right)\right]\right) \tag{13}$$

where n and p are the sizes of $\mathbf{P}^f$and $\mathbf{R}$ respectively. By substituting Eqs. (12) and (13) into Eq. (11) and differentiating with respect to $\mathbf{x}$ 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,

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{Hx}^f) \quad , \quad \mathbf{P}^a = \mathbf{P}^f - \mathbf{KHP}^f \tag{14}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\left[\mathbf{HP}^f\mathbf{H}^T + \mathbf{R}\right]^{-1} \tag{15}$$

The matrix $\mathbf{P}^a$ is the error covariance of the estimated state, $\mathbf{x}^a$. 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. 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 & 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 $\mathbf{R}(t_i)$ becomes diagonal as ex-

pressed by Eq. (9). This allows for an efficient sequential updating of data from different tidal gauge stations within the same time step, (Madsen & Cañizares 1999). The values of the diagonal elements are set, based on reflections on the error sources discussed in Section 2. Estimates of $\mathbf{x}^f(t_i)$ are typically based on the composite hydrodynamic and the AR(1) model in Eq. (6). The error covariance matrix, $\mathbf{P}^f(t_i)$, 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, Raghunath, Wunsch & Haidvogel 1993) to geometric or physical assumptions (Fox, Haines, de Cuevas & Webb 2000), and transient propagation of $\mathbf{P}^f(t_i)$ by the hydrodynamic equations (Verlaan & 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, $\mathbf{Q}(t_i)$. 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 $n_\eta \times n_\eta$ portion of $\mathbf{Q}(t_i)$ is non-zero. The specification of $\mathbf{Q}(t_i)$ 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 ourselves 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 Sections 4.1 and 4.2 the pdf is approximated by a finite ensemble. However, for the Reduced Rank Square Root Kalman filter presented in Section 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, $\boldsymbol{\eta}_i$, to be a zero mean i.i.d. random variable, cf. Eq. (6), then the expectation of $\mathbf{x}(t_{i+1})$is:

$$E\left\{\mathbf{x}(t_{i+1})\right\} = E\left\{M(\mathbf{x}(t_i), \mathbf{u}_D(t_i))\right\} \tag{16}$$

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 & 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, j, of the ensemble of q 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.

$$\mathbf{x}^f_{j,i} = M(\mathbf{x}^a_{j,i-1}, \mathbf{u}_{D,i-1}) + \boldsymbol{\eta}_{j,i-1} \qquad , \qquad j = 1, 2, ..., q \tag{17}$$

where the model error $\boldsymbol{\eta}_{j,i-1}$ is randomly drawn from a Gaussian distribution with zero mean and nxn covariance matrix $\mathbf{Q}_i$ which represents the system noise. An estimate of the state vector (forecast) is calculated as the average of the ensemble members, i.e.

$$\mathbf{x}^f_i = \bar{\mathbf{x}}^f_i = \frac{1}{q}\sum_{j=1}^{q} \mathbf{x}^f_{j,i} \tag{18}$$

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

$$\mathbf{P}^f_i = \mathbf{S}^f_i (\mathbf{S}^f_i)^T \qquad , \qquad \mathbf{s}^f_{j,i} = \frac{1}{\sqrt{q-1}}(\mathbf{x}^f_{j,i} - \bar{\mathbf{x}}^f_i) \tag{19}$$

where $\mathbf{s}^f_{j,i}$ is the jth column in $\mathbf{S}^f_i$.

4.1.2 4.1.2 Update

An ensemble of size q of possible measurements is generated

$$\mathbf{y}^o_{j,i} = \mathbf{y}^o_i + \boldsymbol{\varepsilon}_{j,i} \qquad , \qquad j = 1, 2, ..., q \tag{20}$$

where $\mathbf{y}^o_i$ is the actual measurement vector, and $\boldsymbol{\varepsilon}_{j,i}$ is the measurement error that is randomly generated from a Gaussian distribution with zero mean and covariance matrix $\mathbf{R}_i$.

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 $\mathbf{H}_i$ only has a few non-zero elements and only the columns in $\mathbf{P}^f_i$ that correspond to these non-zero elements in $\mathbf{H}_i$ 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 & Chen 1991),

$$\mathbf{x}^a_{i,k} = \mathbf{x}^a_{i,k-1} + \mathbf{k}_{i,k}\left(\mathbf{y}^o_{i,k} - \mathbf{h}_{i,k}\mathbf{x}^a_{i,k-1}\right) \quad , k = 1, .., p \quad , \mathbf{x}^a_{i,0} = \mathbf{x}^f_i \tag{21}$$

where p is the number of measurements, $\mathbf{h}_{i,k}$ is the k'th row in the measurement matrix $\mathbf{H}_i$, $\mathbf{h}_{i,k}\mathbf{x}^a_{i,k-1}$ is the element in the state vector that corresponds to the measurement $\mathbf{y}^o_{i,k}$, (i.e. $(\mathbf{y}^o_{i,k} - \mathbf{h}_{i,k}\mathbf{x}^a_{i,k-1})$ is the model deviation from measurement k), and $\mathbf{k}_{i,k}$ is a Kalman gain vector corresponding to measurement k. The Kalman gain vector is given by

$$\mathbf{k}_{i,k} = \frac{\mathbf{S}^a_{i,k-1}\mathbf{c}_{i,k}}{\mathbf{c}^T_{i,k}\mathbf{c}_{i,k} + \sigma^2_j} \quad , \mathbf{c}_{i,k} = (\mathbf{S}^a_{i,k-1})^T\mathbf{h}^T_{i,k} \quad , \mathbf{S}^a_{i,0} = \mathbf{S}^f_i \tag{22}$$

where the numerator is the covariance between the measurement k and the state vector and the denominator is the sum of the variance of measurement k 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 $\mathbf{S}^a_{i,k}$ is calculated from the ensemble cf. Eq. (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 $\mathbf{x}_i^f$ has also been implemented for the purpose of calculation of the non-linearity measures discussed in Section 5. This filter is referred to as the Central Ensemble Kalman Filter (CEnKF). A new central state vector, $\mathbf{x}_i^c$, is introduced. At initial time t_0 it is set equal to the mean estimate of $\mathbf{x}_i^a$ and subsequently it is propagated and updated like any other of the ensemble members, i.e:

$$\mathbf{x}_0^{c,a} = \overline{\mathbf{x}_0^a} \tag{23}$$

$$\mathbf{x}_i^{c,f} = M(\mathbf{x}_{i-1}^{c,a}, \mathbf{u}_{D,i-1}) \tag{24}$$

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

$$\mathbf{x}_i^f = M(\mathbf{x}_{i-1}^a, \mathbf{u}_{D,i-1}) \tag{25}$$

$$\mathbf{P}_i^f = \mathbf{F}_i \mathbf{P}_{i-1}^a \mathbf{F}_i^T + \mathbf{Q}_i \tag{26}$$

$$\mathbf{F}_i = \left. \frac{\partial M}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_i^f} \tag{27}$$

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 $\mathbf{S}_{i-1}^a$ the approximation of rank q of the square root of the error covariance matrix $\mathbf{P}_{i-1}^a$. The propagation of the error covariance matrix is then given by

$$\mathbf{S}_i^f = \left[\mathbf{F}_i \mathbf{S}_{i-1}^a \middle| \ \mathbf{Q}_i^{1/2} \right] \tag{28}$$

where $\mathbf{Q}_i^{1/2}$ is the $n \times p$-dimensional square root of $\mathbf{Q}_i$. The matrix $\mathbf{S}_{i-1}^a$ has q columns where q is chosen much smaller than the dimension of the state vector. To calculate the derivatives in $\mathbf{F}_i$ a finite difference approximation is adopted as follows,

$$(\mathbf{F}_i \mathbf{S}_{i-1}^a)_j = \left[M(\mathbf{x}_{i-1}^a + \mathbf{s}_{j,i-1}^a, \mathbf{u}_{D,i}) - M(\mathbf{x}_{i-1}^a, \mathbf{u}_{D,i}) \right] \ j = 1, .., q \tag{29}$$

where $\mathbf{s}_{j,i-1}^a$ is the jth column of $\mathbf{S}_{i-1}^a$. Thus, the propagation of the error covariance matrix requires q model integrations.

The propagation step in Eq. (28) increases the number of columns in the error covariance matrix from q to $q+p$. 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 $\mathbf{S}_i^f$ is applied by keeping only the q leading eigenvectors of the error covariance matrix. The reduction is achieved by an eigenvalue decomposition of the matrix $(\mathbf{S}_i^f)^T \mathbf{S}_i^f$. For full details refer to (Cañizares 1999). For a proper reduction, $\mathbf{S}_i^f$ 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 $(\mathbf{S}_i^f)^T\mathbf{S}_i^f$ 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 q, $\mathbf{S}_i^f$, the error covariance matrix can be calculated as $\mathbf{P}_i^f = \mathbf{S}_i^f(\mathbf{S}_i^f)^T$, 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 $\mathbf{S}_i^f$ directly. In this case the state vector is updated using Eq. (21), and the updated square root covariance matrix is given by (Cañizares 1999),

$$\mathbf{S}_{i,k}^a = \mathbf{S}_{i,k-1}^a - \frac{\mathbf{k}_{i,k}\mathbf{c}_{i,k}^T}{1+\sqrt{\frac{\sigma_j^2}{\mathbf{c}_{i,k}^T\mathbf{c}_{i,k}+\sigma_j^2}}} \quad , \; \mathbf{S}_{i,0}^a = \mathbf{S}_i^f \tag{30}$$

where $\mathbf{k}_{i,k}$ and $\mathbf{c}_{i,k}$ 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 $1/\sqrt{q}$). 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 & 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 $\mathbf{Q}_i$ 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 attempt to minimise the variance assuming no bias, (Dee & da Silva 1998). However, a bias, $\mathbf{b}$, 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 $\mathbf{P}^f + \mathbf{b}\mathbf{b}^T$ instead of $\mathbf{P}^f$ 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 properly selected $\mathbf{A}$ 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 $\mathbf{R}_i$ and $\mathbf{Q}_i$. If $\mathbf{Q}_i$ 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 $\mathbf{Q}_i$

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 $\mathbf{Q}_i$ 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 & 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 nonlinearity 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, although 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 & 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 sources and the initial field. All three measures are time varying spatial L_2-norms. Based on the derivation in (Verlaan & Heemink 2001), the V_2 measure can be written as:

$$V_2(t_i) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{b_j(t_i)}{\gamma_j(t_i)}\right)^2} \tag{31}$$

$$b_j(t_i) = x_j^c(t_i) - x_j(t_i) \tag{32}$$

Here, n is the number of elements in the state vector and $\gamma_j(t_i)$ is standard deviation of the state estimate derived as the square root of the diagonal elements of $\mathbf{P}^a(t_i)$. The bias, $b_j(t_i)$ 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 Vmeasure suggested in (Verlaan & Heemink 2001),

$$V(t_i) = \sqrt{\mathbf{b}^T\mathbf{P}^{-1}\mathbf{b}}, \quad \mathbf{b} = [b_1, \ldots, b_n]^T \tag{33}$$

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 q still being the ensemble size, the s_2-measure is simply the spatial L_2-norm of the skewness, $s_j(t_i)$:

$$s_j(t_i) = \frac{q \sum_{k=1}^{q} \left(x_k(t_i) - \bar{x}(t_i)\right)^3}{(q-1)\,(q-2)\,\gamma_j^3(t_i)} \tag{34}$$

$$s_2(t_i) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(s_j(t_i)\right)^2} \tag{35}$$

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, $k_j(t_i)$:

$$k_j(t_i) = \frac{q\ (q+1) \sum_{k=1}^{q} \left(x_k(t_i) - \bar{x}(t_i)\right)^4}{(q-1)\,(q-2)\,(q-3)\,\gamma_j^4(t_i)} - \frac{3\ (q-1)^2}{(q-2)\,(q-3)} \tag{36}$$

$$k_2(t_i) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(k_j(t_i)\right)^2} \tag{37}$$

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 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 2^{nd} 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, k, the bias measure, β_k, is defined as,

$$\beta_k = \frac{1}{T}\sum_{i=1}^{T} \frac{y^o_{i,k} - \mathbf{h}_{i,k}\mathbf{x}^a}{\sqrt{\mathbf{h}_{i,k}\mathbf{P}^a_i\mathbf{h}^T_{i,k} + \sigma^2_k}} \tag{38}$$

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 & Heemink 2001), (Madsen & 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,

$$RMSE = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{T}\sum_{i=1}^{T}(\zeta_i^{true}(j) - \zeta_i^{pert}(j))^2} \tag{39}$$

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

$$Bias = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{T}\sum_{i=1}^{T}\left(\zeta_i^{true}(j) - \zeta_i^{pert}(j)\right)\right]^2} \tag{40}$$

$$St.dev. = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{T}\sum_{i=1}^{T}((\zeta_i^{true}(j) - \bar{\zeta}_i^{true}(j)) - (\zeta_i^{pert}(j) - \bar{\zeta}_i^{pert}(j)))^2} \tag{41}$$

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 kilometres and the time step is 15 minutes. The northern open boundary has surface elevation 1.0 meter and the eastern has surface elevation 0.0 meters. The bathymetry, which is shown in Figure 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 nonlinearity 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 minutes 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 nonlinearity.

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 Figure

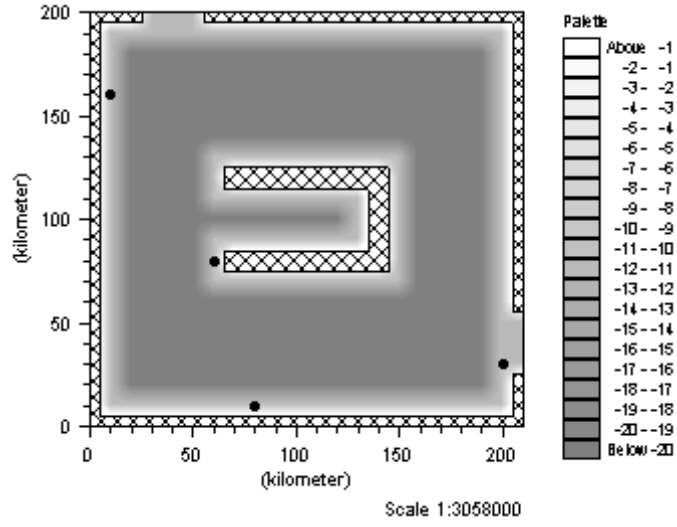

Figure 1: Test case bathymetry [m]. The black dots indicate measurement positions (10 km, 160 km), (60 km, 80 km), (80 km, 10 km) and (200 km, 30 km).

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 12^{th} time step or equivalently every 3 hours. 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 km, 100 km, 495 km 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, Figure 2 shows a time series of water level at the measurement point (60 km, 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 Figures 3 and 4, 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 nonlinearity 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

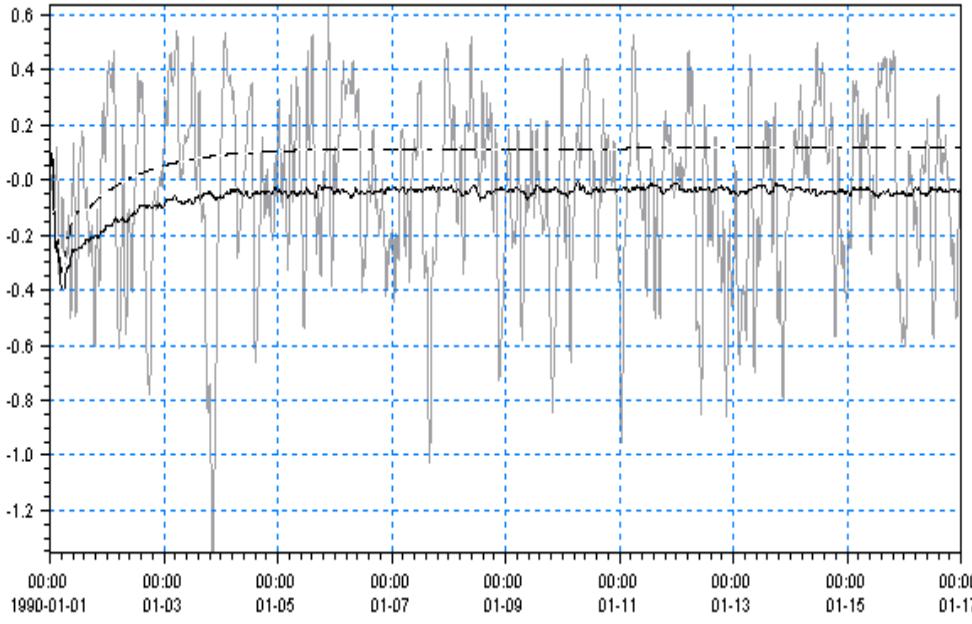

Figure 2: Water levels extracted at (60 km, 80 km). Grey: True run. Black dot-dashed: Central forecast false run. Black: Ensemble forecast false run.

Figures 5 and 6. Note the reduction in bias, while the standard deviation remains literally unaltered.

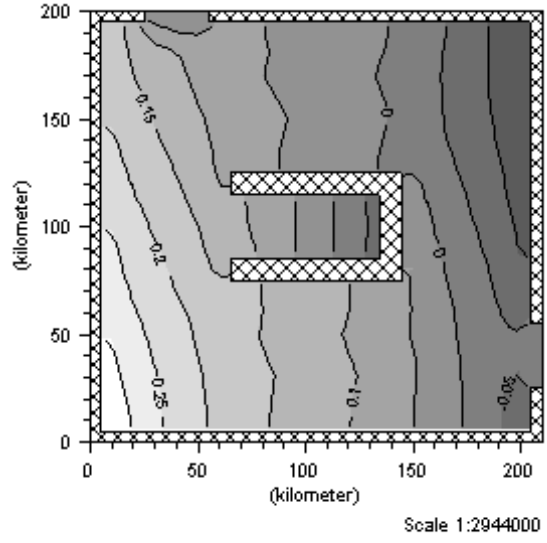

Figure 3: Central forecast NL false run water level bias [m].

7.1.2 General filter performance

The assimilation schemes all improve the rather poor false solution significantly. Figures 7 to 10 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 Figures 3 to 6. 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 11 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 Figure 10. 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

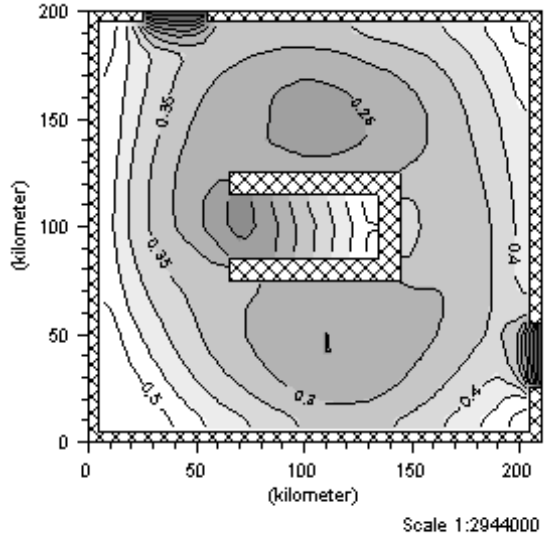

Figure 4: Central forecast NL false run water level standard deviation [m].

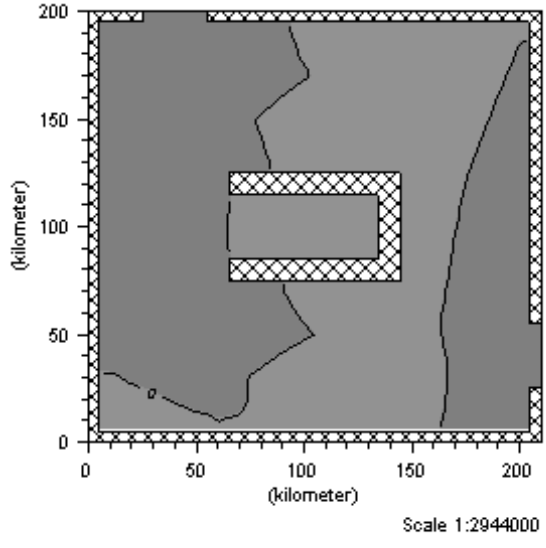

Figure 5: 1000 ensemble forecast NL false run water level bias [m].

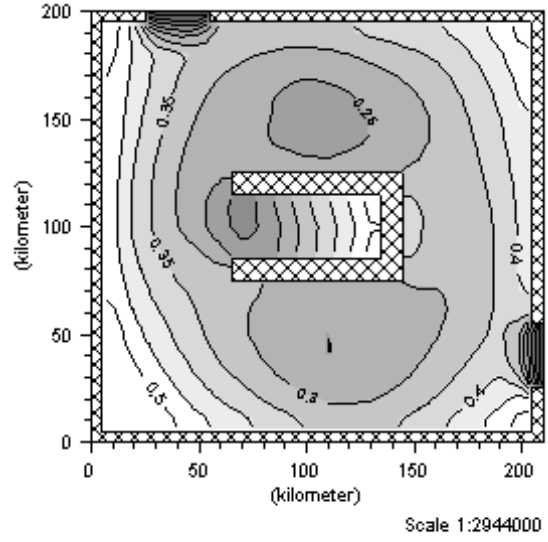

Figure 6: 1000 ensemble forecast NL false run water level standard deviation [m].

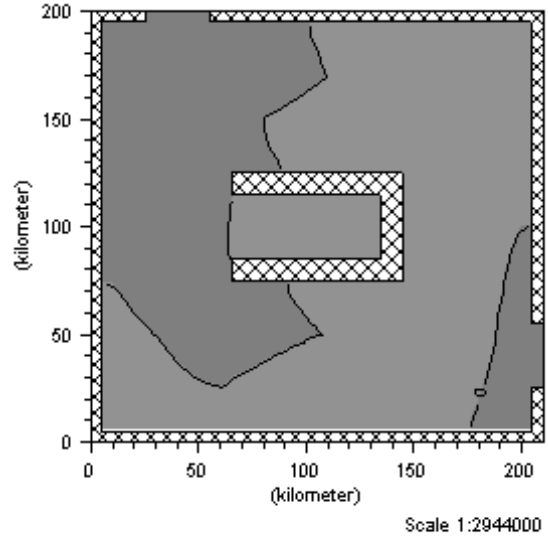

Figure 7: Forecast NL false run water level bias [m] using the RRSQRT.

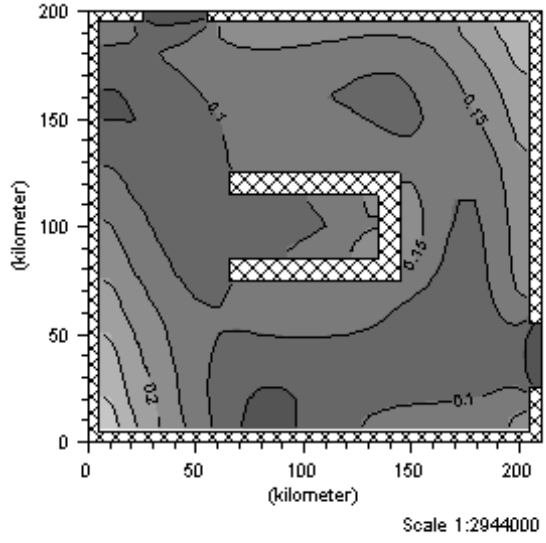

Figure 8: Forecast NL false run water level standard deviation [m] using RRSQRT.

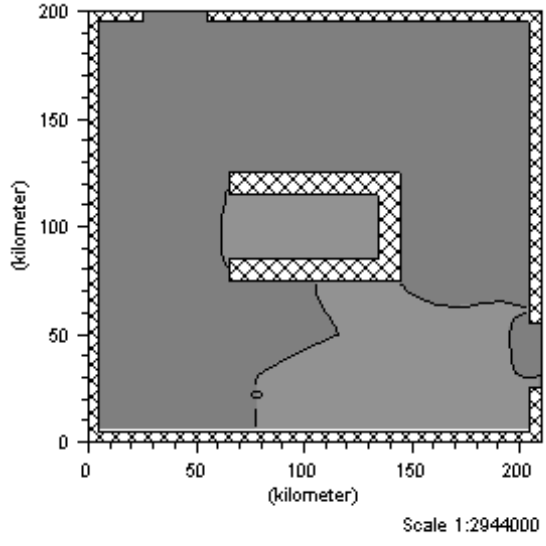

Figure 9: Forecast NL false run water level bias [m] using the 1000 EnKF.

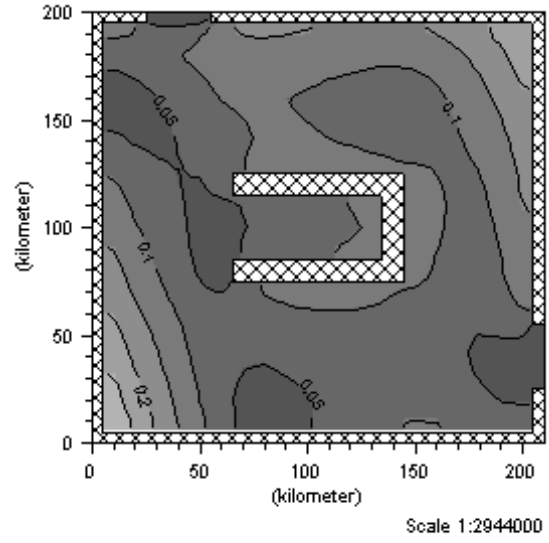

Figure 10: Forecast NL false run water level standard deviation [m] using 1000 EnKF.

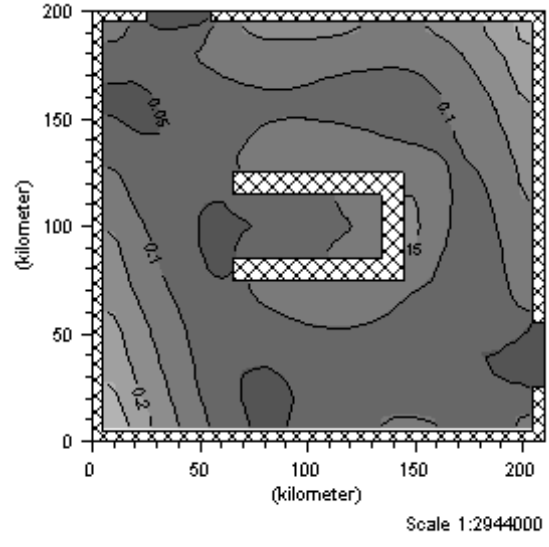

Figure 11: Estimated water level standard deviation [m] by the 1000 EnKF.

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 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∞.

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 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 & 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 CEnKF versus the EnKF based forecast with no assimilation. Figure 3 and 5 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 nonlinearity measure, V_2, defined in Section 5. As can be seen in

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 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∞.

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

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

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

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 interval for the chosen intervals and even in the most linear case (ui=1) the solution seems to be non-Gaussian. There-

fore, 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 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 Figure 12. 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 meters. 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 Figure 12 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.

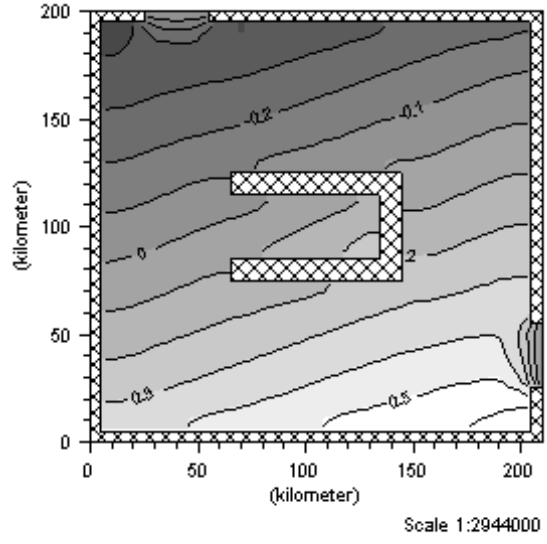

Figure 12: Forecast ES false run water level bias [m].

