Project

Assimilation methodologies

Reference: MyWave-xxxxxxxxx

Project N°: FP7-SPACE-2011-284455	**Work programme topic:** SPA.2011.1.5.03 – R&D to enhance future GMES applications in the Marine and Atmosphere areas
Start Date of project : 01.01-2012	**Duration**: 36 Months

WP leader: Deltares	**Issue:**
Contributors :	
MyWave version scope : <all project versions> or <version x>	
Approval Date :	**Approver:**
Dissemination level: PU	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:			
Distribution			

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
0.1	15-dec-2013	all	First draft of document	M. Verlaan,, S. Caires, G-J. Marseill and J. Stavena	J. S. Stellenfeleth and A.Stoffelen
1.0	10-jan-2014	all	Document finalization	M. Verlaan, S. Caires, G.J. Marseille, K. Wahle, H.G. Günther, J. Staneva	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABBREVIATIONS

3D-Var	Three-dimensional variational analysis
4D-Var	Four-dimensional variational analysis
DA	Data assimilation
EnKF	Ensemble Kalman Filter
HARMONIE	Hirlam Aladin Research on Meso-scale Operational Nwp In Euromed
H_s or H_{m0}	Significant wave height
NN	Neural Network
SWAN	Simulating Waves Nearshore
$T_{m-1,0}$ or T_{m1}	Mean wave period
WAM	WAve Modelling

APPLICABLE AND REFERENCE DOCUMENTS

Applicable Documents

	Ref	Title	Date / Issue
DA 1	MyWave-A1	MyWave: Annex I – "Description of Work	September 2011

Reference Documents

	Ref	Title	Date / Issue
	Altaf et al. (2009)	Altaf, M.U., A.W. Heemink, and M. Verlaan. "Inverse shallow-water flow modeling using model reduction." International journal for multiscale computational engineering 7.6 (2009).	2009
	Bishop (1995)	Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.	1995
	Booij et al. (1999)	Booij, N., Ris, R. C., and L. H. Holthuijsen, 1999: A third generation wave model for coastal regions. Part 1. Model description and validation, *J. Geophys. Res*., 104(C4), 76497666.	1999
	Burgers et al. (1998)	Burgers, G., P. J. van Leeuwen and G. Evensen, 1998:Analysis Scheme in the Ensemble Kalman Filter. *Mon. Wea. Rev*., 126, 1719–1724.	1998
	Courtier et al. (1998)	Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier and M. Fisher, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part 1: formulation. *Quart. J. Roy. Meteor. Soc*., 124, 1783-1807.	1998
	Evensen (1994)	Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res*., 99, pp 10,143-10,162.	1994
	Evensen (2003)	Evensen, G., 2003: The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*. 53, 343-367.	2003

	De Haan et al. (2013)	De Haan, S., Marseille, G.J., de Valk, P., de Vries, J., 2013: Impact of ASCAT Scatterometer Wind Observations on the High-Resolution Limited-Area Model (HIRLAM) within an Operational Context. *Weather and Forecasting*, 28 , pp. 489-503, doi: http://dx.doi.org/10.1175/WAF-D-12-00056.1	2013
	Hasselmann et al. (1997)	Hasselmann, S., P. Lionello, and K. Hasselmann, 1997: An optimal interpolation scheme for the assimilation of spectral wave data. *J. Geophys. Res.*, 102(C7), 15823–15836, doi:10.1029/96JC03453.	1997
	Haykin (1994)	Haykin, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.	1994
	Heemink and Kloosterhuis (1990)	Heemink, A. W., and H. Kloosterhuis. "Data assimilation for non-linear tidal models." International journal for numerical methods in fluids 11.8 (1990): 1097-1112.	1990
	Hersbach (1998)	Hersbach, H., 1998. Application of the adjoint of the WAM model to inverse wave modeling, *J. Geophys. Res*., 103 (C5), pp.10469-10487	1998
	Kalman (1960)	Kalman, Rudolph Emil. "A new approach to linear filtering and prediction problems." Journal of basic Engineering 82.1 (1960): 35-45.	1960
	Kalnay et al. (2007)	Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang and J. Ballabrera, 2007: 4D-Var or Ensemble Kalman Filter? *Tellus A*, 59, 758–773.	2007
	Lionello et al. (1995)	Lionello, P., H. Günther, B. Hansen, 1995: A sequential assimilation scheme applied to global wave analysis and prediction, *J. Mar. Syst.*, 6, 87–107.	1995
	Sakov et al. (2010)	Sakov, P., G. Evensen and L. Bertino, 2010: Asynchronous data assimilation with the EnKF. Tellus A, 62.	2010
	Schiller (2007)	Schiller, H., 2007. Model inversion by parameter fit using NN emulating the forward model - Evaluation of indirect measurements, Neural Networks, 20 (4), Computational Intelligence in Earth and Environmental Sciences, pp. 479-483.	2007
	Voorrips et al. (1997)	Voorrips, A.C., V.K. Makin and S. Hasselmann, 1997: Assimilation of wave spectra from pitch-and-roll buoys in a North Sea wave model. *J. Geophys. Res*., 102 (C3), 5829-5849.	1997
	Voorrips (1998)	Voorrips, A.C., 1998: *Sequential data assimilation methods for ocean wave models*. PhD Thesis, 175 pp.	1998

I INTRODUCTION

I.1 Framework

Work package 2 (WP2) of the MyWave project focusses on increasing the use of earth observations by improving data processing algorithms and data assimilation systems for ocean waves. Aiming at, exploration of new methodologies in data assimilation, improvement of the use of near-shore satellite data and connection of large-scale forecast to near-shore forecasts. Its ultimate goal is to use improved data processing and data assimilation methods to obtain better wave forecasts from regional or coastal high-resolution models.

I.2 Motivation

Data assimilation (DA) techniques can be divided in synchronous and asynchronous (Sakov 2010) techniques. Synchronous methods are also known as sequential methods. In the synchronous techniques the observations are used to correct the model (first guess) data at the analysis moment, without regard for model dynamics between analysis moments. The increased availability of computer power has lead to an operational use of more advanced and generally asynchronous DA techniques. However, the most commonly used DA technique in wave forecasting is still optimal interpolation, a simple, synchronous DA technique, where the model results are corrected using simultaneous observations accounting for both model and observation errors. Although the observations are local, the corrections are spread over a larger area (see e.g. Lionello et al., 1995). In numeric weather prediction (NWP) the most commonly used synchronous DA technique is three-dimensional Variational (3D-Var) which, contrary to optimal interpolation, can handle non-linear observation operators. Asynchronous techniques, such as 4D Variational (4D-Var) and Kalman filtering techniques, not only consider the errors in the observations and model results, they also take into account the dynamics of the models. 4D-Var is the DA technique used at the European Centre for Medium-range Weather Forecast (ECMWF) operational NWP model.

The application of new insights from the development and application of new data assimilation (DA) methods, e.g. asynchronous and variational, mainly from meteorology and oceanography, to wave forecasting can significantly improve the amount of information extracted from observations.

I.3 Objectives

In this report innovative assimilation methodologies being implemented in the framework of the WP2 of the MyWave project with coastal wave forecast goals are

described. The term innovative is used here not to indicate that the assimilation technique is new, rather that its application in the field of wave or non-hydrostatic atmospheric modelling) is novel.

I.4 Contributors to the report

The MyWave WP2 team members that have contributed to this report are Gert-Jan Marseille and Ad Stoffelen from KNMI on assimilation of scatterometer data in a non-hydrostatic atmospheric model (sections II.2 and III.2), Martin Verlaan and Sofia Caires from Deltares on Ensemble Kalman Filter data assimilation on a coastal wave model (sections II.3 and III.3) and Kathrin Wahle, Heinz Günther and Joanna Stavena, from HZG on neural networks data assimilation on a coastal wave model (sections II.4 and III.4).

II METHODOLOGIES

II.1 Introduction

Data assimilation covers a range of methods to incorporate observations into a model to improve the accuracy of the model forecasts. In many ocean and meteorology applications, the main goal is to improve the accuracy of the model initial conditions by making use of the available observations. However, in regional applications the boundary conditions often play an important role and therefore need to be optimized by the data assimilation. Finally, observation biases and uncertain model parameters may also be estimated by the data assimilation.

In the 1950's the weather forecasting models were initialized with interpolated fields taken from observations. Obviously, it is difficult to interpolate observations in areas where there are only few observations. Later, it became clear that it was useful to use the results from the previous forecasts as a starting point for the interpolation of observations. Later this subjective procedure was replaced with an objective analysis based on statistical estimation theory, where the interpolation weights are computed based on assumptions about the accuracy of the model forecast and the accuracy of the observations. Major challenges for this approach were the estimation of the error covariance and the fact that the meteorological model often showed spurious oscillations as a result of an imbalance between the variables of the estimated initial conditions. Numerous adjustments to this basic procedure have been proposed to overcome these issues. Optimal Interpolation and Incremental 3D-VAR are two of the most popular techniques using this approach.

The next major step in data assimilation development was the extension of the estimations from three spatial dimensions in order to include time as a fourth dimension. This allows observations to be included at the time the observation was taken and limits the estimation to those 4D fields that are consistent with the model dynamics. 4D-VAR is the most well-known method of this type. It does, however, require the laborious determination of the adjoint of the model in question (Kalnay et al., 2007).

In 1960 a new method, now known as the Kalman filter, was proposed for the conceptually similar problem of state estimation by Kalman (Kalman, 1960). Although the Kalman filter quickly became popular in engineering, it was not possible to apply the method for data assimilation since the computational load grows too rapidly with the size of the model, which is often of 10^6-10^7 compared to order 10-100 in many engineering problems. Although there are some early applications (Heemink en Kloosterhuis 1990), it was not until the introduction of the Ensemble Kalman Filter or EnKF (Evensen, 1994, 2003) that Kalman filtering became widespread for data assimilation.

It is possible to unify most of the methods mentioned in one common framework that uses Bayesian estimation as its foundation. Variational methods attempt to find the maximum a posteriori probability by computation of the cost function gradient and applying optimization methods to search for the cost function maximum. Note that the most common approach is to minimize, for convenience, minus the logarithm of the a posteriori probability. Combined with the common assumption of Gaussian a priori probabilities this leads to a quadratic minimization problem. Kalman filtering takes another approach. The model and observation operator are assumed linear, so that the optimization can be solved analytically.

Since data assimilation is often quite computationally demanding many approximations have been proposed, such as Proper Orthogonal Decomposition (POD, Altaf et al. 2009). Also aiming at computational efficiency, in this report a method based on Neural Networks is proposed.

In the next section, a 3D-var approach for assimilating scatterometer wind data into a non-hydrostatic atmospheric model is described. In sections II.3 and II.4, EnKF and neural networks approaches for data assimilation in coastal wave models are described, respectively.

II.2 3D-Var assimilation of scatterometer data

Three-dimensional variational analysis (3D-Var, Courtier et al., 1998) is an incremental DA method where the analysis increment is found by iteratively finding the minimum of the cost function J:

$$\begin{aligned} J &= J_b + J_o \\ &= (x_b - x)^{\mathrm{T}} \mathbf{B}^{-1} (x_b - x) + (y - Hx)^{\mathrm{T}} \mathbf{R}^{-1} (y - Hx) \end{aligned} \qquad \text{Eq. II-1}$$

with x_b the vector model background state (also denoted as first guess), obtained from a short-term forward model integration from the previous cycle, y is the vector with all observations, H is the (non-linear) observation operator that relates observed values with the model state x. $\mathbf{B}$ and $\mathbf{R}$ are the positive definite background error covariance and observation error covariance matrices, respectively. The non-linear observation operator is generally considered as an interpolation operator from model grid to observation location but may, for instance, also include highly non-linear relationships between observed satellites radiances and model state temperature and humidity. The first term on the right hand side of Eq. II-1 is denoted the background error and the second term-the observation error. The objective of 3D-Var is to find the model state, x, also denoted as the analysis, x_a, which minimizes the cost function Eq. II-1. Under certain assumptions it can be shown that the analysis is obtained from

$$x_a = x_b + \mathbf{K}\left(y - Hx_b\right), \quad \text{with } \mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}$$

, Eq. II-2

where the optimal weight matrix **K** is also denoted the Kalman gain and **H** the linearization of the non-linear operator *H*. In practice, the incremental formulation is used to solve the optimization problem by solving the increment $\delta x = x - x_b$ from

$$\left(\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)\delta x = \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\left(y - Hx_b\right), \qquad \text{Eq. II-3}$$

where $(y - Hx_b)$ is known as the innovation. It can be shown that $x_a = x_b + \delta x$ is the minimum variance solution of Eq. II-1, with analysis error covariance matrix $\mathbf{A}$ which fulfills the relationship

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} \qquad \text{Eq. II-4}$$

Since **B** and **R** are positive definite matrices it follows from Eq. II-4 that each observation adds information (for non-zero **H**) and thus contributes to a reduction of the analysis error covariance **A**.

In a cycled 3D-VAR experiment the forecast initiated with the analysis of the previous cycle is used as first-guess in the current cycle. The non-hydrostatic atmospheric model used in this study and in which we plan to assimilate scatterometer data is HARMONIE (http://hirlam.org/index.php/documentation/harmonie). The cycling interval of the HARMONIE is 6 hours, implying 4 cycles per day that are centred at 00, 06, 12 and 18 UTC, i.e., the analysis times. The 3D-Var assimilation scheme assumes that all observations are valid at the analysis time. Observations from for instance SYNOP stations are available at one hour intervals, but only those observations closest to the analysis time are used in the analysis.

The text below is extracted from de Haan et al. (2013) but added here for completeness to explain the assimilation of scatterometer ocean wind in the HARMONIE model.

