Data assimilation for hydrodynamical modeling of the Odra lagoon

Laurent BERTINO[1], Julien SÉNÉGAS[1], Hans WACKERNAGEL[1], Hans VON STORCH[2]

[1]Centre de Géostatistique, Ecole des Mines de Paris
35 rue Saint Honoré, F–77305 Fontainebleau, France

[2]Institute for Hydrophysics, GKSS Research Centre
Max-Planck-Straße, D–21502 Geesthacht, Germany

Contact: bertino@cg.ensmp.fr

Abstract

Data assimilation can be defined as the incorporation of measurements into the numerical model of a physical system, to improve the forecasts of this model. Data assimilation has gained increasing popularity in the atmospheric and oceanographic communities over the last two decades. The random functions approach of geostatistics, its multivariate spatial modeling tools and its change-of-support models and the spatial simulation capabilities provide an attractive framework for performing data assimilation.

The study of nutrients in estuaries from the Ebro and Odra rivers in the EC funded PIONEER project allows combination of both techniques for forecasting the physical, biological and chemical state of the estuarine system.

The dynamical evolution of the estuarine variables and corresponding observations are modeled with a state-space model and, since their evolution is given by non-linear functions, an extended Kalman Filter is used for the data assimilation. Different suboptimal schemes such as the Reduced Rank Square Root (RRSQRT) Kalman Filter and simple Data Assimilation methods are compared, focusing attention to computer time and memory requirements. Geostatistical modeling ideas are discussed in the application of these algorithms.

1 Introduction

Data assimilation has been widely used since the sixties in operational weather forecast for melding atmospheric model predictions with independent informations (measurements ...). Since then, they have been applied in oceanography mainly by the means of sequential methods (Evensen [4]). Data Assimilation methods meld a physical model deriving from the system equations with in-situ measurements. These methods lead to a best estimate of the past, present or future state of the system, but they also allow for parameter estimation and initial or boundary conditions estimation. However, Data Assimilation in its actual form doesn't modify the system

equations, nor their discretization.

In the present study, the goal of data assimilation is to give short-term operational predictions of the hydrodynamical state of the Odra lagoon, which allows for forecasting extreme events such as the Odra flood of summer 1997 (Rosenthal [7]). Furthermore, the hydrodynamical state has to be known with high accuracy for efficient ecological monitoring of the lagoon.

The Odra Laggon straddles the Polish-German border. Its surface covers about 600 km^2, its average depth is 5m, so that its volume is about $3 * 10^9$ km^3. Fresh water enters the lagoon from several rivers, with the Odra being the most important. Intermittently salt water intrudes in small amounts through the three narrow entries to the Baltic Sea. Such intrusions are caused by water levels differences between the Baltic Sea and the Lagoon and have a large impact on the whole ecosystem.

The present article first reviews simple Data Assimilation methods and Kalman Filtering methods applied to high-dimensional models. Then an example of data assimilation of the Odra Lagoon water levels is given using the RRSQRT Kalman Filter.

2 The Data Assimilation problem: continuous and discretized

Let $\mathbf{Z}$(x,t) be a multivariate function of space and time. In most sections of this article, we will discretize the spatial dependence of $\mathbf{Z}$, that is, we will consider the multivariate random function $\mathbf{Z}(t)$ where all locations x are different components of the state vector $\mathbf{Z}(t)$, without loss of generality. We will use the continuous spatio-temporal notation $\mathbf{Z}(x,t)$ in section 4.4 when comparing data assimilation to Kriging methods.

The state $\mathbf{Z}$ is supposed to follow the differential equation of the physical model :

$$\frac{\partial \mathbf{Z}}{\partial t} = f(\mathbf{Z}) \tag{1}$$

f is often a nonlinear function of $\mathbf{Z}$ and of the spatial derivatives of $\mathbf{Z}$. Equation (1) is called the *evolution equation* of the system.