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 km, 80 km), (160 km, 10 km), (130 km, 90 km) and (190 km, 190 km). Based on these the L_2-norm of β is 2.1. In all cases the measure shows that the model-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.

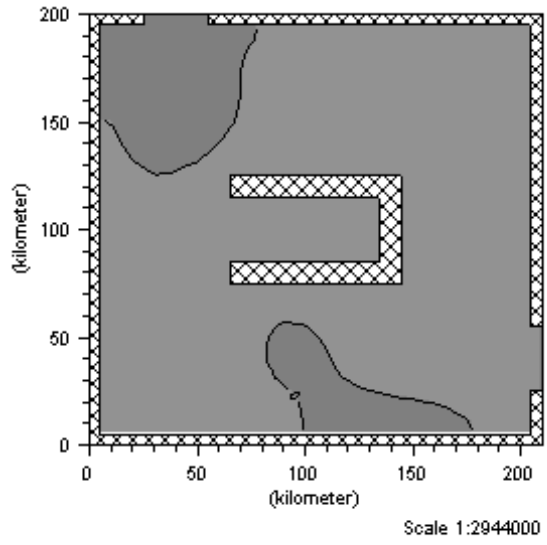

Figure 13: Forecast ES false run water level bias [m] using 100 EnkF with a time constant of 10^6 and a spatial correlation scale of 10,000 km.

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 Figure 13 showing the bias from the 100 EnKF scheme correctly assuming a biased error with a very high spatial correlation of 10.000 kilometres. Alternatively, if the error is assumed to be white, a bias will always remain as shown in Figure 14, still assuming a spatial correlation of 10.000 kilometres. The results are summarised in Tables 4 and 5, 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 seconds, 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 metres. 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.

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

Table 4: The L_2-norm of the bias. 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.

β	**0 km**	**100 km**	**495 km**	**10,000 km**
0 s.	4.13	1.63	1.12	0.84
10^6 s.	3.03	0.64	0.21	0.06

Table 5: 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.

Applying a wrong spatial correlation scale can potentially increase the bias in data sparse areas as demonstrated in Figure 15, which shows the bias for a spatial correlation scale of 0 kilometres and a time constant of 10^6 seconds. Compared to Figure 12 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.

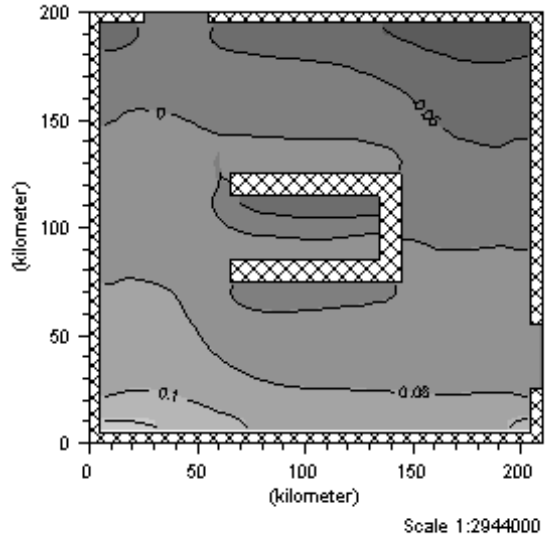

Figure 14: Forecast ES false run water level bias [m] using 100 EnkF with a time constant of zero and a spatial correlation scale of 10,000 km.

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 nonlinearity 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 nonlinearity 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 nonlinearity and nonGaussianity has been verified. The MAP interpretation of the estimated state must be discredited in the case of strong nonlinearities or lack of Gaussian noise input. Finally,

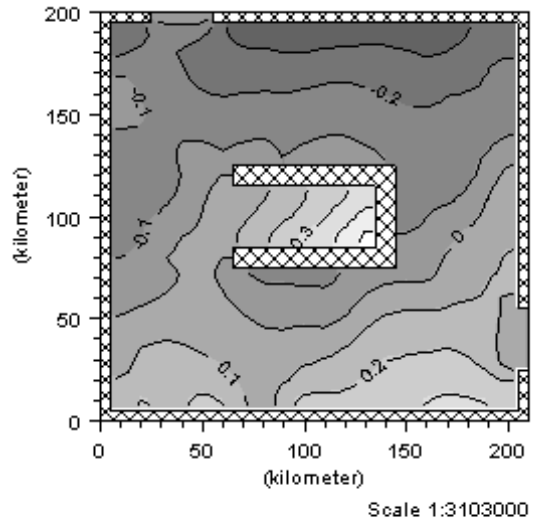

Figure 15: Forecast ES false run water level bias [m] using 100 EnkF with a time constant of 10^6 and a spatial correlation scale of 0 km.

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, P. J. & Evensen, G. (1998), 'Analysis scheme in the ensemble Kalman filter', *Monthly Weather Review* **126**, 1719–1724.

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

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Christakos, G. (2002), 'On the assimilation of uncertain physical knowledge bases: Bayesian and non-bayesian techniques', *Advances in Water Resources* **25**, 1257–1274.

Chui, C. K. & Chen, G. (1991), *Kalman filter with real-time applications*, Vol. 17 of *Springer Series in information sciences*, Springer-Verlag.

Cohn, S. E. & Todling, R. (1996), 'Approximate data assimilation schemes for stable and unstable dynamics', *Journal of Meteorological Society of Japan* **74**, 63–75.

Dee, D. P. (1991), 'Simplification of the Kalman filter for meteorological data assimilation', *Q.J.R. Meteorological Society* **117**, 365–384.

Dee, D. P. (1995), 'On-line estimation of error covariance parameters for atmospheric data assimilation', *Monthly Weather Review* **123**, 1128–1145.

Dee, D. P. & da Silva, A. M. (1998), 'Data assimilation in the presence of forecast bias', *Q.J.R. Meteorological Society* **124**, 269–296.

DHI (2001), *MIKE 3 estuarine and coastal hydrodynamics and oceanography*, DHI Water & Environment.

Erichsen, A. C. & Rasch, P. S. (2002), Two- and three-dimensional model system predicting the water quality of tomorrow, *in* M. L. Spaulding, ed., 'Proceedings of the seventh international conference on estuarine and coastal modeling', American Society of Civil Engineers, American Society of Civil Engineers, pp. 165–184.

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

Fox, A. D., Haines, K., de Cuevas, B. A. & Webb, D. J. (2000), 'Altimeter assimilation in the OCCAM global model. Part i: A twin experiment', *Journal of Marina Systems* **26**, 303–322.

Fukumori, I. & Malanotte-Rizzoli, P. (1995), 'An approximate Kalman filter for ocean data assimilation; an example with an idealised Gulf Stream model', *J. Geoph. Res.* **100**(C4), 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. Geoph. Res.* **104**(C11), 25647–25665.

Fukumori, I., Raghunath, R., Wunsch, C. & Haidvogel, D. B. (1993), 'Assimilation of sea surface topography into an ocean circulation model using a steady-state smoother', *Journal of Physical Oceanography* **23**, 1831–1855.

Heemink, A. W. (1986), Storm surge prediction using Kalman filtering, PhD thesis, Twente University of Technology.

Heemink, A. W., Bolding, K. & Verlaan, M. (1997), 'Storm surge forecasting using Kalman filtering', *Journal of Meteorological Society of Japan* **75**(1B), 305–318.

Ignagni, M. B. (1990), 'Separate bias Kalman estimator with bias state noise', *IEEE Transactions on Automatic Control* **35**, 338–341.

Jazwinski, A. H. (1970), *Stochastic Processes and filtering theory*, Vol. 64 of *Mathematics in Science and Engineering*, Academic Press.

Madsen, H. & Cañizares, R. (1999), 'Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling', *International Journal of Numerical Methods in Fluids* **31**(6), 961–981.

Øresundskonsortiet (1998), The øresund link. assessment of the impacts on the marine environment of the øresund link, update, Technical report, Øresundskonsortiet.

Segers, A. J., Heemink, A. W., Verlaan, M. & van Loon, M. (2000), Kalman filtering for nonlinear atmospheric chemistry models: second (order) experiences, Technical report, Delft University of Technology.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2002), Towards an operational data assimilation system for a three-dimensional hydrodynamic model, *in* R. A. Falconer, B. Lin, E. L. Harris & C. A. M. E. Wilson, eds, 'Proceedings of the fifth International Conference on Hydroinformatics', International Water Association, IWA Publishing.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2003d), Water level forecast skill of a hybrid steady Kalman filter - error correction scheme, Submitted: Ocean Dynamics.

Verlaan, M. & Heemink, A. W. (1997), 'Tidal flow forecasting using reduced rank square root filters', *Stochastic Hydrology and Hydraulics* **11**, 349–368.

Verlaan, M. & Heemink, A. W. (2001), 'Nonlinearity in data assimilation applications: A practical method for analysis', *Monthly Weather Review* **129**, 1578–1589.

Vested, H. J., Berg, P. & Uhrenholdt, T. (1998), 'Dense water formation in the Northern Adriatic', *Journal of Marine Science* **18**, 135–160.

Wunsch, C. (1996), *The ocean circulation inverse problem*, Cambridge University Press.

Paper C

C

Towards an operational data assimilation system for a three-dimensional hydrodynamic model

Originally published as

Jacob V. Tornfeldt Sørensen, Henrik Madsen, and Henrik Madsen. Towards an operational data assimilation system for a three-dimensional hydrodynamic model. In I.D. Cluckie, D. Han, J.P. Davis and S. Heslop, *Proceedings of the Fifth International Conference on Hydroinformatics*, pages 1204–1209, Cardiff, Wales, July 2002.

Towards an operational data assimilation system for a three-dimensional hydrodynamic model

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

Abstract

A data assimilation system is developed for both a three-dimensional and a two-dimensional hydrodynamic model for bays, estuaries, coastal areas and shelf seas. A sequential time varying assimilation method based on the Kalman filter to assimilate water level measurements is implemented. The method utilises an ensemble Kalman filter based on a Monte Carlo approach for propagation of model errors (EnKF). Further approximations, which will speed up the calculations and thus enable operational use of the system, are considered. A simple approach, using time averaged Kalman gains from the time varying filter, is implemented (Steady Filter). The state considered in the approaches consists of the two-dimensional fields of water levels and horizontal depth averaged velocities in each grid point. The three dimensional horizontal velocity field is updated by using a constant predefined profile constrained by maintenance of dynamical balance. The idea of this is to reduce the computational costs by using a simplified one-layer dynamics in the two-dimensional model to propagate the error-covariance matrix.

1 Introduction

Reliable data assimilation methodologies, applicable for low budget computer facilities, are consultancy and management needs of tomorrow. During the past decade data assimilation has emerged in the 3D hydrodynamic modelling community and has been developed into applicable tools for research in coastal and shelf seas. However, only for models that have very low computational costs, is data assimilation usable in practical operational modelling systems without requiring high performance computing (HPC) facilities.

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

For 2D models on the other hand, Kalman filter based data assimilation has been demonstrated to be applicable even for more advanced set-ups (Cañizares et al. 2001). Here, an operational approach based on physically justified error assumptions was devised. A similar approach is also applicable in 3D models, (Sørensen, Madsen & Madsen 2001), but its use is time consuming and has large memory requirements. The price paid in the 2D models is the lack of vertical structure. This paper presents an operational approach to Kalman filter based data assimilation in a 3D model run, by using physical information from the 2D model in the assimilation of data in the 3D model. This makes the set-up of the system endurable and maintains the vertical structure.

2 The hydrodynamic models

The two hydrodynamic models used in the present study, MIKE 21 and MIKE 3 developed at DHI Water and Environment, are applicable in coastal and continental shelf seas. In this study both models are used assuming constant density.

MIKE 21 is a 2D depth averaged model that solves for hydrodynamic flow on a rotating sphere. MIKE 3 solves the same problem, but includes full vertical structure in the calculations. Both models have the same time stepping scheme and horizontal grid including a dynamic two-way nesting facility and can thus be executed on the same bathymetry and forcing inputs. However, the models have lived each their lives and separate codes exist. This means that small differences exist, e.g. in the details of the bottom drag and turbulence closure formulation. Further, the algorithms are optimised after different criteria, which together with a change in computer precision makes MIKE 21 a factor 4 times faster than MIKE 3 in a one-layer setting for the particular set-up of this paper. Similarly, MIKE 21 is 8 times faster than MIKE 3 in a 10-layer set-up.

Both MIKE 3 and MIKE 21 contain surface elevation, ζ, and depth averaged velocity, (U_x,U_y) in their state representation. In MIKE 3 the full 3D velocity field together with surface elevation constitute the state vector needed for the model to be uniquely defined. However, if a correction of the depth averaged velocity is needed, then this correction

can be distributed in the vertical by assuming a uniform distribution. This maintains the vertical structure but shifts the mean. The vertical velocity is calculated from the hydrostatic balance equation. Following this approach a sub space of the state vector in MIKE 3 that contains information about the variability is (ζ, U_x,U_y). If an average estimate of the vertical structure is available, a correction in this sub space can be redistributed to the full MIKE 3 state vector by means of the rule above.

The MIKE models can be expressed as,

$$\mathbf{x}_M(t_{i+1}) = M(\mathbf{x}_M(t_i), \mathbf{u}(t_i))$$

where t_i is time indexed by i, $\mathbf{x}_M$ is the model state vector and $\mathbf{u}$ is the external forcing in terms of open boundary conditions and meteorological forcing.

3 The state estimator

Model and measurements give two independent estimates of the oceanic state. Both are uncertain and are thus best described as stochastic variables. Assuming the distribution of both the model and the observation are known, a best linear unbiased estimate (BLUE) of the model state can be achieved. This is the estimator used in the Kalman filter and is carried over to state estimation in non-linear models as well. Let the observational vector, $\mathbf{y}_i^o$, at time t_i have a linear relation to the state vector and have additive noise,

$$\mathbf{y}_i^o = \mathbf{H}_i \mathbf{x}_M(t_i) + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{\varepsilon}_i$ is an i.i.d. random variable with zero mean and covariance matrix, $\mathbf{R}_i$. Further, assume the error covariance matrix, $\mathbf{P}_i$, of $\mathbf{x}_M(t_i)$ to be known. Then at a given time step, the BLUE estimate of the system state, $\mathbf{x}^a(t_i)$, is,

$$\mathbf{x}^a(t_i) = \mathbf{x}_M(t_i) + \mathbf{K}_i(\mathbf{y}_i^o - \mathbf{H}_i \mathbf{x}_M(t_i))$$

$$\mathbf{K}_i = \mathbf{P}_i \mathbf{H}_i^T \left[\mathbf{H}_i \mathbf{P}_i \mathbf{H}_i^T + \mathbf{R}_i\right]^{-1}$$

4 The Ensemble and Steady Kalman filter

The main bottleneck in large scale Kalman filter based data assimilation techniques is the estimation of the model error covariance, $\mathbf{P}_i$, which needs to be propagated in time. Its dimension is $n \times n$, where n is the size of the state vector, which typically is of the order 10^5-10^7. The main feature of the Ensemble Kalman filter (Evensen 1994) is to estimate $\mathbf{P}$ from an ensemble of q model states, thus reducing the time requirements by a factor $2n/q$, while simultaneously capturing the full non-linear propagation of the system state probability density function. In practical applications q of the order 10^2 is sufficient to provide good results.

When propagating the ensemble, a model noise assumption needs to be made. In the present set-up open boundary level forcing and wind forcing are assumed to be the dominant sources of uncertainty and thus other sources are ignored or rather assumed embedded in the modelled errors. An outline of the implemented ensemble Kalman filter can be found in (Madsen & Cañizares 1999) and (Sørensen et al. 2001).

If the Kalman gains, $\mathbf{K}_i$, based on the time varying Ensemble Kalman filter approach shows fairly little time variation, it is tempting to time average the gains over a sufficiently long period and then use the constant Kalman gains in the future, which completely omits the error covariance propagation (Cañizares et al. 2001). This is how the Steady Kalman Filter works. However, the constant gains need to be generated by a time varying filter, thus still requiring these runs to be feasible for the system to be set up.

5 Dynamical approximations

If it is assumed that the main errors in velocity and surface elevations are barotropic, then MIKE 21 might be as good as MIKE 3 at modelling the error propagation all though it lacks the vertical representation of the complete flow field. Thus, steady Kalman gain vectors can be obtained from a MIKE 21 simulation which is significantly faster than the MIKE

3 because of the simpler dynamics.

Seen in the light of error covariance propagation, the idea of using Ensemble Kalman filter based MIKE 21 gains in MIKE 3, combines rank reduction and dynamical approximation in the generation of the Steady gains and applies the cheap steady assumption for operational use.

6 Experimental design

The purpose of the experimental design is to demonstrate the applicability of using Kalman gains derived from a MIKE 21 run to assimilate water level in MIKE 3 in an idealised bay set-up. The basic idea is to perform a twin test in a well understood idealised bay. First, the set-up for the MIKE 3 base run will be presented. Thereafter, some realistic errors are incorporated to mimic the situation encountered in reality, where both wind and boundary forcing have non gaussian errors (perturbed runs). MIKE 21 is similarly executed in a set-up with the realistic errors. Finally, the different assimilation runs in the study are described.

6.1 Base run

The bathymetry of the idealised bay, in which the twin test is situated, is shown in Figure 1. It covers a 200 km x 200 km area, has a linear variation from land to 100 meters depth and an open northern boundary. MIKE 3 was run with a 12-hour period and one meter amplitude sinusoidal level forcing at the boundary and a west to east moving cyclone with maximum wind speeds of about 35 m/s. The solution is a cyclonically moving Kelvin wave perturbed by the cyclone. The model has a 10 km x 10 km horizontal resolution and a uniform 10 meter vertical resolution. Each time step is 15 minutes. Thus, it is a fast model to run and thus allows for many repetitions of even quite expensive assimilation schemes.

Time series from the base run were extracted from the three points indicated in Figure 1. The base run is taken to be the truth and these three

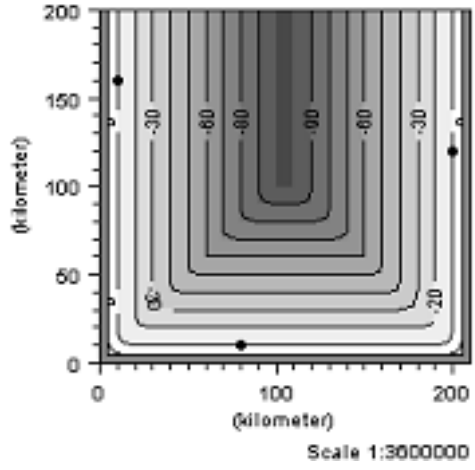

Figure 1: Idealised bay bathymetry. The three dots indicate measurement positions, (1,16), (8,1) and (20,12) counter clockwise.

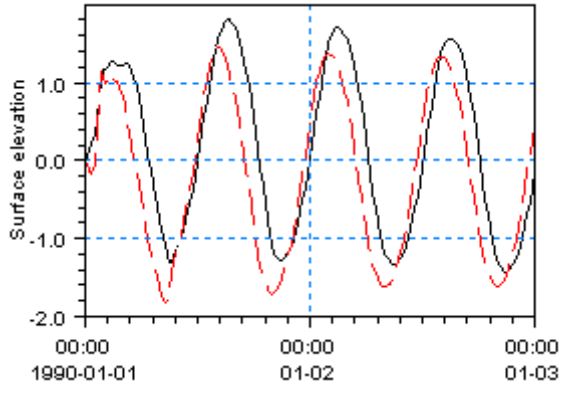

Figure 2: Surface elevation in (1,16) for the base run (solid) and the perturbed run (dashed).

time series will be considered observations and are the only information from the true run in the perturbed runs.

6.2 Perturbed runs

In the perturbed runs all settings are the same apart from the boundary and wind forcing. The sinusoidal boundary forcing has a one-hour phase lag and the cyclone has a time varying error in both strength and direction. This yields a quite severely distorted solution. To exemplify, Figure 2 shows the time series of surface elevation for the base and the perturbed set-up of MIKE 3.

6.3 Description of runs

1. Base run. MIKE 3 with base forcing. This run is taken as the truth. Run (3), (4), (6), (7) and (8) assimilates the three time series of surface elevation extracted from the base run.

2. MIKE 3 perturbed run. MIKE 3 with perturbed forcing and no data assimilation.

3. MIKE 3 Ensemble Kalman filter. MIKE 3 with perturbed forcing and data assimilation. The ensemble Kalman filter is used with 100 ensembles.

4. MIKE 3 Steady Kalman filter. MIKE 3 with perturbed forcing and data assimilation. The steady Kalman filter is used with average Kalman gain matrices calculated from (3).

5. MIKE 21 perturbed run. MIKE 21 with perturbed forcing and no data assimilation.

6. MIKE 21 Ensemble Kalman filter. MIKE 21 with perturbed forcing and data assimilation. The ensemble Kalman filter is used with 100 ensembles.

7. MIKE 21 Steady Kalman filter. MIKE 21 with perturbed forcing and data assimilation. The steady Kalman filter is used with average Kalman gain matrices calculated from (6).

8. MIKE 3 with Steady Kalman filter based on MIKE 21. MIKE 3 with perturbed forcing and data assimilation. The steady Kalman filter is used with average Kalman gain matrices calculated from (6).

7 Results and discussion

Since the entire state of the truth is known, the performance of the various runs can be measured by a root mean square error estimate calculated for the last 24 hours and averaged over the entire domain,

$$RMSE = \frac{1}{N_{Water}} \sum_{\{j,k\,|pos(j,k)\ is\ water\}} \frac{1}{N} \sum_{i=1}^{N} (\zeta_i^{true}(j,k) - \zeta_i^{pert}(j,k))^2$$

N_{Water} is the number of water points in the domain and N is the number of time steps over the last 24 hours. Only the surface elevation is included. The superscript *true* indicates base run water levels and superscript *pert* indicates the various perturbed run water levels. The RMSE measure captures both the variance and the bias. The results for run (2) to (8) are shown in Figure 3.

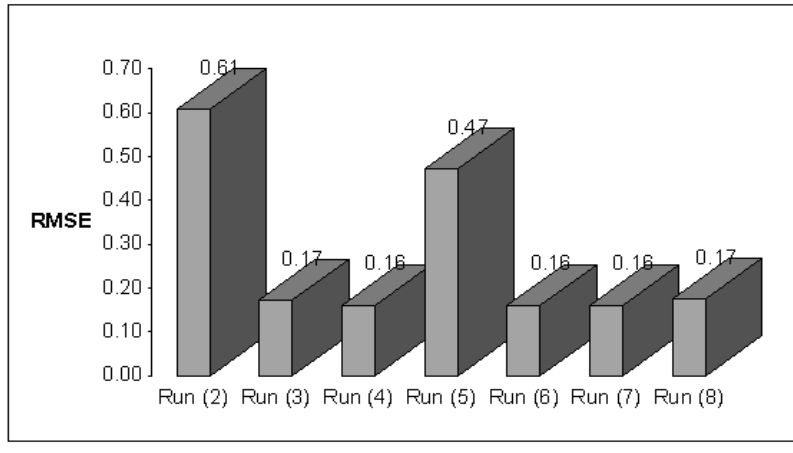

Figure 3: RMSE's for run (2) – run (8). Values are shown as well.

A number of remarks need to be made. As expected the RMSE of run

(2) and run (5), with no data assimilation is significantly larger than in all other runs. Run (5) on the MIKE 21 engine gives a smaller RMSE than the similar run (2) on the MIKE 3 engine. This is mainly due to a smaller sensitivity to the perturbations. The Ensemble Kalman filter runs, (3) and (6) both give significant reductions in RMSE of surface elevation. The Steady runs, (4) and (7) both maintain or even decrease this low value despite the fact that further approximations are made. This is due to the fact that the optimal gains in the test case actually are quite constant in time. Hence, the steady formulation acts to stabilise the filter, by reducing the stochastic noise through the time averaging.

Finally, pay attention to run (8). This run is the cheapest of all the 3D assimilation runs and yet it gives an equally low RMSE, while estimating the full three-dimensional solution. Using MIKE 21 to generate the gains makes the approach operational, since it only requires about 1/10 of the time needed in MIKE 3. The set-up of any operational data assimilation system requires some tuning and occasional recalculation. Hence, the factor of 10 becomes even more essential.

MIKE 21 and MIKE 3 do not in general produce the same results because of their somewhat different formulations. It is interesting to note, however, that the sensitivity to the forcing, which is expressed in the Kalman gains shows sufficient similarity for the MIKE 21 gains to be applied in the MIKE 3 assimilation scheme. Further tests are required to see whether this carries over to shallow water applications with more complicated bathymetries where the model differences play a more important role.

8 Conclusions

A 2D model has been used to calculate Kalman gains that relate a difference between model and measurement estimate of surface elevation at a given point to a correction of surface elevation and depth integrated velocities at all points in the model domain. These have been averaged in time with the purpose of using them in an assimilation scheme that assumes time constant gains. A scheme for extrapolating these gains to the partly similar variables of a 3D model has been implemented based

on the assumption of a barotropic error. Finally, it was assumed that a 2D model together with the extrapolation scheme gives a sufficiently good representation of the 3D model Kalman gains. The entire combined approach gives very good results for a simple bay test case. These results are promising for application in complicated state-of-the-art operational systems of coastal seas without the use of HPC-facilities. The approach can thus be applied for a range of low-budget purposes. Future work will test the applicability of the approach in an operational system of the North Sea/Baltic Sea system.

References

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

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

Madsen, H. & Cañizares, R. (1999), 'Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling', *International Journal of Numerical Methods in Fluids* **31**(6), 961–981.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2001), Data assimilation of tidal gauge data in a three-dimensional coastal model, *in* 'Proceedings 4^{th} DHI Software Conference', DHI Water & Environment.

Paper D

Data assimilation in an operational forecast system of the North Sea - Baltic Sea system

D

To appear as

Jacob V. Tornfeldt Sørensen, Henrik Madsen, Henrik Madsen, Henrik René Jensen, Peter Skovgaard Rasch, Anders C. Erichsen, and Karl Iver Dahl-Madsen. Data assimilation in an operational forecast system of the North Sea - Baltic Sea system. In H. Dahlin and N.C. Flemming and K. Nittis and S. E. Petersson *Building the European Capacity in Operational Oceanography, Proceedings of the Third EuroGOOS Conference*, Athens, Greece, December 2002.

Data assimilation in an operational forecast system of the North Sea - Baltic Sea system

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], Henrik Madsen[2], Henrik René Jensen[1], Peter Skovgaard Rasch[1], Anders C. Erichsen[1], and Karl Iver Dahl-Madsen[1]

Abstract

The operational service the "Water Forecast" gives daily forecasts for the North Sea, Baltic Sea and interconnecting waters. The basic computational units include a 3D hydrodynamic module, a 3D environmental module and a wave module. An ongoing development is focused on data assimilation of tidal gauge and SST data. A cost-effective Kalman filter based procedure that uses a regularised constant Kalman gain is applied for the tidal gauge data. For assimilation of SST data a simplified Kalman filter procedure is adopted. The combined approach gives an acceptable computational overhead for operational applications. Performance of the modelling system is evaluated.

1 Introduction

During the last decades a number of complimentary developments within oceanographic modelling and monitoring have been taking place. Numerical modelling has advanced to the stage where operational systems are now run on a routine basis, predicting an ever-increasing number of physical and biogeochemical properties (Pinardi & Woods 2002), (Erichsen & Rasch 2002). Simultaneously, a growing amount of observations of a wide range of these properties in the shelf and coastal seas are becoming available in real or near-real time. Hence with the advance of data assimilation schemes suitable for shelf and coastal seas, the potential of an integrated approach has become clear. It is now possible to estimate the state of the sea as a composite of on-line observations and model results through the use of data assimilation techniques. In this way, the

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

relatively precise but sparse data can in essence be interpolated by the theoretical knowledge embodied in the physically consistent model.

DHI Water and Environment is operating a forecast system of the North Sea – Baltic Sea system called the Water Forecast (WF). This contribution demonstrates the application of cost-effective data assimilation schemes for assimilation of tidal gauge and SST data into the high-resolution model, which provides the computational component of the WF.

