Data assimilation in TELEMAC using optimal interpolation with application to the North Sea

Kai Chu[1], W. Alexander Breugem[1]
kai.chu@imdc.be, Antwerp, Belgium
[1]: IMDC NV

Abstract – **Ensemble Kalman Filtering is a powerful technique for performing data assimilation. However, the disadvantage of this technique is that it is computationally expensive, thus making the application in TELEMAC-2D and especially TELEMAC-3D difficult. Therefore, in the present paper, an alternative methodology was applied, namely Optimal Interpolation. In this technique, the Kalman Gain, used for data assimilation using a Kalman filter is parametrized and precomputed, rather than computed during an Ensemble Kalman Filtering simulation. Therefore, only a single computation needs to be performed, meaning that the computational cost of a simulation with data assimilation is comparable to the computational cost of a simulation without data assimilation. This method was implemented in TELEMAC and applied in the test cases in TELEMAC-2D. Here, measured water level data are assimilated into IMDC's continental shelf model of the North Sea (iCSM), where it is shown that in a reanalysis, the root mean square error (RMSE) in the water levels decreases by a factor of two with data assimilation.**

Keywords: TELEMAC-2D, TELEMAC-3D, Data-assimilation, Kalman Filter, optimal interpolation, reanalysis, continental shelf model.

I. INTRODUCTION

In order to increase the predictive power of operational models, data assimilation is often used. Further, data assimilation is often applied to perform a reanalysis, in which a combination of observed data and model simulation is adopted to generate an accurate high resolution dataset, which can be for example used to obtain boundary condition data for nesting smaller scale models.

Ensemble Kalman Filtering [1] [2] is a powerful technique for performing data assimilation. However, this technique is computationally expensive, thus making the application in models like TELEMAC-2D and especially TELEMAC-3D difficult. Therefore, in the present paper, an alternative methodology was applied, namely Optimal Interpolation [3] [4]. In this technique, the Kalman Gain, used for data assimilation with Kalman filter is parametrized and precomputed. Therefore, only a single computation needs to be performed, meaning that the computational cost of a simulation with data assimilation is comparable to the computational cost of a simulation without data assimilation.

The objective of this paper is to implement data assimilation using Optimal Interpolation in TELEMAC by means of a generic module, which can then be used in any of the TELEMAC modules. Further it is the objective to test the data assimilation for hindcast simulations in TELEMAC-2D using IMDC's Continental Shelf Model of the North Sea (iCSM), in which measured water level data are assimilated into the model.

The structure of the paper is as follows. First the data assimilation methodology is explained, including details on the implementation in TELEMAC. The iCSM model is presented next, and it is described how data assimilation is applied in this model. The results of the validation calculation are shown in a subsequent section. An outlook on future activities with respect to data assimilation is given thereafter. The paper is ended with some conclusions.

II. DATA ASSIMILATION

A. Optimal interpolation methodology

Data assimilation using optimal interpolation is performed using the following equation to update a model variable x_{mod}, using measurement data x_{meas}:

$$\vec{x}_{update} = \vec{x}_{mod} + K(\vec{x}_{meas} - H^T\vec{x}_{mod})$$

Here, x_{mod}, x_{meas} and x_{update} are vectors. The x_{mod} and x_{update} have the same size, which in TELEMAC-2D equal to the number of nodes in the mesh (NPOIN × 1). X_{meas} has a different size, namely the number of observation points that are used for data assimilation (NOBS × 1). H is an operator that maps the modelled data to the location of the observations. In case linear interpolation is used, H can be written as a matrix of size NOBS × NPOIN. Finally K is the Kalman Gain (a matrix of size NPOIN × NOBS), which determines how the difference between model and observations (at the location of the observations) is used to update the model prediction.

The Kalman Gain K is determined from:

$$K = P_f H^T \left(H P_f H^T - R\right)^{-1}$$

Here $P_f H^T$ is the covariance between the model data (size NPOIN × NOBS) at the location of the observation and the model data anywhere in the model. R is the correlation matrix (size NOBS × NOBS), which prescribes the uncertainty of the measurements. When using Ensemble Kalman Filtering, multiple simulations are performed using perturbed model data and/or observations as a kind of Monte-Carlo simulations, to determine $P_f H^T$ using statistical calculations.

In Optimal Interpolation, K is not determined during the simulation, but precomputed. There are often two approaches to do so:

- Perform a prior calculation using Ensemble Kalman filtering, and store the calculated values of K, for using in later simulations. Typically, in such calculations, K is averaged in time [5].