The component terms in Eq. II-1 are quadratic forms expressing the ‘distance’ between the analysis state and the prior or background state and observations respectively. The cost function J_o comprises the contribution of individual observation types, i.e.,

$$J_o = J_{o,SYNOP} + J_{o,ASCAT} + \cdots \qquad \text{Eq. II-5}$$

For ASCAT ocean surface winds from the Metop-A/B satellites the cost function is defined as

$$J_{o,ASCAT} = \sum_{j}^{N_{obs}} \left(\sum_{i=1}^{N_j} J_i^{-p} \right)^{-1/p} \qquad \text{Eq. II-6}$$

where

$$J_i = \left(\frac{u - u_i}{\sigma_{o,ASCAT}} \right)^2 + \left(\frac{v - v_i}{\sigma_{o,ASCAT}} \right)^2 - 2 \ln P_i \qquad \text{Eq. II-7}$$

is the cost of the *i*-th ambiguity. N_j is the number of ambiguities in observation *j*, (u, v) and (u_i, v_i) are the analysis and ASCAT wind vector ambiguity components respectively, σ_0, ASCAT is the expected standard deviation of the error in the ASCAT wind components with a value of 1.8 ms^{-1}, P_i is the a-priori solution probability and p is an empirical weight factor for the ambiguities which currently has the value of four. This weight factor emphasizes the discrimination between the ambiguities and makes the expression for the cost function behave more as an 'if'-statement.

II.3 EnKF data assimilation in SWAN

II.3.1 Introduction

SWAN (Booij et al. 1999, http://swanmodel.sourceforge.net/) is a wave model currently used by the Dutch Government for North Sea wave forecasts. Although, currently no data are assimilated into the model, the assimilation of data in the model using EnKF is under consideration. Next we briefly describe EnKF data assimilation and how it has been implemented for SWAN in openDA (http://www.openda.org).

II.3.2 EnKF data assimilation

The formulation of the EnKF (Evensen 1994 and 2003) is quite similar to the 3D-Var method introduced in the previous section. The main difference is that the static background error covariance (**B** in equation II-1) is replaced by the sample covariance. This sample covariance is computed from an ensemble of model forecasts in a procedure very similar to Monte Carlo methods.

Starting from an initial ensemble of model states $\xi_i^a(t_0)$ the model is used to compute a forecast for each ensemble member:

$$\xi_i^f(t_{k+1}) = M \xi_i^a(t_k) + w_i(t_k), \qquad \text{Eq. II-8}$$

where $w_i(t_k)$ denote the system noise, used to model uncertainties in the model. From this one can compute the sample mean as

$$x^f(t_k) = 1/n \sum_{i=1}^{n} \xi_i^f(t_k) \qquad \text{Eq. II-9}$$

and covariance

$$\mathbf{P}^f(t_k) = 1/(n-1) \sum_{i=1}^{n} (\xi_i^f(t_k) - x^f(t_k))(\xi_i^f(t_k) - x^f(t_k))' . \qquad \text{Eq. II-10}$$

Similar to the previous section the Kalman gain is expressed as

$$\mathbf{K}(t_k) = \mathbf{P}^f(t_k)\mathbf{H}'(\mathbf{H}\mathbf{P}^f(t_k)\mathbf{H}' + \mathbf{R})^{-1}, \qquad \text{Eq. II-11}$$

where $\mathbf{H}$ denotes the observation operator that maps the model state to values that match the observations. $\mathbf{R}$ is the error covariance of the observations at time t_k.

The analysis or measurement-step of the EnKF uses a perturbation of the observations $v_i(t_k)$ and a separate analysis for each of the ensemble members to obtain a consistent ensemble of states that incorporate the observations $y(t_k)$,

.

$$\xi_i^a(t_k) = \xi_i^f(t_k) + \mathbf{K}(t_k)(y(t_k) - H\xi_i^f(t_k) - v_i(t_k))$$

Eq. II-12

If required one can obtain the mean and covariance of the model state after the analysis, that can be computed from

$$x^a(t_k) = 1/n \sum_{i=1}^{n} \xi_i^a(t_k), \qquad \text{Eq. II-13}$$

and

$$\mathbf{P}^a(t_k) = 1/(n-1) \sum_{i=1}^{n} (\xi_i^a(t_k) - x^a(t_k))(\xi_i^a(t_k) - x^a(t_k))' . \qquad \text{Eq. II-14}$$

Note that the classical EnKF formulation requires the model simulations to stop each time an observation is available. In the asynchronous EnKF (Sakov et al., 2010) this restriction is relaxed. The observations are accumulated over a predefined time interval $]t_k, \ldots, t_{k+m}]$. During the model forecast for each member the matching values $\mathbf{H}\xi_i^f(t_k), \ldots, \mathbf{H}\xi_i^f(t_{k+m})$ are collected. Finally the observations are assimilated all at

once as if they occurred at time t_{k+m}, but with the predicted values that were collected at the appropriate times.

II.3.3 Model uncertainty or system noise for spectral wave models

Since spectral wave models are stable forced systems it is crucial to include system noise or model uncertainty. Without system noise or covariance inflation the Kalman filter will diverge, i.e., the error covariance matrix $\mathbf{P}^f(t_k)$ will become smaller and smaller and the observations will effectively be ignored.

Two likely sources of uncertainty in a spectral wave model are the uncertainty in the wind forcing and uncertainty for the wave parameters that are specified at the open-boundary. For the wind forcing we have assumed that errors in *x* and *y* directions are independent. For each component the errors are assumed to be spatially and temporally correlated with an exponential decay with distance and time-difference. For the open-boundary only an exponential temporal correlation is applied. The parameters are interpolated in space from a limited number of support points to the other grid cells at the boundary. It may be necessary to include a spatial correlation in the future.

II.3.4 Implementation of SWAN in OpenDA

OpenDA is a generic toolbox for data assimilation. It includes, among several other algorithms, an implementation of the EnKF with the option for asynchronous filtering as described above. The easiest and most flexible way to connect a model to OpenDA is with what is called a black-box wrapper. The characteristics of a black-box wrapper are that the model remains a separate executable with interaction by the input and output files of the model, see Figure II.1. For this purpose one needs to supply subroutines for reading and writing of these model specific file formats. These routines have been implemented for SWAN and are made available through the official OpenDA release (http://www.openda.org). The examples provided there should run out of the box on windows and linux platforms.

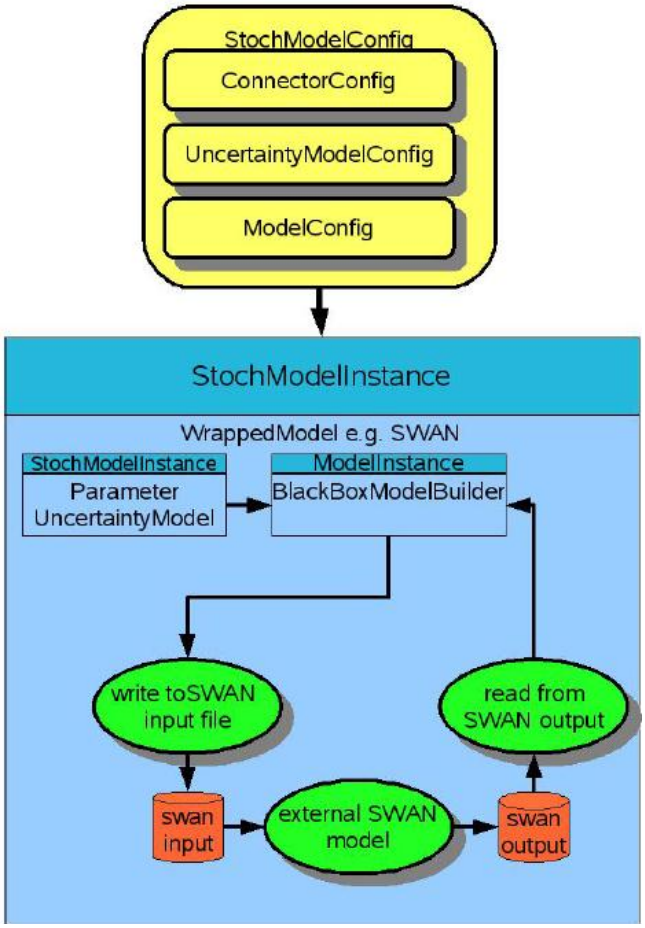

Figure II.1 Schematic diagram of the black box connection between SWAN and OpenDA.

In the EnKF, the model forecasts for each of the n ensemble members, with n typically of the order 100, is quite computationally demanding. Fortunately each of the model runs is independent and can thus easily be computed in parallel. OpenDA provides this functionality models with a black box wrapper. The model runs are distributed over a specified number of nodes using a round robin schedule, see Figure II.2. The processing of observations in the analysis of the EnKF cannot be performed in parallel yet in OpenDA. In the experiments shown in this document the analysis is computed on a separate computational node. It is likely that this will limit the scalability of the parallel computation.

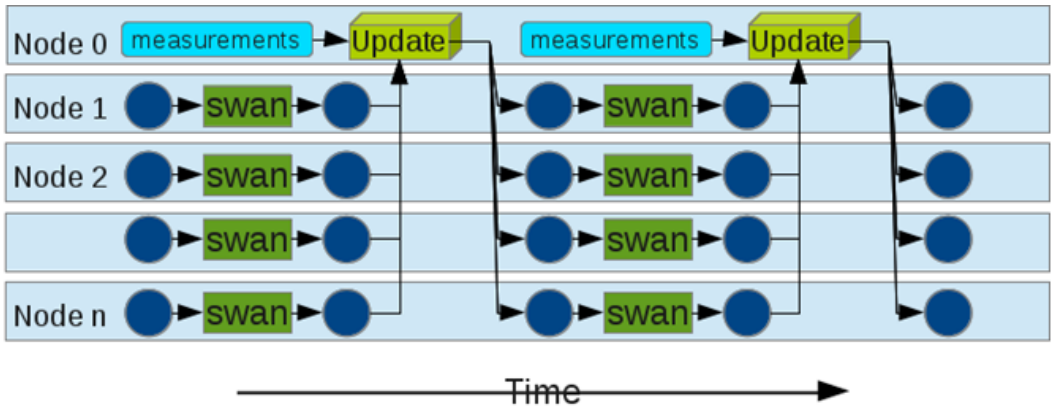

Figure II.2 Currently available computational schedule of EnKF for SWAN in OpenDA.

II.4 Assimilation using Neural Networks

II.4.1 Basic Idea

The neural networks (NNs) approach is a novel method for data assimilation into the wave models. In general the NN aims to explore an extensive parallel network of simple elements in order to obtain result in a very short time and, at the same time, with insensitivity to loss and failure of some of the elements of the network. These unique properties make possible to use the NN in a wide range of applications, e.g. remote sensing (e.g. Schiller, 2007), financing, engineering, image processes, recognition of patterns, etc. Detailed description of the NN method can be found in Haykin (1994) and Bischop (1995)

Neural Networks can be used to approximate an arbitrary non-linear function that maps a vector of input variables to a vector of output variables. It is also possible to use previous outputs of the NN as input for a next step of the computation, but these recursive NN are not considered here. The application of a NN can be divided into a training phase and a forecasting phase. During the training phase a large dataset of input and output vectors are used to train the NN, i.e. to estimate the coefficients and structure of the NN. The training phase consists of adjusting the weights for the best performance of the network in establishing the mapping of many input/output vector pairs.

Contrary to physically based models, with the NN it is not necessary that the relation between inputs and outputs is causal, a statistical relation like correlation is sufficient. This gives an additional freedom to decide which variables are inputs and which variables are outputs. This unique property of NN makes possible to perform data-assimilation by simply changing the input and output variables for the NN. Where physically based wave models, such as WAM (http://www.hzg.de/institute/coastal_research/structure/system_analysis/KSD/topics/developments/003136/index_0003136.html), require wind and boundary conditions as inputs and provide

computed wave parameters on the grid-points as outputs, a NN can accept observed wave parameters as inputs and wind and boundary conditions as outputs. The technical changes for this are very small. The challenge is to select the right input and output variables that work well, since a NN will always provide an answer, but some choices can result in much more accurate results than others.

To understand the performance of NNs for data assimilation in the wave models it is difficult to estimate both wind and boundary conditions together. One first has to use the new methodology either for optimizing the wind forcing or the boundary conditions. In this report the estimations are limited to optimizing boundary conditions.

NNs provide a statistical estimation procedure and thus have similar properties to e.g. multiple linear regression methods. For example, if too many input variables are selected, with a limited set of training data, then the training may overfit the data. It is therefore necessary to reserve part of the available data for validation. The most obvious indication for overfitting is when the NN has a much higher accuracy for the training data than for the validation data. Another property is generalization, i.e. a NN may sometimes generate good estimates for new inputs (i.e. data with properties not well captured by the training data set), but there is no guarantee.

An extreme example is if the training data only covers calm weather, than the NN may perform poorly for storms. An important technique to reduce overfitting is to reduce the number of inputs. One way to do this is with Principal Component Analysis (PCA), sometimes also called POD, POP or EOF. Another way is to lump a variable e.g. for a whole boundary instead of allowing spatial variation.

A NN is usually trained for each scalar output variable separately. This makes it cumbersome to compute output for many output variables. The approach proposed here is to use an 'inverse' NN to estimate the wave parameters at the open boundary of the wave model from the observations. Next, these estimated boundary conditions can be used as input for a run with a physically based model, here WAM. It is also possible to train a forward NN to generate output for a limited number of output locations.

The combined procedure that we are developing works as follows: First train a forward NN, with wind forecasts and boundary conditions from a larger scale model, e.g. significant wave height (H_s), mean wave period, etc. at a number of locations along the open-boundary. To reduce the number of inputs for wind and boundary conditions, a PCA is used.