2 The Water Forecast operational system

In 1999 the development of an end-user oriented web based operational modelling system of the North Sea - Baltic Sea was initiated at DHI Water & Environment, (Jensen, Møller & Rasmussen 2002). Since June 2001, the system has produced operational forecasts. The model area is depicted in Figure 1. It includes two open boundaries in the North Sea and stretches to cover the entire Baltic Sea. An area of particular interest is defined, which surrounds Denmark and southern Sweden as shown in Figure 1. The basic computational engine is composed of a two-way dynamically nested 3D baroclinic hydrodynamic module (MIKE 3 HD), a 3D environmental module (MIKE 3 EU) as well as a 2D wave module (MIKE 21 SW). Every 12 hours a 4 days forecast is provided, predicting a range of physical and environmental parameters. These include water level, currents, salinity, temperature, wave height, period, spectra and swell as well as chlorophyll-a, oxygen and algae growth. A thorough description of the system can be found in (Erichsen & Rasch 2002).

3 The data

In principle all data, which can be assimilated at an acceptable cost and yet provide an improvement to the ocean state estimation skill, ought to be considered. For the WF system two data sources are considered initially: Tidal gauge water level observations and satellite sea surface temperature (SST) observations from the Ocean Pathfinder AVHRR sen-

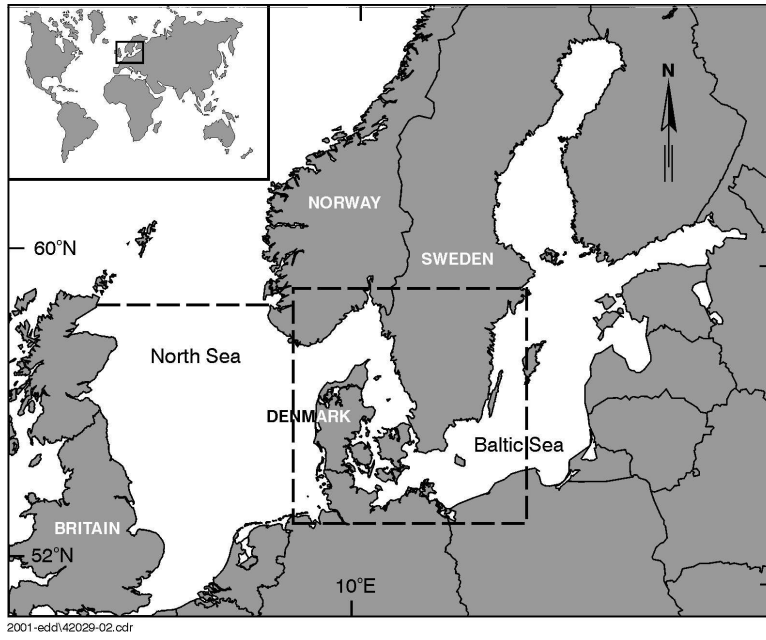

Figure 1: The WF model area. The dotted line between Scotland and Norway indicates an open boundary whereas the dashed square shows the area in focus.

sors.

3.1 Tidal gauge data

Tidal data from 14 stations in the focus area have been selected for the present study. Eight of these will be assimilated and six used for validation. Figure 2 shows the positions of these stations and whether each station acts as a measurement (M) or a validation (V) station. Data are provided from the Danish Meteorological Institute and the Danish Coastal Authority.

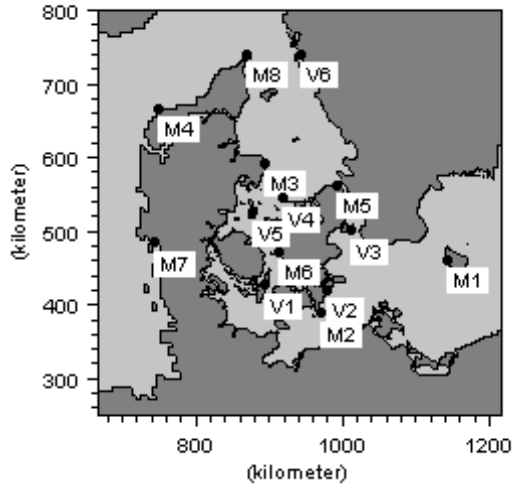

Figure 2: Tidal gauge measurement (M) and validation (V) stations.

3.2 SST data

The Pathfinder AVHRR SST data was obtained from Collecte Localisation Satellites (CLS), who has pre-processed the data into 10-day interval products as part of the EU funded project, GANES. For the purpose of assimilation it is essential to notice that the SST fields are derived from composit images and are therefore not snapshots in time. Data from 22/9, 2/10 and 12/10 1994 was used for assimilation. The SST field

from the 22/10 1994 was used for validation. Figure 3 shows the data coverage by the SST data. Further, all available temperature data from the ICES database in the given period was used for validation. These are highly sparse in time and space. Their spatial distribution is shown in Figure 4.

4 The data assimilation approach

The final aim is to utilise data assimilation techniques that can have a widespread use in engineering and scientific applications and thus it is essential to develop schemes, which are both cost-effective and robust. We define this as model execution time and memory requirement less than five times that of a pure model run and preferably below a factor of two. The following two approaches comply with these constraints while retaining a robustness and effectiveness caused by the advanced data assimilation approaches on which they build and the corresponding physical error assumptions. Thus both are applicable for operational systems.

4.1 Tidal gauge assimilation

The implemented assimilation scheme for tidal gauge data is a hybrid scheme that combines the ensemble Kalman filter approach (Evensen 1994) with a barotropic dynamical approximation (Sørensen et al. 2002), a steady gain assumption (Cañizares et al. 2001) and a regularisation of the gain matrix. A one-layer barotropic version of the three-dimensional hydrodynamic model is run over a three days period with an ensemble Kalman filter using 100 ensembles. Errors are assumed to originate solely from the open boundaries and the wind field. The time varying Kalman gain is averaged over the last two days of the run and saved for application in the steady Kalman filter approach. This two-dimensional Kalman gain basically assumes the model errors to be barotropic and hence for application of the gain in the three-dimensional baroclinic model the same assumption can be followed to relate the full velocity field to the depth averaged velocities. Thus, an update of the full velocity field based

on a depth averaged gain will merely shift the mean of the vertical profile, not the structure. For further detail refer to (Sørensen et al. 2001).

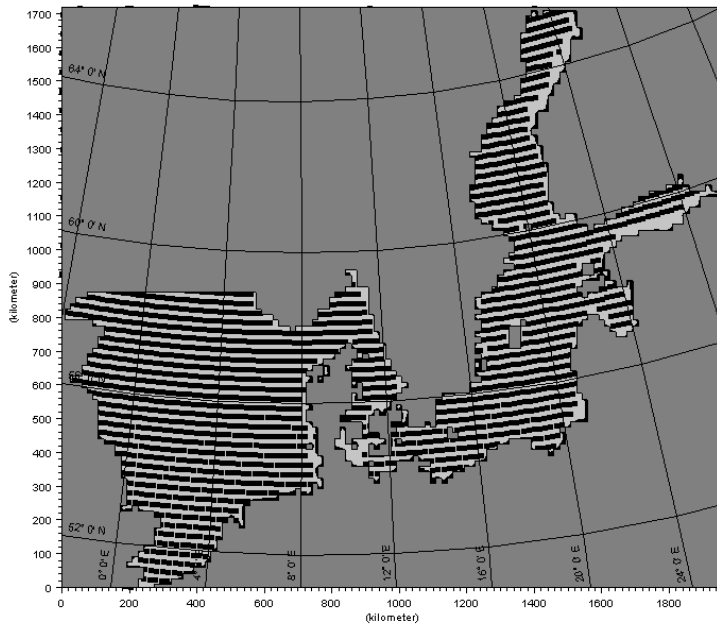

Figure 3: Data coverage for every 10 day period of the AHVRR SST data.

Due to spurious correlations in the ensemble Kalman filter, which have not diminished in the averaging process, rather large Kalman gain values can be observed in data sparse regions even when such correlations have no physical interpretation. Also the correlation between water level and velocity is dominated by noise in large parts of the area. Thus, in order to ensure robust results a rough manual regularisation of the gain is performed. This practically sets velocity gains to zero and cuts off water level gains at the 0.01 contour. More advanced regularisation, which allows the velocity to re-enter and significant negative correlations to remain must be considered as a future improvement. However, the present implementation is an important first step.

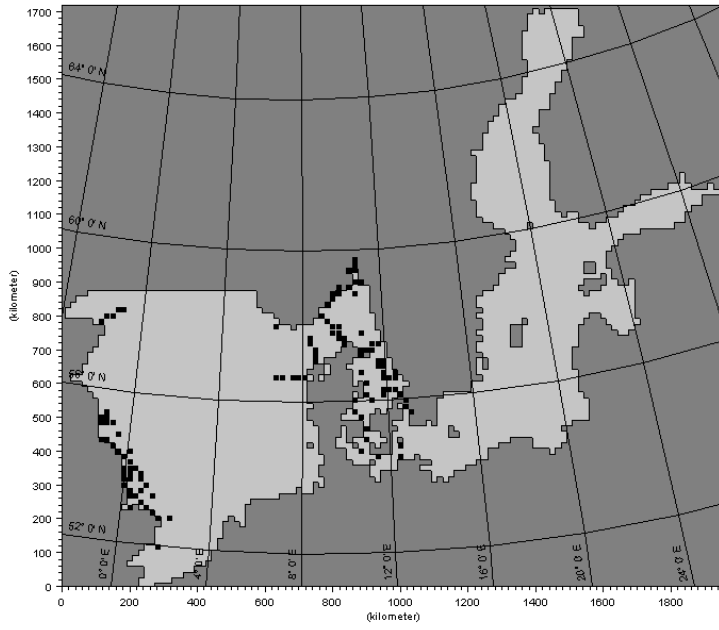

Figure 4: Positions of in-situ temperature measurements from the ICES database during the validation period.

4.2 SST assimilation

A module for cost-effective data assimilation of SST data (Annan & Hargreaves 1999) has been implemented in MIKE 3 HD. Based on a few simple dynamical assumptions imposed on the Kalman filter approach the data assimilation module is able to correct the temperature field above the mixed layer. It is assumed that horizontal correlations are small enough to be ignored. Further, it is assumed that the areas above and below the mixed layer are each well mixed. This yield a one-dimensional gain vector for each SST data point approximating the Kalman gain. The SST data are interpolated in time to provide an observation at every time step. This represents the fact that the SST fields are averages over a longer period of time. The base of the mixed layer can be defined in a range of different ways. Note that the mixed layer is merely a theoretically constructed concept. In the present approach, it is taken to be the highest grid point with a diffusion coefficient of 10^{-4} or lower. However, the exact threshold value is a calibration factor.

5 Results and discussion

For the purpose of testing the developed water level data assimilation schemes, the period 3 to 27 February 2002 was selected, whereas a period between 20 September and 22 October 1994 was considered for the SST assimilation. The different periods were chosen to match the available data for assimilation and validation. In both periods the model was run both with (assimilation) and without (model) the appropriate assimilation scheme.

5.1 Tidal gauge results

The performance of the model with assimilation of tidal gauge data is compared to a pure model execution in Figures 5 and 6, which show the root mean square error of the validation points and the measurement points respectively. The mean values are also shown. There is a clear improvement in all stations with an average of a 35% increased

performance in validation points and 58% in measurement points. This significant improvement can be obtained at an overhead in execution time of less than a factor of two. All stations here are located in data dense regions. In more data sparse regions, performance converges to that of a normal model execution. When the regularised gain is not used a significant bias can be introduced from spurious correlations in the ensemble Kalman filter. Thus, an assimilation scheme has been implemented which meets the constraints of fast execution and robust improvements.

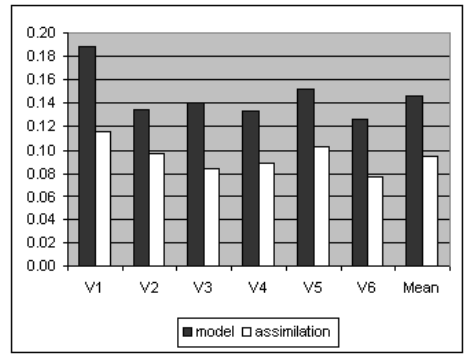

Figure 5: Root mean square error of water level results in validation points with (white bars) and without (dark bars) assimilation.

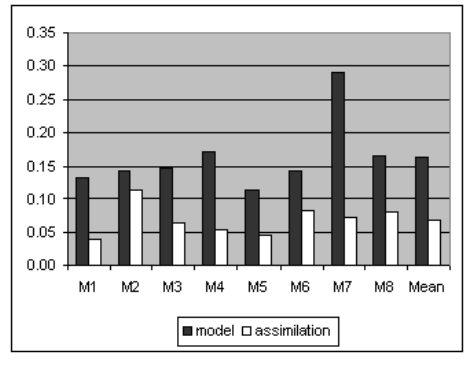

Figure 6: Root mean square error of water level results in measurement points with (white bars) and without (dark bars) assimilation.

5.2 SST results

The performance of the model with assimilation of SST data is compared to a pure model execution in Table 1, showing the root mean square error of the 10 days forecast and the validation SST field from the 22 October 1994. Also shown is the root mean square error compared to the in-situ measurements from the ICES database. The latter is divided into two bins above and below 20m in an attempt to roughly look at results above and below the thermocline. The assimilation scheme clearly improves the results, where expected. It was assumed that no information was available below the thermocline and thus it is consistent to maintain the performance below 20m. Above 20m we see an 18% reduction of the RMSE. However a more significant reduction must be expected to be obtainable if the SST data was assimilated timely. The 10 days forecast gives a 30% RMSE reduction. These are encouraging results, but also in the forecast statistics, a further improvement must be expected when the validation is done timely, since movements of fronts and rapidly changing atmospheric conditions will be more accurately captured.

RMSE (m)	CLS 10 days forecast	ICES Above 20m	ICES Below 20m
Model	0.74	0.66	1.27
Assimilation	0.52	0.54	1.28

Table 1: Root mean square error (RMSE) of temperature results with (Assimilation) and without (Model) assimilation.

6 Conclusions and future work

The successful assimilation of tidal data and satellite derived SST data have been demonstrated in a model of the North Sea – Baltic Sea for operational use. For the assimilation of water level data a proper regularisation of the Kalman gain will be considered for further improvement. The SST assimilation scheme will be developed to use data timely in cloud free areas leaving the propagation of the information to the model

dynamics. In the near future simple optimal interpolation schemes for the assimilation of chlorophylla and dissolved oxygen will also be implemented.

References

Annan, J. D. & Hargreaves, J. C. (1999), 'Sea surface temperature assimilation for a three-dimensional baroclinic model of shelf seas', *Continental Shelf Research* **19**, 1507–1520.

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Erichsen, A. C. & Rasch, P. S. (2002), Two- and three-dimensional model system predicting the water quality of tomorrow, *in* M. L. Spaulding, ed., 'Proceedings of the seventh international conference on estuarine and coastal modeling', American Society of Civil Engineers, American Society of Civil Engineers, pp. 165–184.

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

Jensen, H. R., Møller, J. S. & Rasmussen, B. (2002), 'Operational hydrodynamical model of the danish waters. danish national programme for monitoring the water environment', *Operational Oceanography* **66**, 87–97.

Pinardi, N. & Woods, J., eds (2002), *Ocean forecasting. Conceptual basis and applications*, Springer-Verlag.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2001), Data assimilation of tidal gauge data in a three-dimensional coastal model, *in* 'Proceedings 4^{th} DHI Software Conference', DHI Water & Environment.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2002), Towards an operational data assimilation system for a three-dimensional hydrodynamic model, *in* R. A. Falconer, B. Lin, E. L. Harris & C. A. M. E. Wilson, eds, 'Proceedings of the fifth International Conference on Hydroinformatics', International Water Association, IWA Publishing.

Paper E

Efficient Kalman Filter Techniques for the Assimilation of Tide Gauge Data in Three-Dimensional Modelling of the North Sea and Baltic Sea System

Accepted by *Journal of Geophysical Research*.

Efficient Kalman Filter Techniques for the Assimilation of Tide Gauge Data in Three-Dimensional Modeling of the North Sea and Baltic Sea System

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

Abstract

Data assimilation in operational modeling systems is a discipline undergoing a rapid development. Despite the ever increasing computational resources, it requires efficient as well as robust assimilation schemes to support on-line prediction products. The parameter considered for assimilation here is water levels from tide gauge stations. The assimilation approach is Kalman Filter based and examines the combination of the Ensemble Kalman Filter with spatial and dynamic regularisation techniques. Further, both a steady Kalman gain approximation and a dynamically evolving Kalman gain is considered. The estimation skill of the various assimilation schemes is assessed in a four week hindcast experiment using a setup of an operational model in the North Sea and Baltic Sea system. The computationally efficient dynamic regularisation works very well and is to be encouraged for water level nowcasts. Distance regularisation gives much improved results in data sparse areas, while maintaining performance in areas with a denser distribution of tide gauges.

1 Introduction

Marine operational forecasting systems are being increasingly applied for a number of engineering and public service purposes, e.g. (Pinardi & Woods 2002), (Erichsen & Rasch 2002). The products are valuable for both hindcast, nowcast and forecast situations and in all cases there is a need for higher precision simulation of the physical variables. In order to increase the predictive skill, the numerical models have been continuously improved during the past decades. Better numerical methods have been

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

developed, smaller scales resolved and improved parameterizations implemented. The developments in the numerical models have been carried over to the operational systems as robustness has been proved. Along with this development, attention has been paid to including an increasing number of physical variables in the models. Hence the portfolio of products has been expanded from the hydrodynamic and thermodynamic parameters to include estimation and prediction of waves, biogeochemical parameters and sediments. All together these developments have taken us to the stage we have reached today.

A number of model errors remain despite the clear improvements of the predictive skill that the present operational systems have experienced during their life time. However, the measurements that demonstrate this error are often available on-line and can potentially be used to update the estimation of the ocean state in real-time. Methods that pursue this line of thought are referred to as operational data assimilation techniques. Data assimilation is a cross disciplinary field with a range of uses, e.g. the engineering community and meteorological sciences have a long history of successful applications. Data assimilation in ocean models for hindcast studies has also been rather widespread during the past decade.

However, the methodologies are computationally demanding and hence, the use of assimilation approaches has only been applied to a lesser extend in the operational modeling community. Examples are the MERCATOR project, (Bahurel, Mey, Provost & Traon 2002), and the MFSTEP project, (Pinardi, Auclair, Cesarini, Demirov, Umani, Giani, Montanari, Oddo, Tonani & Zavatarelli 2002). In common for these and similar developments is the accessibility of high performance computational resources and assimilation of a large range of satellite and in-situ measurements into three-dimensional regional or global models. For more widespread application, techniques must be applicable on the moderate computational resources available to project engineers and scientists working in applied modeling. For the assimilation of tide gauge data, operational storm surge forecasting has been one of the targets. Here smaller geographical areas and simpler two dimensional models have often been considered, which gives some reduction in the required computational resources. Simultaneously, cheap assimilation methods have also been proven successful, hence encouraging its implementation, e.g. (Vested et al. 1995), (Gerritsen et al. 1995) and (Cañizares et al. 2001).

A less computationally demanding assimilation approach based on the steady solution of the Riccati equation and subsequent use in a Kalman filter was suggested by (Heemink & Kloosterhuis 1990). This approach reduces the computational demands to the same order of magnitude as a standard model execution. (Verlaan & Heemink 1997) suggested the improved reduced rank square root (RRSQRT) extended Kalman filter, with successful application for storm surge prediction along the Dutch coast. (Bertino, Evensen & Wackernagel 2002) similarly used a RRSQRT Kalman filter for water level assimilation in the Odra Lagoon. (Madsen & Cañizares 1999) demonstrated an implementation of the RRSQRT and the Ensemble Kalman filter (EnKF), (Evensen 1994), in an idealised bay. They showed that the two schemes have similar computational demands and performance. However, computational times are of the order 10^2 times greater than a standard model execution. (Cañizares et al. 2001), demonstrated a successful application in the North Sea and Baltic Sea system of a steady Kalman filter using a gain obtained as a time average of the gain produced by the EnKF. Interestingly, spurious correlations caused the results to get worse in data sparse regions, showing the limitation of the Kalman filter approach. Based on ideas from (Houtekamer & Mitchell 1998), (Hamill, Whitaker & Snyder 2001) discussed this artifact of the EnKF and suggested that a distance function can be used to control the effect of uncertain ensemble estimates. (Evensen 2003) argued that such an approach should be avoided because it no longer generates updated ensembles as linear combinations of the forecast ensembles.

The main computational issue in Kalman filter based data assimilation is the propagation of the system error covariance matrix. The EnKF and RRSQRT schemes along with e.g. the SEEK, (Pham et al. 1997), and the SEIK filter, (Pham et al. 1998), attempt to save computational resources by constructing a low rank approximation of the model error covariance. The Steady filter assumes no time variation, but still requires a solution of the Riccati equation or a more elaborate scheme for the generation of the gain. (Dee 1991) suggested using a simpler dynamical model propagator for the error propagation. (Fukumori & Malanotte-Rizzoli 1995) presented a scheme which employed a coarser grid for the error propagation, hence reducing the dimension of the state space but simultaneously simplifying the dynamics.

The objective of the present study is to investigate the possibility of

combining a range of approximate Kalman filter based techniques to the assimilation of tidal gauge data in the North Sea and Baltic Sea system. The techniques are selected in order to provide an optimally efficient scheme for this case, but their nature is discussed in a general regularisation perspective. This framework acknowledges the violation of underlying assumption in the elaborate assimilation schemes and enables the incorporation of prior independent knowledge in the estimation of the ocean state.

Within the regularisation framework we describe four approximations to the EnKF. These are temporal smoothing of the Kalman gain, the steady Kalman gain, a barotropic model error approximation and a distance dependence of the Kalman gain. The performance of the techniques is presently examined in a hindcast scenario of the North Sea and Baltic Sea system, but the goal is to develop schemes that can be used in an operational forecast setting.

In Section 2 the two- and three-dimensional hydrodynamic models employed in this study are presented along with the available tide gauge measurements. Section 3 provides the theoretical basis of the assimilation approaches considered. This encompasses a general discussion of the estimation technique used in Kalman filtering along with a discussion of model and measurement uncertainties. The EnKF is also described in Section 3 as are four regularisation techniques leading to a Kalman gain smoothing, a Steady Kalman filter as well as a barotropic and a distance regularisation. Section 4 presents the design of the numerical experiments. The results are shown and discussed in Section 5, while Section 6 concludes the paper. The nomenclature suggested in (Ide et al. 1997) is followed throughout this work, where applicable.

2 Description of Models and Measurements

The Water Forecast is an operational forecasting system covering a large part of the North Sea, the Baltic Sea and the interconnecting waters, (Erichsen & Rasch 2002). The hydrodynamic model has run operationally as part of the Water Forecast service since June 2001. While the system provides four-day forecasts of hydrodynamic, water quality

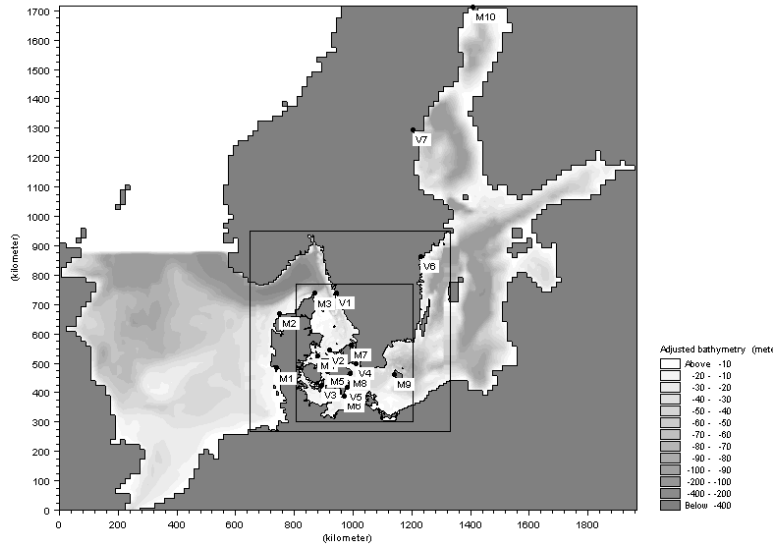

Figure 1: Bathymetry and available tide gauge stations, including 10 measurement stations (M1-M10) and 7 validation stations (V1-V7).

and wave parameters every 12 hours, this study restricts attention to water levels in a hindcast setting.

The hydrodynamic model of the forecast system is the three-dimensional MIKE 3, (DHI 2001), which handles free surface flows. It solves the primitive equations making the hydrostatic and the Boussinesq approximations. The turbulence closure scheme adopted is the k-ϵ model in the vertical and Smagorinsky horizontally. The area covered by the model is shown in Figure 1. Tidally varying water levels are prescribed at the two open boundaries, which are situated in the English Channel and in the Northern North Sea between Stavanger in Norway and Aberdeen in Scotland. Wind fields and sea surface pressure are derived from the Vejr2 commercial weather service, (Rogers, Black, Ferrier, Lin, Parrish & DiMego 2001), and force the momentum equations at the sea surface. The vertical resolution is two meters within the top 80 meters. Larger depths are contained in the model bottom layer. The model is nested as displayed in Figure 1 and the horizontal resolution varies from 9 nautical miles to 1 nautical mile in the inner Danish waters and one third nautical

mile in a few narrow straits. A two-way nesting technique is employed securing a dynamic exchange of mass and momentum between grids.

The numerical model mentioned above attempts to express the true state of system in discrete space and time. The model space is spanned by water level, l, velocity, $\mathbf{v}$, temperature, T, and salinity, S, averaged over spatial volumes at discrete times. Let $\mathbf{x}_{M3}(t_{i-1}) \in \mathbf{R}^{n_{M3}}$ be the model estimate of the true state at time t_{i-1}. Hence, the one-time-step-ahead model operation, M_{M3} provides the model estimate at time t_i as,

$$\mathbf{x}_{M3}(t_i) = M_{M3}(\mathbf{x}_{M3}(t_{i-1}), \mathbf{u}(t_{i-1})) \tag{1}$$

where $\mathbf{u}(t_{i-1})$ is a vector containing the model forcing.

For the purpose of this study the barotropic two-dimensional model MIKE 21, (DHI 2002), was setup on the same horizontal grid and with the same forcing. The model has a smaller state space dimension and simpler dynamics excluding density variations and collapsing the vertical axis. Only water level, l, and depth averaged velocities, $\mathbf{V}$, enters the state, $\mathbf{x}_{M21}(t_{i-1}) \in \mathbf{R}^{n_{M21}}$. The model propagator, M_{M21} provides the model estimate at time t_i as,

$$\mathbf{x}_{M21}(t_i) = M_{M21}(\mathbf{x}_{M21}(t_{i-1}), \mathbf{u}(t_{i-1})) \tag{2}$$

Both in terms of state space dimension and execution times the barotropic model is significantly cheaper.

For the purpose of this study, 17 tide gauge stations were selected. These are displayed in Figure 1. All stations are situated in Danish or Swedish waters. The 10 tide gauge stations used for assimilation will be referred to as measurement stations and indicated by an 'M'. The 7 stations used for performance assessment will be referred to as validation stations and indicated by a 'V'. The stations in each of these groups are numbered consecutively and the corresponding station names can be read from Figures 2 and 3.

A much better data coverage than what is used in this constribution is needed for improved storm surge predictions in the North Sea. However, the Water Forecast model does not have storm surge modeling as a sole objective. The objective also lies in transports as well as ecosystem parameters and the aim is to apply a unified consistent model for

all purposes. Hence, we employ a three-dimensioanl model and it would be appropriate to also validate results against the velocity. However, very little representative velocity data has been at our disposal in the considered period and we follow a more traditional storm surge model validation approach in a restricted area, which should be regarded only a partial validation for the full purpose of the system. In this respect, it is important that there are validation stations (V6 and V7) far from assimilation stations to examine aspects of the consistency of the employed techniques.