The system state $\mathbf{Z}$(x,t) is observed through a measurement vector z of m observations. The observed variables are not necessarily the state variables of $\mathbf{Z}$ but they are linked through the *observation equation* (2) :

$$\begin{aligned} z &= \mathcal{L}(\mathbf{Z}(x,t)) \\ \forall \alpha = 1 \ldots m,\, z_\alpha &= \mathcal{L}_\alpha(\mathbf{Z}(x,t)) \end{aligned} \tag{2}$$

Contrarily to the evolution equation (1), this operator is linear.

The system of both equations (1) and (2) is called a *state space model*. This system is ill-conditioned and generally has no solution since every measurement contradicts the model predicted state, but the introduction of random errors in both equations allows for finding an optimal solution that satisfies best both constraints.

When discretized, the above equations (1) and (2) become for time step n :

$$\begin{aligned} \mathbf{Z}_{n+1} &= F(\mathbf{Z}_n) \\ z_n &= \mathcal{L}(\mathbf{Z}_n) \end{aligned}$$

where $\mathbf{Z}_n$ is a vector of length *number of state variables* $*$ *number of spatial grid cells* and z_n a vector of length m. Within the discretized framework, it should be noted that measurements and model state values have different spatio-temporal supports : a measurement is an average of the variable of interest on a small spatio-temporal interval, while model state values are averages on a whole grid cell during a model step which are both much larger. Therefore, measurements have much higher variability than model output and they have to be averaged in order to compare model output with observations that have similar variability.

3 General solution to the nonlinear problem

The evolution equation (1) is perturbed by a random model noise dw_t, classically considered a multivariate white noise and becomes an Itô stochastic differential equation.

$$d\mathbf{Z}(t) = f(\mathbf{Z}, t)dt + g(\mathbf{Z}, t)dw_t$$

When defined, the pdf $\phi_t(\mathbf{Z})$ of $\mathbf{Z}$(t) follows the *Fokker-Planck equation* :

$$\frac{\partial \phi_t}{\partial t} = -\nabla.(f(\mathbf{Z}, t)\phi_t) + \sum_{i,j} \frac{\partial^2}{\partial \mathbf{Z}_i \partial \mathbf{Z}_j}(G/2)_{ij}\phi_t$$

with $G = gg^T$. This equation is a deterministic advection-diffusion equation having additional constraints of positivity of ϕ_t and of unit sum of probabilities $\int \phi_t = 1$.

The Fokker-Planck equation could be solved as a partial differential equation with more complexity due to both constraints quoted above, Miller [6] found a solution to the double-well problem, a nonlinear monovariate problem. But for high-dimensional problems such as ocean models no solution has been found yet.

4 Sequential methods

In the sequential approach, one can estimate the state of the system at time step n only from the estimated state at time step $n-1$ and from in-situ data at time step n. The unknown true state remains $\mathbf{Z}_n$, the model forecast state at time n is noted Z_n^f and the analyzed state after assimilation of the measurements Z_n^a.

4.1 Primitive methods

These methods are simple and economical. They provide a basis for assessment of more expensive methods.

Direct Insertion The model-forecast value is replaced by an observation at all grid cells where measurements are available. As a first remark one can notice that extending a quasi-punctual measurement to a whole grid cell is statistically incorrect.

Nudging Also called *Newtonian relaxation*, this method aims at a dynamical model relaxation towards the observations :

$$Z_n^f = F(Z_{n-1}^a) \tag{3}$$

$$Z_n^a = Z_n^f + K_n(z_n - \mathcal{L}(Z_n^f)) \quad (4)$$

Equation (3) is called *Time Step* and equation (4) the *Measurement Step*.

An example where K_n is obtained by cost function minimization is given by Zou and Le Dimet [11].

Optimal Interpolation This technique is also know as *Statistical Interpolation* in meteorology and commonly used for numerical weather prediction (Daley [3]). The Time and Measurement Step equations are the same as above, but the gain matrix K_n is generally obtained by classical kriging techniques (Wackernagel [10]).