The outputs of the forward model are given by the outputs of the wave model corresponding to the actual observations. This implies that the forward model will mimic the behaviour of the physically based wave model.

An additional inverse NN is trained with the same data, but with a reversed role. Here WAM-output matching the observations as input of the inverse-NN and boundary and wind PCA values are obtained as outputs. Note that the experiments with a preliminary version in this report use winds as an input for the inverse-NN. Note that

the training procedure does not require any real observations, but it does require model output for a reasonably long WAM run. In general, it is also possible to use the same training procedure, but with real observations.

After training of the forward-NN and inverse-NN one can perform a forecast by first running the inverse-NN with real observations, which results in an estimate for the open-boundary (and wind forcing). The forward-NN and/or the WAM model can then be used to compute the forecast. To forecast more than a few hours ahead the boundary-conditions and wind fields are complemented with forecasted boundary-conditions from a larger model and wind fields from a meteorological model.

The experiments in this report are performed with a synthetic dataset. The first-guess uses $H_s = 0$ at the open-boundary and the 'truth' model, that is used to generate training data and synthetic observations, uses wave parameters from the large scale North Sea WAM model for the boundary conditions.

This is a rather extreme case, which was chosen to clearly show the impact of the data-assimilation procedure. It is important to mention that after training NN has a lower computational cost than extended and linear KF, variational method, and particle filter.

Within the MyWave Project a new methodology based on the use of NN for data assimilation in the wave models is developed. The following assimilation schemes have been developed:

1. apply inverse WAM NN for each (point) measurement → ensemble of boundary values (and / or wind ensemble)

2. apply forward WAM NN for each ensemble member → emulated measurements in each point error (quality) estimate

3. from these 'ensemble' of boundary values chooses the best one in terms of error.

The idea for the data assimilation based on NN is schematically presented in Figure II.3.

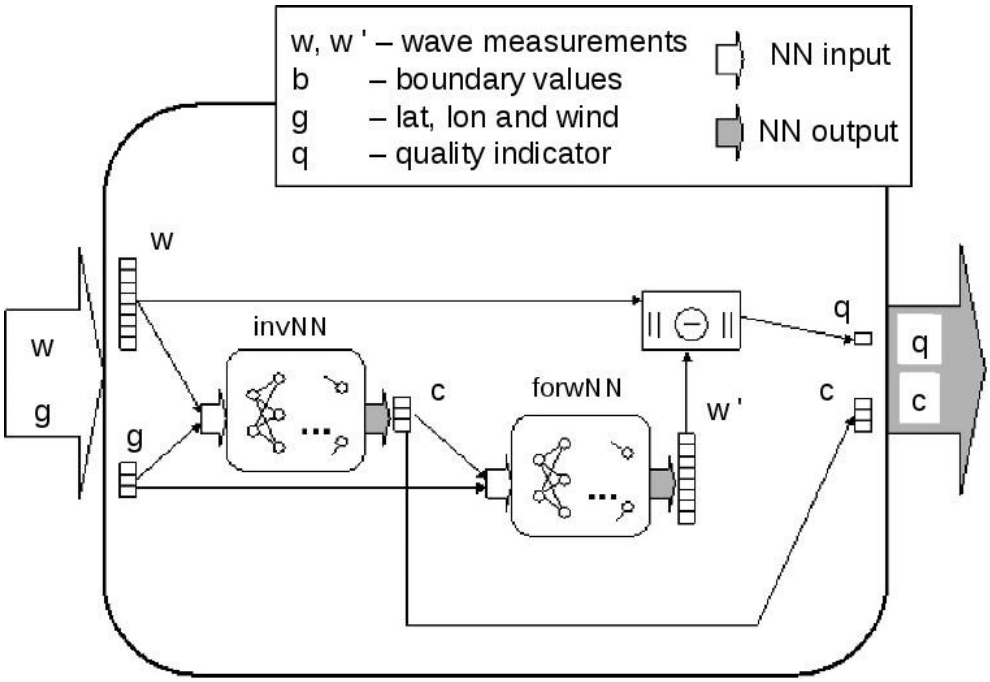

Figure II.3 Idea of assimilation scheme, w, w' are the wave measurements, b-boundary values, g-latitude, longitude and wind values, c-other parameters, q-quality indicator.

One big advantage of using NNs in data assimilation is its computational speed: once you have the NNs trained its further application costs little time. Additionally, the NNs scheme implies a quality/ out of scope check. These aspects make the use of NNs attractive compared to direct inversion of numerical models, like AdWAM by Hersbach (1998).

The methodology and preparation of data for the NNs data assimilation is described in the next two sections.

II.4.2 Preparation of data and training and testing the NN

The analyses have been done for the German Bight area as a test case. The WAM output files for the German Bight for the period from September 1st 2012 to June 30th 2013 has been used to train and test the Neural Networks. It is taken from the German Bight WAM version 4.5.3 model simulations (detailed information about the model set-up can be found in the pre-operational wave forecast system for COSYNA, www.cosyna.de)

As a first step the dimensionality (German Bight = 258x203 pixels) of the input/output data is reduced by performing principal component analysis for wind and boundary data. In either case the first PC describes well above 90% of the variance (see Table II.1). Therefore, we decided to take only the first 2 PC's into account.

	wind u comp.		wind v comp.		H_s bound north		H_s bound west	
PC 1	93.2%		91.0%		96.1%		95.4%	
PC 2	3.5%	(96.7%)	4.1%	(95.2%)	3.6%	(99.7%)	3.7%	(99.1%)
PC 3	1.1%	(97.8%)	1.9%	(97.0%)	0.2%	(99.9%)	0.6%	(99.7%)
PC 4	0.5%	(98.4%)	0.7%	(97.8%)				
PC 5	0.3%	(98.7%)	0.4%	(98.1%)				

Table II.1 Leading 5 (3) principal components (PC's) for approx. one year WAM data of wind and boundary significant wave height in the German Bight.

The wave data are available every three hours. For each time step the 'measurement' region was sampled randomly and approx. 20% of the maximum available number is usually been retained. Large values (three times higher than the year mean in this point) were always kept and even duplicated. (during the study period there were not many strong storms over the German Bight region). In this way a large training / testing table was compiled containing:

- date
- boundNorthPCA1 @ t=0, -3, -6, -9, -12h
- boundNorthPCA2 @ t=0, -3, -6, -9, -12h
- boundWestPCA1 @ t=0, -3, -6, -9, -12h
- boundWestPCA2 @ t=0, -3, -6, -9, -12h
- windNorthPCA1 @ t=0, -3, -6, -9, -12h
- windNorthPCA2 @ t=0, -3, -6, -9, -12h
- windEastPCA1 @ t=0, -3, -6, -9, -12h
- windEastPCA2 @ t=0, -3, -6, -9, -12h
- boundNorth_tm1 (*) @ t=0, -3, -6, -9, -12h
- boundWest_tm1 (*) @ t=0, -3, -6, -9, -12h
- boundNorth_thq (*) (cos and sin) t=-6h
- boundWest_thq (*) (cos and sin) t=-6h
- lonIdx
- latIdx
- hs@lonlat @ t=0, -3, -6h
- tm1@lonlat @ t=0, -3, -6h
- thq@lonlat (cos and sin) @ t=0, -3, -6h

(*) wave period (Tm1) and wave direction (thq) at the boundaries do not vary much with space; therefore information of Tm1 and thq was kept in the data table as mean value over boundaries and no PCA was done on these variables.

The wave travelling time through the German Bight area is 6 to 12 hours. Therefore in order to predict present wave height in the 'measurement' region past 6-12 hours of wind and boundary values are fully relevant.

From this large database a subset of approx. 90% (>600,000) was randomly chosen for training the various NNs. The remaining 10% (75,000) was used as an independent testing data set.

II.4.3 Forward WAM NN

Input to the Forward WAM NN are the northern and western boundary values (first two PC's of H_s, Tm1 and thq at one location) reaching 3 to 12 hours back in time, wind (first two PC'a of u- and v component) reaching 0 to 12 hours back in time and the location (lat-, lon index) of the wave 'measurement'. The needed output includes wave integrated parameters significant wave height, mean wave period (Tm1), and mean wave direction (thq) at the same location at the present time and reaching up to 6 hours back in time.

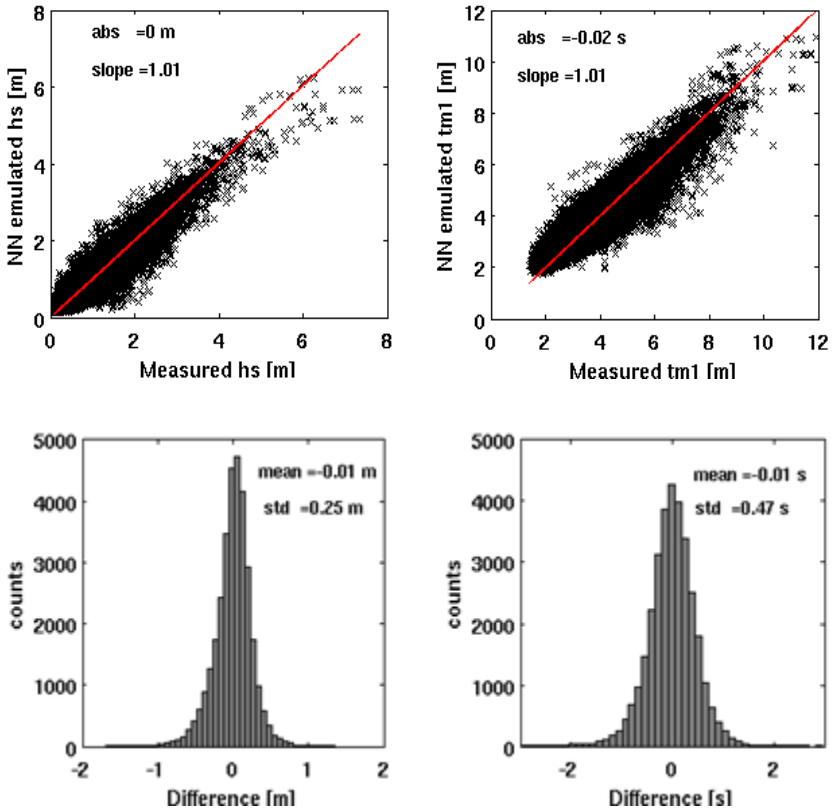

Figure II.4 Performance of forward WAM NN when applied to test data. Left: significant wave height. Right: mean wave period.

The performance of the NN for H_s and Tm1 (t=0) for all points in the testing set is presented on Figure II.4 .The root mean square error for both variables in each of the 'measurement' locations is shown on Figure II.5.

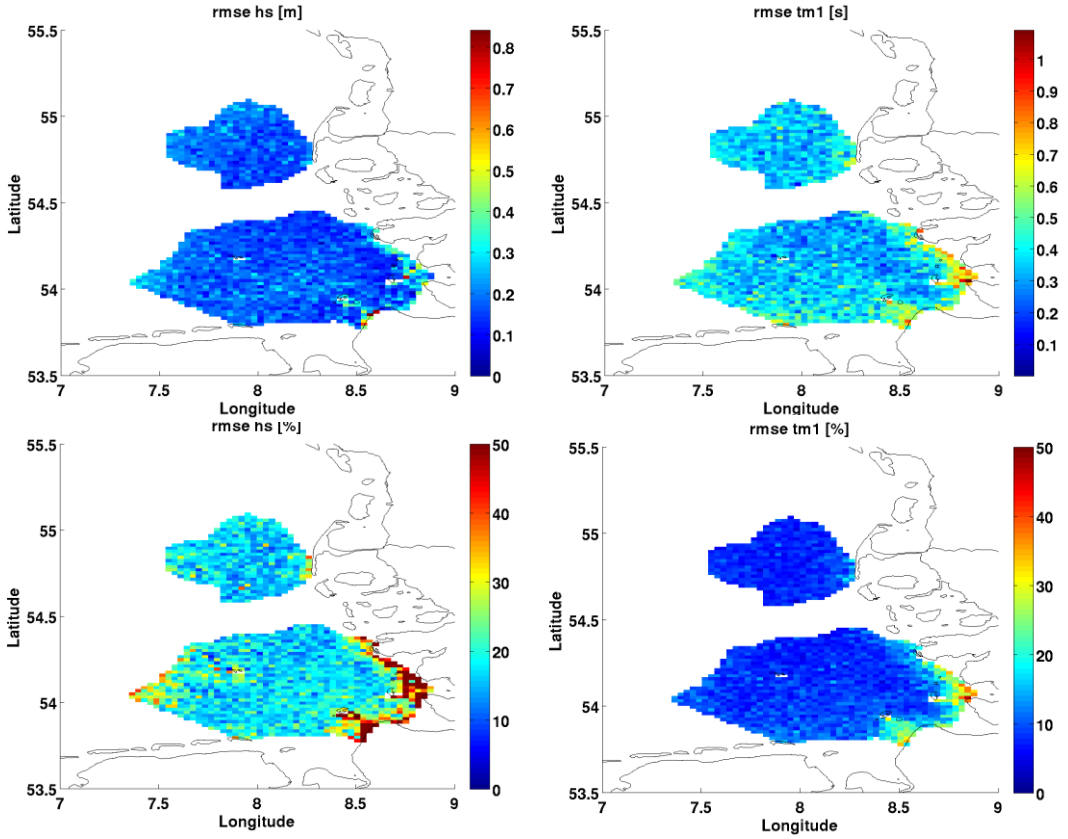

Figure II.5 Same as Figure II.6 but mean error at each 'measurement' location.

From the analyses it can be seen that: (i) there is a very good overall performance of the forward NN (Figures for t=-3 and -6h are very similar) (ii) the performance in the very shallow area (south east corner) is much weaker than the one in the off-shore area, which is probably due to the wave breaking processes.