3 Assimilation Approach

The schemes used for the assimilation of water level data in the present study can be categorized as sequential estimation techniques. The theory can basically be divided into two parts. One is an estimation of the true state based on the distributions of the model and measured variables, respectively. The standard approach is to assume no bias and use the best estimator in a minimal variance sense. This estimator is presented in Section 3.1. The other part is a specification and a subsequent propagation of the stochastic model state in between measurement times. The observational error also needs to be quantified. The specification of an error model for the numerical model as well as for the observations is build on a number of assumptions. A discussion of these and a description of the error models employed in the present study is presented in Section 3.2. In a dynamical model the model error is continuously altered by the model dynamics and hence the error description needs to be propagated in time. A Markov Chain Monte Carlo approach is followed here leading to the Ensemble Kalman Filter (EnKF) described in Section 3.3.

The ensemble approach is an efficient way of making the work load of the model error propagation tractable, by reducing the degrees of freedom in the description dramatically. However, the resulting scheme is still too expensive for many operational systems, which are typically pushed close to the limit in terms of computational resources in order to resolve as many processes as possible. Further, the EnKF scheme may introduce spurious correlations in data sparse regions due to an inaccurate model error description and the stochastic nature of the scheme. Hence, despite

risking introducing non-dynamical modes in the system, various forms of regularisation of the gain is proposed. Section 3.4.1 presents a simple temporal smoothing of the Kalman gain matrix, Section 3.4.2 describes a Steady Kalman gain approach, while Section 3.4.3 presents a dynamic regularisation based on the assumption that the errors are barotropic. Finally Section 3.4.4 describes a distance regularisation technique.

3.1 BLUE Estimator

In this section the estimation of the state of the system is under consideration. This is often referred to as the analysis step. The state is essentially a multivariate continuous four-dimensional field. Observations are noisy samples from this field and are typically integral measures over some spatial and temporal scale. Similarly, model variables represent per definition spatial averages of the true state. The spatial and temporal representation of the three-dimensional model is taken as a common reference frame and the state is described in terms of its projection onto it, $\mathbf{x}^t(t_i) \in \mathbf{R}^n$. Here t_i denotes time indexed by i and $\mathbf{x}^t(t_i)$ is further restricted to include the model variables, water level, l_i and velocity, $\mathbf{v}_i$, hence excluding temperature and salinity. This approximation is due to later time savings in the EnKF error propagation and in order to facilitate the barotropic regularisation in Section 3.4.3. Next, let the prediction by a numerical model, $\mathbf{x}^f(t_i)$, describe the first moment of the stochastic state and assume its error covariance matrix, $\mathbf{P}^f(t_i)$, known.

Now, let the observation at time t_i, $\mathbf{y}_i^o \in \mathbf{R}^p$ be related to $\mathbf{x}^t(t_i)$ through the linear measurement equation,

$$\mathbf{y}_i^o = \mathbf{H}_i \mathbf{x}^t(t_i) + \boldsymbol{\epsilon}_i \tag{3}$$

The matrix $\mathbf{H}_i \in \mathbf{R}^{p \times n}$ is a linear operator projecting the state representation onto the measurement space and the measurement noise term $\boldsymbol{\epsilon}_i$ is assumed to be an i.i.d. random process. Assume the first and the second moments of this noise to be known, respectively, $\mathbf{0}$ and $\mathbf{R}_i$.

When information from the true system becomes available in the form of measurements an improved state estimate can be obtained. One procedure for doing this is to assume a linear combination of the unbiased model prediction and the observation that gives the minimum variance

estimate, $\mathbf{x}^a(t_i)$. This approach is called the Best Linear Unbiased Estimate (BLUE). A derivation can be found in (Jazwinski 1970) yielding the following estimator,

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}^f(t_i)) \tag{4}$$

The Kalman gain matrix, $\mathbf{K}_i \in \mathbf{R}^{n\times p}$, is given by,

$$\mathbf{K}_i = \mathbf{P}^f(t_i)\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}^f(t_i)\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \tag{5}$$

The error covariance, $\mathbf{P}^a(t_i)$, of $\mathbf{x}^a(t_i)$ will always be less than or equal to $\mathbf{P}^f(t_i)$ and can be calculated as,

$$\mathbf{P}^a(t_i) = \mathbf{P}^f(t_i) - \mathbf{K}_i\mathbf{H}_i\mathbf{P}^f(t_i) \tag{6}$$

The set of equations (4) and (5) supplies the variance minimizing analysis among the class of linear equations and (6) the a posteriori error covariance, if $\mathbf{R}_i$ and $\mathbf{P}^f(t_i)$ indeed were the real a priori error covariances.

3.2 Error Descriptions

The optimality of the BLUE estimator for the analysis relies on a correct specification of the model and measurement error covariances. Hence, any misspecification of these will make the scheme suboptimal. This section takes a closer look at model and observation errors and how they are quantified in the scheme.

3.2.1 Measurement Error

The error, $\boldsymbol{\epsilon}_t$, in the measurement equation (3) includes both an instrumental error and a representation error and is thus properly refered to as a measurement constraint error as suggested by (Fukumori et al. 1999). Depending on the observation considered, either instrumental or representation error can be the dominating source. The instrumental error refers to the actual error in measurement of the physical property under consideration. Often such statistics can be assessed rather precisely. However, the instruments may be badly calibrated and electronic

or mechanical malfunctioning may induce systematic errors. Hence, an elaborate quality check on the data must be performed.

The other contribution to the measurement constraint error is due to the fact that the state estimation is done in the model space. (Fukumori et al. 1999) provide a nice discussion of this, arguing that the resulting measurement representation error contributes to the measurement constraint error. A measurement typically represents a physical property averaged over a different spatial and temporal scale than the model representation. As an example, the measured water level in a corner of a harbour needs not be representative of the water level averaged over an area of $1 \times 1\ km^2$. If we had retained the continuous reality as the space in which we estimate the state, then the spatial discretisation of the model should have been described as a model error. However, the adaption of the discrete model space as the state space moves the error to the measurement equation, now expressing that the measurement only approximately represents the spatial average adopted by the model.

Similarly, observed signals that are caused by processes not included in the numerical model can be described as a representation error. The only difference is that we now consider the dynamical subspace spanned by the model rather than the spatial subspace. It must be stressed that the adoption of the model space as the state space is a choice. We could have chosen another projection of reality, but important in either case is that the error process description is formulated according to this choice.

When using the state definition discussed above, a white noise error does not provide a good description of the expected error and hence the entire premises of the Kalman filter is violated. The representation error must be expected to be colored and it should be described as such. This can be done by augmenting the state space by its colored components of a suitable measurement error model. This is however a difficult task and the rather crude approximation of merely increasing the white measurement error standard deviation is taken here, as suggested by (Fukumori et al. 1999).

The measurement error is usually given some predefined value, by considering instrumentation errors and representation errors as discussed above. The measurement error can easily be time varying if justified by

such considerations. In the present implementation, the error at a given tide gauge station is assumed independent to all other stations. For the instrumental errors this is true, but for representation error this might be violated.

3.2.2 Model Error

Let $\mathbf{x}_M^t(t_i)$ define the true state represented in the space of the model at time t_i. A system equation can then be formulated as,

$$\mathbf{x}_M^t(t_i) = M_M(\mathbf{x}_M^t(t_{i-1}), \mathbf{u}(t_{i-1})) + \boldsymbol{\eta}_{M,i} \tag{7}$$

Thus, the model error, $\eta_{M,i}$, describes the error imposed by the model operator, M_M at time t_i. This error must be described along with the error covariance of the state at an initial point in time in order to provide a stochastic description of the system.

The description of model error is a complex task. The exclusion of processes at the very level of the definition of the mathematical model and the spatial discretisations used in the state description are model errors, but described in terms of the representation error as discussed in Section 3.2.1. Errors in the mathematical formulation of processes we wish to describe (including feedback from undescribed processes!) and the numerical methods used to solve the equations as well as numerical truncation errors and parameter specifications, all impose errors in the model simulation. Finally, incorrect forcing terms are potentially major sources of model error. The model error has a complex spatially varying structure and is dynamically altered throughout its propagation in time. It is thus presently intractable to describe accurately. However, an approximate second order description of its statistical properties is not out of reach.

When looking at the sources of model error in a well calibrated hydrodynamic model of a coastal area, it is a good first approximation to assume that the main error source at each time step comes from the forcing terms. The system is quite strongly driven by its forcing and these are known to be inaccurate. Atmospheric forcing is provided by meteorological forecast or hindcast models and open boundary water levels are typically described by a model of harmonic constituents. In the present

implementation it is assumed that all other model errors are neglectable and hence, a model error description can be provided if the error sources in the forcing terms can be propagated throughout the system. The errors in the forcing terms are assumed to be colored processes described by an autoregressive model with a spatially co-varying error driving it, i.e.

$$\boldsymbol{\xi}_i = M_{AR(1)}(\boldsymbol{\xi}_{i-1}, \boldsymbol{\eta}_i) \;=\; \mathbf{A}\,\boldsymbol{\xi}_{i-1} + \boldsymbol{\eta}_i \tag{8}$$

where $\mathbf{A} = diag(\boldsymbol{\alpha})$. For the sake of simplicity, the noise process $\boldsymbol{\eta}_i$ is assumed Gaussian with zero mean and error covariance matrix, $\mathbf{Q}_i^{\eta} \in \mathbf{R}^{r\times r}$. Hence, $\mathbf{x}^t(t_i)$ is augmented with the open boundary water level and wind velocity error description and an extended operator, $M = (M_M, M_{AR(1)})^T$, is introduced. This leads to a system equation with additive noise, which will be used in the remainder of this work,

$$\begin{bmatrix} \mathbf{x}_M^t(t_i) \\ \boldsymbol{\xi}_i \end{bmatrix} = \mathbf{x}^t(t_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_{i-1}), \boldsymbol{\eta}_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_{i-1})) + \begin{bmatrix} 0 \\ \boldsymbol{\eta}_i \end{bmatrix} \tag{9}$$

The error covariance of $\begin{bmatrix} 0 \\ \boldsymbol{\eta}_i \end{bmatrix}$ is $\mathbf{Q}_i = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{Q}_i^{\eta} \end{bmatrix}$. The determination of the error covariance $\mathbf{Q}_i^{\eta}$ is based on experience and theoretical considerations.

3.3 Ensemble Kalman Filter

The Kalman filter based data assimilation schemes used today are all based on the BLUE estimator. They differ mainly in the way they propagate the stochastic state representation. The foundation of the Ensemble Kalman Filter (EnKF) is to approximate the propagation of the full pdf using a Markov Chain Monte Carlo technique, (Evensen 1994). While the deterministic model in Eq. (1) or Eq. (2) propagates the state assuming the model and forcing to be perfect, the EnKF takes the stochastic nature of the model prediction and the non-linearities explicitly into account.

An ensemble of q state realizations is defined at an initial point in time. In the approach presented here, the same initial state defines all ensembles with zero variance at the beginning of a spin-up period. During this period the forcing errors are propagated throughout the system to provide the initial mean state estimate and model error covariance matrix.

Each ensemble member is propagated according to,

$$\mathbf{x}_j^f(t_i) = M(\mathbf{x}_j^a(t_{i-1}), \mathbf{u}(t_{i-1}), \boldsymbol{\eta}_{j,i}), \quad j = 1, ..., q \tag{10}$$

The model error, $\boldsymbol{\eta}_{j,i}$ is randomly drawn from a predefined distribution with zero mean and covariance, $\mathbf{Q}_i^\eta \in \mathbf{R}^{r\times r}$. With each ensemble propagated by Equation (10), the mean state estimate and model error covariance estimate are provided by the following equations,

$$\hat{\mathbf{x}}^f(t_i) = \frac{1}{q}\sum_{j=1}^{q} \mathbf{x}_j^f(t_i) \tag{11}$$

$$\mathbf{P}^f(t_i) = \mathbf{S}_i^f(\mathbf{S}_i^f)^T, \quad \mathbf{s}_{j,i}^f = \frac{1}{\sqrt{q-1}}(\mathbf{x}_j^f(t_i) - \hat{\mathbf{x}}^f(t_i)) \tag{12}$$

The vector, $\mathbf{s}_{j,i}^f \in \mathbf{R}^n$, is the $j'th$ column of $\mathbf{S}_i^f \in \mathbf{R}^{n\times q}$. The update can be performed by Equations (4) and (5), when given the proper interpretation in an ensemble setting. For computational efficiency an algebraically equivalent set of equations are used.

In order to maintain correct statistical properties of the updated ensemble, each ensemble member must be updated rather than the ensemble state estimate. For the same reason an ensemble of measurements must be generated and used for each ensemble member update accordingly rather than the measurement itself, (Burgers et al. 1998). Hence,

$$\mathbf{y}_{j,i}^o = \mathbf{y}_i^o + \boldsymbol{\epsilon}_{j,i}, \quad j = 1, ..., q \tag{13}$$

Randomly generated realizations, $\boldsymbol{\epsilon}_{j,i}$, of $\boldsymbol{\epsilon}_i$ are added for each member. The update scheme presented here specifically uses the uncorrelated measurement structure to assimilate simultaneous measurements sequentially. The updating algorithm for every ensemble member, j, reads, (Chui & Chen 1991),

$$\mathbf{x}_{j,m}^a(t_i) = \mathbf{x}_{j,m-1}^a(t_i) + \mathbf{k}_{i,m}(y_{j,i,m}^o - \mathbf{h}_{i,m}\mathbf{x}_{j,m-1}^a(t_i)), \quad m = 1, ..., p \tag{14}$$

and $\mathbf{x}_{j,0}^a(t_i) = \mathbf{x}_j^f(t_i)$. In (14) $y_{j,i,m}^o$ is the $m'th$ element in $\mathbf{y}_{j,i}^o$ and $\mathbf{h}_{i,m}$ is the $m'th$ row of $\mathbf{H}_i$. Treating one measurement at a time the Kalman gain is a vector, $\mathbf{k}_{i,m}$, given by,

$$\mathbf{k}_{i,m} = \frac{\mathbf{S}_{i,m-1}^a \mathbf{c}_{i,m}}{\mathbf{c}_{i,m}^T \mathbf{c}_{i,m} + \sigma_{i,m}^2}, \quad \mathbf{c}_{i,m} = (\mathbf{S}_{i,m-1}^a)^T \mathbf{h}_{i,m}^T \tag{15}$$

The $m'th$ diagonal element in $\mathbf{R}_i$ is denoted $\sigma^2_{i,m}$. The matrix $\mathbf{S}^a_{i,m}$ in (15) is calculated as

$$\mathbf{S}^a_{i,m} = [\mathbf{s}^a_{1,i,m} ... \mathbf{s}^a_{q,i,m}], \quad \mathbf{s}^a_{j,i,m} = \frac{1}{\sqrt{q-1}}(\mathbf{x}^a_{j,m}(t_i) - \hat{\mathbf{x}}^a_m(t_i)) \tag{16}$$

for $m = 1, ..., p$ and $\mathbf{S}^a_{i,0} = \mathbf{S}^f_i$. Now, (14), (15) and (16) provides the update equations of all ensemble members, one measurement at a time. The time consumption of the EnKF is of the order of q times a standard model execution.

3.4 Regularisation

In the EnKF most of the computational effort is used for providing an estimate of the Kalman gain matrix, $\mathbf{K}_i$. This matrix contains $n \times p$ elements which are calculated based on q ensembles. This leads to uncertain estimates which in particular can have an unwanted effect in data sparse regions with large model variability. Such areas are susceptible to erroneous updates from spurious correlation estimates, (Hamill et al. 2001). However, even for $q = \infty$, the gain estimate will only have a limited accuracy because of the simplistic nature of the models used to describe measurement and model error. Propagating an approximate error source gives an approximate error covariance matrix.

Regularisation methods allow the expression of a prior knowledge about the elements in $\mathbf{K}_i$ and their interdependence to be taken into account, (Hastie, Tibshirani & Friedman 2001). The techniques can usually be cast in a Bayesian framework, e.g. if a prior information about the model error covariance, $\mathbf{P}^{prior}$, is available for $\mathbf{P}^f$, then the posterior estimate, $\mathbf{P}^{posterior}$, is

$$(\mathbf{P}^{posterior})^{-1} = (\mathbf{P}^{prior})^{-1} + (\mathbf{P}^f)^{-1} \tag{17}$$

Such an approach is not tractable in the high-dimensional state space under consideration. However, the line of thought can still provide a useful angle at Kalman filtering. Is there knowledge about the model error covariance that clearly conflicts with the estimates provided by e.g. the EnKF? Regularisation methods deliberately makes biased estimates in order to lower the variance of the estimated elements. Because of the approximate error models and structural model errors the estimates of

the $\mathbf{K}_i$-elements will typically be biased anyhow, so it makes sense to express this in order to lower the total prediction error of the elements, which is a sum of squared bias and variance.

3.4.1 Smoothing of Kalman Gain

Not to be confused with Kalman smoothing for the state vector estimate, a temporal smoothing factor, s, is introduced. It is used to regularise the EnKF derived gain matrix in (5) and implemented in (15). With the instantaneous Kalman gain still being denoted $\mathbf{K}_i$, a smoothed Kalman gain, $\mathbf{K}_i^S$, which replaces $\mathbf{K}_i$, is obtained as,

$$\mathbf{K}_i^S = (1-s)\mathbf{K}_{i-1}^S + s\mathbf{K}_i, s \in [0;1] \tag{18}$$

This approach reduces the stochastic variability of the gain estimate at the cost of leaving out high frequency signals in the gain as well as introducing a phase error. In general the use of a smoothing factor gives a good performance even for insufficient ensemble sizes, (Sørensen, Madsen & Madsen 2003b). Thus, it allows a smaller q to be chosen for the same performance, implicitly saving computational time. This proves the need for regularisation techniques for efficient filtering. It can be regarded as an intermediate method in between the ensemble Kalman filter and the Steady Kalman filter described subsequently.

3.4.2 Steady Approximation

The Steady Kalman filter can be regarded as an ad-hoc regularisation method. Instead of calculating the Kalman gain at every measurement time, it can be assumed that the state and measurement error covariances are the same at every update, which yields a constant Kalman gain. This gain is calculated as a long time average of Kalman gains estimated by the EnKF. Since the gain actually is varying, this introduces a bias in the gain, but the time averaging that creates the steady gain smoothes the gain and lowers the variance. This variance reduction possibly lowers the prediction error of the gain elements if the time varying bias indeed is not too big. Using a snapshot of the gain from the EnKF would similarly

be expected to make the estimate worse, since it still would result in an increased bias without lowering the variance.

The Steady Kalman filter uses (9) for the model propagation with $\boldsymbol{\eta}_i = 0$

$$\mathbf{x}^f(t_i) = M(\mathbf{x}^a(t_{i-1}), \mathbf{u}(t_{i-1})) \tag{19}$$

Subsequently (14) is used for the analysis, where $\mathbf{k}$ is calculated as a time average from an execution of the EnKF, $p = 1$ and $y^o_{j,i,m} = y^o_{j,i}$. The time consumption of the Steady Kalman filter is of the order of a standard model execution.

The Kalman gain is calculated as a long term average of the gain from an EnKF, where the error has been assumed to lie in the wind velocity. The noise is thus included as a quadratic term in the momentum equations. On average this leads to an overestimation of the values in the Steady Kalman gain, but in periods with strong winds it is underestimated. This is the approximation made by the steady assumption.

3.4.3 Barotropic Approximation

The method described in this section belongs to the group of methods that apply simplified dynamics for calculating the model error covariance and hence the Kalman gain. The idea is that since the water level response to variations in tides and wind forcing is mainly barotropic, its error covariance due to errors in open boundary conditions and wind velocity can be well modelled by a depth averaged barotropic model. The forecast step is composed of an ensemble forecast step using the 2D model and a single forecast of the 3D model according to (19).

The first component of the analysis step consists of the EnKF analysis for the 2D model. The other component is to update the full 3D forecast based on the ensemble statistics from the 2D model. The augmented $AR(1)$ error model part of the state space has the same size and interpretation in both model spaces and hence it can be carried directly over from 2D to 3D. Depth averaged velocity, $\mathbf{V}$ and water level l also have similar interpretation and hence in the corresponding subspace the 2D Kalman gain can be applied directly in 3D.

However, temperature, T, salinity, S, and the three-dimensional velocity, $\mathbf{v}$, are not included in the 2D state space and thus additional assumptions must be imposed. The error covariances between T and S and water levels in the measurement points are all assumed to be zero. This means that the thermodynamic variables are unaffected by the analysis. The velocity, on the other hand needs to be updated. When $\mathbf{V}$ is updated, then for consistency $\mathbf{v}$ must be updated as well, since $\mathbf{V}$ is the depth average of $\mathbf{v}$. A vertical structure, $s(z_k)$, must be chosen under the constraint that its depth average is not zero. Let V_x^a be the updated depth averaged x-velocity component in the 3D model. Let $v_x^f(z_k)$ and $v_x^a(z_k)$ be the forecast and analysis of the x-velocity in the 3D model at depth z_k. Now, the updated full velocity field can be found by solving the following equations for $\mathbf{v}_x^a = (v_x^a(z_1), ..., v_x^a(z_{K_{max}}))$ and the zonal structure scaling parameter θ.

$$V_x^a = (\mathbf{v}_x^a)^T \mathbf{dz} \tag{20}$$

$$\mathbf{v}_x^a = \mathbf{v}_x^f + \mathbf{s}\theta \tag{21}$$

where $\mathbf{v}_x^f = (v_x^f(z_1), ..., v_x^f(z_{K_{max}}))$, $\mathbf{s} = (s(z_1), ..., s(z_{K_{max}}))$ and $\mathbf{dz} = (dz(z_1), ..., dz(z_{K_{max}}))$ is a vector of layer depths.

The solution to (20) and (21) is,

$$\theta = \frac{V_x^a - (\mathbf{v}_x^f)^T \mathbf{dz}}{\mathbf{s}^T \mathbf{dz}} \tag{22}$$

$$\mathbf{v}_x^a = \mathbf{v}_x^f + \mathbf{s}\frac{V_x^a - (\mathbf{v}_x^f)^T \mathbf{dz}}{\mathbf{s}^T \mathbf{dz}} \tag{23}$$

A similar set of equations can be solved for the other velocity component. The vertical velocity is updated by the mass conservation equation. In the present study $\mathbf{s}$ was chosen to be $\mathbf{s} = (1, ..., 1)$. This corresponds to simply moving the entire forecast velocity profile to match the updated depth averaged velocity.

In the light of regularisation, the scheme assumes all elements in $\mathbf{K}_i$ that are used for updating T and S to be zero and elements for updating the velocity components to be related through (23). Again, this certainly may introduce a bias, if the assumption of no correlation to the observed water levels or the vertical interdependence of the velocity errors

break down, but a much lower variance has been obtained. And most importantly, a huge time reduction has been gained in the model error propagation.

(Sørensen et al. 2002) compared the barotropic approximation applied to the Steady Kalman filter to the standard EnKF as well as the Steady Kalman filter with no barotropic approximation in an idealised bay setup. All methods showed similar performance, but the barotropic approximation has the lowest computational requirements both in terms of time and memory demand.

3.4.4 Distance Regularisation

The use of distance regularisation comes down to a trade-off between accepting inaccurate elements in the Kalman gain and introducing spurious or non-dynamical modes in the analysis. The model error covariance is modelled dynamically by assuming errors in the forcing terms. There are many good reasons for doing so, but it may lead to correlations in the error of the state that conflicts with our prior knowledge and hence a regularisation can be performed taking this into account.

The distance regularisation is an ad-hoc procedure for expressing that we do not believe any tide gauge observation should be used for updating state variables that are positioned far away. This is implemented by constructing a vector, with coefficients between 0 and 1, which are a Gaussian function ϕ of their geographical distance, d_m to observation, m, according to,

$$\phi(d_m) = exp(-\frac{d_m^2}{2D^2}) \tag{24}$$

The parameter, D specifies the spatial decorrelation scale. This regularisation can be used in either the EnKF or the Steady Kalman filter ($p = 1$) presented above, by modifying the analysis equation (14) according to

$$\mathbf{x}^a_{j,m}(t_i) = \mathbf{x}^a_{j,m-1}(t_i) + \hat{\mathbf{k}}_{i,m}(y^o_{j,i,m} - \mathbf{h}_{i,m}\mathbf{x}^a_{j,m-1}(t_i)), \quad m = 1, ..., p \tag{25}$$

$$\hat{\mathbf{k}}_{i,m} = \begin{pmatrix} \hat{k}_{i,m}(1) \\ \vdots \\ \hat{k}_{i,m}(n) \end{pmatrix} = \begin{pmatrix} k_{i,m}(1)\phi(d_m) \\ \vdots \\ k_{i,m}(n)\phi(d_m) \end{pmatrix}$$

4 Description of Experiments

The main objective of this study is to demonstrate the hindcast performance of the time efficient barotropic approximation for a Steady Kalman filter in the Water Forecast model. Further, the impact of applying a time varying EnKF with barotropic and distance regularisation is investigated. No comparison is made with an ensemble Kalman filter in the full three-dimensional setting, because it is not operationally feasible. Further, (Sørensen et al. 2002) demonstrated that the dynamic regularisation has similar performance to the full three-dimensional implementation of time varying EnKF in a simple test case.

All experiments span the period: 00:00 January 1 to 00:00 January 29, 2002. The initial state is obtained from the data base of the operational system. The steady gain employed in the study is based on the period 00:00 January 2 to 00:00 January 6, 2002. Figure 8 shows that this period includes a single storm surge event and average winter conditions the rest of the time.

The results will be compared to a reference run, which is obtained from a hindcast execution of the Water Forecast system. All assimilation runs make use of the barotropic approximation and hence do not have higher demands to the computational hardware than the Water Forecast itself and has operational execution times less than 2.5 times that of the reference run. The model runs can be summarized as:

- **Reference run**. Standard 3D Water Forecast model execution with no use of data assimilation.
- **Steady**. 3D Model execution with the Steady Kalman filter. The gain is obtained from the 2D model using the EnKF with temporal smoothing.
- **Steady Dist**. 3D Model execution like Steady, but with distance regularisation used in the 3D environment.
- **EnKF**. 3D Model execution using a time varying gain is obtained from the 2D model employing temporal smooting. Distance regularisation is enabled in the 3D execution.

When adopting the barotropic regularisation in all assimilation runs, it is implied that the central model forecast (19) is employed for the 3D model, while the ensemble forecast (10) and (11) is employed for the calculation of 2D steady and time-varying Kalman gains.

In any assimilation approach it is important first to correct the measurement datum such as to approximately represent model datum in order to allow proper inter-comparison between observed and modelled quantities. Model datum is determined by the open boundary levels and a long term average dynamical balance. In order to assess the model datum, the water levels at all measurement stations were extracted from a one year simulation spanning all of 2002. The time average was calculated for each station, and the corresponding measurement was adjusted to match this average. Note that the model error may have a seasonal dependence, and hence the datum corrected measurements may still contain an offset in January, where the study is performed. The measurements were adjusted to the model datum for both the 2D and the 3D model.

A number of parameters need to be specified in the filtering schemes. The assimilation system is too complex for statistically based parameter estimation and hence first guesses based on experience and theoretical considerations are used. For tide gauges, the measurement representation error is in general dominating over the instrumentation error. The water level readings can be expected to measure the truth projected onto the model space with an accuracy around 0.05 m. Hence, the tidal gauge measurement errors are assumed to have mutually uncorrelated, unbiased Gaussian distributions with a standard deviation of 0.05 m. However, sometimes less trust is put in the measurement in order to constrain the model less. This is the case in M1, Esbjerg, M9, Rønne and M10, Kalix, where the standard deviations were assumed to be 0.15, 0.08 and 0.15, respectively.