- Parametrize $P_f H^T$.

In the present study, experiments were initially performed with the first method, in which ADAO (Data Assimilation and Optimization) [6] was used to determine K. However, these experiments were unsatisfactory. The main reason was that in ADAO it is needed to prescribe the full matrix P_f (size NPOIN × NPOIN) to have perturbations to the model data that vary smoothly in space, which is too large to be practically possible. Therefore, method two was used. The following parametrization was used, which used the correlation function with the distance between the location of the observations $\vec{x}_{obs}$ and the model location of each node in the model $\vec{x}$ [7]:

$$P_f H^T = \sigma_{mod}^2 e^{\left(\frac{-|\vec{x}-\vec{x}_{obs}|}{L}\right)}$$

Here σ_{mod} is the standard deviation of the model data (i.e a measure of the uncertainty in the model), and L is the correlation length scale over which the model data is correlated.

The matrix R is parametrized as:

$$R = \begin{pmatrix} \sigma_{meas}^2 & 0 \\ 0 & \sigma_{meas}^2 \end{pmatrix}$$

Here σ_{meas} is the standard deviation in the observations (i.e. a measure of the uncertainty in the observations, assumed to be 0.1 m throughout the study).

B. Implementation in TELEMAC

In order to assimilate the optimal interpolation in TELEMAC, a new module named OPTIMAC was programmed. This module was written in FORTRAN to be easily integrated in the rest of the codes in TELEMAC, such that it can be used in any kind of calculations (with or without TelApy, serial and in parallel). In this module, a precomputed Kalman Gain (which is stored in a SELAFIN file) is used in combination with an ASCII input file containing the data to be assimilated. This file contains x, and y coordinates of the location of the observations (and the z coordinate in case of data assimilation in TELEMAC-3D), as well as time series of the observed data. The module consists of three functional parts:

- **Initialization**. Here, data is allocated for the necessary arrays. The precomputed Kalman Gain, is read from a SELAFIN file. Then, the location of the measurements are read from the ASCII file. The coefficients of the matrix H (interpolation matrix) are determined using linear interpolation, for use in all future interpolation steps.

- **Application**. For every time step during the simulation, it is checked whether observation data is available. In case new data are available, a model variable is updated. A no data value (-999) is used to handle gaps in the time series. A threshold for the water depth is used, to prevent the use of data assimilation in areas with very shallow water depths, in order to prevent instabilities due to the combination of wetting and drying and data assimilation. Note that no validation of the observation data is performed inside the calculation. It is assumed that data validation has been performed previously (before the start of the TELEMAC simulation). This validation is very important, because the algorithm readily assimilates wrong data in the model, leading to an incorrect result of the simulation.

- **Finalization**: The allocated arrays are cleaned and internal variables are reset for the use in a new calculation.

It should be noted that this methodology was developed particularly for the case, where the number of observations is rather low, which occurs for example when a couple of point measurements are used in the data assimilation. It is not very suited for cases where large fields of data need to be assimilated, such as happens for example when using data assimilation on satellite data.

C. Determination of the distance

The parametrized equation for the Kalman gain depends on the distance between the location of the observations and the nodes in the mesh. However, this distance should take into account the presence of the coastlines in the model. Therefore, the distance was defined as the shortest distance between two nodes following the edges of the model. This distance was calculated in MATLAB using Dijkstra's algorithm [8], by converting the mesh to a graph. For the location of the observation, the closest node in the mesh is used (Figure 1). Note that using this method, the distance is expected to be somewhat larger than in reality. However, because the parameter L is used as a tuning parameter in this study, this assumption shall not lead to any problems.

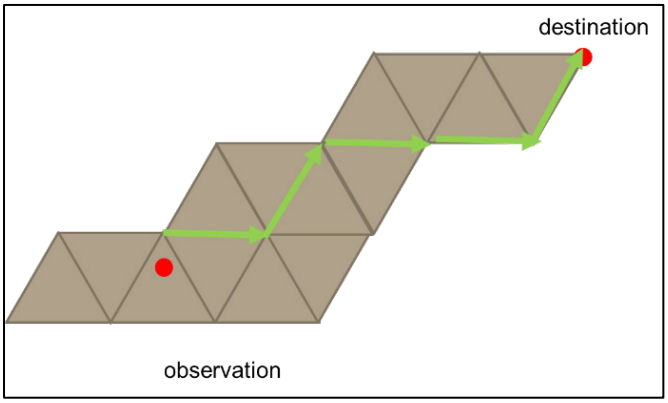

Figure 1. Determination of distance between the observation point and all nodes in the mesh using Dijkstra's algorithm.