II.4.4 Inverse WAM NN

As a first attempt we perform the exact inverse of what is described in Section 3 for the forward WAM NN, namely: as input to specify wave integrated parameters H_s, Tm1, and thq at a location (given by lat-, lon index) at the present time and reaching up to 6 hours back in time. The output are the northern and western boundary values (first two PC's of H_s, Tm1 and thq at one location) reaching 3 to 12 hours back in time, wind (first two PCs of u- and v component) reaching 0 to 12 hours back in time and the location (lat-, lon index) of the wave 'measurement'.

However, the results from this inverse NN showed errors that are of one order of magnitude larger than the corresponding forward NN experiment (Figure II.6a). The reason for these large errors is that the model is not invertible for this experiment: one and the same sea state might have been caused by different combinations of swell and wind sea. Thus we decided to change the methodology and to reduce the complexity of the inversion problem. In this way we thereby improve the NN performance by giving additional input information to the NN. Figure II.6b shows the errors of the inverse WAM NN when wind information was given as an additional input (= NN derives only boundary values). The performance of this experiment is considerably improved compared to the one of forward NN. Still, it is interesting to note that not only the PC's of boundary data of H_s fit well but also the reconstructed wave heights in the entire basin compare well with the target values (see Figure II.7).

The results demonstrate that the performance of the simplified inverse WAM NN is very promising. Additionally a second inverse WAM NN for deriving wind parameters (with boundary values as additional input) will be trained and these two NNs will be part of a future assimilation experiment.

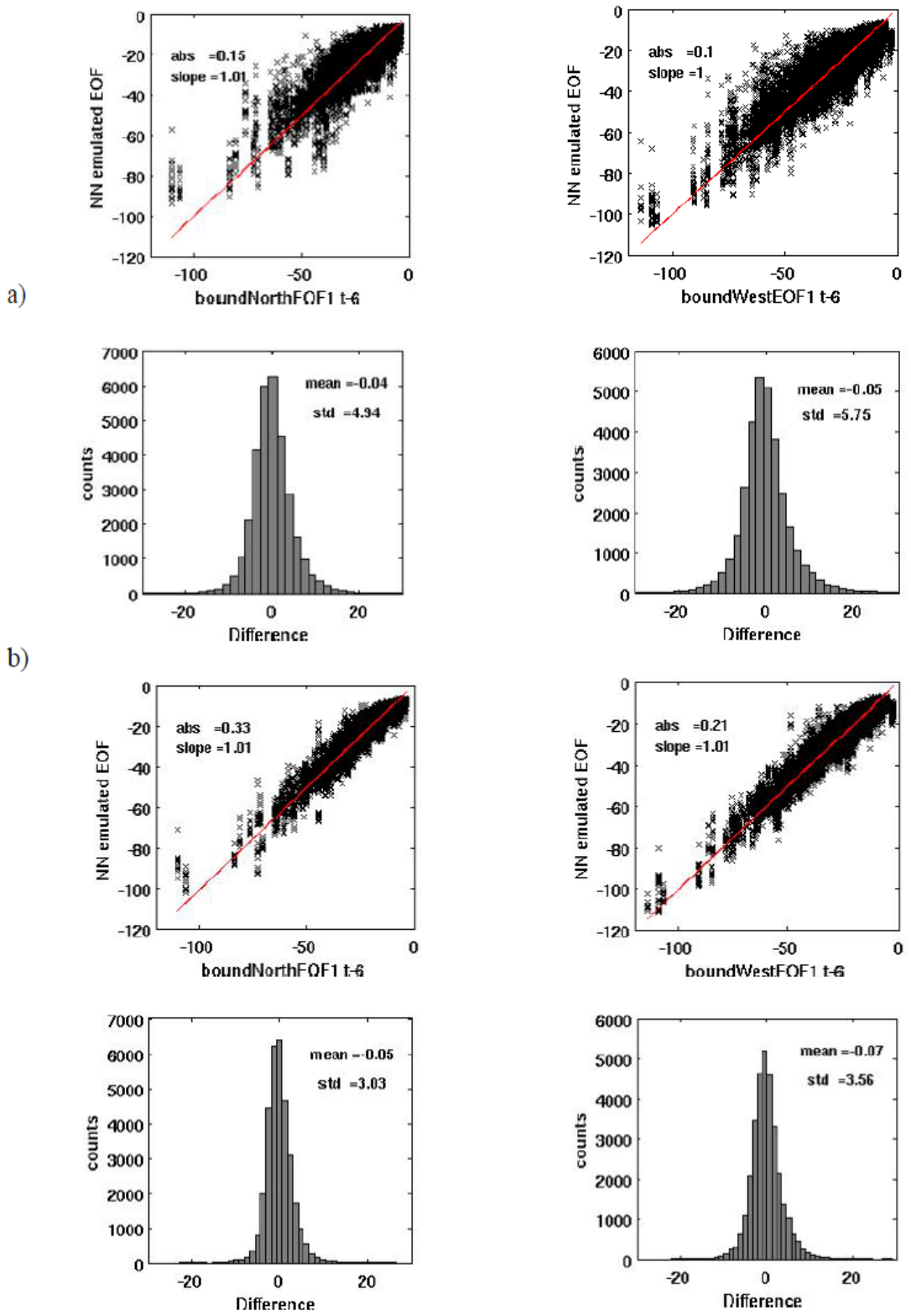

Figure II.6 Performance of inverse WAM NN when applied to test data. Left: dominant PC of northern boundary H_s, right: same for western boundary. (a) corresponds to first attempt of inverting the model (full inverse), (b) inverse WAM NN for emulating only boundary values.

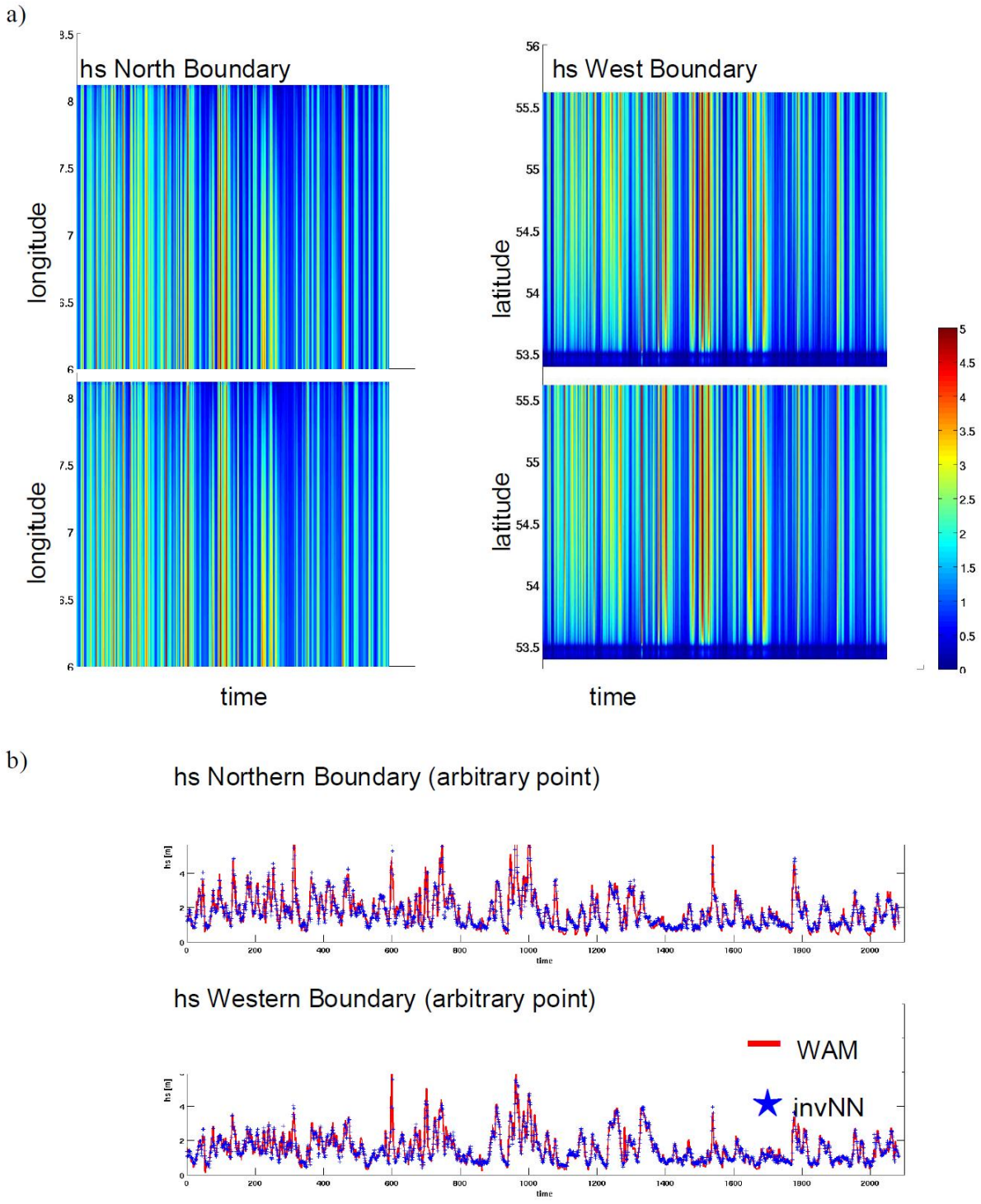

Figure II.7 Same as Figure II.6 but for reconstructed significant wave height. (a) along each boundary as function of time (top: WAM, bottom: inverse NN); (b) for a single point at each boundary.

III EXAMPLES

III.1 Introduction

In this chapter we present a number of trials carried out with the data assimilation techniques presented in the previous chapter. Most of these have been carried out using synthetic data.

III.2 3D-Var

Single observation experiments have been conducted with the HARMONIE model. Such experiments provide a good indication how the assimilation system spatially (both horizontal and vertical) spreads the information content of an observation in the model domain. Figure III.1 shows the resulting increment, also denoted structure function, from a 1 degree temperature innovation. The increment amplitude is maximum at the observation location with a value of about 0.35 degrees and decays by a factor of 7 at about 200 km distance in North-South direction and about 300 km distance in West-East direction from the observation location.

In a variational data assimilation system like 3D-Var not only the model temperature is corrected by a temperature observation but for instance also the flow is corrected following geostrophic balance equations, inherent in the background error covariance matrix. Figure III.2 shows the corresponding wind component increments resulting from a 1 degree temperature innovation at location 51 degrees latitude, 3 degrees longitude at 500 hPa pressure level.

Figure III.3 shows an example where satellite measured ocean surface winds deviate from the model first-guess at some locations. The resulting wind increment at the lowest model level (10m above the surface) in Figure III.4 clearly demonstrates the correction of the wind field of several ms^{-1} from assimilating scatterometer winds.

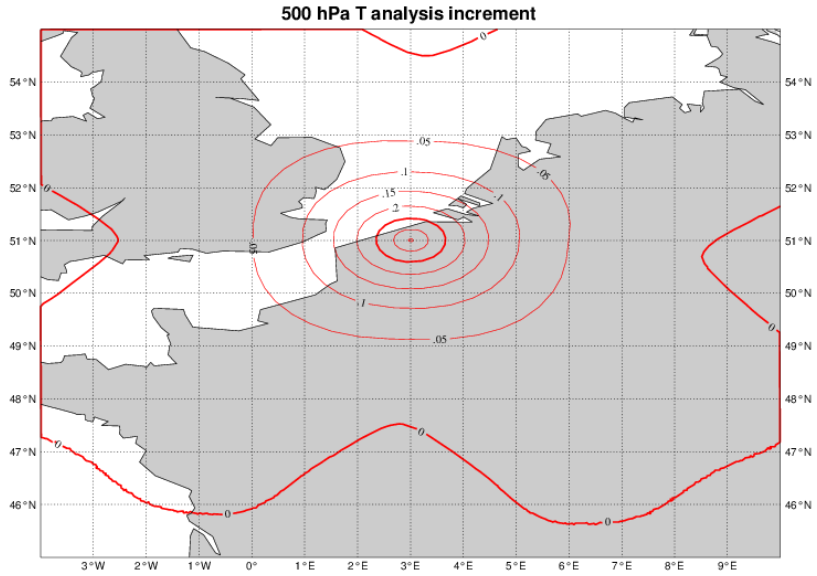

Figure III.1. Temperature analysis increment (Celsius) at 500 hPa resulting from a temperature innovation of 1 degree Celsius at location (lat,lon,pressure) = (51 degrees, 3 degrees, 500hPa).

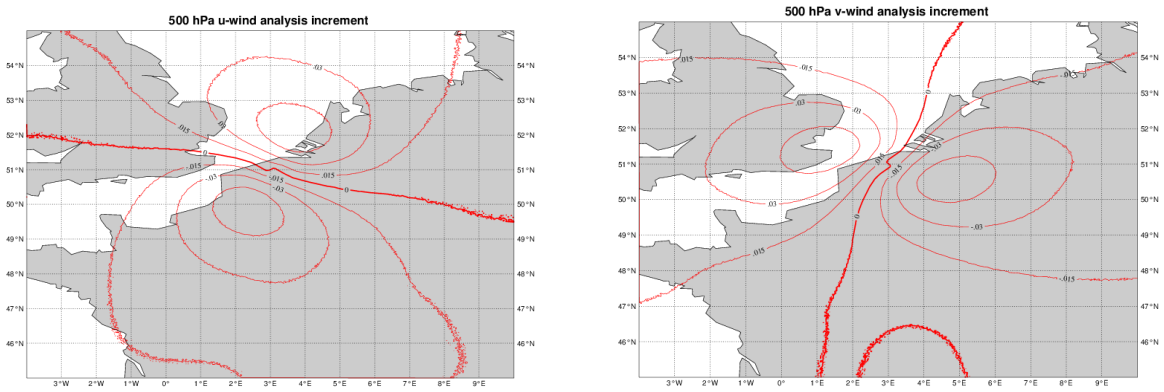

Figure III.2. Zonal wind (left) and meridional wind (right) increment (ms^{-1}) at 500 hPa resulting from a 1 degree temperature innovation at location (lat, lon, pressure) = (51 degrees, 3 degrees, 500hPa).