The model wind error was assumed to have a temporal correlation scale of 5.7 hours and a spatial correlation scale of 300 km. The standard deviation of the white noise in (8) was assumed to be 0.3 m/s leading to a standard deviation of 3.0 m/s for the wind. The model error in open boundary water levels was assumed to have a temporal correlation scale of 1.7 hours and a spatial correlation scale of 95 km. The standard deviation of the white noise in the boundary error was assumed to be

0.05 m. This leads to a standard deviation of 0.27 m for the water level.

An ensemble size of 100 and a smoothing factor of 0.05 was used for the EnKF runs of the study. Measurements were available every 30 minutes. These were linearly interpolated in time and the model was updated from the interpolated time series every 10 minutes. The distance parameter, D, in the distance regularisation was set to 250 km.

As a measure of the filter performance, the root mean square error ($RMSE$) of water levels, l, calculated over the 28 day simulation period for each measurement and validation station is used,

$$RMSE = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(l_{obs}(t_i) - l_{predicted}(t_i))^2} \tag{26}$$

5 Results & Discussion

The performance of the reference run and the three assimilation runs are summarized in Figures 2 and 3 for measurement and validation stations respectively. For the reference run, $RMSE$ is in the range $0.10\,m$ to $0.15\,m$ for most stations with M1-Esbjerg peaking above $0.20\,m$.

The $RMSE$ can be decomposed into a standard deviation and a bias component. Such an analysis shows that the datum correction method equating one-year averages discussed in Section 4 leaves a variability at monthly time scales with biases in the range -0.12 to 0.07 for the reference run. This might be due to long term variability in meteorological error (the boundaries can not explain such long term variability) or long term error components in the model (biases in annual cycle of the density modeling etc.). However, at present, the bias is accepted as the working conditions, adhering to the bias correction properties of filters using a colored noise implementation, (Sørensen et al. 2004a).

Figure 2 also shows that all assimilation runs significantly reduce the $RMSE$ in measurement points. The remaining error is in fair agreement with the standard deviation of 0.05 m assumed in most stations. Figure 3 shows similar good performance in stations close to measurement

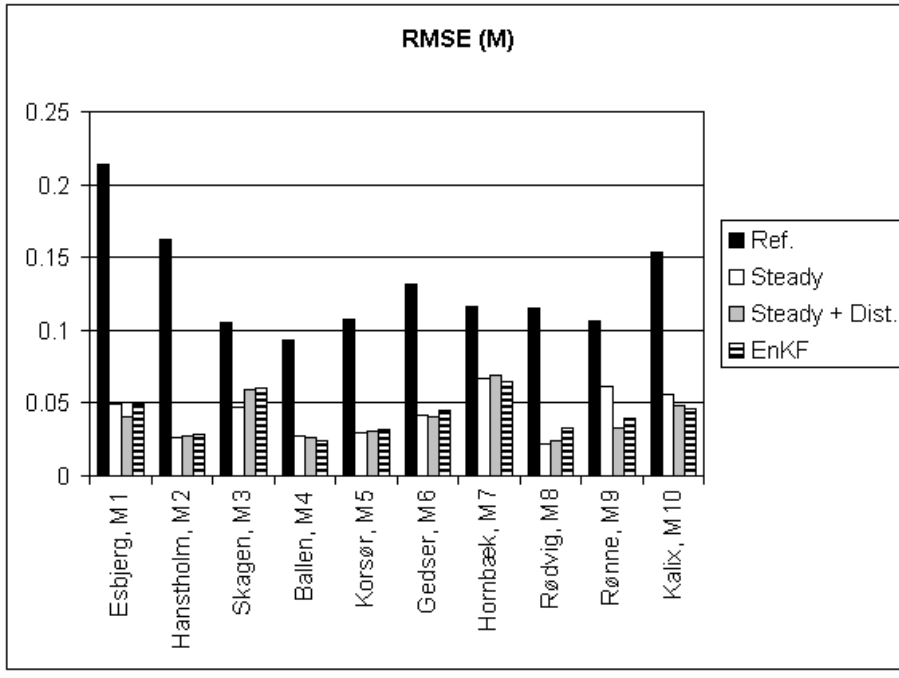

Figure 2: RMSE performance in measurement stations

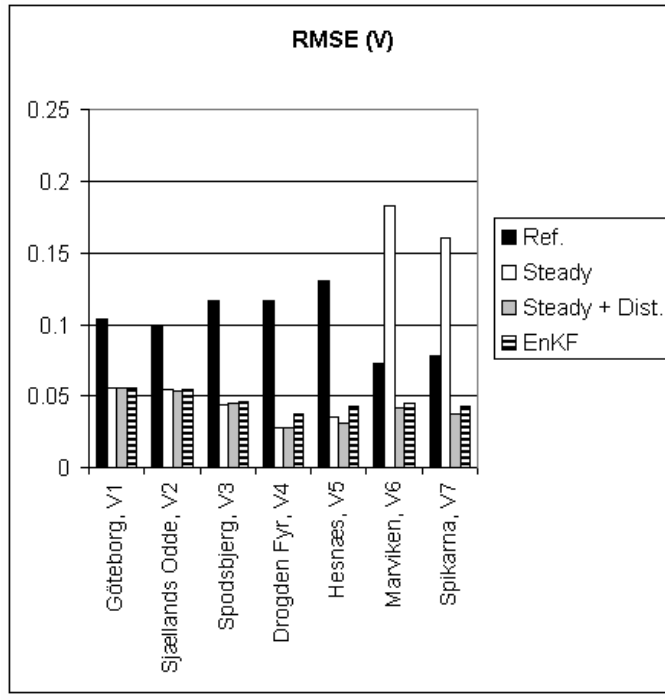

Figure 3: RMSE performance in validation stations

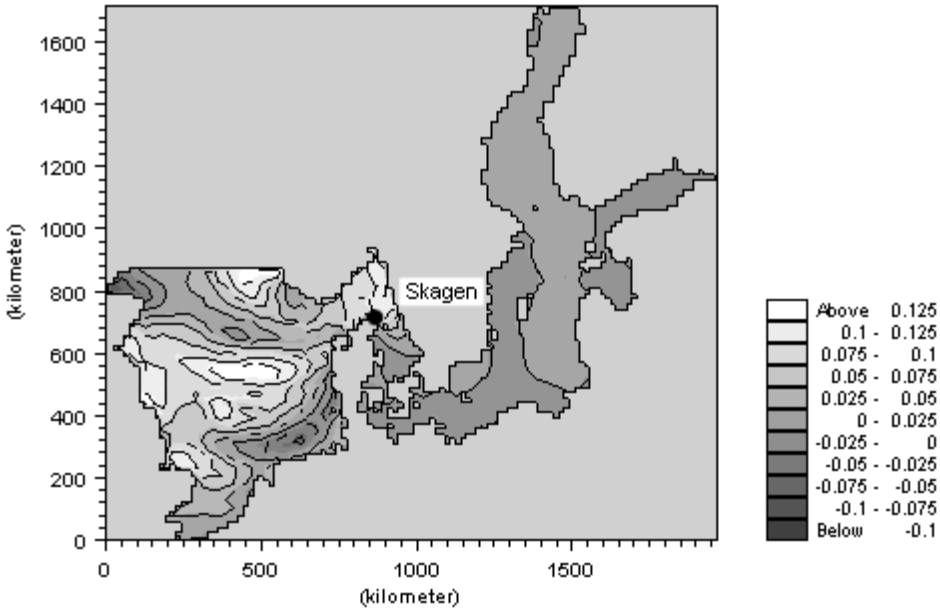

Figure 4: Water level part of the Steady Kalman gain for M3-Skagen

points. However, in the Baltic Sea (V6 and V7) far from measurements, the Steady Kalman filter without distance regularisation significantly degrades the results.

Figures 4 to 7 show examples of the water level part of the Steady Kalman gain for the measurement stations, M3-Skagen and M9-Rønne, with and without the distance regularisation imposed. In the M3-Skagen station the gain is clearly affected by the error assumed in the tidal signal, while the M9-Rønne station is dominated by the wind driven dynamics. In this latter case the gain without distance regularisation shows large corrections in the entire southern part of the Baltic. This can in part be explained by the assumed wind error model, which has spatial correlation scale of 300 km. The distance regularisation of the M3-Skagen station effectively filters the gain structure in the North Sea, which constrasts our prior understanding of the system and our modelling capability.

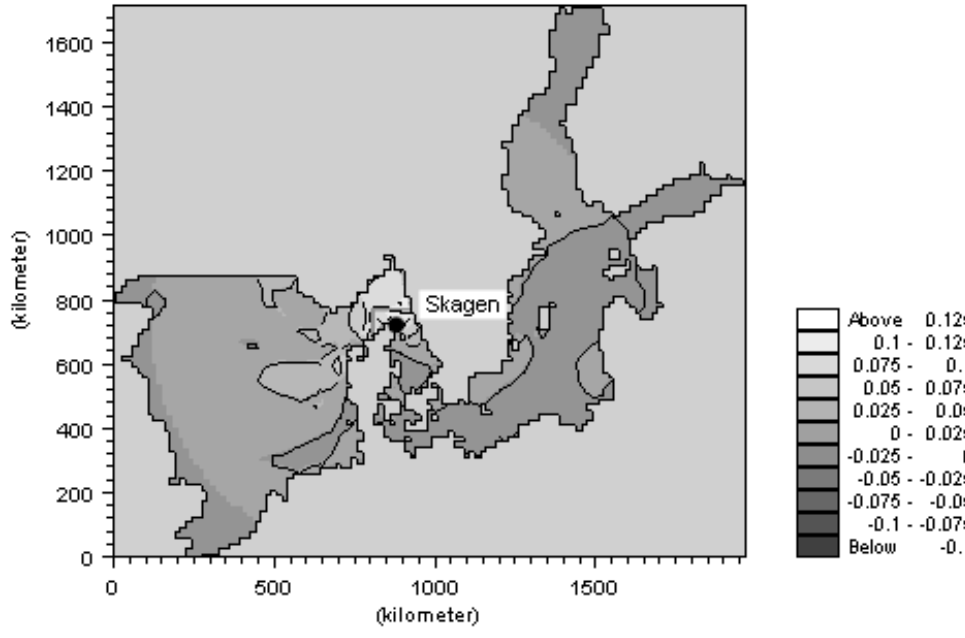

Figure 5: Water level part of the distance regularised Steady Kalman gain for M3-Skagen

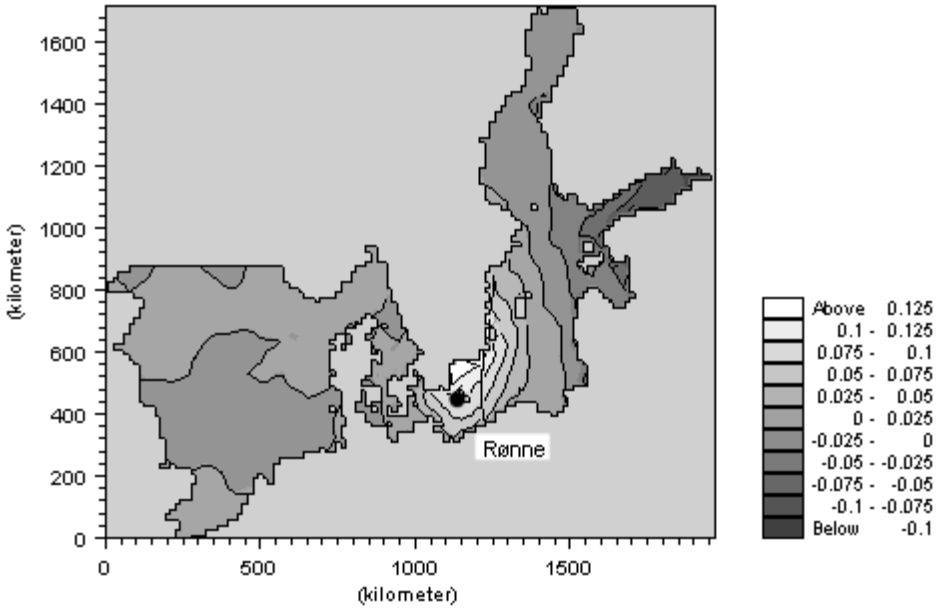

Figure 6: Water level part of the Steady Kalman gain for M9-Rønne

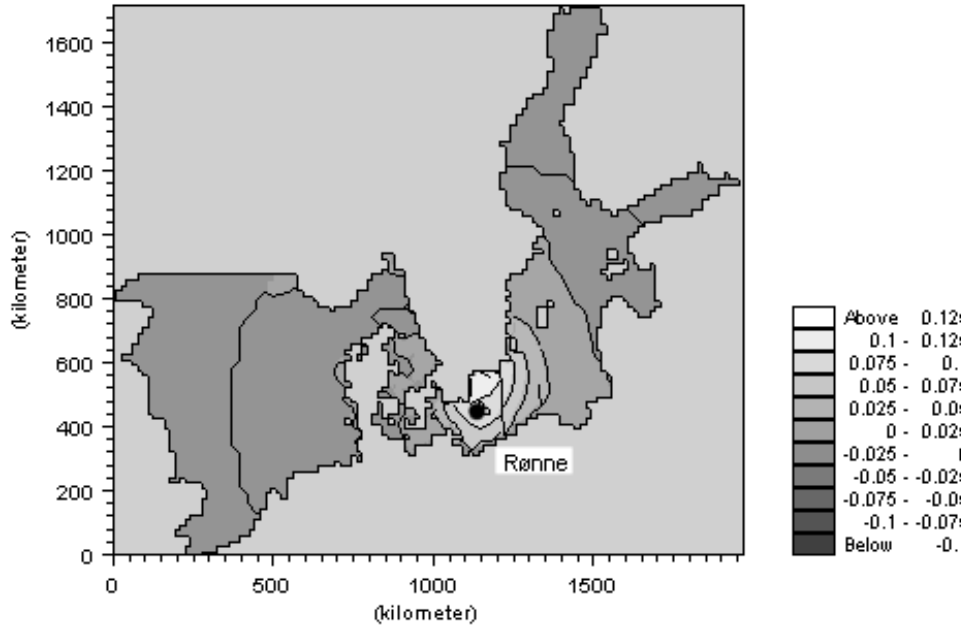

Figure 7: Water level part of the distance regularised Steady Kalman gain for M9-Rønne

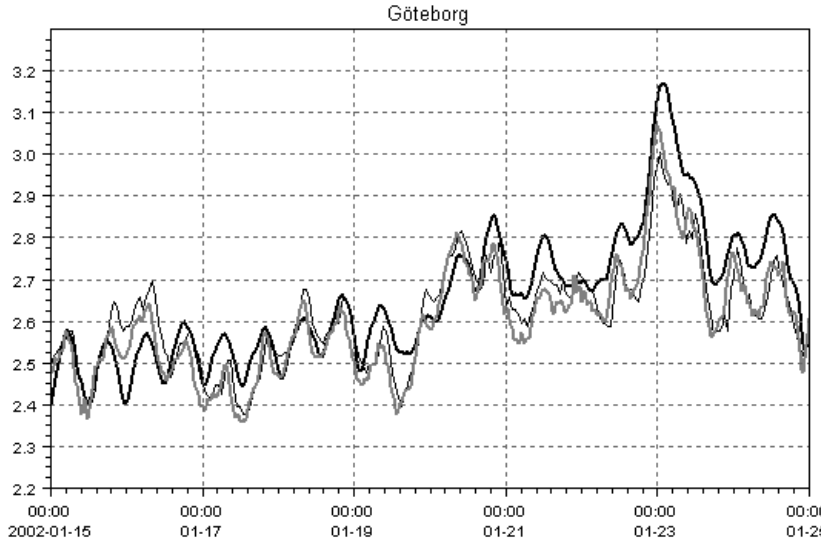

Figure 8: Time series of water level in V1-Göteborg. The thin black line is the measured level. The thick black line shows the reference solution. The thick gray line shows the solution with the barotropic and distance regularised Steady Kalman filter.

The model water level has a large variance far from observations in the Baltic, and hence even a small covariance will provide an impact of the update in this area. The stochastic variability of the gain is filtered out in the Steady approach and thus does not contribute significantly to the gain structure. The distance regularised gain structure dampens the effect of distant correlations by imposing the assumption that such error correlations does not exist despite its prediction by the filter. As is evident in Figure 3 this significantly improves the results in the data sparse Baltic. The distance and barotropic regularised Steady Kalman filter adds significant state estimation skill in all measurement and validation points at a very low computational cost both for the generation of the gains and for execution, enabling use in an operational setting.

A time series plot of measured water level in the validation station, V1-Göteborg, is displayed in Figure 8 along with estimates by the reference run and the distance and barotropic regularised Steady Kalman filter. This plot shows the good performance previously expressed by the sta-

tistical $RMSE$ measure in a visually interpretative form.

The use of a time varying gain from the EnKF with the barotropic approximation and distance regularisation was compared with the successful Steady approach. Figures 2 and 3 show the results as the last horizontally striped bar. The performance is similar to that of the Steady distance regularised scheme. However, its implementation is more demanding on computational time, although the 2D EnKF execution with 100 ensemble members has a similar speed as a single 3D model execution, and hence still can be applied in operational settings.

Figures 9 and 10 show the variance of the EnKF derived gain for the water level portion of the gain for the M9-Rønne station with and without smoothing in the ensemble run. The variability of the non-smoothed Kalman gain shows its maximum values far from the station itself indicating spurious correlations. This also explains the model problems that has been encountered when applying the standard EnKF without distance regularisation. In this case, the analysis may impose a state estimation which is not a likely outcome in the real system despite the fact that the EnKF always produces its analysed ensemble members as linear combinations of the forecast members (an example is pulling water out of a shallow region until a water point is dried out). A spurious correlation can last over a dynamically significant length of time due to the colored noise implementation.

The variability near M9-Rønne is quite similar with and without smoothing in the EnKF gain calculation and in both cases the variability is small compared to the actual size of the Steady gain in Figure 6. This small Kalman gain variability in regions where the update is also largest explains the similar performance of the Steady filter and the EnKF.

6 Conclusions

The water level estimation problem has been discussed and the well known Ensemble Kalman Filter technique presented for solving the problem. In this sequential setting the estimation of the water levels requires an estimation of the elements of the Kalman gain matrix as an interme-

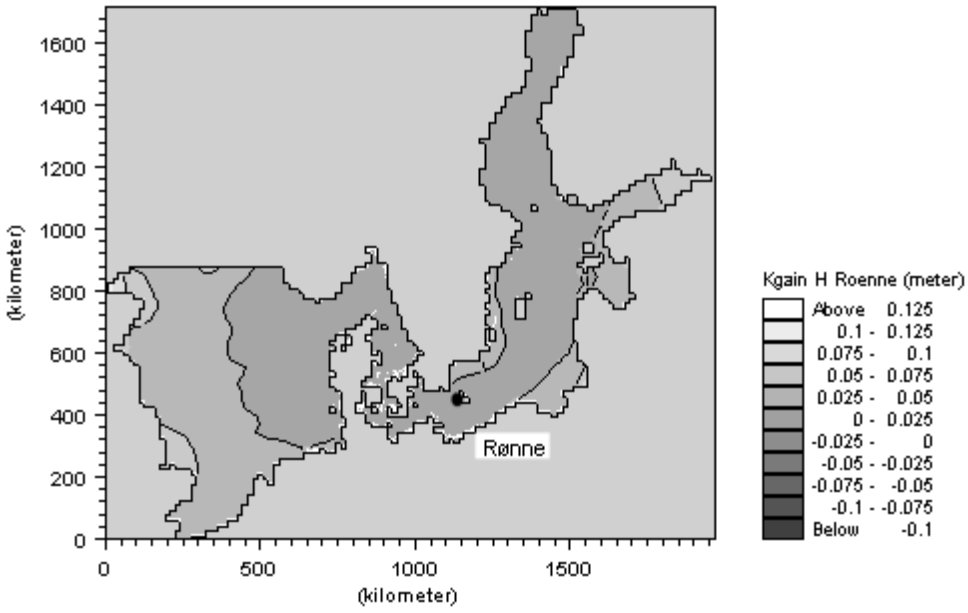

Figure 9: Standard deviation of M9-Rønne water level part of the Kalman gain derived using smoothing.

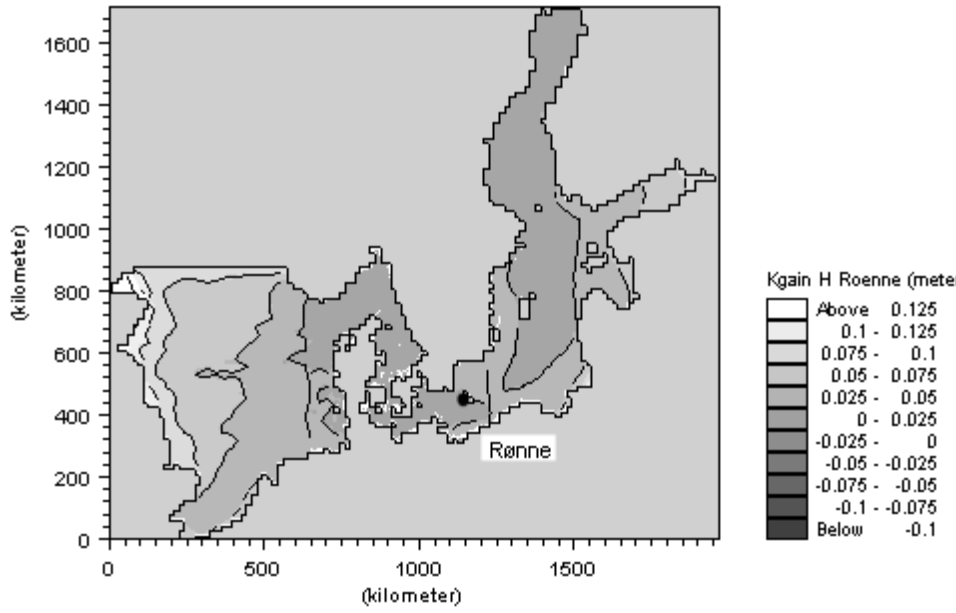

Figure 10: Standard deviation of M9-Rønne water level part of the Kalman gain derived without using smoothing.

diate step, which is important for understanding the behaviour of the scheme. The estimate of the gain elements possess both a bias and a variance, because of inaccurate measurement and model error descriptions and the stochastic variability in the EnKF. This uncertainty is discussed from the viewpoint of regularisation techniques and a Kalman gain smoothing, a Steady Kalman filter, a barotropic approximation and a distance regularisation is discussed in this light.

These techniques are combined and tested for the assimilation of water levels in the Water Forecast operational system. The Steady and the barotropic approximations show the best performance at the lowest cost. The use of distance regularisation has been demonstrated to be important for data sparse regions, while maintaining performance in areas with denser data coverage. The difference in the $RMSE$ of the various filter algorithms is moderate in the Inner Danish Waters and it must be kept in mind that the sensitivity to parameter values is likely on the same scale.

The distance and barotropic regularised Steady Kalman filter has a good estimation skill in all areas of the model. Further, its low computational cost enables easy operational implementation.

Future developments will investigate the use of regularisation techniques for controlling the bias-variance trade-off together with attempting to improve model error description. Also, more work needs to be done on estimating the bias of the measurements when applied in a model datum frame work. Most important in an operational setting is the forecast skill. This will be addressed in a future study.

Acknowledgments

This research was carried out jointly at DHI Water and Environment and the Technical University of Denmark under the Industrial Ph.D. Programme (EF835). Contribution of tide gauge data from the Danish Meteorological Institute, the Royal Danish Administration of Navigation and Hydrography and the Swedish Meteorological and Hydrological Institute is acknowledged.

References

Bahurel, P., Mey, P. D., Provost, C. L. & Traon, P. Y. L. (2002), A godae prototype system with applications - example of the mercator system, *in* CNES, ed., 'En route to GODAE', pp. 137–142.

Bertino, L., Evensen, G. & Wackernagel, H. (2002), 'Combining geostatistics and Kalman filtering for data assimilation in an estuarine system', *Inverse Problems* **18**, 1–23.

Burgers, G., van Leeuwen, P. J. & Evensen, G. (1998), 'Analysis scheme in the ensemble Kalman filter', *Monthly Weather Review* **126**, 1719–1724.

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Chui, C. K. & Chen, G. (1991), *Kalman filter with real-time applications*, Vol. 17 of *Springer Series in information sciences*, Springer-Verlag.

Dee, D. P. (1991), 'Simplification of the Kalman filter for meteorological data assimilation', *Q.J.R. Meteorological Society* **117**, 365–384.

DHI (2001), *MIKE 3 estuarine and coastal hydrodynamics and oceanography*, DHI Water & Environment.

DHI (2002), *MIKE 21 coastal hydraulics and oceanography*, DHI Water & Environment.

Erichsen, A. C. & Rasch, P. S. (2002), Two- and three-dimensional model system predicting the water quality of tomorrow, *in* M. L. Spaulding, ed., 'Proceedings of the seventh international conference on estuarine and coastal modeling', American Society of Civil Engineers, American Society of Civil Engineers, pp. 165–184.

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

Evensen, G. (2003), The ensemble Kalman filter: Theoretical formulation and practical implementation, In print: Ocean Dynamics.

Fukumori, I. & Malanotte-Rizzoli, P. (1995), 'An approximate Kalman filter for ocean data assimilation; an example with an idealised Gulf Stream model', *J. Geoph. Res.* **100**(C4), 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. Geoph. Res.* **104**(C11), 25647–25665.

Gerritsen, H., de Vries, H. & Philippart, M. (1995), The Dutch continental shelf model, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 19, pp. 425–467.

Hamill, T. M., Whitaker, J. S. & Snyder, C. (2001), 'Distance-dependent filtering of background-error covariance estimates in an ensemble kalman filter', *Mon. Wea. Rev.* pp. 2776–2790.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The elements of statistical learning : data mining, inference, and prediction*, Springer-Verlag.

Heemink, A. W. & Kloosterhuis, H. (1990), 'Data assimilation for non-linear tidal models', *Int. J. Num. Methods in Fluids* **11**, 1097–1112.

Houtekamer, P. L. & Mitchell, H. L. (1998), 'Data assimilation using an ensemble kalman filter technique', *Monthly Weather Review* **126**, 796–811.

Ide, K., Courtier, P., Ghil, M. & Lorenc, A. C. (1997), 'Unified notation for data assimilation: Operational, sequential and variational', *Journal of Meteorological Society of Japan* **75**(1B), 181–189.

Jazwinski, A. H. (1970), *Stochastic Processes and filtering theory*, Vol. 64 of *Mathematics in Science and Engineering*, Academic Press.

Madsen, H. & Cañizares, R. (1999), 'Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling', *International Journal of Numerical Methods in Fluids* **31**(6), 961–981.

Pham, D. T., Verron, J. & Gourdeau, L. (1998), 'A singular evolutive Kalman filter for data assimilation in oceanography', *C. R. Acad. Sci. Paris* **326**, 255–260.

Pham, D. T., Verron, J. & Roubaud, M. C. (1997), 'Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography', *Journal of Marine Systems* **16**, 323–340.

Pinardi, N. & Woods, J., eds (2002), *Ocean forecasting. Conceptual basis and applications*, Springer-Verlag.

Pinardi, N., Auclair, F., Cesarini, C., Demirov, E., Umani, S. F., Giani, M., Montanari, G., Oddo, P., Tonani, M. & Zavatarelli, M. (2002), Towards marine environmental predictions in the Mediterranean Sea coastal areas: A monitoring approach, *in* N. Pinardi & J. Woods, eds, 'Ocean forecasting. Conceptual basis and applications', Springer-Verlag, chapter 17, pp. 339–376.