III. MODEL SETUP OF THE CONTINENTAL SHEL MODEL OF THE NORTH SEA

A. *Model setup for iCSM*

The IMDC Continental Shelf Model (iCSM) is a 2D barotropic tidal surge model developed in-house in TELEMAC-2D [9] [10], focusing on the continental shelf of the North Sea (Figure 2). The model is built in a spherical Mercator projection with Coriolis effect included. The computational mesh consists of approximately 150,000 nodes and 292,000 elements. The unstructured mesh is refined near the coastal zones, e.g. with a minimal resolution of 500 m at Belgian coast. Mesh refinement is also applied along the coastlines of the UK, France and The Netherlands as well as in the Wadden Sea and the English Channel.

The bathymetry in iCSM is derived from the European Marine Observation and Data Network (EMODnet) which is referenced to Mean Sea Level (version 2020). The model includes the most dominant physical processes in the North Sea, such as inverse barometer correction, which accounts for an isostatic response of the oceans to atmospheric pressure. The self-attraction and loading [11] due to three effects is also taken into account: the deformation of the seafloor under the weight of the water column; the redistribution of Earth mass and its corresponding changes in the gravitational field; the gravitational attraction induced by the water body on itself. It has a well-acknowledged impact on the tidal phases and therefore is included in the iCSM using a beta (β) approximation approach. The internal tidal dissipation considers the dissipation of tidal energy through generation of internal tides which is the dominant mechanism when tides propagate over steep topography in deep stratified waters. This is important in the Bay of Biscay and is also included in iCSM.

To account for the effect of bottom friction, a spatially-varying roughness field of Nikuradse value was automatically calibrated [12] on bottom friction using ADAO with three-dimensional variational assimilation (3D-Var). This automated optimization tool allows to find the best possible parameter set for the model.

In its present form, the model reproduces the hydrodynamics in the European Continental Shelf accurately. For instance, the Root-Mean-Squared-Error (RMSE) of water levels along the Belgian coast is in the order of 10 cm. The RMSE of stationary velocity magnitude in the Belgian Coastal Zone is of the order 0.1 m/s, which is considered as top-of-range numerical model accuracy.

The details of the model setup and its performance are referred to [12], thus will not be elaborated here.

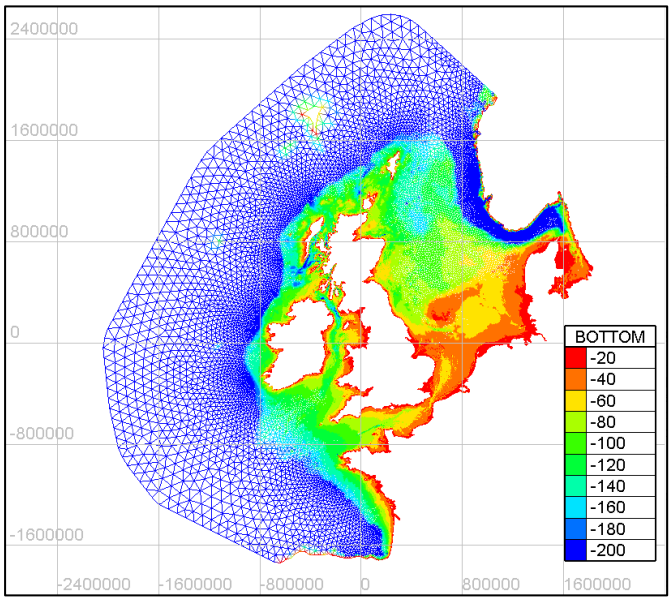

Figure 2. iCSM model mesh and bathymetry (horizontal system: Spherical Mercator projection. Vertical datum: MSL).

B. *Model setup for assimilation of measured water levels*

Measurement of water level data at 29 stations in the North Sea are used in this study (Figure 3). Table 1 presents the model simulations carried out in this study.