4.2 Principles of the Kalman Filter

Kalman filtering applied to oceanography is described in detail by Bennett [1]. Discretization is required in our applications for resolving the advection-diffusion equations. The crude Kalman filter is designed for linear systems such as :

$$\mathbf{Z}_{n+1} = F_n \mathbf{Z}_n + q_n$$

q_n is the *Model Error*, a random noise vector of variance-covariance matrix Σ_m. This matrix can be made time-dependent without any incidence on the ongoing calculations.

The observation equation (2) also has a random additive error r_n, called the *Measurement Error* of variance-covariance matrix Σ_o :

$$z_n = \mathcal{L}(\mathbf{Z}_n) + r_n$$

The classical hypotheses for the Kalman Filter are that both q_n and r_n are independent centered Gaussian white noise processes :

- $E(q_n) = E(r_n) = 0$.
- $\forall p \neq n$, $cor(r_p, r_n) = cor(q_p, q_n) = 0$
- $\forall (n, p), cor(r_n, q_p) = 0$

The spatial structure of the measurement error is often modeled by a Nugget Effect, and the random functions $\mathbf{Z}$ and z are generally implicitly assumed stationary of order 2 while computing the covariance matrices Σ_o and Σ_m.

4.3 Equations of the Kalman Filter

The Time Step equation is linear :

$$Z_n^f = F_{n-1} Z_{n-1}^a$$

and the Measurement Step equation is equation (4). Z^f and Z^a are estimates of the true state $\mathbf{Z}$, Z^f being a first guess given by the model and Z^a the optimal estimate after assimilation of

the measurements. The estimation errors $\epsilon^f = \mathbf{Z} - Z^f$ and $\epsilon^a = \mathbf{Z} - Z^a$ have respectively C^f and C^a for covariance matrices. In Kalman Filtering these matrices evolve according to :

$$\begin{aligned} C^f_{n+1} &= F_n C^a_n F_n^T + \Sigma_m & (5)\\ K_{n+1} &= \mathcal{L}^T(C^f_{n+1})(\mathcal{L}\mathcal{L}^T(C^f_{n+1}) + \Sigma_o)^{-1} & (6)\\ C^a_{n+1} &= C^f_{n+1} - K_{n+1}\mathcal{L}(C^f_{n+1}) & (7) \end{aligned}$$

Where it can be shown under the above hypotheses that K_{n+1} minimizes the variance of the estimation error (see Maybeck [5]).

$$J(K_{n+1}) = \mathrm{Tr}(E[\epsilon^a_{n+1}(\epsilon^a_{n+1})^T])$$

4.4 Formal equivalence with Cokriging

Given a time step n, we demonstrate equation (6) by **Simple Kriging** of the random function $(\mathbf{Z} - Z^f)$: the prediction error $(\mathbf{Z}(x) - Z^f(x))$ is estimated by a combination of the observation misfits $z - \mathcal{L}(Z^f)$ at a given location x. The demonstration is carried out in the monovariate case in order to simplify the notations. In the multivariate case, similar equations lead to the result that the Kalman Filter Measurement Step is equivalent to a Simple Cokriging.

When applied to a function of two positions such as the covariance $C(x_1, x_2)$, the one-position function obtained by measurements of the first (resp. the second) position is noted $\mathcal{L}(C)(x_2)$ (resp. $\mathcal{L}^T(C)(x_1)$), then, the $m * m$ matrix obtained by composition of both $\mathcal{L}$ and $\mathcal{L}^T$ is noted $\mathcal{L}\mathcal{L}^T(C)$. For $\alpha = 1 \ldots m$, $\mathcal{L}_\alpha$ is the function giving the α measurement.