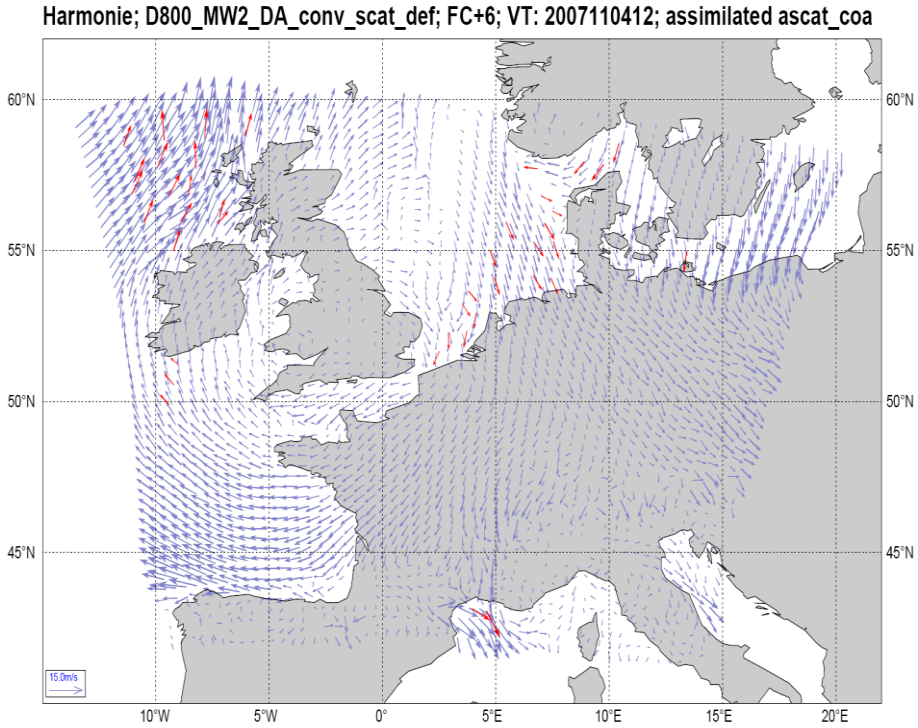

Figure III.3. HARMONIE 10m wind (purple) and assimilated ocean surface satellite winds from the ASCAT scatterometer on Metop-A (red).

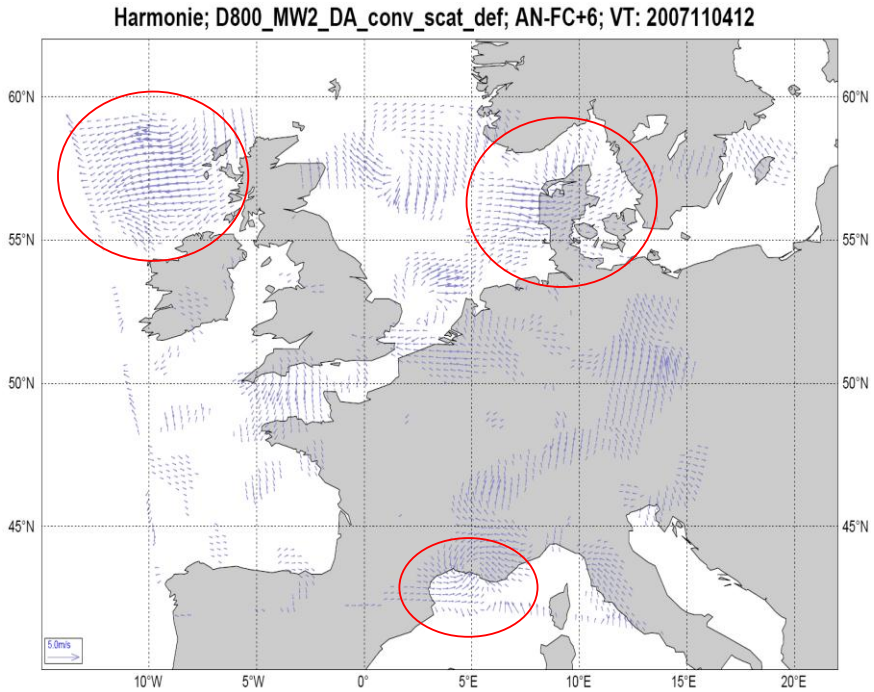

Figure III.4. 10m wind analysis increment from assimilated observations from radiosondes, aircraft, synop (ground) stations and ASCAT scatterometer. Increments inside the red circles are mainly from assimilated ASCAT winds, see also Figure III.3.

III.3 EnKF

In this section the EnKF for SWAN, as described in Section II.3, is applied to a number of experiments using small SWAN wave models for the North Sea. The considered experiments are *twin experiments,* meaning that the same model is used to create the observations and the model results and the differences between the observations and the model results are solely due to (known) differences in the model input or parameters. The aim of these North Sea twin experiments is to find the appropriate settings for this application and to gain insight into the sensitivities.

In the next section the general EnKF parameters are described. The twin experiments were carried out using simplified 1D and 2D SWAN models; these are described in sections III.3.2 and III.3.3, respectively. In Section III.3.4 a few conclusions are drawn.

III.3.1 Model parameters and settings

The ensemble Kalman filter is sensitive to a number of parameters, such as:

- number of ensemble members;
- assimilated data and their uncertainty; and
- uncertainty specification for wind and boundary (control variables).

Unfortunately, these parameters can only be tuned experimentally and there are also some interactions between them.

As a way of reducing computational time, OpenDA supports asynchronous filtering, where the analysis times are specified by the user instead of the times given by the observations. A larger interval between subsequent analyses reduces the computation time, but increases the analysis increments. This may deteriorate the accuracy, but it is uncertain to what extent.

Given that the presented experiments are twin experiments, a proper tuning of the noise parameters cannot be obtained, since it depends on the actual quality of observations and model.

In the twin experiments small models will be considered, these allow for many experiments because the runs are fast. On the other hand these models are a simplification of a real operational model. It is often efficient to use a small model to learn the behaviour of the system and then use this to reduce the number of experiments with larger models.

III.3.2 1D twin experiment

The SWAN spectral wave model can be run in a 1D mode, where the other direction is considered homogeneous. A convenient curve for a 1D trial of EnKF data

assimilation in SWAN in the Dutch North sea coast is the arc between the locations North Cormorant and Europlatform, see Figure III.5.

The considered SWAN grid follows a great arc at roughly 172 to 174 degrees with respect to North. This arc is divided in 25 grid points of 43.6063km. Several observation locations are available near this curve, see Table III.1.

Location	Longitude (°)	Latitude (°)	Distance to boundary (km)
North Cormorant	1.166	61.340	0
Anasuria	0.786	57.261	454
D151	2.934	54.325	787
K13	3.220	53.218	911
Europlatform	3.275	51.999	1047

Table III.1 Coordinates of the observation stations.

At each grid cell a wave spectrum with 32 frequencies and 36 wave directions is computed. The model time step is 1 hour.

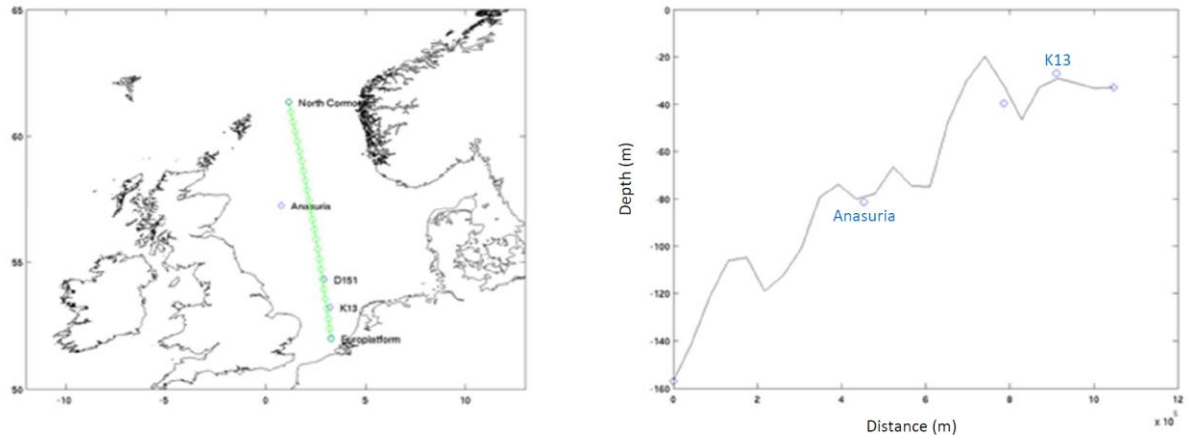

Figure III.5 Location (left) and depth schematization (right) of the used 1D SWAN model and observation stations.

III.3.2.1 Kalman filter settings

The Kalman filter setup includes uncertainty for the open boundary. Here an AR(1) model with exponential temporal correlation with scale 6 hours and standard deviation 0.4m for the significant wave height (H_s) is used. Similar perturbations can

be added to other boundary parameters, such as the wave-period, but this is not used for now.

Control variables/noise is also used for the wind forcing. The temporal correlation has a scale of 12 hours and a magnitude of 1.0m/s and the spatial correlation is of 500km. The two wind components are treated independently. The noise is specified on the same grid as the wind input, which may be different from the computational grid.

Observations of H_s are assimilated every hour at the 5 locations. The standard deviation for errors in the observations is set to 0.2m.

The considered test case is an identical twin experiment, that is the observations are generated with the same model, but with different model input. The purpose is to show that the data-assimilation works and that the observations are able to recover a significant part of the model errors.

The twin experiment is composed of 4 parts:

1. swell that is present at the boundary for the truth (observations), but not in the reference model

2. swell that is present in the reference model, but not in reality/observations

3. strong wind (and wind sea) that is present in the observations, but not in the reference model. The wind forcing is uniform over the domain.

4. strong wind (and wind sea) present in the reference model, but not in the observations.

The true and first guess model forcing are shown in Figure III.6. As can be seen in the figure, part 2. and part 4. are shifts in time of part 1. and part 3., respectively. We note that these simulated differences between the first guess and the observations are much larger than what we expect in reality. Nevertheless, we consider such large deviations in order to demonstrate the EnKF data assimilation impact more clearly and to test the stability of the method.

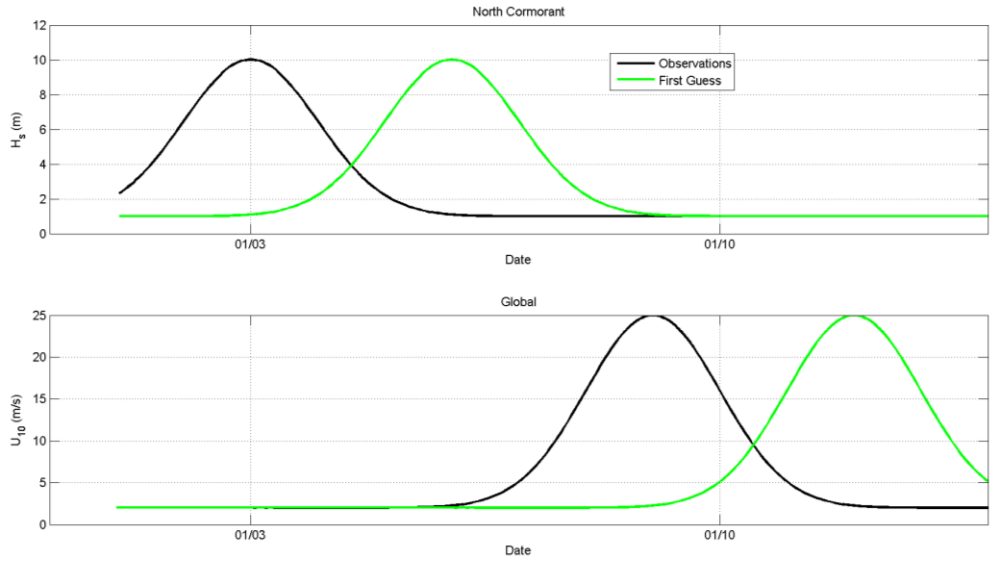

Figure III.6 Time series of boundary significant wave height and global wind speed first guess and observations.

III.3.2.2 Results

We have carried out three distinct data assimilation experiments, considering

1. Uncertainty only in the open boundary;
2. Uncertainty only in the forcing wind;
3. Uncertainty in both the open boundary and the forcing wind.

In all cases we have considered 80 ensembles in the EnKF.

Figure III.7, Figure III.8 and Figure III.9 show for the considered 5 observation stations (cf. Figure III.5), the first guess, observations and EnKF analysis time series of the significant wave height, mean wave period, $T_{m-1,0}$ [1], and wind speed, U_{10}, respectively.

We start by describing the differences between the first guess and the observations. At all locations there is in the observations time series a single U_{10} peak at the start of the period that is shifted forward in time in the first guess time series (cf. Figure III.9). At the North Cormorant boundary location there is in the observations time series a single H_s and $T_{m-1,0}$ peak at the start of the period (as imposed at the boundary) that is shifted forward in time in the first guess time series (cf. top left

[1] There are several parameters for describing the sea state period. One of these is $T_{m-1,0} = (m_1 / m_0)^{-1}$ where m_n, the *n* order spectral moment, is $m_n = \int_0^\infty f^n S(f) df$. Using different moments other period parameters can be defined.

panels of Figure III.7 and Figure III.8). As one moves in the direction of the coast, due to the wind forcing in the model domain a second H_s and $T_{m-1,0}$ peak appears (cf. Figure III.7 and Figure III.8). This wind sea peak gains importance in relation to the swell peak in the direction to the coast due to the larger wind fetch and the dissipation of the swell imposed at the boundary.