Run01 is the base run without data assimilation, which is used to compare to all the experiments using data assimilation. Run02 assimilates measured water level at all the 29 stations as shown in Figure 3, the purpose of this run is to test the module developed for this study (§II). Run03 is a validation run which investigates whether data assimilation could improve water level predictions at stations where no data assimilation is applied directly. The validation stations are selected in a way to have at least one station per coastal zone/country (Figure 5). More sensitivity runs using different validation stations will be considered in future studies. Run04 to Run07 are sensitivity runs with different values of standard deviation of the model data (σ_{mod}) and the length scale over which the model data is correlated (L). Note that an attempt was made to determine L directly from the results of a separate model simulation without data assimilation by determining the autocorrelation of the modelled water levels around the different measurement stations. These data suggested that the shape of $P_f H^T$ might look more like a Gaussian function than an exponential function, but give correlation length in the order of 100 km, used in this study. This is something that will be studied further in future.

Table 1: Summary of model simulations carried out in this study.

Run ID	σ_{mod}	L	Description
Run01	-	-	Base run without data assimilation, same as [12].
Run02	0.1m	100 km	Full data assimilation using measured water level at all 29 stations.
Run03	0.1 m	100 km	Idem Run02, but with 7 validation stations and 22 assimilation stations.
Run04	0.25 m	100 km	Idem Run03
Run05	0.05 m	100 km	Idem Run03
Run06	0.1 m	50 km	Idem Run03
Run07	0.25 m	150 km	Idem Run03

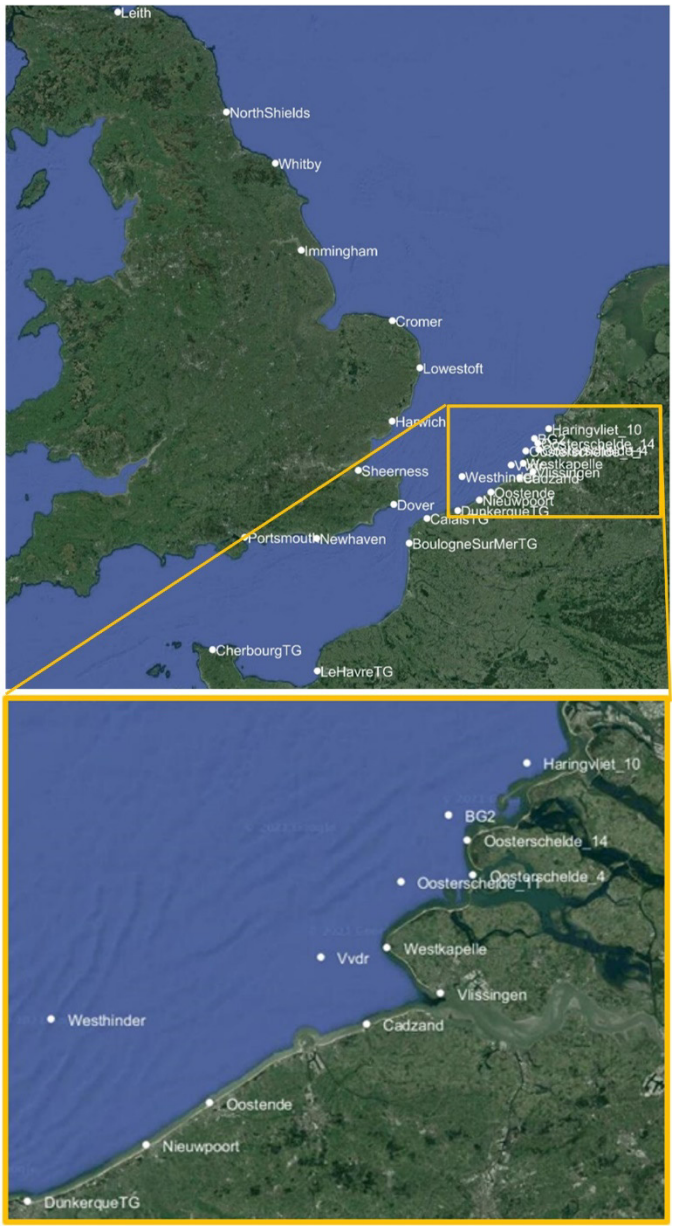

Figure 3: Top: Observation stations of water level used for data assimilation in the North Sea. Bottom: Close up of the Belgian and Dutch coast.

The Kalman Gain is firstly computed for Run02 using $\sigma_{meas} = 0.1$ m; $\sigma_{mod} = 0.1$ m; L = 100 km. Note that the Kalman gain has a size of NPOIN × NOBS (hence 29 maps). The 29 maps of the Kalman Gains are assembled into one graph (Figure 4) for readability, by presenting the maximum value of Kalman Gain at each computational node.