Rogers, E., Black, T., Ferrier, B., Lin, Y., Parrish, D. & DiMego, G. (2001), 'Changes to the ncep meso eta analysis and forecast system: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3dvar analysis', *NWS Technical Procedures Bulletin*.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2002), Towards an operational data assimilation system for a three-dimensional hydrodynamic model, *in* R. A. Falconer, B. Lin, E. L. Harris & C. A. M. E. Wilson, eds, 'Proceedings of the fifth International Conference on Hydroinformatics', International Water Association, IWA Publishing.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2003b), Parameter sensitivity of three Kalman filter schemes for the assimilation of tide gauge data in coastal and shelf sea models, Submitted: Ocean Modelling.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2004a), Data assimilation in hydrodynamic modelling: On the treatment of nonlinearity and bias, Accepted: Stochastic Environmental Research and Risk Assessment.

Verlaan, M. & Heemink, A. W. (1997), 'Tidal flow forecasting using reduced rank square root filters', *Stochastic Hydrology and Hydraulics* **11**, 349–368.

Vested, H. J., Nielsen, J. W., Jensen, H. R. & Kristensen, K. B. (1995), Skill assessment of an operational hydrodynamic forecast system for

the North Sea and Danish belts, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 17, pp. 373–396.

Paper F

Water level forecast skill of a hybrid steady Kalman filter - error correction scheme

F

Submitted to *Ocean Dynamics*. Revised version.

Water level forecast skill of a hybrid steady Kalman filter - error correction scheme

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

Abstract

The forecast skill of a Steady Kalman filter alone and combined with a simple autoregressive error correction model is demonstrated in the North Sea, Baltic Sea system. Any practical Kalman filter estimate will in general provide a sub-optimal state estimate, if for no other reasons, then because of approximate model and measurement error descriptions. The state estimation problem is reviewed with a close consideration of the interconnection between representation error and error modeling. This leads to an interpretation of innovation auto-covariance and suggests a hybrid data ASsimiation - Error correction Prediction scheme (ASEP) for the forecast. In this scheme a modified system equation is assumed with an autoregressive model predicting the innovation, which is assimilated with a Steady Kalman filter. Compared to a hydrodynamic forecast the ASEP scheme gives an improved prediction skill for 10-11 hours on average. Recently, a distance regularised Kalman gain has been shown to significantly improve filtering performance in areas with sparse data coverage. The forecast skill of a distance regularised Steady Kalman filter is tested. This gives an improvement for several days, but the ASEP scheme no longer gives a significant further improvement.

1 Introduction

A large part of the world's population lives close to the ocean and is affected by the coastal environment. Storm surges, toxic algae and oil spill pollution are a few examples of events with harmful consequences for people and whose prediction can help take proper action to minimize human harm and economic expense. Therefore, forecasting of key parameters in the coastal ocean has been on the agenda for decades and in

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

many countries warning systems are being operated for selected key parameters, e.g. (Vested et al. 1995), (Gerritsen et al. 1995) and (Erichsen & Rasch 2002).

More recently, with continual improvements of the modeling systems encompassing forecasts of an increasing number of physical and biogeochemical parameters, the operational services have matured to a degree, where more and more needs of governmental agencies and private enterprises are met at an affordable price. For most forecast products, the forecast skill is of prime importance. Since numerical modelling is only slowly improving and has fundamental limitations, the present on-going development also focuses on the on-line assimilation of available data.

The basic idea in most assimilation systems with a forecasting objective, is to provide the best possible estimate of the ocean state at the time of forecast. Such an approach was implemented by (Heemink 1986) in a storm surge model for the Dutch coast. He used a Steady Kalman filter and showed an improved skill relative to a standard forecast model at both a three and six hour forecast horizon. (Vested et al. 1995) and (Gerritsen et al. 1995) also investigated the forecast skill in the Southern North Sea. They similarly found that Kalman filter based initialisation improves the forecast skill at short time scales. However, at longer time scales the skill deteriorates for a while before converging to that of the standard forecast model. (Cañizares et al. 2001) applied the Steady Kalman filter for assimilating tide gauge data in the North Sea, Baltic Sea system, where they showed a good filtering performance in areas of fairly dense data coverage. However, far from observations the filtering skill was degraded. This problem was treated in (Sørensen, Madsen & Madsen 2004) and a regularisation technique (distance regularisation) introduced to solve it. The effect on forecast skill of applying distance regularisation is investigated in this paper.

A different approach to provide improved forecast skill is to use error correction schemes. These provide an off-line forecast of a vector of numerical model residuals in points of observation by constructing data driven models that relate the residuals to the residuals at previous time steps, state variables predicted by the numerical model and forcing variables. (Babovic, Cañizares, Jensen & Klinting 2001) demonstrated a successful application of a neural network based error correction scheme

for the prediction of current speeds in Øresund strait between Sweden and Denmark. Another error correction approach was taken by (Babovic & Keijzer 2001), who used a Gaussian process model to improve forecast skill of water level in a model of the Venice lagoon. All of these methods, however, only provide local predictions. Any spatial distribution of the predicted residuals must rely on assumptions about spatial correlations, (Babovic & Fuhrman 2002).

When using data assimilation to improve the initial field at time of forecast, the improved prediction skill is limited to a time horizon, at which initial conditions are washed out. Error correction methods do not have this constraint, but they only provide forecasts in measurement points. Spatial distribution of the predicted correction rely on simple statistical assumptions rather than system dynamics. Note that error correction is a post processing procedure. It does not influence the forecast made by the numerical model.

In this study a hybrid data ASsimilation Error correction Prediction (ASEP) scheme is suggested, where the error correction methodology is applied as an integrated part of a sequential model state update using the Kalman filter. It is shown that data assimilation methods have a colored innovation sequence in practice. Error correction is applied to forcast the one-step-ahead errors of the numerical model in measurement points and subsequently distributing this forecast in space according to the Kalman filter update scheme.

Section 2 reviews the system description of the ocean and the classical augmentation approach for model construction in the case of colored system noise. Also included in Section 2 is a discussion of the choice of state space definition when using numerical models for the state propagation and its relation to system error and representation error included in the measurement equation. Section 3 investigates the nature of the innovation in a filtering scheme for correct and incorrect model and measurement error descriptions, while the adapted filters (Ensemble and Steady Kalman filters) are briefly outlined in Section 4. Section 5 describes the hybrid approach taken for data assimilation, while Section 6 describes the setup of the North Sea, Baltic Sea forecasting system and the assimilation schemes. The results of the numerical experiments are presented and discussed in Section 7 and Section 8 concludes the work.

2 System description and state estimation

2.1 System

The physical system under consideration is the ocean in general. Our focus is on coastal and continental shelf areas with the North Sea, Baltic Sea and interconnecting waters providing our test grounds. The properties of this system are composed of a wide range of interacting, dynamically evolving components usually classified as hydrodynamic, water quality, wave and sediment variables. In the theoretical treatment we retain the full system in a rather general discussion about system models. However, in the numerical experiments, we restrict our attention to the mass and momentum distribution and hence the pressure and velocity components.

Our inability to perfectly observe and predict the ocean naturally leads us to put error bars on our predictions, and more elaborately to make the state variables and system evolution stochastic. However, reality only provides a single realisation of the actual state trajectory. Thus, the perception of the actual ocean state having a probability distribution is just a theoretical device - at least in the realm of Newtonian physics. It is however, a very useful device, which allows us to express our imposed imperfections and thus estimate the ocean state based on several sources of information.

Let $\mathbf{x}^t(t_{i-1}) \in \mathbf{R}^n$ be the true ocean state defined in discrete space and time t_{i-1}. Let M be a one time step ahead model propagator of this state. Now assume that the error imposed in this ocean state propagation is an unbiased white noise process, $\boldsymbol{\eta}_i$, with covariance $\mathbf{Q}_i$. This gives the following *system equation*:

$$\mathbf{x}^t(t_i) = M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_{i-1})) + \boldsymbol{\eta}_i \tag{1}$$

where $\mathbf{u}(t_i)$ is the external forcing.

Let $\mathbf{y}_i^o \in \mathbf{R}^p$ be a an observation of a subspace of the state. This measurement is assumed to be unbiased and its error, $\boldsymbol{\epsilon}_i$, to be white noise with covariance $\mathbf{R}_i$. This gives the following *measurement equation*:

$$\mathbf{y}_i^o = \mathbf{H_i}\mathbf{x}^t(t_i) + \boldsymbol{\epsilon}_i \tag{2}$$

where $\mathbf{H_i}$ is a linear operator, which projects the state space onto the measurement space.

2.2 State estimation

Now assume the ideal case that (1) and (2) constitute the correct stochastic descriptions of the state propagation and the measurements. If we know the probability density function (pdf) of the state at some initial time t_0, then the nonlinear model operator, M, and the error distribution, $\boldsymbol{\eta}_i$, can be used to predict the a priori pdf of the state at a subsequent time t_i. Denote the prior expectation of this state $\mathbf{x}^f(t_i)$ and the prior covariance $\mathbf{P}^f(t_i)$. If a measurement becomes available at time t_i, the prior information embedded in the measurement can be combined with the model prior to give an improved a posteriori estimate. The Minimal Variance (MV) estimator is given by,

$$\mathbf{x}^{MV}(t_i) = E\{\mathbf{x}^t(t_i)|\mathbf{y}_i^o\} \tag{3}$$

The prior model state estimate does not need to be Gaussian for (3) to hold. Nor does the measurement error. For a nonlinear model and hence a non-Gaussian distribution of $\mathbf{x}^t$, the estimate of $\mathbf{x}^{MV}$ does not need to be linear in $\mathbf{y}^o$. However, assuming this and solving the minimisation problem under this constraint, gives the Best Linear Unbiased Estimate (BLUE), $\mathbf{x}^a$, as

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}^f(t_i)) \tag{4}$$

where

$$\mathbf{K}_i = \mathbf{P}_i^f(t_i)\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}_i^f(t_i)\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \tag{5}$$

The error covariance of the updated state estimate, $\mathbf{x}^a$, is given by,

$$\mathbf{P}^a(t_i) = \mathbf{P}^f(t_i) - \mathbf{K}_i\mathbf{H}_i\mathbf{P}^f(t_i) \tag{6}$$

2.3 Model and state space

For now it is assumed that some perfect technique is used for propagating the pdf of the state and estimating $\mathbf{P}_i^f$ for a given M. A numerical model,

M_N, is chosen for the prediction of the dynamical evolution of the model state, $\mathbf{x}_N$. This enables us to use our basic theoretical knowledge about the ocean dynamics for constructing the model. However, in practically all real ocean models, the model operator contains several error sources, which can not be described as white noise processes like assumed in (1). Thus, the starting point when a numerical model is used for the state propagation, is the presence of colored noise in the system equation.

When using a numerical model as the corner stone of the predictor, the model state space is simultaneously adopted as the carrier of the system state space. We have to make a choice about the exact space in which we wish to estimate the state and define our error processes according to that. (Fukumori et al. 1999) provide a nice discussion of representation error and emphasises that the error description in the measurement equation is a measurement constraint error rather than a measurement error. Their basic assumption is that the filtered reality, which we adopt as the state space, is defined as the model range both in terms of projection onto spatial and temporal averages as well as dynamical projections.

Adapting the model state space allows a decomposition of the true continuous ocean state into this model solution space, Σ_M, and the corresponding null-space, $\Sigma_M^{\perp}$. It means that the limitation of the model definition is embedded in the state space. They do not give a rigorous definition of the meaning of such a dynamical projection, but emphasises that model errors are based on 'truncation and/or approximation in physics'. Such a state representation makes sense. We do not want to reinitialise the model with a state, whose further propagation is unsupported. In this way we suppress signals, which are not solutions to the model operator, at least in the ideal filtering case.

However, the definition of the dynamical model projection is a matter of definition. No matter what dynamical projection is chosen within the same spatial and temporal averages, then as long as the errors in the resulting system equation are well described, it provides a stochastic model within the chosen projection of the system. A good definition leaves small and uncorrelated model errors. To obtain this, the processes deliberately left out in the model definition should be represented as representation error, while feedback from these processes back into the

modelled processes should be described as model errors together with unwanted approximations such as numerical errors as well as uncertain parameters and forcing.

Measurements are usually averages over different length and time scales than the numerical model. Also, signals from all physical processes that remain after the averaging operations employed in the measurement acquisition are present. Such incompatibilities are usually treated in geostatistics through the use of so called change of support models, (Bertino et al. 2002). However, a more rudimentary approach is to increase measurement error by considering the representation error. Despite the fact that it is the model, which is unable to represent the true system state, then having adapted the model state space projection, the portion of the observation that lies in the null-space must be accredited to measurement representation noise. With this noise description, the measurement constraint error can easily be colored and will generally depend on the state. The measurement matrix that projects the model space onto the measurement space is denoted, $\mathbf{H}_M$.

It is important to note in this respect that most validation procedures, including the one employed in this paper, look for the best fit with measurements and hence the model is validated in measurement space, while the estimation was performed in model space. Hence, if spatio-temporal correlations exist among the null-space projections of the measurements, an overly fit state estimate may yield better prediction skill than the ideal filter. An overly fit model attempts to pull the state into the model null-space. Particularly, close to observations, this may provide a state estimate closer to reality despite the erroneous error assumptions. However, the subsequent propagation of the estimated state is no longer guaranteed to be well described by the model operator. This may either be due to neglected dynamics or due to numerical schemes, where the valid regimes are exceeded by the state estimate.

With colored noise components, the BLUE estimator no longer provides the optimal state estimate. The classical way to solve this problem in time series modeling is to augment the model state with variables that descrobe the autocorrelation and leave an error that fulfills the white noise assumption in (1). So far, we have simply assumed some white model error with known mean and covariance. However, analysis of the

model equations suggests that this is not the case, e.g. dominant model errors can be expected to come from a colored error in the forcings or slowly varying parameters. Errors derived from the numerical methods will generally be state dependent and hence colored, e.g. unwanted numerical damping.

This knowledge about the error sources can be used to develop a better model of the system. This is done by augmenting the state space and model operator with elements, $\mathbf{x}_\eta$, and functions, M_η, needed to model the error of the numerical model, which we assume to know the properties of. The error models need to be expressed as Markov processes having white noise errors themselves. Time dependencies with lags greater than one can be included through an additional state augmentation. We introduce an extended numerical model operator, M_M, which includes the numerical model operator, M_N, and the forcing of the model state $\mathbf{x}_N$ with the errors, $\mathbf{x}_\eta$.

Similarly the state can be augmented with the colored measurement errors, $\mathbf{x}_\epsilon$, the time evolution of which is described by M_ϵ. All together this leads to the following system equation,

$$\mathbf{x}^t(t_i) = \begin{pmatrix} \mathbf{x}_N^t(t_i) \\ \mathbf{x}_\eta^t(t_i) \\ \mathbf{x}_\epsilon^t(t_i) \end{pmatrix} \tag{7}$$

$$= \begin{pmatrix} M_M(\mathbf{x}_N^t(t_{i-1}), \mathbf{u}(t_{i-1}), \mathbf{x}_\eta^t(t_{i-1})) + \mathbf{G}_M \boldsymbol{\eta}_{M,i-1} \\ M_\eta(\mathbf{x}_\eta^t(t_{i-1})) \quad + \quad \boldsymbol{\eta}_{\eta,i-1} \\ M_\epsilon(\mathbf{x}_\epsilon^t(t_{i-1})) \quad + \quad \boldsymbol{\eta}_{\epsilon,i-1} \end{pmatrix} \tag{8}$$

$$= M(\mathbf{x}^t(t_{i-1}), \mathbf{u}(t_{i-1})) + \mathbf{G}\boldsymbol{\eta}_{i-1} \tag{9}$$

where $\mathbf{G}_M$ is a transformation matrix of the state space, in which the white noise model is defined, onto the numerical model state space and

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_M & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{10}$$

is the general noise-to-model-space transformation operator. The generalised model error $\boldsymbol{\eta}_i$ is composed of the white numerical model error, $\boldsymbol{\eta}_{M,i}$, the error in the colored model error description, $\boldsymbol{\eta}_{\eta,i}$, and the error

in the colored measurement and representation error description, $\boldsymbol{\eta}_{\epsilon,i}$. The measurement equation now reads,

$$\mathbf{y}_i^o = \mathbf{H}_{i,M}\mathbf{x}_N^t(t_i) + \mathbf{x}_\epsilon^t(t_i) + \boldsymbol{\epsilon}_i = \mathbf{H}\mathbf{x}^t(t_i) + \boldsymbol{\epsilon}_i \tag{11}$$

The two equations (7) and (11) outline the standard framework for handling colored model and measurement errors, by augmentation of the space to include the colored components.

3 Innovation autocorrelation

In this section the whiteness of the innovation sequence for perfect error assumptions and a linear model, $\mathbf{M}$, will be demonstrated leading to an expression for the auto-covariance when approximated error assumptions are used. The information in the one-step ahead forecast error of the observation is contained in the innovation, $\mathbf{d}_i = \mathbf{y}_i^o - \mathbf{H}\mathbf{x}^f(t_i)$. Let $\mathbf{x}^f(t_i) = \mathbf{x}^t(t_i) + \boldsymbol{\delta}_i^f$ and consider the expected lag-one auto-covariance of the innovation sequence,

$$\begin{aligned}
& Cov(\mathbf{d}(t_{i+1}), \mathbf{d}(t_i)) \\
= \;& Cov(\mathbf{y}_{i+1}^o - \mathbf{H}\mathbf{x}^f(t_{i+1}), \mathbf{y}_i^o - \mathbf{H}\mathbf{x}^f(t_i)) \\
= \;& Cov(\boldsymbol{\epsilon}_{i+1} - \mathbf{H}\boldsymbol{\delta}_{i+1}^f, \boldsymbol{\epsilon}_i - \mathbf{H}\boldsymbol{\delta}_i^f) \\
= \;& Cov(\mathbf{HM}\left[\boldsymbol{\delta}_i^f + \mathbf{K}_i(\boldsymbol{\epsilon}_i - \mathbf{H}\boldsymbol{\delta}_i^f)\right] + \mathbf{H}\boldsymbol{\eta}_i, \mathbf{H}\boldsymbol{\delta}_i^f - \boldsymbol{\epsilon}_i) \\
= \;& \mathbf{HM}(\mathbf{I} - \mathbf{K}_i\mathbf{H})\mathbf{P}_i^{f,t}\mathbf{H}^T - \mathbf{HMK}_i\mathbf{R}_i^t \\
= \;& \mathbf{HM}\Big[\mathbf{P}_i^{f,t}\mathbf{H}^T\left[\mathbf{H}\mathbf{P}_i^{f,t}\mathbf{H}^T + \mathbf{R}_i^t\right]^{-1} - \\
& \hat{\mathbf{P}}_i^f\mathbf{H}^T\left[\mathbf{H}\hat{\mathbf{P}}_i^f\mathbf{H}^T + \hat{\mathbf{R}}_i^t\right]^{-1}\Big]\left[\mathbf{H}\mathbf{P}_i^{f,t}\mathbf{H}^T + \mathbf{R}_i^t\right]
\end{aligned} \tag{12}$$

$\mathbf{P}_i^{f,t}$ and $\mathbf{R}_i^t$ are the true model and measurement error covariance matrices, while $\hat{\mathbf{P}}_i^f$ and $\hat{\mathbf{R}}_i$ are those assumed and derived from the assimilation scheme. In order to simplify the expression, we have used the assumption that $\boldsymbol{\epsilon}_{i+1}$ and $\boldsymbol{\eta}_{i+1}$ are uncorrelated with previous measurement and model errors, or at least that these terms are much smaller than the other terms.

Equation (12) shows that the autocorrelation is only zero in the ideal case with correct error assumptions and error propagation, i.e. $\hat{\mathbf{P}}_i^f = \mathbf{P}_i^{f,t}$

and $\hat{\mathbf{R}}_i^t = \mathbf{R}_i^t$. In all other cases the auto-covariance of the innovation sequence is given by the propagation of the difference between the estimated and the true Kalman gain matrices projected onto the measurement space and multiplied by the true covariance of the innovation. This means that in most real case applications of approximate Kalman filters using suboptimal error modeling, the innovation sequence has a predictable component. Further, it can be used to test for correct error assumptions.

The innovation sequence is potentially a result of the composite model-measurement error signal. If the colored components are due to pure model error then the model error description should be improved. The assumption then is that the model structure is capable of producing results that overlap the measurements, but due to improper error description the analysis fails to do so. In this respect a suggestion for error model improvement is to use the innovation sequence to construct a model, M_η, of the model error, $\mathbf{x}_\eta$ in measurement space. This can then be distributed to the entire model space by imposing proper covariance assumptions. A simple first guess is to use the Kalman gain for this purpose, all though this assumes that the previously neglected error component has the same covariance as the modelled error component. This corresponds to M_M in (8) with no measurement error model to be given by,

$$M_M(\mathbf{x}_N^t(t_i), \mathbf{u}(t_i), \mathbf{x}_\eta^t(t_i)) = M_N(\mathbf{x}_N^t(t_i), \mathbf{u}(t_i)) + \mathbf{K}_i \mathbf{x}_\eta^t(t_i) \tag{13}$$

Once the model is changed in this way, the resulting innovation series will change as well and there is no guarantee that the process leads to an improved model error description and hence a more efficient estimate, but it is an obvious way to attempt to implement the information contained in the original innovation sequence to improve the system description.

If the colored components are due to pure measurement constraint error, i.e. lies in the model null space, then the measurement error model should be improved. Remember, that such measurement error models also encompass models of the structural model errors. Thus, if the model imposed measurement error dominates, a subsequent forecast of the measurement error actually predicts the true physical system and not an instrumental error. This estimate, which is defined in measurement space can be used to estimate the state of the null space through imposing additional error covariance assumption, hence leaving to an improved

estimate of the true state composed of the sum of the estimates in Σ_M and $\Sigma_M^{\perp}$.

4 Assimilation scheme

So far we have merely assumed perfect propagation of the first and second moments of the model probability distribution. Obviously, this is not possible and much work has been put into deriving proper propagation schemes for data assimilation. In this study we use the Steady Kalman filter approximation with the steady Kalman gain calculated off-line by the Ensemble Kalman Filter, (Evensen 1994). Its description and details about its implementation in the present setup can be found in (Sørensen et al. 2004), which also describes the distance regularisation based on ideas in (Houtekamer & Mitchell 1998). The adoption of a suboptimal scheme and particularly the steady approximation imposes a number of errors in the Kalman gain estimation, and hence a colored component in the innovation sequence according to (12). However, the misspecification of model and measurement errors most likely imposes equally large errors and their impact has been much less studied. Hence, the theoretical discussion has focused on the latter, while acknowledging the importance of proper state and error covariance propagation.

5 Hybrid prediction scheme

It is well known that model residuals can be modelled very well by data driven models, e.g. (Babovic et al. 2001), and thereby improve the predictive power. This is usually done off-line by superimposing error forecasts to the model simulation and allows the use of information contained in measurements to be used for the prediction of observed variables. In a data assimilation system this additional information is used to update the entire system state, hence leaving no information in the innovation. However, as discussed above, this is not the case for suboptimal filtering and a reminiscent signal is left in the innovation sequence. This can be modelled by a suitable data driven model, M_d, which subsequently can be used to forecast the innovation. This approach, where the presence of

imperfection in the assimilation approach is acknowledged and modelled, is referred to as the hybrid data ASsimilation Error correction Prediction scheme (ASEP).

The filter analysis step (4) can be rewritten,

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i\hat{\mathbf{d}}_i + \mathbf{K}_i(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}^f(t_i) - \hat{\mathbf{d}}_i) \tag{14}$$

where $\hat{\mathbf{d}}_i$ is the predictable part of the innovation in an imperfect assimilation scheme. Hence, the analysis is split into a term assimilating only the unpredictable projection of the measurement and an update based on the predictable part of the innovation. The closest approximation to this in a forecast setting is to maintain the first two terms instead of just the first term. Hence, in forecast the model propagation is changed to,

$$\mathbf{x}^f(t_{i+1}) = M_N(\mathbf{x}^f(t_i), \mathbf{u}(t_i)) + \mathbf{K}_i\hat{\mathbf{d}}_i \tag{15}$$

$$\hat{\mathbf{d}}_{i+1} = M_d(\hat{\mathbf{d}}_i) \tag{16}$$

Following (15), $\mathbf{K}_i$, is the Kalman gain used in the estimation of the error model. However, this is based on assimilating measurements and not predicted innovations, which will be more uncertain. Hence, alternatively the $\mathbf{K}_i$ to be used in the forecast setting can be estimated by considering the innovation estimate as an additional source of information and modeling its error statistics. This would result in smaller corrections. The approach adopted in the present study is to use a steady Kalman gain both for hindcast and for forecast.

In the present study a very simple data driven model was used for the innovation prediction, namely a univariate AR(1) model. Much more elaborate models such as artificial neural networks, local linear modeling or genetic programming can be derived, but for the present purpose of demonstrating the feasibility of the approach, the simple autoregressive model was chosen.

$$\hat{\mathbf{d}}_{i+1} = M_d(\hat{\mathbf{d}}_i) = \mathbf{A}\hat{\mathbf{d}}_i \tag{17}$$

where $\mathbf{A} = diag(\alpha_1, ..., \alpha_p)$ and $|\alpha_i| < 1, \quad i = 1, ..., p.$

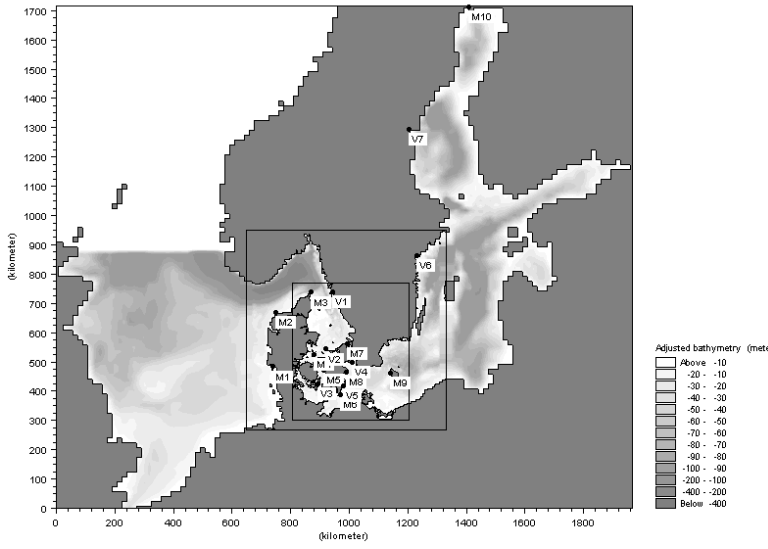

Figure 1: Bathymetry and available tide gauge stations, including 10 measurement stations (M1-M10) and 7 validation stations (V1-V7).

6 Design of experiments

The area under consideration is the North Sea, Baltic Sea and interconnecting waters. We restrict our attention to the barotropic hydrodynamics and hence employ the depth averaged numerical model, MIKE 21, developed at DHI Water & Environment, DHI (2001). The area and bathymetry is shown in Figure 1 with the available tidal gauge measurement points indicated. The gauges were divided into measurement stations (M) used in the assimilation and validation stations (V), which were only used for performance assessment. The spatial resolution varies from 9 to 1/3 nautical miles through a two-way dynamic nesting technique. The temporal resolution is 2.5 minutes and measurements are available every 30 minutes. The measurements are linearly interpolated and assimilated every 10 minutes, i.e. every fourth model time step.

The period of January 2002 was used in the study. A steady Kalman gain was estimated as an average of the gain calculated in an execution of the EnKF in a three day period from 1 January to 4 January. All measurements were adjusted to have the same average as a standard

model prediction in January 2002 to diminish datum problems.

The experiments were designed to test the forecasting performance of five prediction schemes:

- A standard model execution
- A Steady Kalman filter
- A Steady Kalman filter until time of forecast and the ASEP thereafter
- A Steady Kalman filter using distance regularisation
- A Steady Kalman filter using distance regularisation until time of forecast and the ASEP thereafter

Twenty forecasts were performed with one day intervals. Each model run included one day of hindcast and a four day forecast. Hindcast wind fields were used for forecast.