Let us consider the following linear combination

$$\begin{aligned} \epsilon^a(x) &= (\mathbf{Z}(x) - Z^f(x))^* & (8)\\ &= \sum_\alpha \lambda^\alpha (z_\alpha - \mathcal{L}_\alpha(Z^f)) \end{aligned}$$

Since the observation error is supposed to be stationary of order 2, the estimation error has a finite variance. The error covariances are :

$$\begin{aligned} \mathrm{Var}(\epsilon^f(x)) &= C^f(0) & (9)\\ \forall \alpha = 1 \ldots m, \mathrm{Cov}(\epsilon^f(x), z_\alpha - \mathcal{L}_\alpha(Z^f)) &= \mathcal{L}^T_\alpha(C^f)(x)\\ \forall \alpha, \beta, \mathrm{Cov}(z_\alpha - \mathcal{L}_\alpha(Z^f), z_\beta - \mathcal{L}_\beta(Z^f)) &= (\Sigma_o)_{\alpha,\beta} + \mathcal{L}_\alpha \mathcal{L}^T_\beta(C^f) \end{aligned}$$

Using the independence between Model and Measurement Errors.

Considering the Kalman Filter assumptions $E(r) = E(q) = 0$ for all previous time steps,

$$E[\epsilon^f(x)] = 0$$

which means that the mean of $\mathbf{Z} - Z^f$ is known, equal to zero and therefore that we are performing a **Simple Kriging**.

Lastly, at a given location x the corresponding line of the Kalman gain matrix K is called $\Lambda = (\lambda^\alpha)_{\alpha=1\ldots m}$.

$$J(\Lambda) = \mathrm{Var}(\epsilon^f(x) - \sum_\alpha \lambda^\alpha (z_\alpha - \mathcal{L}_\alpha(Z^f)))$$

$$= C^f(0) - 2\sum_\alpha \lambda^\alpha \mathcal{L}_\alpha(C^f)(x) + \sum_{\alpha,\beta} \lambda^\alpha \lambda^\beta ((\Sigma_o)_{\alpha\beta} + \mathcal{L}_\alpha \mathcal{L}^T_\beta(C^f)) \quad (10)$$

In order to minimize $J(\Lambda)$, the derivatives of the above expression along all $\lambda^\alpha, \alpha = 1 \ldots m$ are set equal to zero, then the following matrix products appear:

$$\begin{array}{rrcccl} \forall \alpha : \frac{\partial J(\Lambda)}{\partial \lambda^\alpha} = 0 & \Leftrightarrow & \Lambda(\Sigma_{o\alpha} + \mathcal{L}_\alpha \mathcal{L}^T(C^f)) & = & \mathcal{L}_\alpha(C^f)(x) \\ & \Leftrightarrow & \Lambda(\Sigma_o + \mathcal{L}\mathcal{L}^T(C^f)) & = & \mathcal{L}(C^f)(x) \\ & \Leftrightarrow & \Lambda & = & \mathcal{L}(C^f)(x)(\Sigma_o + \mathcal{L}\mathcal{L}^T(C^f))^{-1} \\ \forall x \Rightarrow & K = & \mathcal{L}^T(C^f)(\Sigma_o + \mathcal{L}\mathcal{L}^T(C^f))^{-1} & & \end{array}$$

When varying the estimation location x we find back the Kalman Gain expression (6). The same result can be found in the multivariate case and demonstates the equivalence between the Measurement Step of the Kalman Filter and Simple Cokriging.

4.5 Extension of the KF to the non-linear case

Hydrodynamical models are often non-linear and the Kalman Filter method has to be modified to the *Extended Kalman Filter*, in which the system evolves according to the following equation

$$\mathbf{Z}_{n+1} = F(\mathbf{Z}_n, q_n)$$

and the covariance matrix is propagated by first order derivatives of F. The measurement step is performed by the means of a linearization of the model. Truncation to higher orders can be necessary in some - exceptional - cases, see Verlaan [9] for a discussion.

For coastal modeling application, even with first order development, the propagation of the covariance matrix and the computation of the Kalman Gain are such time and memory demanding tasks that the implementation of the EKF is not realistic. The propagation of the covariance matrix by itself runs forward the model as many times as there are columns in the covariance matrix C^f, usually about 10^5 times. Therefore many *Sub-Optimal Schemes* have been developed for faster and more practical computations, but whose solutions do not strictly minimize the estimation error variance. A few are listed here:

Stationary filter The Gain is user-assigned and constant. This method is interesting if the gain K_n converges when $n \to +\infty$.