The Kalman Gain is recomputed for Run03 using 22 measurement stations, see the map in Figure 5.

Figure 6 shows the comparison of computed Kalman Gain with different values of σ_{mod} of 0.1 m (Run03), 0.25 m (Run04) and 0.05 m (Run05). It clearly shows the trend that higher σ_{mod} leads to higher values of the Kalman Gain. It essentially means that the updated water level will be based more on measurement data than on the model.

Figure 7 shows the comparison of the computed Kalman Gain with different values of L of 100 km (Run03), 50 km (Run06) and 150 km (Run07). It clearly shows the trend that larger L leads to higher values of Kalman Gain. It essentially means that the updated water level will be based more on the measurement data than on the model results since the Kalman Gain has a larger area of influence.

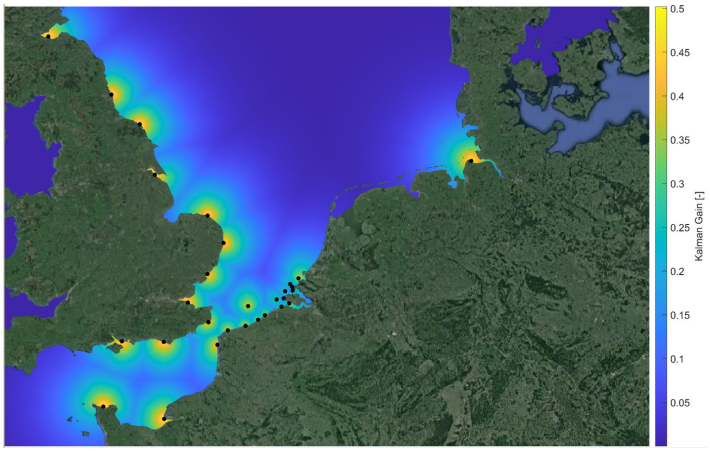

Figure 4: Kalman Gain with 29 measurement stations (Run02).

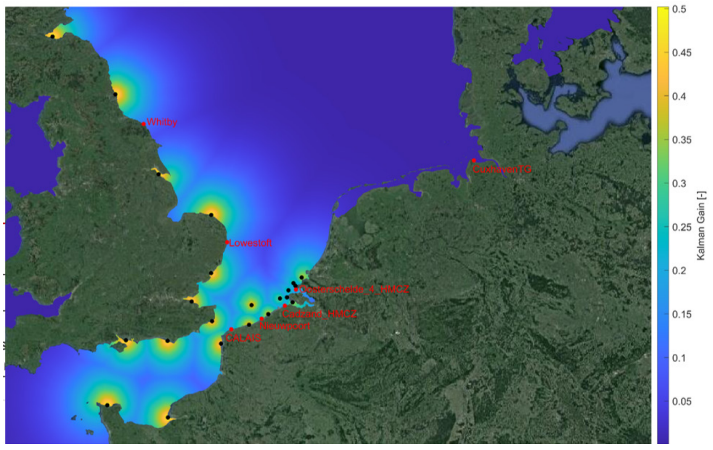

Figure 5: Kalman Gain with 22 measurement stations and 7 stations (marked in red) for validation (Run03).