In the assimilation schemes, the model error was assumed to derive solely from errors in the wind velocity and open boundary water level forcing terms. These errors were assumed to be colored with temporal correlation scales of 5.7 and 1.7 hours, respectively, and to have spatial correlation scales of 300 and 95 km. All measurements constraint errors were assumed to have a standard deviation of 0.05 m. See (Sørensen et al. 2003b) for a more detailed description of the Kalman filter settings and their effect on hindcast performance.The spatial decorrelation scale of the distance regularisation was set to 250 km.

The steady Kalman filter was run for the full month of January both with and without distance regularisation and the resulting innovation sequences were used to construct their respective AR(1) model in each measurement point. The autoregressive parameters was set to the lag one autocorrelation of the corresponding innovation series. The parameters are shown in Table 1. The innovation sequences look quite different in different locations indicating a heterogeneous model error structure. Figure 2 and 3 show two examples of the innovation time series without using regularisation in M4-Ballen and M6-Gedser, respectively. As

expected from their relatively high and low auto-correlation values, M4-Ballen has a rather smooth variability, while the M6-Gedser innovation sequence is more high frequent. M6-Gedser also seems to have a time varying variance. Clearly, in an elaborate implementation of the ASEP approach, a more complex error correction model must be developed.

Station	No regularisation	Distance regularisation
Esbjerg (M1)	0.95	0.96
Hanstholm (M2)	0.84	0.88
Skagen (M3)	0.96	0.97
Ballen (M4)	0.95	0.96
Korsør (M5)	0.98	0.99
Gedser (M6)	0.87	0.85
Hornbæk (M7)	0.94	0.92
Rødvig (M8)	0.93	0.94
Rønne (M9)	0.96	0.93
Kalix (M10)	0.98	0.998

Table 1: Estimated innovation lag one auto-correlations in January 2002 for a standard Steady Kalman filter (no regularisation) and imposing distance regularisation

The performance of the schemes were assessed as root mean square errors $RMSE$ of the $A = 20$ forecasts for each forecast horizon, t_i, and tidal gauge station, s,

$$RMSE(t_i, s) = \sqrt{\frac{1}{A}\sum_{a=1}^{A}(y_i^o(s) - H(s)x^a(t_i))^2} \tag{18}$$

For each station the time horizon, over which the assimilation approaches improve the forecast according to the $RMSE$-measure, was derived by visual inspection. Further, bulk performance measures were constructed as averages of measurement and validation stations.

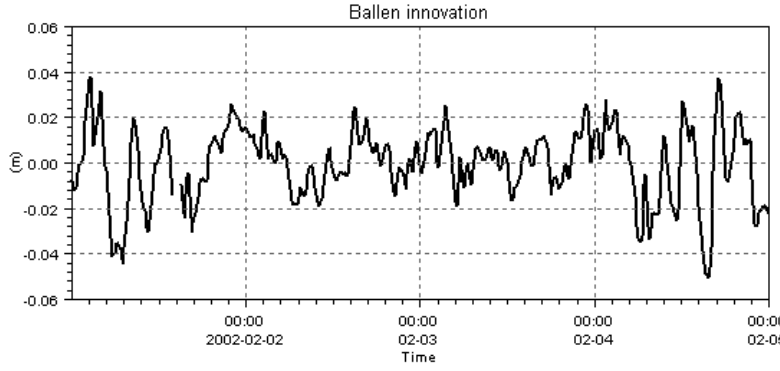

Figure 2: Time series of the innovation in M4-Ballen for the first four days of February 2002

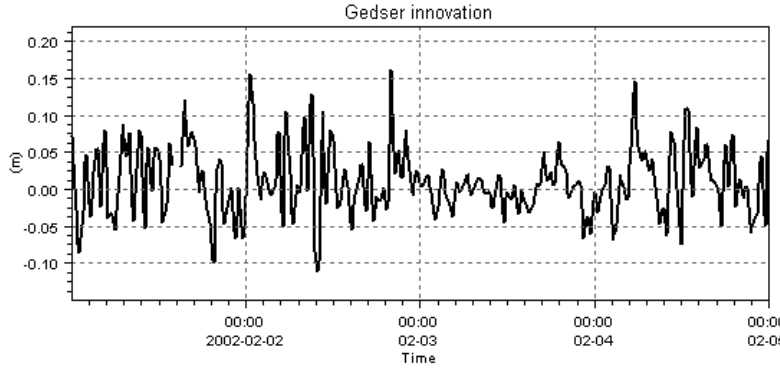

Figure 3: Time series of the innovation in M6-Gedser for the first four days of February 2002

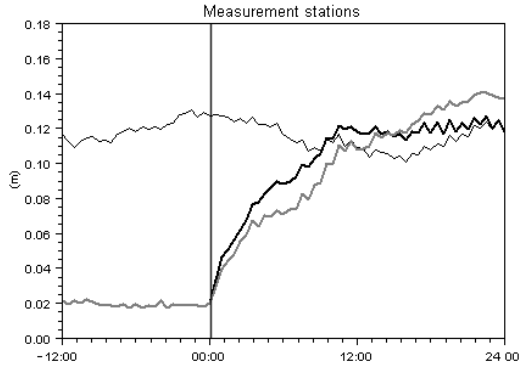

Figure 4: Aggregated RMSE of the reference run (thin black), the Steady run (thick black) and the ASEP run (thick grey) in all measurement points. The horizontal axis is time (h) relative to time of forecast.

7 Results and discussion

In the case of the non-regularised steady Kalman filter, the bulk RMSE statistics are shown in Figures 4 and 5 for measurement and validation stations respectively. The overall picture is that while the data assimilation clearly improves the state estimate in hindcast, this improved skill on average only lasts 6-8 hours without the hybrid scheme and 10-11 hours with it. After this period of improved predictive skill a period follows with degraded water level predictions.

In order to understand this behaviour, the spatial distribution of improved predictive horizon is shown in Figure 6 with values and relative ASEP improvement listed in Table 2. Large differences exist and it is clearly evident that M1-Esbjerg in the Southern North Sea and stations south of the Danish straits have the worst performance, while stations in the Skagerak and Kattegat improve prediction 18-36 hours and even up to 55 hours for the hybrid scheme.

The areas with short prediction horizon lies close to areas of sparse data coverage. The Steady Kalman filter has been shown to provide bad state estimates in these areas, (Sørensen et al. 2004), and when data no longer is available to constrain the solution, the errors in these areas

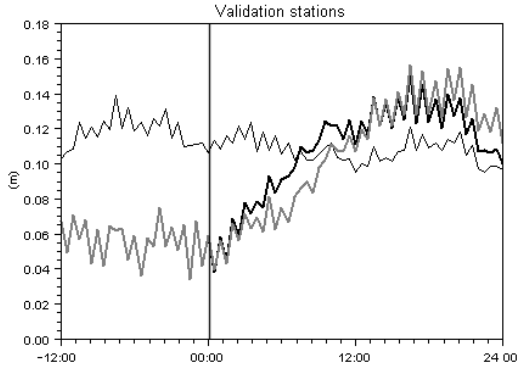

Figure 5: Aggregated RMSE of the reference run (thin black), the Steady run (thick black) and the ASEP run (thick grey) in all validation points. The horizontal axis is time (h) relative to time of forecast.

Station	Steady (hours)	ASEP (hours)	Improvement (%)
Esbjerg (M1)	2	3	52
Hanstholm (M2)	18	18	0
Skagen (M3)	28	31	11
Göteborg (V1)	32	55	72
Sjællands Odde (V2)	30	34	13
Ballen (M4)	28	37	32
Korsør (M5)	15	17	13
Spodsbjerg (V3)	14	16	14
Gedser (M6)	7	10	43
Hornbæk (M7)	31	34	10
Drogden (V4)	5	8	60
Rødvig (M8)	5	8	60
Hesnæs (V5)	5	9	80
Rønne (M9)	5	8	60
Marviken (V6)	0	0	0
Spikarna (V7)	0	0	0
Kalix (M10)	8	13	63
Average	14	18	29

Table 2: Forecast horizon (hours) in all tide gauge stations

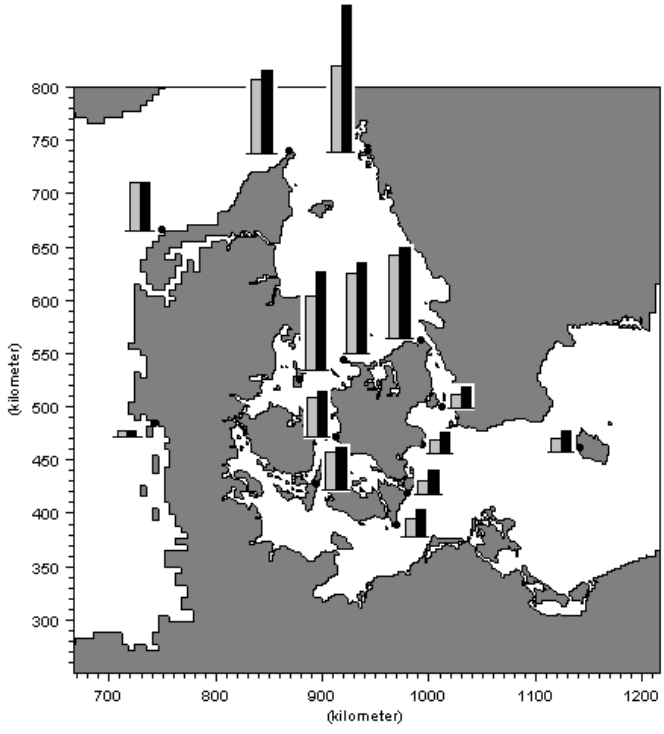

Figure 6: Schematic view of the spatial distribution of the forecast horizons for the Steady initialisation (grey) and the ASEP approach (black) listed in Table 2.

flush into the observed regions. This dynamic propagation of the errors is damped, reflected or redirected before it reaches Skagerak and Kattegat where prediction performance hence is good. An example of water level hindcast and forecast for the validation station, V2-Sjællands Odde is shown in Figure 7.

The average $RMSE$ of Skagerak and Kattegat stations is shown in Figure 8. An important observation is that there is no subsequently degraded performance in these stations, where good performance is observed. Thus, when an assimilation scheme is used, which minimizes the estimation errors everywhere in hindcast, then the assimilation based forecast is also to be trusted even on longer horizons.

It turns out that the ASEP improvement over the Steady lasts the longest

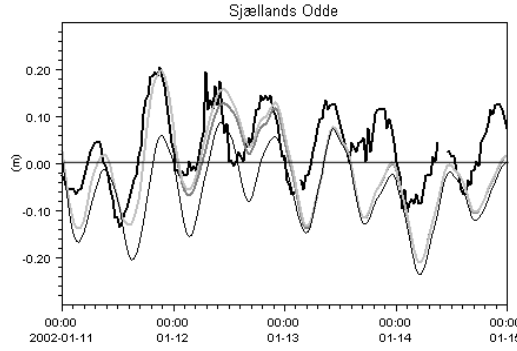

Figure 7: Example of improved forecast skill of the reference run (thin black), the Steady run (thick grey) and the ASEP run (thick light grey) in the validation station V2- Sjællands Odde in the Southern Kattegat. Observations are in thick black. Time of forecast is at 00:00 2002-01-12.

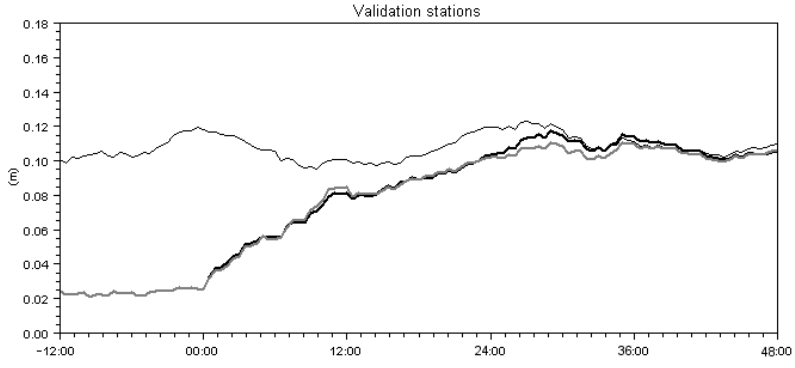

Figure 8: Aggregated RMSE in of the reference run (thin black), the Steady run (thick black) and the ASEP run (thick grey) the Skagerak and Kattegat stations. The horizontal axis is time (h) relative to time of forecast.

in stations, where the Steady performs rather well by itself. It is the same gain that is used to assimilate the observations and the forecast innovation and hence the predicted innovation in the ASEP is distributed according to a gain, which provides good forecast and thus a longer lasting effect. On the other hand, the best relative improvement is found in the stations with modest prediction skill because of the influence by the erroneous state estimate in the Baltic. This implies larger actual corrections relative to error imposed in every forecast. In other words, the error correction method corrects more, when there is more error to correct.

A closer analysis reveals that the innovation sequence has both periodic and higher order autoregressive components. Thus, an AR(1) model will give phase errors and an improved error correction model should be applied. However, this present study is meant as a first step in the direction of applying hybrid data assimilation - error correction schemes in a forecasting scenario. In actual implementations a decision must also be made about the purpose of the model. For example, if storm surge prediction is the ultimate goal of the scheme, then the error correction component of the scheme should be trained under such conditions.

Many future challenges and developments remain. A first step would be to include a stochastic error in the error correction model during the forecast. The predicted innovation does not have the same error as the real innovation. Hence it should be modelled and a corresponding gain calculated for merging the predicted innovation and the numerical model estimate of the state.

Alternatively, the presence of the innovation time correlation suggests an improved system model. Such an improvement could be inspired by the innovation sequence itself, hence trying to use the hybrid model presently employed in the forecast in hindcast as well. This changes the entire state space and system description, but may nevertheless provide an improved model. Such an approach was pursued for bias estimation in (Dee & da Silva 1998).

Now we consider the results of the experiments using distance regularisation. Figures 9 and 10 show aggregated $RMSE$ measures for measurement and validation stations respectively. The most striking feature

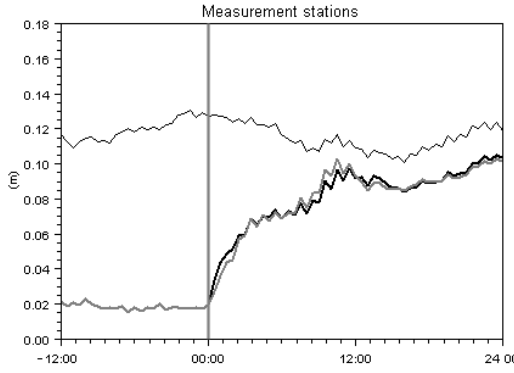

Figure 9: Aggregated RMSE of the reference run (thin black), the Steady distance regularised run (thick black) and the corresonding ASEP run (thick grey) in all measurement points. The horizontal axis is time (h) relative to time of forecast.

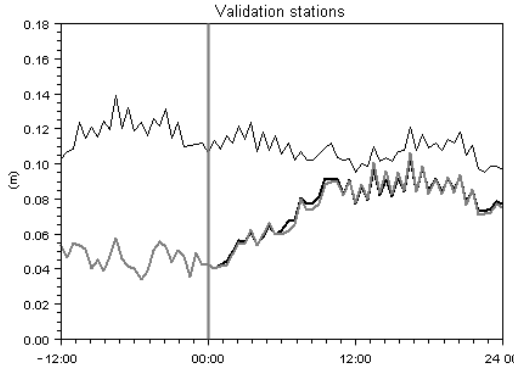

Figure 10: Aggregated RMSE of the reference run (thin black), the Steady distance regularised run (thick black) and the corresonding ASEP run (thick grey) in all validation points. The horizontal axis is time (h) relative to time of forecast.

here is the impressive effect of distance regularisation on the water level forecast skill, which now exceeds 24 hours. The $RMSE$ measure levels out to that of the reference run after 2-3 days and no deterioration is observed at any forecast horizon for any station. In this case the ASEP scheme does not improve results further. A smaller error is left to correct by the hybrid scheme and hence no significant improvement is observed. The distance regularisation makes an improved global state estimate and hence no erroneous signals are set free to propagate in the domain at time of forecast.

8 Conclusion

This paper has highlighted the general problem of properly estimating model and measurement errors in sequential data assimilation schemes. Since the imposed error model typically contains an error in real data assimilation applications in coastal seas, then any scheme will be suboptimal for this reason alone. We have shown that sub-optimality leaves predictability in the innovation series and suggested a combined data assimilation - error correction scheme for prediction in this case. The scheme contains the predictable part of the combined forecast-analysis in the assimilation scheme.

The performance of the scheme was investigated in an operational model of the North Sea, Baltic Sea and interconnecting waters. Forecast initialisation by the Steady Kalman filter gave an improved prediction for a period of 5-32 hours in the Region stretching from Skagerak to the Western Baltic. The hybrid scheme improves this improved forecast horizon to 8 to 55 hours.

An interesting observation is the lack of a subsequent degradation of the performance on longer time scales in the Skagerak and the Kattegat. In areas, which are dynamically influenced by regions of sparse data coverage and poor state estimates, the longer time scale prediction skill of the Steady Kalman filter initialisation is hampered by the propagation of these errors. This longer time scale degradation is slightly more pronounced when using the Hybrid scheme.

The use of distance regularisation significantly improves the forecast skill and is to be encouraged for operational forecasting purposes.

References

Babovic, V. & Fuhrman, D. R. (2002), 'Data assimilation of local model error forecasts in a deterministic model', *International Journal for Numerical Methods in Fluids* **39**, 887–918.

Babovic, V. & Keijzer, M. (2001), A Gaussian process model applied to prediction of the water levels in Venice Lagoon, *in* 'Proceedings of the XXIX Congress of International Association for Hydraulic Research', International Association for Hydraulic Research.

Babovic, V., Cañizares, R., Jensen, H. R. & Klinting, A. (2001), 'Neural networks as routine for error updating of numerical models', *Journal of Hydraulic Engineering* **127**, 181–193.

Bertino, L., Evensen, G. & Wackernagel, H. (2002), 'Combining geostatistics and Kalman filtering for data assimilation in an estuarine system', *Inverse Problems* **18**, 1–23.

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

Dee, D. P. & da Silva, A. M. (1998), 'Data assimilation in the presence of forecast bias', *Q.J.R. Meteorological Society* **124**, 269–296.

Erichsen, A. C. & Rasch, P. S. (2002), Two- and three-dimensional model system predicting the water quality of tomorrow, *in* M. L. Spaulding, ed., 'Proceedings of the seventh international conference on estuarine and coastal modeling', American Society of Civil Engineers, American Society of Civil Engineers, pp. 165–184.

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

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. Geoph. Res.* **104**(C11), 25647–25665.

Gerritsen, H., de Vries, H. & Philippart, M. (1995), The Dutch continental shelf model, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 19, pp. 425–467.

Heemink, A. W. (1986), Storm surge prediction using Kalman filtering, PhD thesis, Twente University of Technology.

Houtekamer, P. L. & Mitchell, H. L. (1998), 'Data assimilation using an ensemble kalman filter technique', *Monthly Weather Review* **126**, 796–811.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2003b), Parameter sensitivity of three Kalman filter schemes for the assimilation of tide gauge data in coastal and shelf sea models, Submitted: Ocean Modelling.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2004), 'Efficient kalman filter techniques for the assimilation of tide gauge data in three-dimensional modeling of the north sea and baltic sea system', *J. Geophys. Res.* **109**(C3), C03017.

Vested, H. J., Nielsen, J. W., Jensen, H. R. & Kristensen, K. B. (1995), Skill assessment of an operational hydrodynamic forecast system for the North Sea and Danish belts, *in* D. R. Lynch & A. M. Davies, eds, 'Quantitative Skill Assessment for Coastal Ocean Models', American Geophys. Union, chapter 17, pp. 373–396.

Paper G

Parameter estimation in a hydrodynamic model of the North Sea and Baltic Sea

G

Originally published as

Jacob V. Tornfeldt Sørensen, Henrik Madsen, Henrik Madsen. Parameter estimation in a hydrodynamic model of the North Sea and Baltic Sea. Technical Report, DHI Water & Environment, DK-2970 Hørsholm, Denmark, 2004.

Parameter estimation in a hydrodynamic model of the North Sea and Baltic Sea

Jacob V. Tornfeldt Sørensen[1,2], Henrik Madsen[1], and Henrik Madsen[2]

Abstract

This report briefly reviews the initial work done on parameter estimation and calibration in DHI's hydrodynamic models of the marine environment. In a setup of MIKE 21 in the Water Forecast area covering the North Sea, Baltic Sea and Inner Danish Waters, an optimisation of bed friction maps and wind friction parameters, which minimises the standard deviation of model residuals is performed. This is done in an approximate weak constraint formulation, where the model assimilates tide gauge data and in a strong constraint formulation, where a standard deterministic model is employed. Generally, a rather poor performance is obtained in a validation run. However, the approximate weak constraint formulation using data assimilation in the parameter estimation gives a better performance, because this attempts to estimate the parameters, while acknowledging model errors. Part of the generally poor performance can be ascribed to the fact that the calibration was only done over a five day period. Also, bathymetry variations were not included as controls in the optimisation. The weak constraint formulation further imposed its own errors. One of the tide gauge stations that enter the optimisation objective lies in an area that is erroneously updated by the Kalman filter, thus severely affecting the optimisation.

1 Introduction

Throughout the advance of numerical models in ocean sciences, the problem of determining the model parameter values has been an integrated part of every application. To this date, the most widespread technique in ocean and coastal sea applications is a manual trial-and error ap-

[1]DHI Water & Environment, DK-2970 Hørsholm, Denmark
[2]Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

proach. This allows the modeller to employ a complex physical understanding based on theoretical considerations and experience. However, every modeller will have a different approach and criteria for good performance. The objective of the modeller is typically rather vaguely to make the model trajectory come close to observations. Further, in many cases the nonlinear nature of the parameter estimation problem makes the calibration task very difficult.

Another approach is to cast the problem in an optimisation framework. This mathematical inverse problem was presented and discussed rather elaborately by (Evensen et al. 1998). Here, a weak constraint formulation is suggested, which leaves a well posed inverse problem. If the formulation is relaxed to be strong constraint, the problem may be ill-posed and leave non-unique and noisy parameter estimates. In the strong constraint formulation, a cost function of model residuals is minimised with respect to a set of chosen control parameters. In the weak constraint formulation, penalty functions are added to the cost function for the deviation of the parameters from an initial parameter guess as well as deviation from the model propagator, the boundary condition and the initial conditions, respectively.

The size of the state vector in oceanographic studies using numerical models is large (10^5-10^7). Hence, the most tractable approach of solving this problem is in a variational setting, employing a gradient based numerical optimisation algorithm together with an implementation of the adjoint equations. In the case of multiple local optima, the adjoint solution should be combined with a global optimisation algorithm in order to increase chances of finding the global optimum. Generally this is not pursued and a local quasi-Newton or conjugate gradient method is used, (Heemink et al. 2002).

The combination of local, gradient based optimisation algorithms and solving the adjoint equations has been employed in a number of previous studies for two- and three-dimensional numerical hydrodynamic models. (Lardner, Al-Rabeh & Gunay 1993) used this approach to calibrate bottom friction coefficients and water depth in a barotropic tidal model of the Arabian Gulf over a 29-day period. In order to stabilise the method, they introduced a penalty for strong parameter variations. The calibration was consistent in a 100-day validation run. In (ten Brummelhuis,

Heemink & van den Boogaard 1993) an extended approach was applied to calibrate a numerical model over a one-day period while applying a Steady Kalman filter for assimilating tide gauge data assuming uncertain boundary conditions. This provided smoothed estimates of boundary elevation corrections, which were applied iteratively in an optimisation of water depth and bed friction. (Heemink et al. 2002), applied the adjoint approach for estimating tidal constituents, depth, bed friction and viscosity parameters in a three-dimensional model solving the shallow water equations on the European Continental Shelf. They demonstrated significant improvements in areas of prime interest.

The success of the gradient based studies above proves that gradients are well defined. However, the assumption that a local optimisation technique is sufficient, has mainly been confirmed by the relatively realistic values of the obtained parameter estimates. The non-linear nature of the problem may give a cost function with several local minima. In all cases above the cost functions are based on square error measures in the observational space. Other alternatives, such as phase error measures may be more relevant in certain cases. A major drawback for adjoint techniques is the need for an adjoint model operator. Despite the development of automatic adjoint compilers, (Giering & Kaminski 1998), most models do not have adjoint codes developed and no automatically generated adjoint codes are known to have been compiled for shallow water flow computations in coastal areas.

An alternative parameter estimation approach was taken by (Heemink 1986), who demonstrated the use of a Kalman filter to estimate model parameters in a one-dimensional along-coast setting. With the advance of sequential estimation techniques capable of handling non-linearities, (Evensen 1994) and (Verlaan & Heemink 1997), a similar approach should be applicable even in two and three dimensional models. However, no such application has been found in literature for coastal ocean modelling despite its appeal.

In hydrological literature the parameter estimation problem has been more extensively studied. Here, non-linearities often renders gradient based local optimisation insufficient and hence, focus has been largely on developing global methods that are not gradient based. Among the developments are the Shuffled Complex Evolution (SCE) algorithm, which

has been proven successful in handling strong non-linearities and multiple equilibria, (Madsen 2003). However, compared to efficient Quasi-Newton or conjugate gradient schemes the SCE has a slow convergence. The is an expression of the general trade-off between efficiency and effectiveness. Further, the method is not gradient based and hence, it can not take advantage of the adjoint technique.

In any optimisation scheme, the inverse problem itself is defined by the cost function and hence the calibration data and model parameterisation. In a particular case, the cost function should reflect the application of the model. Thus, a great deal of knowledge is nevertheless required to select these parameters of the problem in a manual way. This must be done such as to allow identifiability of the the parameters and typically results in a spatial grouping of a field of parameters. This is in fact an ad hoc regularisation of the problem and more elaborate regularisation techniques might be explored in this context.

The objective of the present study is to explore the application of parameter estimation techniques, which do not require the development of an adjoint solver, to a depth averaged hydrodynamic model. Using a simplex optimisation algorithm, bed and wind friction coefficients of a depth averaged two-dimensional model are estimated for optimal performance in the inner Danish waters. The parameter estimation is performed in both a strong constraint and a quasi-weak constraint context. The results are validated in both settings as well.

2 Parameter estimation framework

The definition of and the solution to a parameter estimation problem consists of a number of elements. First the model of the system under consideration must be specified. The model can either be deterministic or stochastic in nature. Next, the notion of model performance must be quantified in a cost function expressing the fit or misfit of a model estimate with observations. This includes selecting a calibration period and data as well as functional relationships. It is well known that hydrodynamic models are typically over parameterised and thus not identifiable by the observational data. Therefore, a subset of the parameter space

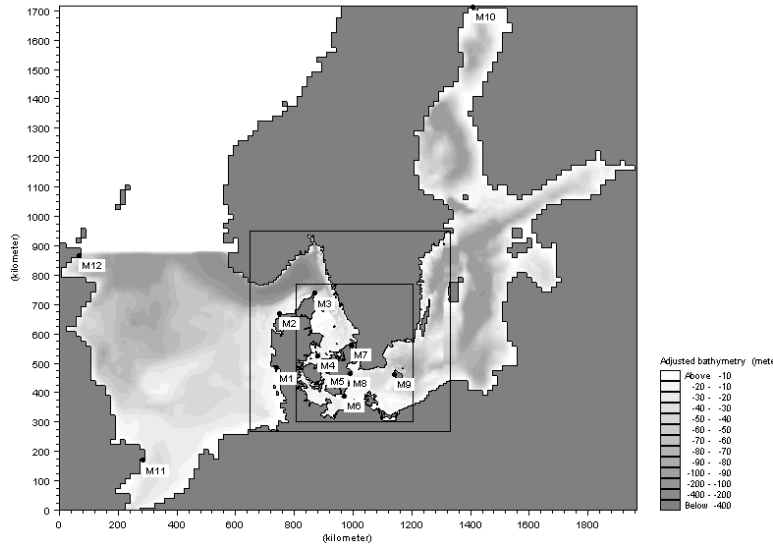

Figure 1: Bathymetry and tide gauge stations, including 12 measurement stations (M1-M12)

needs to be selected as control parameters. Finally, having defined the inverse problem, an optimisation algorithm must be chosen to provide a solution.