Coarse grid approximation The covariance is defined on a coarser grid as the state grid.

Singular Values Partial KF The linearized model evolution matrix is truncated along its highest singular values for the error covariance propagation.

Model reduction The state vector is constrained to a subspace determined by measurement error minimization.

RRSQRT Kalman Filter (Verlaan [9]) Reduced Rank Square Root Kalman Filter : The covariance matrix is simplified by an eigendecomposition and a truncation on the strongest eigenvalues at every assimilation step. As the main time consuming step is generally the

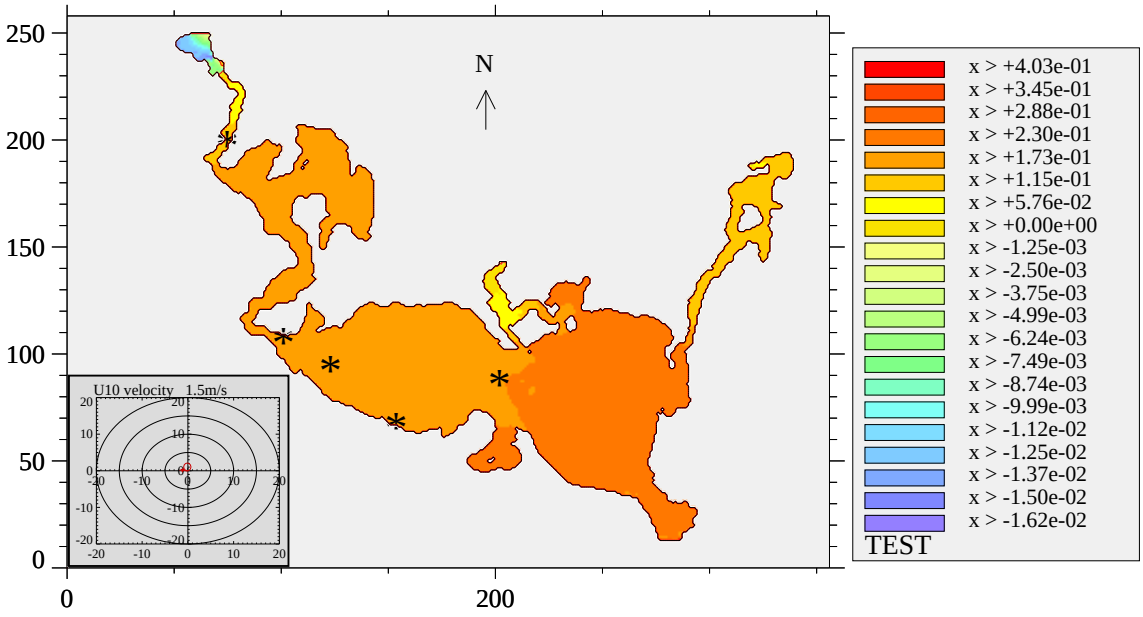

Figure 1: DA by RRSQRT KF : map of the water level assimilated on August the 4^{th}, 1997 at 12:00, the wind direction in the frame indicates a low wind at that time. The Odra inflow comes at the bottom right of the graph and the three channels to the Baltic Sea are on the top left middle and right of the domain. Measurement stations are marked by an asterisk.

forward propagation in time of the reduced square root covariance matrix, the computation time of the crude model is approximately multiplied by the number of eigenmodes kept for rank reduction.

Ensemble KF (Evensen [4]) The state and covariance matrix are estimated by a Monte-Carlo method on an ensemble of simulated states which should be large enough to reproduce the non-Gaussian true pdf. Here, the filter multiplies the computation time by the number of simulated states.

5 Case of the Odra Lagoon hydrodynamics

The TRIM3D model is a 3D numerical model for hydrodynamics (Casulli [2]). It solves the Navier-Stokes equations for free-surface flows under hydrostatic hypothesis. A semi-implicit finite difference scheme is carried out on an Arakawa grid; see Rosenthal et al. [7] and Sénégas [8] for numerical simulations of the Odra Lagoon using TRIM3D.