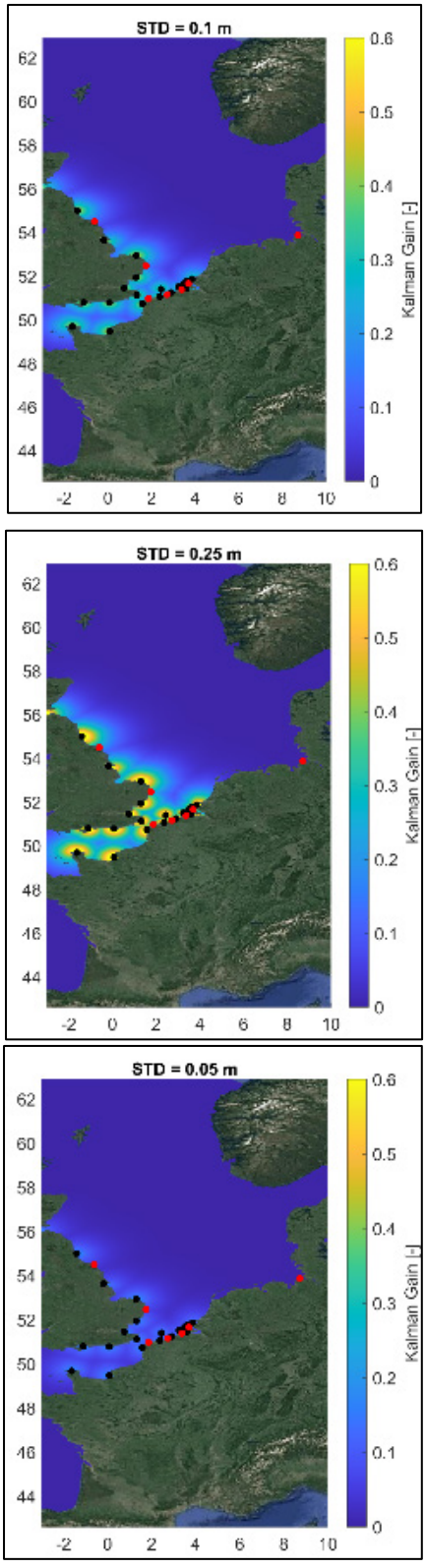

Figure 6: Comparison of Kalman Gain with 22 measurement stations and 7 stations (marked in red) for validation with different values of σ_{mod} (Top:Run03; Middle: Run04; Bottom: Run05).

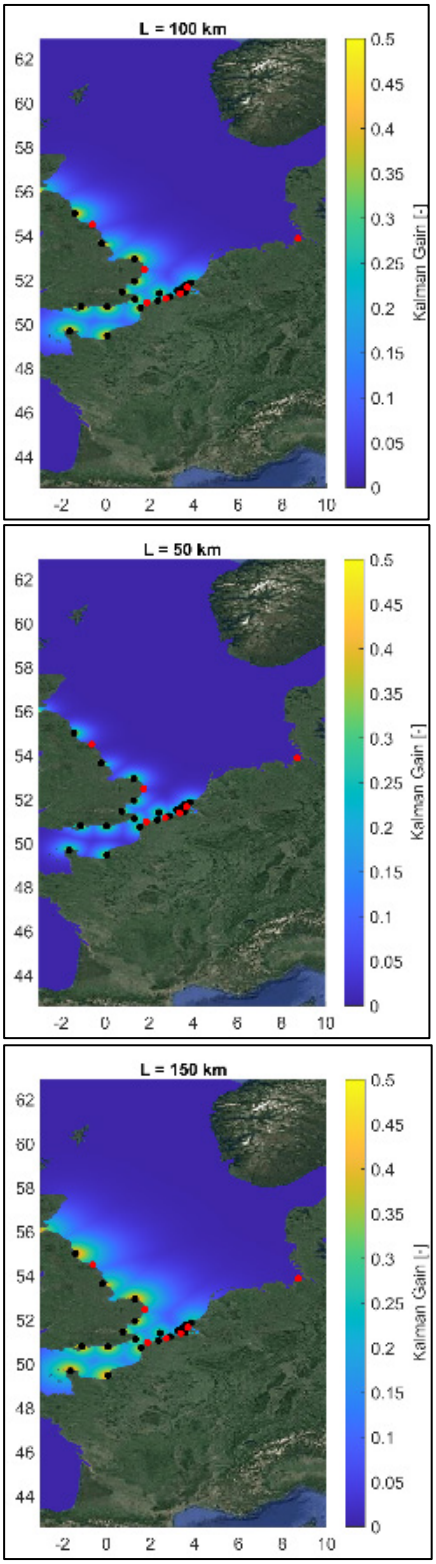

Figure 7: Comparison of Kalman Gain with 22 measurement stations and 7 stations (marked in red) for validation with different values of L (Top:Run03; Middle: Run06; Bottom: Run07).

IV. MODEL RESULTS

Figure 8 shows the comparison of the RMSE of the water levels at the coastal stations from Run01, 02 and 03. The RMSE

is lowered by 50% on average using data assimilation. The RMSE at the validation stations (Run03) is slightly higher than the RMSE with direct data assimilation at those stations (Run02), but it is still lower than the RMSE without data assimilation (Run01). It should be noted that if data assimilation is not applied at Cuxhaven in the German Bight, the RMSE becomes substantially larger than when data assimilation is applied locally. This is reasonable since the data assimilation at the remaining stations are too far away to have a substantial impact on the water level at Cuxhaven.