When considering only uncertainty in the boundary H_s (red lines in Figure III.7 and Figure III.8) the EnKF analysis is able to move the H_s peak at North Cormorant to the period in the observations. However, it does not seem to be able to move the $T_{m-1,0}$ peak and also produces a large and spurious wave height and period peak at the end of the period. These lead to large differences between the EnKF analysis and the observations at all locations. We note that this bad performance of the method may be due to instabilities since the differences between the EnKF analysis and the observations are lower if more ensembles are considered (not shown).

When considering only uncertainty in the forcing wind (magenta lines in Figure III.7 to Figure III.9) at the boundary North Cormorant location the differences between the analysis and the observed wave heights and periods are large, but at the other locations the analysis time series are much closer to the observations than the first guess time series. That the assimilation of wind only fails to produce the right results at the boundary is as expected given that in order to produce the right results at the other 4 locations the wind fetch should be none at the boundary. The analysis wind fields, which also need to counterbalance the differences between the boundary first guess and observation waves, differ more than the first guess winds from the observations, but do lead to analysis waves that are closer to the observations. In other words, observation of H_s alone is insufficient to fully distinguish between errors in boundary and wind forcing.

When considering uncertainty in both the forcing wind and the boundary H_s (blue lines in Figure III.7 to Figure III.9) the analysis (the same parameters that were adjusted when producing the first guess and the observations), the H_s (and $T_{m-1,0}$ to a lesser extent) is at all locations rather close to the observations. On the other hand, the differences between the analysis and observed winds can still be large, with the analysis time series showing wind speed peaks during the observed swell and first guess wind event. However, these do not affect the quality of the analysed waves.

In conclusion, when the chosen control variables are those with associated uncertainties then the assimilation of the wave observations is rather successful.

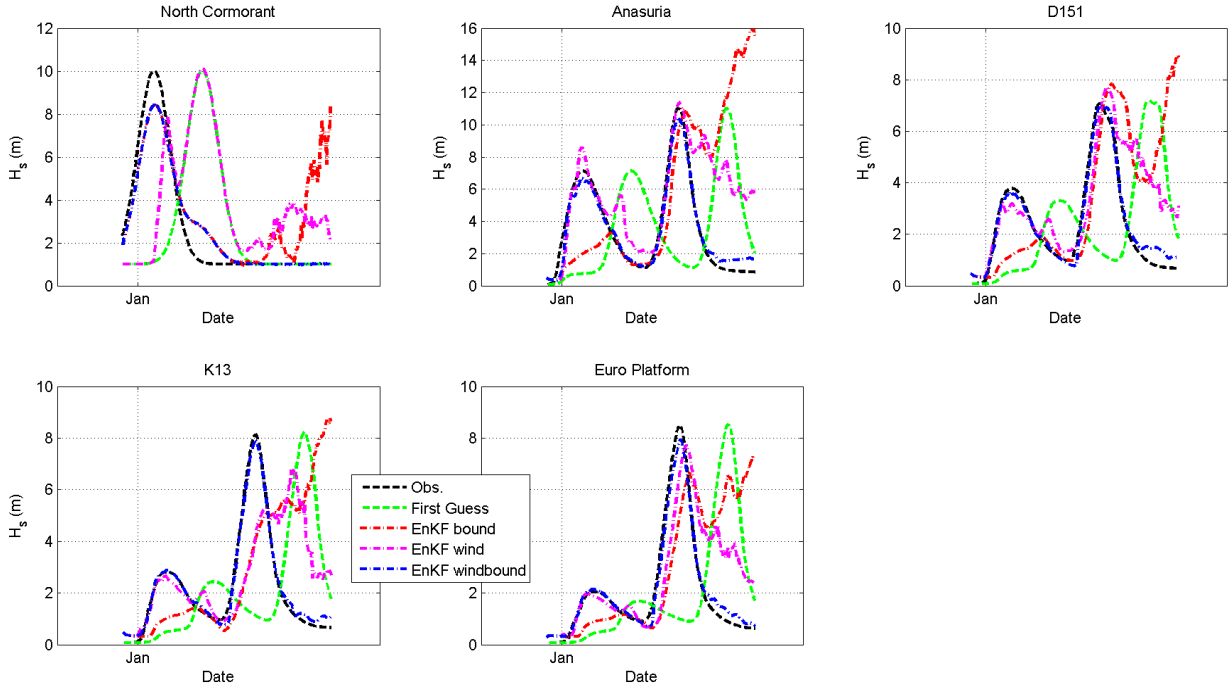

Figure III.7 Time series of 1D twin experiment significant wave height at the five North Sea observation stations.

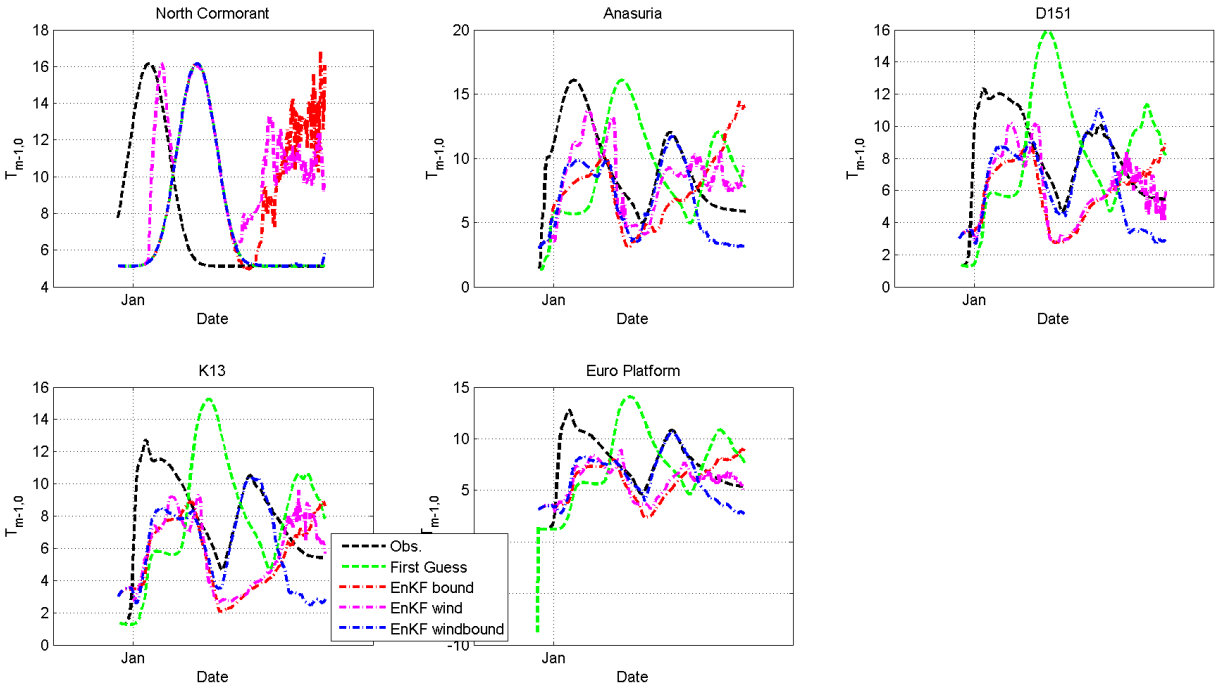

Figure III.8 Time series of the 1D twin experiment mean wave period at the five North Sea observation stations.

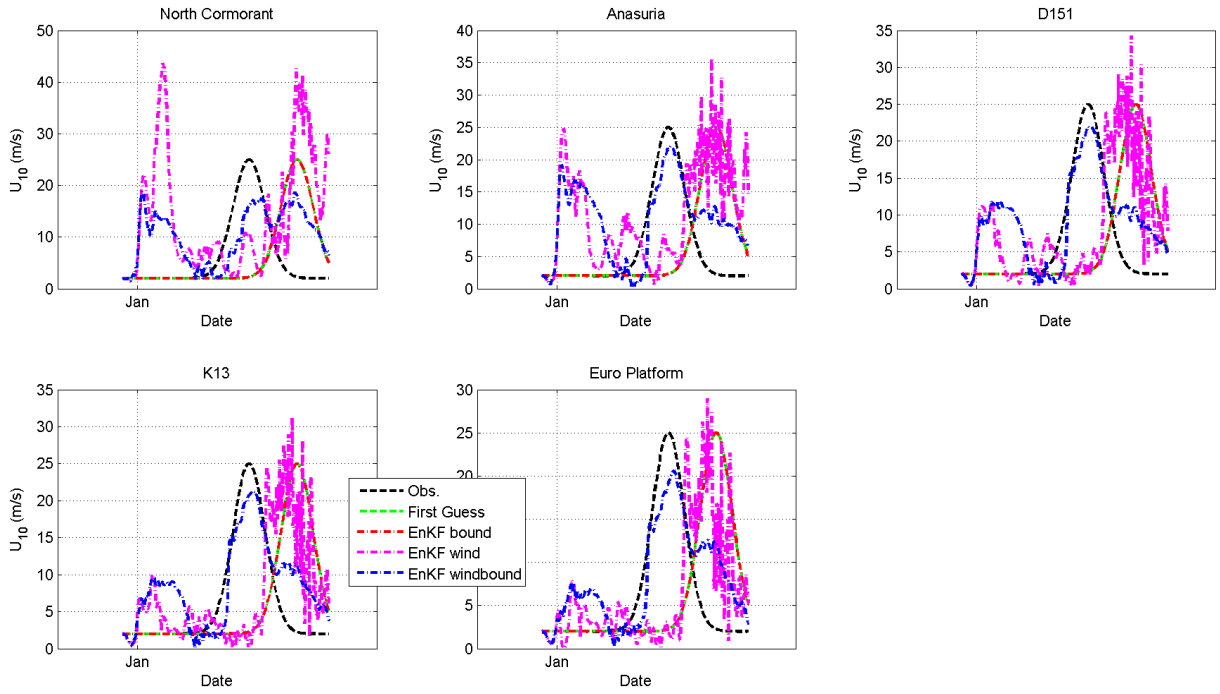

Figure III.9 Time series of the 1D twin experiment wind speed at the five North Sea observation stations.

III.3.3 2D twin experiment

For the purpose of extending the 1D model studied above and testing the performance of parallel computing and asynchronous filtering, a somewhat realistic albeit very coarse model has been created. Several versions of this model with the same domain, but a different resolution were made to test the scaling properties of SWAN and the Kalman filter. The coarsest version shown in Figure III.10 has a resolution of 0.5 degrees in both directions, which results in 27x21 grid cells. The bathymetry for the model has been interpolated from the NOOS bathymetry (http://www.noos.cc/), that has a resolution of 1/40° x 1/60°. Two somewhat finer SWAN models with resolutions of 1/4° x 1/4°and 1/8° x 1/8° are used to study the changes to performance for larger models. All three models have a spectral resolution of 32 frequencies and 36 directions, so the model has 32x36 dynamic model variables at each grid-cell.

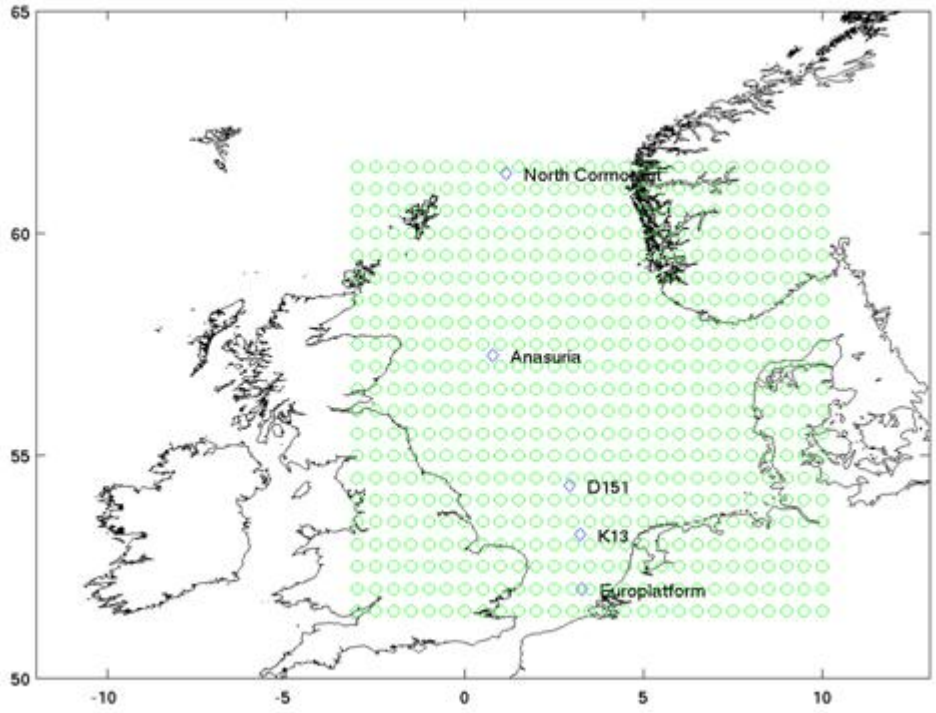

Figure III.10 Grid of the 2D SWAN model at 0.5° resolution and location of the observation stations.