In the following case study, the Odra Lagoon grid has a 250m horizontal resolution and spreads itself on 357*259 square nodes among which 16053 are wet. Three vertical layers are considered which finally gives around 48000 active nodes. The initial state of the model is such that all variables are set up to zero.

The assimilated observations are water levels. Hourly measurements are sampled in five pile stations located in the "Kleines Haff" and near the coast.

Implementing a Kalman Filter prerequisites the knowledge of both observation and model error covariance matrices Σ_o and Σ_m. In our case these matrices are unknown. Observation

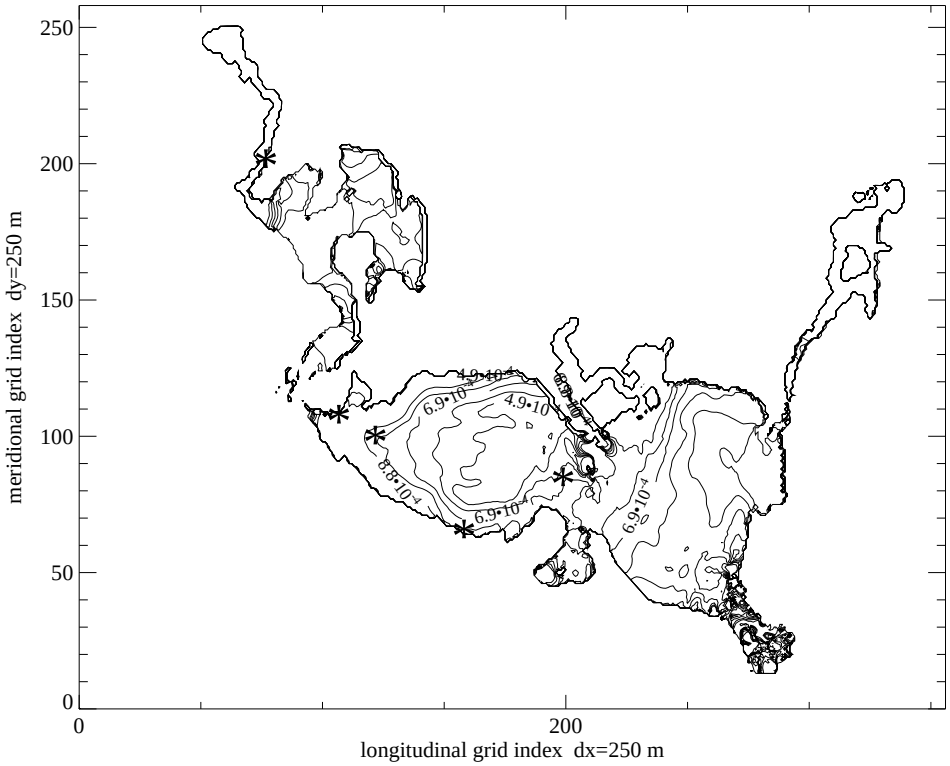

Figure 2: DA by RRSQRT KF : error variances of the previous estimation, variances vary between 0.0004 and 0.002 m^2 which means that the water levels are known with a precision between 2 and 5 cm. Measurement stations are marked by an asterisk.

errors can be considered as spatial white noise because they depict a lack of accuracy in the measurement protocol and are seldom correlated from one device to another. The variance of this white noise has been set to the nugget effect part of the variograms of the measurements time series, which means that this observation variance is overestimated.

Since in this case study the main causes for imprecision in the model are errors in the boundary conditions, the model errors are limited to a nugget effect in the wind field (of amplitude 0.5m/s) and a correlated noise in the Baltic interface water levels, for which the covariance matrix was computed by structural analysis of historical water level data in the Baltic sea.

The number of retained eigenmodes is set to 50 considering the shape of the eigenvalues diagram and a previous simplified 1D study (Sénégas [8]).