2.1 Model setup and data

The numerical model employed in this study, is the widely used MIKE 21 developed at DHI Water and Environment, (DHI 2002). The model solves the depth averaged mass and momentum conservation equations on a staggered grid using finite differences. Thus, it simulates the evolution of the water level, η_i, and horizontal fluxes, $V_{x,i}$ and $V_{y,i}$, discretised in time (t_i indexed i) and space. It is setup in a region covering the North Sea - Baltic Sea system with bathymetry shown in Figure 1.

The model is forced by winds and surface pressure from the Vejr2 weather service, (Rogers et al. 2001), as well as tidal surface elevations at the Northern and Southern boundaries in the North Sea. Initial water levels

Station	Cost function	Assimilation
Esbjerg (M1)	+	+
Hanstholm (M2)	+	+
Skagen (M3)	+	+
Ballen (M4)	+	+
Korsør (M5)	+	+
Gedser (M6)	+	+
Hornbæk (M7)	+	+
Rødvig (M8)	+	+
Rønne (M9)	+	+
Kalix (M10)	-	+
Lowestoft (M11)	+	-
Wick (M12)	+	-

Table 1: Tide gauge stations and their usage in the optimisation

and fluxes are obtained from the Water Forecast Service, (Erichsen & Rasch 2002). The model is calibrated and validated against a number of tide gauge stations, providing water level measurements, $\eta_{obs,i}$. These are listed in Table 1 and their position plotted in Figure 1. The stations have the densest distribution in the inner Danish water, which is the area of prime focus in this study. However, we maintain a few North Sea stations in an attempt to avoid parameters that give unrealistic results there. No attempt has been made here to scale the contribution from the North Sea stations with the much larger tidal range in these positions. The Baltic is not similarly constrained. Ideally, an independent set of tide gauge stations should be used for the validation procedure. This is not done at the present stage of development.

2.2 Steady Kalman filtering

In the deterministic model above, the model equations themselves are considered perfect and the optimal parameters will be sought for under that assumption. This will provide parameter estimates which are fit to compensate for other model errors as demonstrated by (ten Brummelhuis et al. 1993). Alternatively, the model simulation can attempt to correct model errors unrelated to parameter values by using a Kalman filter,

which takes these errors into account.

In a previous study of (Cañizares et al. 2001) the Steady Kalman filter was tested in a similar model set-up of MIKE 21 in the North Sea, Baltic Sea area for assimilating tidal gauge data. In the present study, uncertain boundary and wind forcing is assumed to be the sole sources of model error. Table 1 lists the tide gauges used for assimilation. The Steady Kalman Filter is described in detail in (Sørensen et al. 2004). The distance regularisation proposed in that paper is not used in this study and hence regions far from measurements might have a deteriorated performance.

If the approximations of the Steady Kalman Filter are ignored for the sake of simplicity, then the filtered model solution is the state estimate with the lowest variance within the stochastic model adopted. Hence, assuming errors in the open boundaries and the wind velocities, the Kalman filter provides the solution corresponding to the variance minimizing realisation of the error sources, assuming the model parameters to be perfect. The parameter estimation using the Steady Kalman Filter then ideally seeks parameters that minimise the cost function while allowing the filtered solution to estimate the independent errors sources assumed by the filter.

2.3 Cost function

Defining the cost function is a crucial step in defining the inverse problem itself. In a given application, the modeller will attempt to express a cost function that reflects the aim of the model calibration. It could be a balanced trade-off of all error types, a bias measure, a standard deviation or a phase error measure. In fact any kind of measure which express a deviation from observations or prior knowledge can be included in the cost function.

In the present set-up, a composite measure of standard deviations of model residuals over a five days period ranging from 1 January to 6 January is chosen. All stations are given equal weight. Hence, more emphasis is put on improving performance in areas with dense observation coverage. Due to difficulties of assessing model datum, the standard de-

viation rather than the mean square error is used in the measure. A very simple penalty, J_p, is added for parameter values diverging from the initial guess. Parameter values outside an initially defined acceptable range are given infinite penalty, while, parameters inside the range have zero contribution to the cost function. Thus, the cost function without using the Steady Kalman filter is,

$$J = \sqrt{\sum_{i=1}^{T_c}((\eta_{obs,i} - \overline{\eta_{obs}}) - (\eta_i - \overline{\eta}))^2} + J_p \tag{1}$$

where T_c is the number of observations in the calibration period. The optimisation with and without the Steady Kalman filter can be regarded as solutions to a strong and a quasi-weak constraint inverse problem formulation. Let over-line denote a time average. The cost function using an ideal filter is

$$J = \sqrt{\sum_{i=1}^{T_c}((\eta_{obs,i} - \overline{\eta_{obs}}) - (\eta_i^a - \overline{\eta^a}))^2} + J_p \tag{2}$$

where η_i^a is the element of the model state space which is the expectation of the model state conditioned on all previous measurement. The Steady Kalman imposes a large number of approximations such as a given error model, stationarity of model correlations at time of update and a linear estimator, but without these simplifications, the optimisation would become intractable.

2.4 Calibration parameters

Any model simulation requires a number of parameters to be set and hence their value to be assessed in some way. A coastal ocean model, like MIKE 21, has many free parameters. These include bathymetry, spatial bed friction maps, viscosity parameters and wind frictions parameters and can easily count of the order 10^5 parameters. They are however not all independent - two neighbouring bed friction values will most likely be almost equal. Hence, the effective dimension of the parameter space is much smaller.

In the present study the optimised parameters are wind stress parameters and bed friction maps according to predefined patterns that strongly

reduces the dimension of the parameter space. The wind friction, $\tau = (\tau_x, \tau_y)$, is calculated from the the two wind components, $\mathbf{W} = (W_x, W_y)$, according to,

$$\tau = C_w \frac{\rho_{air}}{\rho_{water}} |\mathbf{W}| \mathbf{W} \tag{3}$$

where ρ_{air} and ρ_{water} are air and water density respectively. The parameter C_w is a piecewise linear function of $\mathbf{W}$ as shown in Figure 4 and is specified by four parameters that are assumed constant in space. Default values are $C_{w,min} = 0.0016$, $C_{w,max} = 0.0026$, $W_{cw,min} = 0 \quad m/s$ and $W_{cw,max} = 24 \quad m/s$ yielding the solid line in Figure 4.

The bed friction map (of Manning numbers) consists of approximately 10^5 points and a strong space reduction is required for many optimisation algorithms in Section 2.5 to be tractable. The approach taken here is in an ad hoc way to let the Manning map be a sum of a limited number of independent contributions. These are constituted by

- a spatially constant background value
- A depth dependent map
- A Gaussian bell centered at Wick with spatial decorrelation scale of 300 km.
- A Gaussian bell centered at Lowestoft with spatial decorrelation scale of 300 km.
- A Gaussian bell centered at Esbjerg with spatial decorrelation scale of 300 km.
- A Gaussian bell centered at Hanstholm with spatial decorrelation scale of 300 km.

The depth dependent map has a zero crossing at 30 m and is a non-linear function of depth with -800 m and 0 meters being of equal size but opposite sign. The Gaussian bells at Wick, Lowestoft, Esbjerg and Hanstholm were chosen at measurement point to allow a local impact and a smooth variation. Traditional alternatives are to divide the domain into zones or use triangularisation. In (Heemink et al. 2002) a triangularisation was made based on a parameter sensitivity estimate by the adjoint of an

initial guess. This is not feasible in our case and the Gaussian bells were adopted.

All together 10 parameters were calibrated in the optimisation.

2.5 Optimisation method

The previous subsections have defined the inverse problem to be solved. The objective is to find the set of parameters defined in Section 2.4, that minimises the cost function (1) or (2) for the numerical model implementation and set-up described in Subsection 2.1.

The solution to this inverse problem can be found using a number of techniques with each having their advantages. The most frequent approach in hydrodynamical modelling is to combine a gradient based local search algorithm, such as a quasi-Newton technique or conjugate gradient, with a solver of the adjoint equations belonging to the inverse problem. The great advantage of this approach is the efficient calculation of the gradients of the cost function with respect to parameter values provided by the adjoint solution. This advantage increases with the number of parameters to be optimised. The disadvantage is the costly development and maintenance of the adjoint solver. Further, for potentially rough (highly variable) cost functions a gradient based algorithm is undesirable in any case.

So far the adjoint, gradient based approach has seemed like the only viable solution to the problem for large scale ocean models. However, by taking full advance of recent developments in parallel high performance computing (HPC), optimisation techniques that do not require the implementation of the adjoint gain increasing interest because of their highly parallisable algorithmic structure.

Previous studies have shown successful application of gradient based optimisation algorithms and hence these will be used in a future implementation. However, due to the code history for the optimisation package at DHI, these techniques are still at the stage of implementation and the existing Simplex and Shuffled Complex Evolution (SCE) algorithms are used in the present study. Neither of these are gradient based. The Sim-

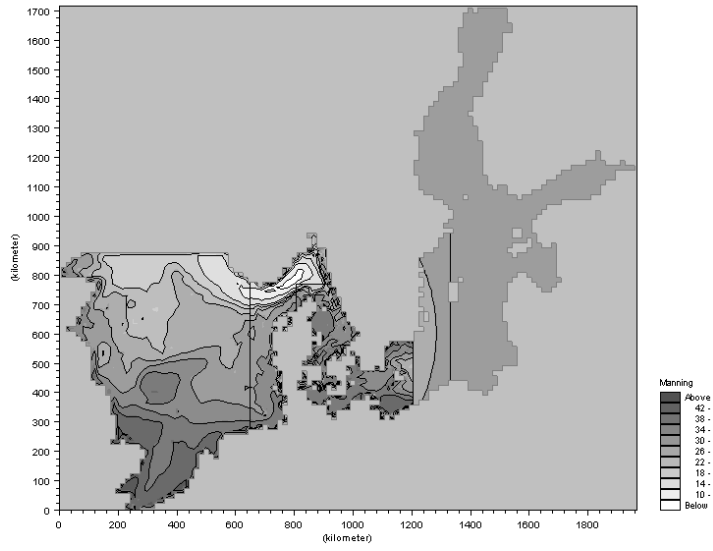

Figure 2: Optimised bed friction Manning map for the non-assimilating model operator (noDA)

plex is a local search algorithm, while SCE is a global search algorithm, (Duan, Sorooshian & Gupta 1992). A parallel implementation have been developed for the SCE optimisation algorithm. A coarse parallisation strategy was taken, in which the performance of a heterogeneous cluster was optimised.

3 Results and discussion

Two parameter estimation procedures are tested in the setup of MIKE 21 described in Section 2.1 with particular focus on the Inner Danish Waters. The bed and wind friction parameters are estimated in both the strong constraint and quasi-weak constraint data assimilating formulation. In either case the calibration is performed in the five-day period from 1 January to 6 January 2002. This is a short period, but enables the technique to be tested despite the lack of a parallel cluster at present and it does provide some insight into the performance of the estimations.

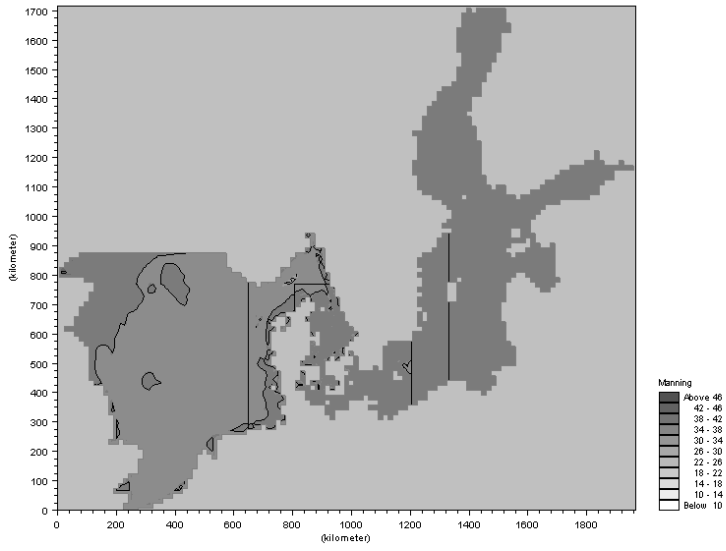

Figure 3: Optimised bed friction Manning map for the assimilating model operator (DA)

The cost function, J, is presented in Table 2 for the calibrated and reference runs in the case of both quasi-weak (DA) and strong (noDA) constraint optimisation. Figures 5 and 6 provides a detailed view of the contribution to the cost function from each tide gauge station. First of all, note that the calibrated parameters do not improve the model in all stations. Actually, in both calibration runs the main absolute improvements are in the North sea - Esbjerg for noDA and Lowestoft for DA. Closer inspection further reveals that the reference using assimilation is actually worse than with no assimilation in Lowestoft. This station is situated quite far from assimilation points and poor performance in data sparse regions has previously been noticed, (Sørensen et al. 2004). In a parameter optimisation context this actually means that we are optimising parameters to minimize the error introduced by the assimilation.

The estimated Manning maps and Wind friction coefficient function are shown in Figures 2, 3 and 4. Clear differences are evident, stressing the different nature of the optimisation in a quasi-weak and strong constraint formulation. Also included in Figure 4 is the default wind drag coefficient

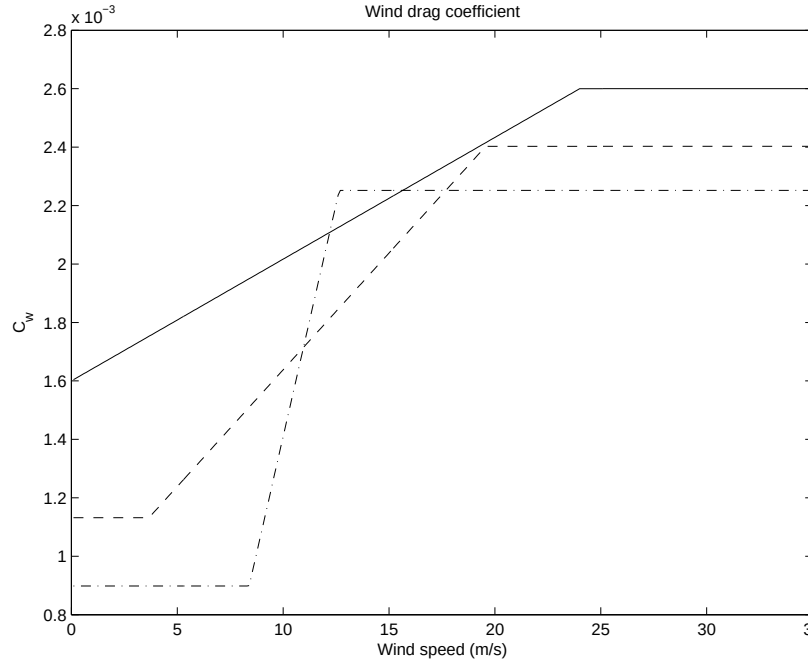

Figure 4: Wind drag coefficient. Solid line: Default values used in reference run. Dot-dashed line: Optimised values in the noDA case. Dashed line: Optimised values in the DA case.

	Reference	Optimised
no DA	1.31	1.25
DA	0.58	0.49

Table 2: Cost function values in the calibration period for the optimised parameter sets and the reference parameters excluding and including the Steady Kalman filter, respectively

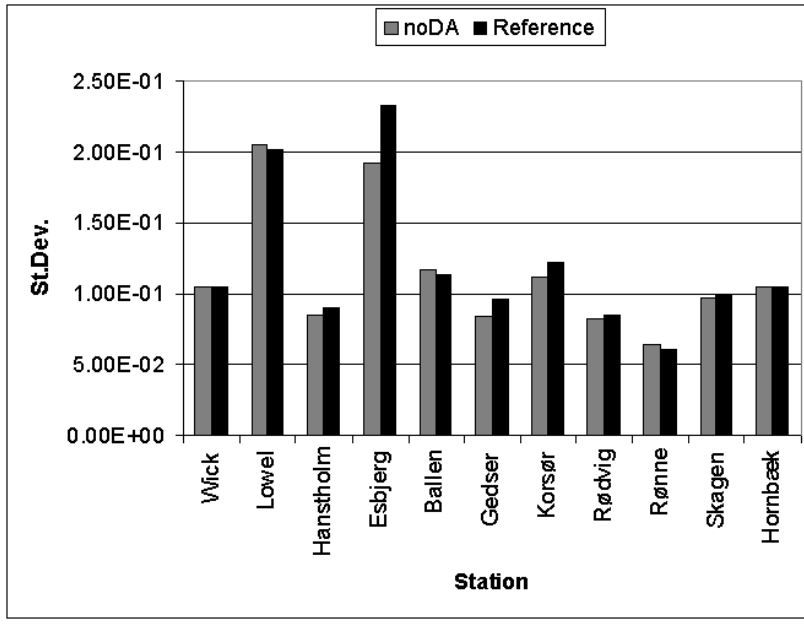

Figure 5: Cost function contributions, Strong constraint (noDA) calibration

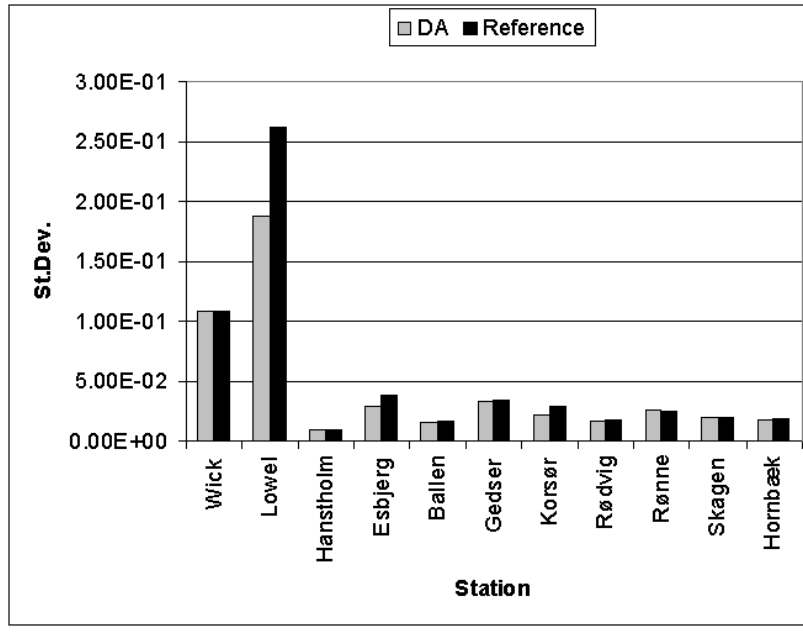

Figure 6: Cost function contributions, Weak constraint (DA) calibration

	Reference	Opt. no DA	Opt. DA
no DA	2.14	2.17	2.08
DA	1.09	1.25	1.06

Table 3: Cost function values in the validation period for parameter sets optimised with and without assimilation and the reference parameters. Each parameter set is validated in both the no assimilation the assimilation setting. the Steady Kalman filter, respectively

for the reference run. The default Manning number is 32 throughout the model. These express an experienced based good first guess in a arbitrary model application.

The performance of the two estimated parameter sets are assessed in a one month model execution spanning February 2002. Table 3 contains the validation performance statistics for an assimilating and non-assimilating February model execution respectively. Figures 7 and 8 provides a detailed view of the contribution to the cost function from each tide gauge station. Wick should be disregarded due to erroneous data in the validation period. It does not affect the relative performance.

In the strong constraint setting, the performance is worse than the reference. Even performance in the Esbjerg station, which had the dominating reduction in the calibration period is now degraded. It must be concluded that the parameters have been tuned to correct other error sources over an insufficient length of time. In the quasi-weak constraint setting the assimilation error correction is still dominating. Only the the North Sea stations, Lowestoft and Esbjerg, affected by this have a notably improvement in the validation period. However, despite this a better performance is obtained in the noDA validation run, indicating that the weak constraint is a more robust approach than strong constraint.

Obviously, a number of changes should be made to the optimisation setup.

- The bathymetry should be included in the optimisation. This has been noted to be one of the most important parameters in previous studies, (ten Brummelhuis et al. 1993).

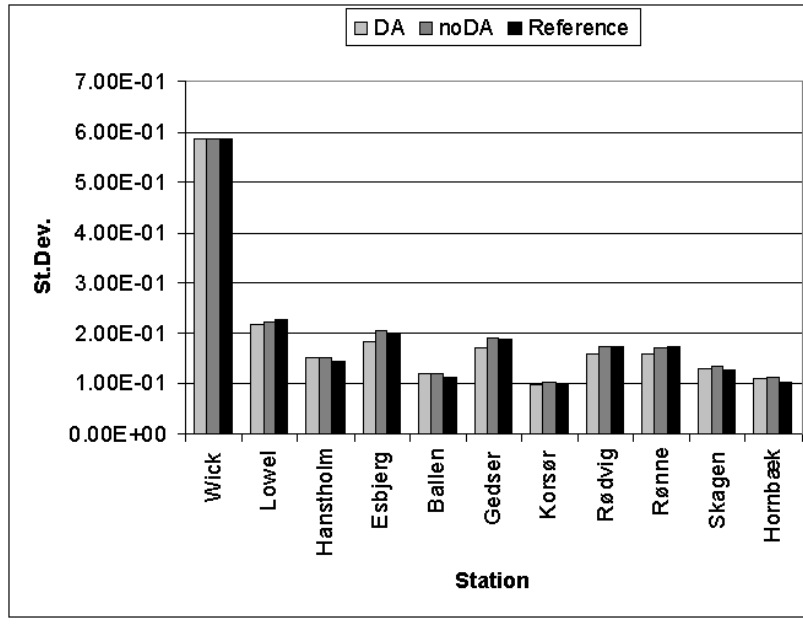

Figure 7: Cost function contributions. Strong constraint (noDA) validation

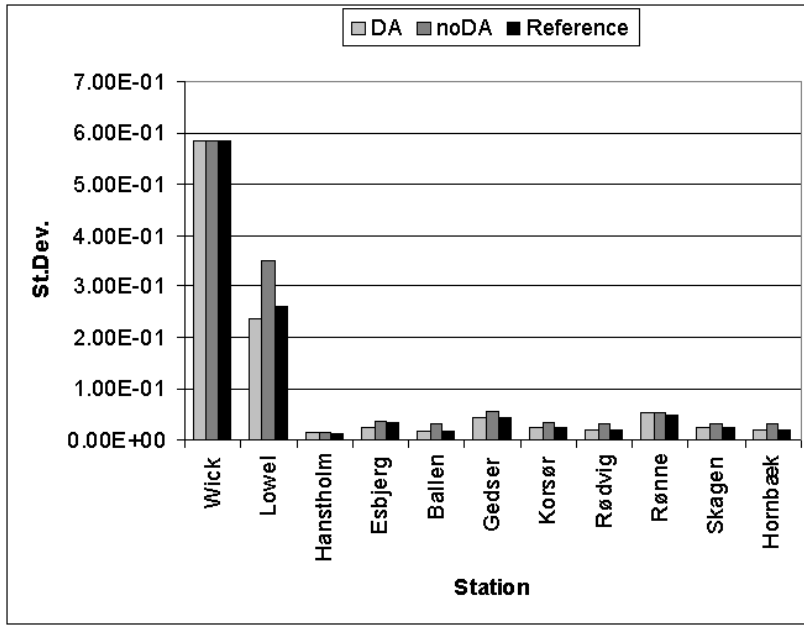

Figure 8: Cost function contributions. Quasi-weak constraint (DA) validation

- A more thorough selection of calibration stations and the cost function definition should be considered. E.g. a relative improvement of standard deviation as compared to the reference run standard deviation would put more emphasis on the stations in the Inner Danish Waters.
- A quasi-Newton optimisation scheme should be used and applied in a parallel cluster allowing a longer calibration period.
- The distance regularisation scheme devised in (Sørensen et al. 2004) must be applied if stations far from assimilation stations are to be included in the optimisation.
- Different space reductions of spatially varying parameters should be investigated.
- A two step calibration procedure should be investigated. Step one estimates the bathymetry, bed friction and possibly tidal constituents in a purely tidal setting. Step two estimates wind friction parameters.

4 Conclusion

This study has presented a parameter estimation framework for two-dimensional hydrodynamic models of coastal and shelf sea areas. The main difference from previous work in the field is to perform the optimisation without requiring the development of an adjoint code. The approach supports current developments in parallel cluster technology, which exploit the computational resources of PC's in an office grid. The optimisation framework included both a strong constraint and a quasi-weak constraint formulation and was demonstrated to estimate bed and wind friction parameters in a model application in the North Sea, Baltic Sea and Inner Danish Waters. The quasi-weak constraint setting seemed more robust, but a combination of factors calls for further experiments for firm conclusions to be drawn.

The use of a formalised parameter estimation framework in DHI's hydrodynamic models has been initiated and fair results obtained considering this being a first attempt with many improvements all ready identified.

References

Cañizares, R., Madsen, H., Jensen, H. R. & Vested, H. J. (2001), 'Developments in operational shelf sea modelling in Danish waters', *Estuarine, Coastal and Shelf Science* **53**, 595–605.

DHI (2002), *MIKE 21 coastal hydraulics and oceanography*, DHI Water & Environment.

Duan, Q., Sorooshian, S. & Gupta, V. (1992), 'Effective and efficient global optimization for conceptual rainfall-runoff models', *Water Resource Research* **28**, 1015–1031.

Erichsen, A. C. & Rasch, P. S. (2002), Two- and three-dimensional model system predicting the water quality of tomorrow, *in* M. L. Spaulding, ed., 'Proceedings of the seventh international conference on estuarine and coastal modeling', American Society of Civil Engineers, American Society of Civil Engineers, pp. 165–184.

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

Evensen, G., Dee, D. & Schröter, J. (1998), Parameter estimation in dynamical models, *in* E. P. Chassignet & J. Verron, eds, 'Ocean Modeling and Parameterizations', NATO ASI, Kluwer Acad. Pub.

Giering, R. & Kaminski, T. (1998), Using tamc to generate efficient adjoint code: Comparison of automatically generated code for evaluation of first and second order derivatives to hand written code from the minpack-2 collection, *in* C. Faure, ed., 'Automatic Differentiation for Adjoint Code Generation', INRIA, pp. 31–37.

Heemink, A. W. (1986), Storm surge prediction using Kalman filtering, PhD thesis, Twente University of Technology.

Heemink, A. W., Mouthaan, E. E. A., Roest, M. R. T., Vollebregt, E. A. H., Robaczewska, K. B. & Verlaan, M. (2002), 'Inverse 3d shallow water flow modelling of the continental shelf', *Continental Shelf Research* **22**, 465–484.

Lardner, R. W., Al-Rabeh, A. H. & Gunay, N. (1993), 'Optimal estimation of parameters for a two-dimensional hydrodynamical model of the Arabian Gulf', *Journal of Geophysical Research* **98**(C10), 18,229–18,242.

Madsen, H. (2003), 'Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives', *Advances in Water Resources* **26**, 205–216.

Rogers, E., Black, T., Ferrier, B., Lin, Y., Parrish, D. & DiMego, G. (2001), 'Changes to the ncep meso eta analysis and forecast system: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3dvar analysis', *NWS Technical Procedures Bulletin*.

Sørensen, J. V. T., Madsen, H. & Madsen, H. (2004), 'Efficient kalman filter techniques for the assimilation of tide gauge data in three-dimensional modeling of the north sea and baltic sea system', *J. Geophys. Res.* **109**(C3), C03017.

ten Brummelhuis, P. G. J., Heemink, A. W. & van den Boogaard, H. F. P. (1993), 'Identification of shallow sea models', *International Journal of Numerical Methods in Fluids* **17**, 637–665.

Verlaan, M. & Heemink, A. W. (1997), 'Tidal flow forecasting using reduced rank square root filters', *Stochastic Hydrology and Hydraulics* **11**, 349–368.