Figure 9 compares the RMSE of water level from Runs 01, 03, 04, and 05. In general, different values of σ_{mod} lead to comparable RMSE values, which are always lower than those obtained without using data assimilation. Lower value of σ_{mod} (0.05 m) result in slightly higher values of the RMSE, implying that the updated water level is dependent more on model predictions than the measurements when assuming the model error is lower (0.05 m). However, this is less justified along the British coast where the original model (Run01, no data assimilation) already produces higher RMSE values.

Figure 10 compares the RMSE of the water level from Runs 01, 03, 06, and 07. In general, different values of L lead to comparable RMSE values, which are always lower than those obtained without using data assimilation. It is noted that the lower value of L = 50 km results in slightly higher values of the RMSE at Leith, North Shields and Whitby in UK, implying that the updated water level is dependent more on the model predictions than on the measurement data with smaller Kalman Gain (Figure 7). Again, this is less justified along the British coast, where the original model (Run01, no data assimilation) already produces higher RMSE values. Interestingly, the RMSE at Immingham and Sheerness are less sensitive to L even when the original model (Run01, no data assimilation) produces a RMSE up to 25 cm. This probably suggests that a station dependent L could be investigated for future studies.

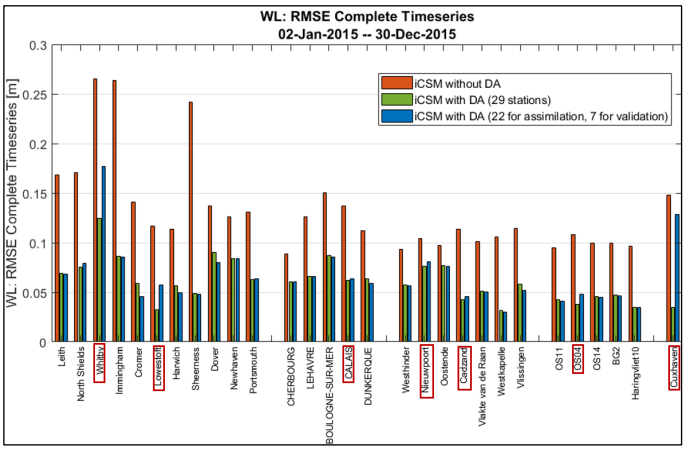

Figure 8: RMSE of water levels from Run01, Run02 and Run03. The stations with red outbox are the seven validation stations.

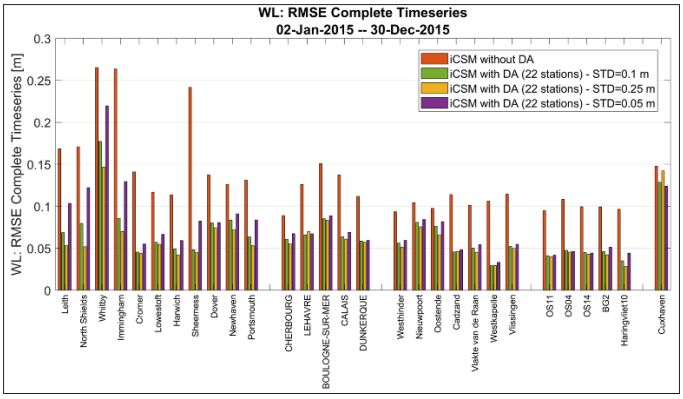

Figure 9: RMSE of water levels using different values of σ_{mod} from Run03, Run04 and Run05.

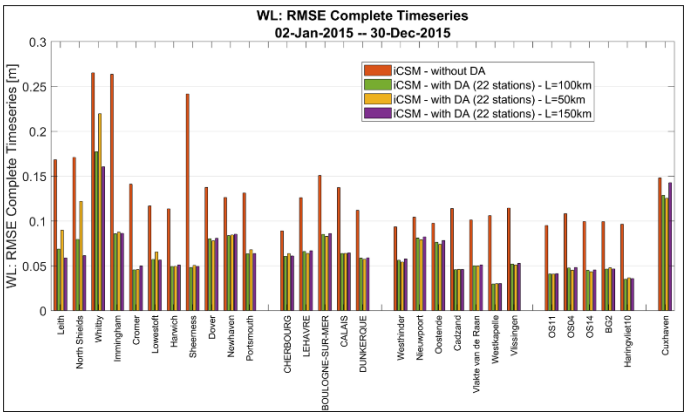

Figure 10: RMSE of water levels using different values of L from Run03, Run06 and Run07.