The RRSQRT KF run simulates the period from the 4^{th} of August 1997 at 00:00 to the 19^{th} of August 1997.

On the assimilated water levels map in Figure 1 the RRSQRT Kalman Filter proves to have realistic smooth state estimates after twelve hours of assimilation, while a corresponding run without assimilation takes two days to erase the marks of the arbitrary initial state. Since the Measurement Step is also an interpolation method, it also diminishes the effect of erroneous initial and boundary conditions.

The error variance map in Figure 2 shows that the error expected by the RRSQRT KF is reasonable (a few centimeters standard deviation) and remains stable. It also shows that the system noise, although it is only introduced at the boundary conditions, is spread on the whole system and has maximum variance near the basin coasts. When following the evolution of the

error variance from one step to another, the error seems to be reflected on the basin coasts like waves in a closed cavity.

It was also observed on the variance maps that during a 15 days DA run the covariance matrix does not converge towards a stable state of the filter, which is an effect of both the time-dependence of the hydrodynamical model and of the variable system noise.

6 Conclusions

When applied to high-dimensional systems such as the hydrodynamics of the Odra Lagoon, Kalman Filtering becomes excessively demanding in CPU and disk space and leads to practical problems.

In this case study, the Kalman Filter based on the suboptimal RRSQRT scheme provides its own analysis of the prediction errors as it interpolates the errors by the means of a *Simple Cokriging* and propagates forward the error statistics in the model.

The inputs of Geostatistics in Data Assimilation methods are the description of support effects while comparing in-situ measurements with model forecasts, improvements of the interpolation technique and computation of the error covariances. For this last topic, another practical problem is found with sparse measurement locations.

The RRSQRT Kalman Filter can handle these difficulties and in future work this Data Assimilation algorithm will be put in competition with other suboptimal schemes.

Acknowledgments

This work was carried out within the EC funded MAST III project PIONEER (description: http://pioneer.geogr.ku.dk).

References

[1] Bennett, A. F. (1992) *Inverse methods in physical oceanography*, Cambridge University Press, Cambridge, UK.

[2] Casulli, V.; Cattani, E.; *Stability, Accuracy and Efficiency of a Semi-Implicit Method for Three-Dimensional Shallow Water Flow*, Computers Math. Applic. (1994), Vol. 27, No. 4, pp. 99-112.

[3] Daley, R. (1991) *Atmospheric Data Analysis,* Cambridge University Press, Cambridge, UK.

[4] Van Leeuwen, P.J.; Evensen, G. *Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation,* Monthly Weather Review (1996), Vol. 124, pp. 2898-2913.

[5] Maybeck, P.S. (1979) *Stochastic models, estimation, and control,* Volume 141-1 of *Mathematics in Science and Engineering*, Academic Press, New York, USA.

[6] Miller, R. N.; Carter, E. F. Jr.; Blue, S. T. *Data Assimilation into Nonlinear Stochastic Models,* Tellus (1999), Vol. 51A, pp. 167-194.

[7] Rosenthal, W.; Wolf, T.; Witte, G.; Buchholz, W.; Rybaczok, P. *Measured and Modelled Water Transport in the Odra Estuary for the Flood Period July/August 1997*, German Journal of Hydrography (1998), Vol. 50, No 2/3, pp. 215-230.

[8] Senegas, J. (1999) *Hydrodynamical modeling and data assimilation within the Odra estuary,* External report GKSS 99/E/42, GKSS Research Center, D-21494 Geesthacht, Germany.

[9] Verlaan, M. (1998)*Efficient Kalman filtering algorithms for hydrodynamic models*, Doctoral thesis at TU Delft, The Netherlands.

[10] Wackernagel, H. (1998) *Multivariate Geostatistics*, Springer Verlag, Berlin.

[11] Zou, X.; Navon, I.M.; Le Dimet, F.X.(1992) *An optimal nudging data assimilation scheme using parameter estimation,* Q. J. R. Meteorol. Soc. (1992) 118, pp. 1163-1186.