Since the SWAN wave model is a stable and forced model mostly affected by the wind-forcing, in the 2D experiments the system noise for the Kalman filter is added only to the wind forcing. The temporal correlation has scale 12 hours, magnitude 1.0m/s and spatial correlation with scale 1000km. The two wind components are treated independently. The noise is specified on the same grid as the wind input, which may be different from the computational grid in SWAN.

Observations of H_s are assimilated for 5 locations with an hourly time-step and a standard-deviation setting of 0.1m in the ensemble Kalman filter. The Observations are generated with the same model, but with very different boundary conditions and wind forcing. Again, the magnitude of the differences is much larger than expected in reality, but this is only intended to see how the ensemble Kalman filter behaves for extreme over prediction or under prediction of wind forcing.

The true and first-guess wind-fields are uniform in space and have a Gaussian shape in time. The peak value is 25m/s from the north, but the timing of the first-guess is completely off (3 days difference). Note that realistic forecast errors of 10-meter wind have an RMSE of around 2m/s, while the peak errors for the wind-forcing here are 25m/s.

We start by looking at the model results when using the coarser model. As can be seen in Figure III.11 to Figure III.13, although the differences between the observed and analysis winds can still be large (Figure III.13), the analysis wave heights (Figure III.11) and periods (Figure III.12) are quite close to the observations.

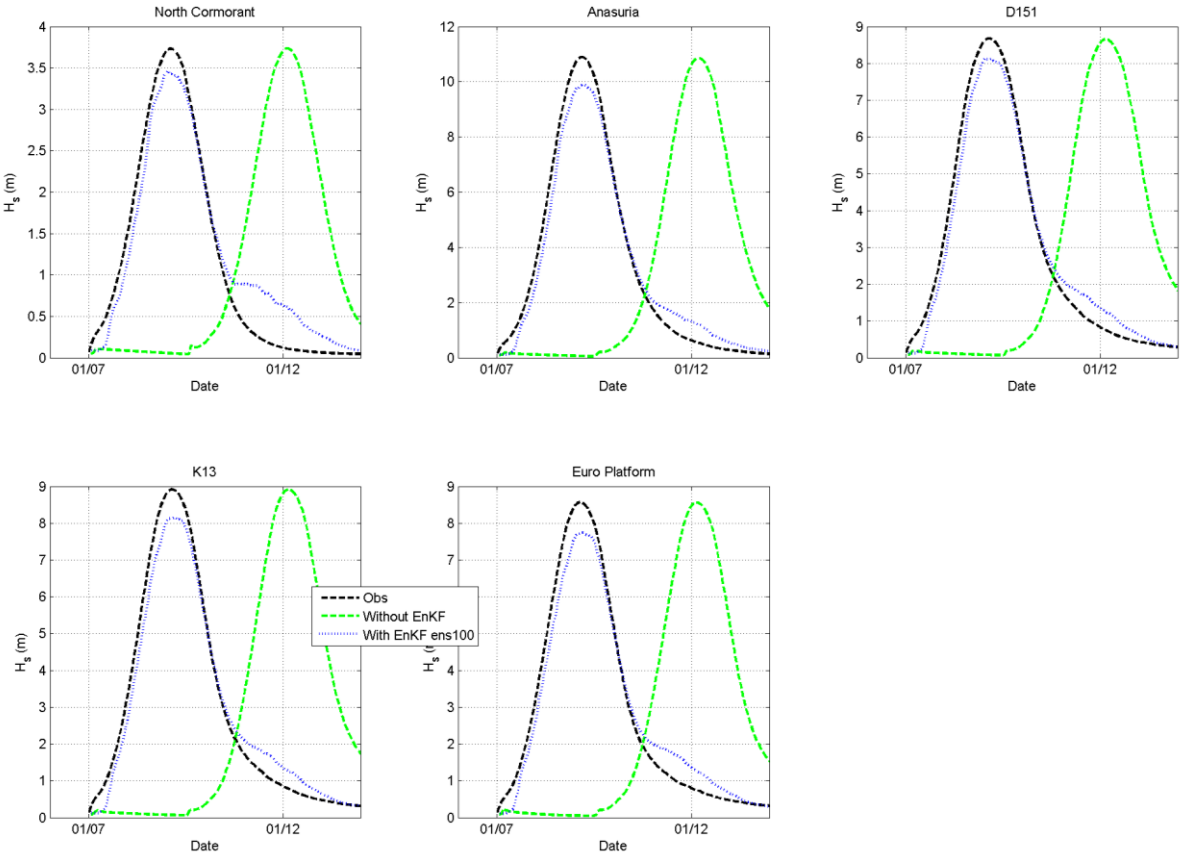

Figure III.11 Time series of 2D twin experiment significant wave height at the five North Sea observation stations.

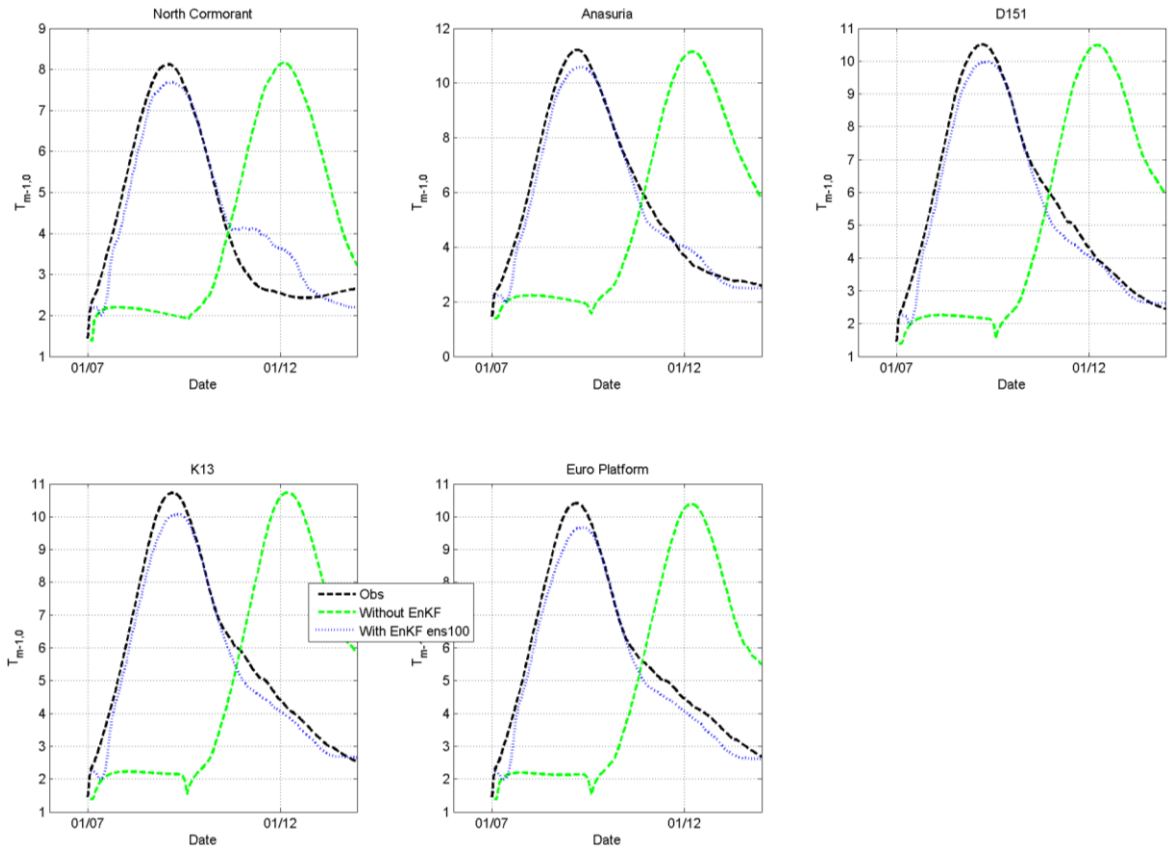

Figure III.12 Time series of the 2D twin experiment mean wave period at the five North Sea observation stations.

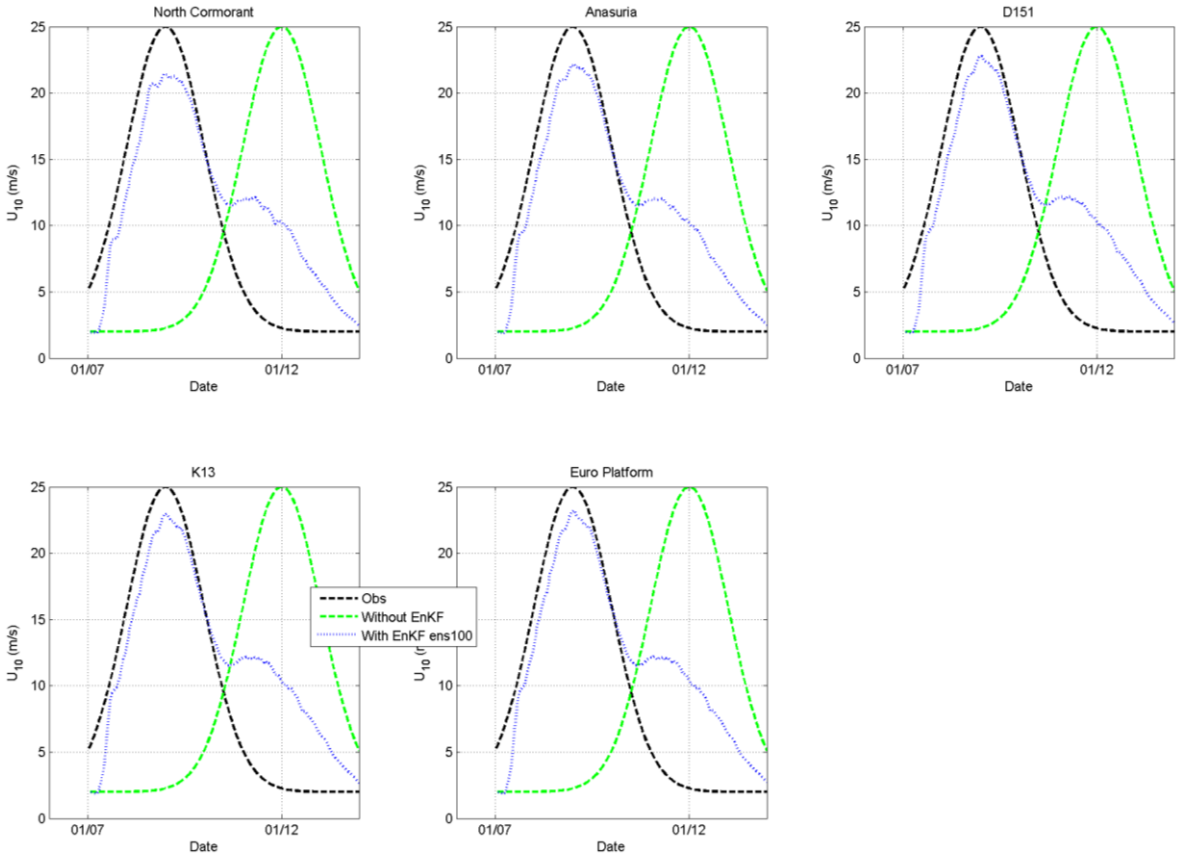

Figure III.13 Time series of the 2D twin experiment wind speed at the five North Sea observation stations.

III.3.3.1 Asynchronous filtering

We now present the results of an experiment analysing the impact of asynchronous filtering. Analysis intervals of 1, 3 and 6 hours were used. In Figure III.14 one can clearly see the larger steps that occur for larger analysis intervals. Larger intervals have a negative impact on the accuracy, but also require less computation time, because the number of model initializations is reduced by a factor of 3 or 6 compared to the 1 hour analysis interval. The asynchronous EnKF still retains a large part of its positive impact for 3 and even for 6 hour analysis intervals.

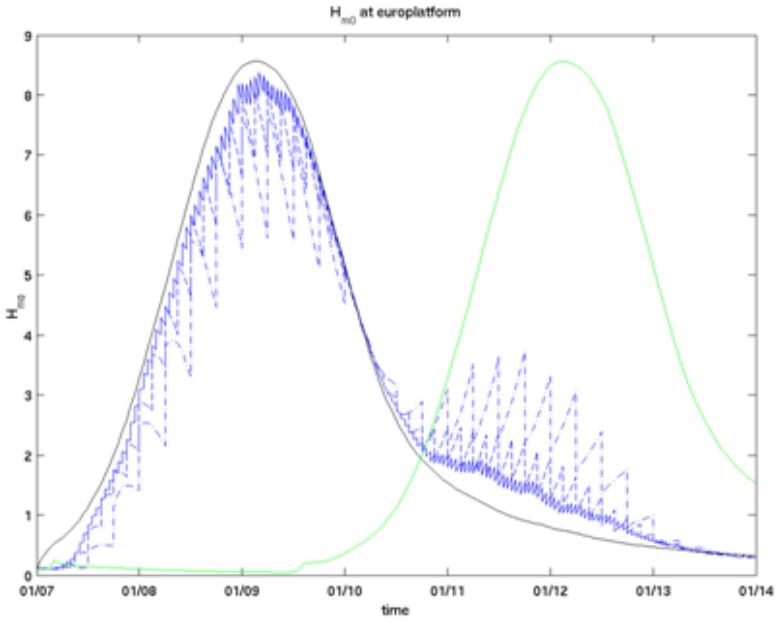

Figure III.14 Time series of the 2D twin experiment with asynchronous filtering significant wave height at the Euro platform.

III.3.3.2 Parallel computing

Next we performed some performance tests on the Lisa cluster of SARA (https://www.surfsara.nl/nl/systems/lisa/description). The nodes used had Intel Xeon L5520 processors with clock-speed 2.26GHz and 8 cores per node. The first set of experiments were for the coarse model with 0.5° resolution and 27x21 cells, so the model state will contain 27x21x32x36=163,184 values.

The table below shows the results for a simulation of one week. Clearly using more nodes for the computation speeds up the simulation until the communication overhead reduces and eventually halts further improvement. In these computations the measurement update is performed sequentially, thus according to Amdahl's law this sequential part will become dominant. By increasing the time interval between updates from 1 to 3 and 6 hours the sequential part is roughly reduced by a factor 3 and 6.

#Nodes	#Cores	Timing 1hr [min]	Timing 3hr [min]	Timing 6hr [min]
1+1	16	586	254	201
2+1	24	298	178	106
4+1	40	187	89	59
8+1	72	133	61	38
16+1	136	101	46	27
32+1	264	106	37	21
64+1	520	117	42	25

Table III.2 Wall clock times [min] for asynchronous EnKF with coarse 0.5° resolution model.

The speed increase by parallel computing allows us to increase the size of the model and still run with realistic computation times. To test the impact of a larger model the runs of the previous experiment were repeated with models of 0.25° and 0.125° resolution instead of 0.5°. By increasing the model size we expect that both computation time and communication time scale nearly linear with the grid size. In the table below one can also see that the measurement update requires much internal memory, since storage of the ensemble of 64 model states grows to 5.0Gb for the 0.125° x 0.125° resolution. Apparently the present implementation keeps multiple copies during computation of the update, since the total memory use during this step was nearly 20Gb. This will be a severe limitation for further increasing the model resolution.

resolution	Number of grid cells	State dimension	Size of ensemble
0.5° x 0.5°	27x21x32x36	653,184	0.3Gb
0.25° x 0.25°	52x42x32x36	2,612,736	1.2Gb
0.125° x 0.125°	104x84x32x36	10,450,944	5.0Gb

Table III.3 Model sizes.

As expected the computation times grow nearly linear with the number of grid cells in the model. At the same time the communication time and the update step also grow linearly, so the speed up is not affected much. For further speed up the communication and scalar part of the computation must be improved.

#Nodes	#Cores	Timing 0.5° (min)	Timing 0.25° (min)	Timing 0.125° (min)
1+1	16	254	733	
2+1	24	178	404	1246
4+1	40	89	221	667
8+1	72	61	164	523
16+1	136	46	122	484
32+1	264	37	125	476
64+1	520	42	135	

Table III.4 Wall clock times for asynchronous EnKF with 3 hour updates.

III.3.4 Conclusions

When the chosen control variables are those with the true associated uncertainties then the assimilation of the wave observations using EnKF in a simplified North Sea SWAN models can be quite successful.

The black-box wrapper of OpenDA works fine in combination with the SWAN model and parallel computing. Experiments show that the combined effect of using OpenMP for SWAN on one node, parallel model runs managed by OpenDA and asynchronous filtering together amount an accumulated time-saving that allows one to use much larger models than without these options. All these options are already provided by OpenDA and SWAN, so no additional programming effort was needed, just a modification of the configuration files.

We were able to let the resolution of a simple SWAN model of the North Sea grow to 1/8°x1/8° or 104 x84 cells. Although this is not enough for the new operational model, it is already higher than the previous generation operational model and enough to perform experiments for further tuning of the configuration. Moreover, the experiments have pointed towards a number of possible further improvements:

- The main barrier for a further increase of the model resolution is the memory use of the OpenDA program on node 0. Although one may be able to reduce the memory use by a small factor, it is clear that only storage of the ensemble of states will quickly fill available memory when the model resolution is increased further.

- The communication between the models in the ensemble and the measurement update is significant and was implemented with a network disc in the experiments reported here. Future developments should explore the use

of faster communication methods that make better use of the available hardware.

- The measurement update now still forms a considerable sequential part of the computation. A parallel implementation of this part of the code can help the speed up for larger number of processors considerably.

- The restart-files are now in ASCII format. Replacing them with a binary format, such as netCDF may increase the speed at which these files are read and written.

- Finally, also changes to the EnKF algorithm or configuration can deliver a further speed-up of the computations. Mostly these aspects and improvements of the parallel computation can be developed simultaneously and can be used in combination.

In summary, the twin and scaling experiments described in this report have been very useful both to deliver a configuration that can be used for further developments of the data assimilation and they have pointed towards a number of issues that can be improved further.

III.4 Assimilation using Neural Networks

The quality of the existing HF wave radar data for the German Bight is still insufficient to make use of in a real assimilation experiment. Therefore, the above described NN assimilation method is tested using synthetic HF radar wave data.

In the first experiment, the 'measurement' data are taken from a WAM output for the German Bight for July 2013. These data have neither been used for testing nor for training of the NN. Additional data to apply the algorithm (wind data) and to validate it (boundary data) are also taken from this run (the 'truth' run).

As a second step, the model run is repeated without any boundary values, i.e. no waves entering the German Bight ($H_s = 0$). This second run served as 'first guess'.

The 'measured' values of the wave parameters (H_s,Tm1 and thq) within the HF radar area (**n** ~ 1,000 grid points) together with the wind data have been given as an input to the inverse NN. As an output an ensemble of boundary values p for each point in time **n** is being produced. Feeding these values into the forward NN and comparing the output with the 'measurements' gives **n** error values (one for each of the **n** grid points). The error is calculated as relative error between 'measured' and forward NN emulated significant wave height. The error distribution in a typical situation is demonstrated in Figure III.15. The 'best' boundary values are calculated as the mean of the 50 with the smallest error. The chosen 'best' boundary values together with the 'truth' for the whole assimilation period is shown in Figure III.16. A comparison of first guess and (NN approximated) assimilation error for significant wave height and mean

wave period in the HF radar region for the German Bight (region 'known to' NN) is demonstrated in Figure III.17

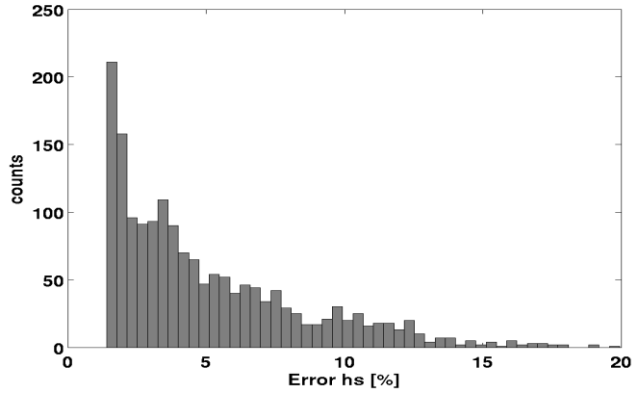

Figure III.15 Error in NN derived H_s (after consecutive applying invNN and forwNN) compared with 'measured' values.

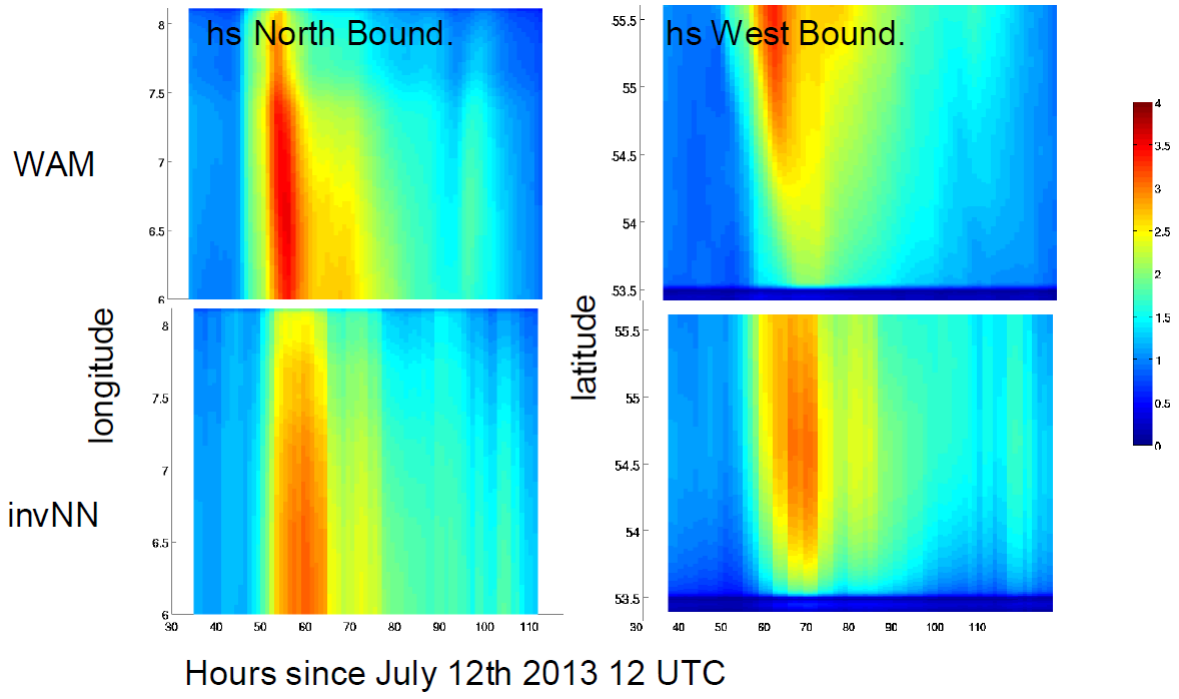

Figure III.16 Boundary values taken from a model run (top) and as emulated by the inverse WAM NN for the time period of the assimilation experiment.

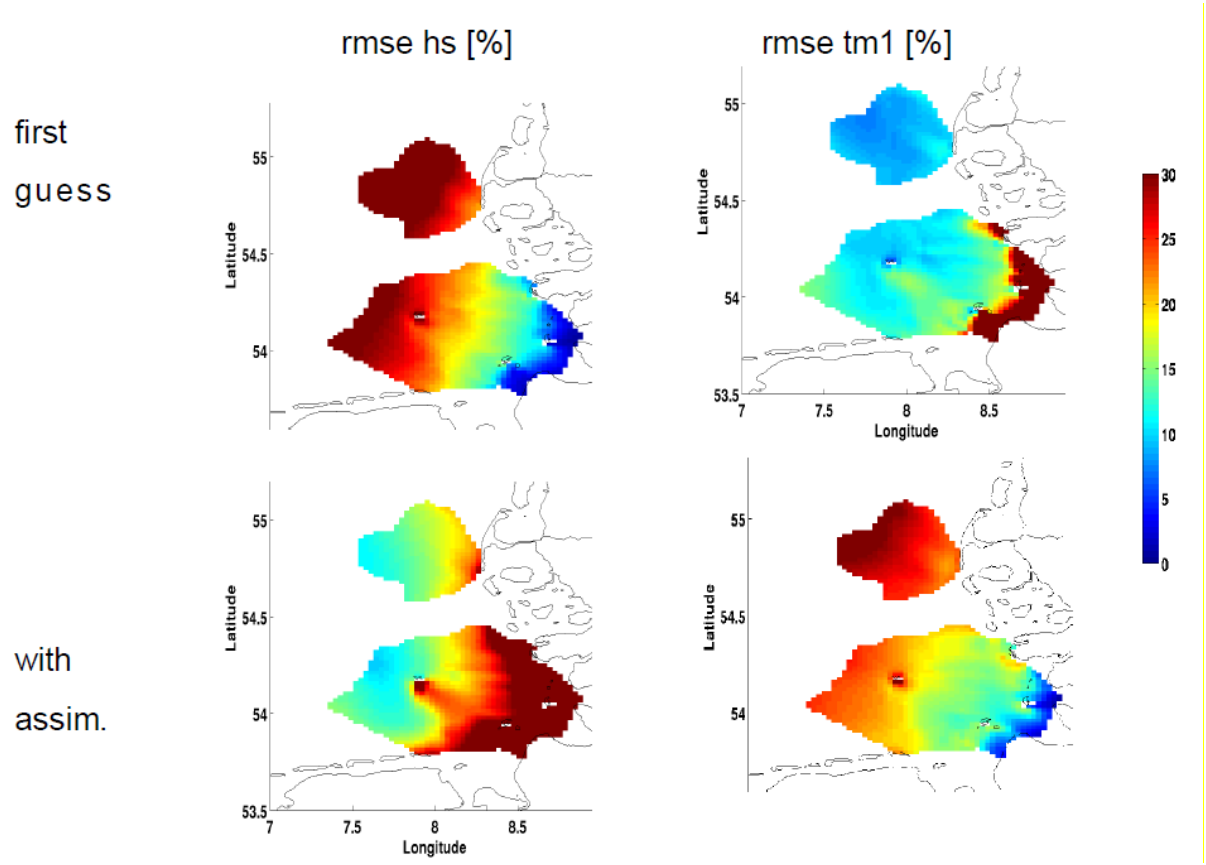

Figure III.17 Comparison of first guess and (NN approximated) assimilation error for significant wave height and mean wave period in the HF radar region (region 'known to' NN).

From the first assimilation experiments we can conclude that:

(i) As expected from the first twin experiment, the first guess error is small in shallow water regions close to the coast where wave state is dominated by local wind. (ii) For the values emulated with the forward NN the opposite is true (see section forward NN) (iii) it can be supposed that running WAM with the NN derived boundary values will give better results in the shallow areas and by definition the impact of the assimilation affects the whole German Bight region (iv) in summary, the novel assimilation scheme based on NN gives very promising results.

IV FINAL REMARKS

In the framework of the MyWave project we have applied innovative data assimilation techniques with the aim of improving nearshore North Sea wave forecasts. The considered approaches were a) 3D-VAR assimilation of coastal scatterometer winds in HARMONIE; b) EnKF assimilation of wave observation in SWAN and c) NN assimilation of wave observation in WAM. As reported, the results of the first trials using mainly synthetic data have led to promising results. In accordance with the project planning, we shall now move on to applying these techniques for assimilation of real wind and wave data considering a number of relevant North Sea storms.