V. OUTLOOK

The application of data assimilation in iCSM shows a decrease of the RMSE of a factor two, meaning that more accurate boundary conditions for the IMDC's Scheldt model are obtained. Tests of using these updated boundary conditions for the nested Scheldt model will be investigated in the future.

The data assimilation module was implemented in a generic way, meaning that it can easily be applied within other TELEMAC modules. Tests, in which the methodology is applied to assimilate salinity data in IMDC's TELEMAC-3D Scheldt model has already been started and show promising results. Note hereby that in this case, the data assimilation is applied to a three dimensional field rather than a two dimensional one. Due to the generic way the module is setup, this was possible without any changes to the data assimilation code.

It is the intention to apply the presented data assimilation module OPTIMAC to IMDC's North Sea wave model in TOMAWAC [13], in order to improve the predicted wave conditions in the Belgian Coastal Zone. The challenge for this is that in TOMAWAC, wave energy density spectra are the principle variable, which at each location in time depend on the direction and frequency of the wave components. This means that the data assimilation needs to be performed on a four-dimensional variable. The preferred data assimilation methodology depends on the availability of the data. In case

measured wave spectra are available, it is preferred to assimilate these directly. However, often only integrated parameters (significant wave height, peak period) or simplified (1D) spectra are available. In this case, it is needed to develop some extra code that parametrizes the effect of these simplified variables on the full two-dimensional wave spectrum. To perform well, it is likely that the Kalman Gain needs to be different for each frequency and especially for each direction in the spectrum.

VI. SUMMARY AND CONCLUSIONS

In this paper, OPTIMAC, a generic module that performs data assimilation in TELEMAC using Optimal Interpolation, was presented. The module was tested by performing data assimilation in IMDC's continental shelf model of the North Sea. In this model, observed water level data were assimilated in a one year hindcast simulation. The RMSE of water level in the North Sea is significantly reduced by 50% with data assimilation. Sensitivity analysis on standard deviation of the model data (σ_{mod}) and the correlation length scale (L) used in Optimal Interpolation have been carried out. In general, different values of σ_{mod} and L lead to relatively comparable RMSE values, which are always lower than those obtained without using data assimilation.

The computational time with and without data assimilation is rather similar, meaning that the Optimal Interpolation algorithm developed for TELEMAC in this study is computational efficient.

ACKNOWLEDGEMENT

The authors want to acknowledge the internal research and innovation program of IMDC for its financial support.

REFERENCES

[1] M. Katzfuss, J. R. Stroud and C. K. Wikle, Understanding the Ensemble Kalman Filter, The American Statistician, 70:4, 2016, pp. 350-357, DOI: 10.1080/00031305.2016.1141709.

[2] G. Evensen, Data Assimilation: The Ensemble Kalman Filter. Germany: Springer Berlin Heidelberg, 2009.

[3] J.L. Hoyer and J. She, Optimal interpolation of sea surface temperature for the North Sea and baltic sea. J Mar Syst. 65, 2007, pp. 176–189.

[4] P.R. Oke, G.B. Brassington, D.A. Griffin and A. Schiller, "Ocean data assimilation: a case for ensemble optimal interpolation". Aust. Meteorol. Oceanogr. J., 59, 2009, pp. 67-76.

[5] F. Zijl, J. Sumihar and M. Verlaan, "Application of data assimilation for improved operational water level forecasting on the northwest European shelf and North Sea". Ocean Dynamics 65, 2015, pp. 1699–1716. DOI: https://doi.org/10.1007/s10236-015-0898-7.

[6] J.P. Argaud, User documentation, in the SALOME 9.3 platform, of the ADAO module for "Data Assimilation and Optimization", Technical report 6125-1106-2019-01935-EN, EDF / R&D, 2019.

[7] C.D. Rodgers, Inverse methods for atmospheric sounding: theory and practice. Series on Atmospheric, Oceanic and Planetary Physics, V2.World Scientific, Singapore, 2000.

[8] T.H. Cormen, H.Thomas, C.E. Rivest, L. Ronald and C. Stein, "Section 24.3: Dijkstra's algorithm". Introduction to Algorithms (Second ed.). MIT Press and McGraw–Hill, 2001, pp. 595–601. ISBN 0-262-03293-7.

[9] W.A. Breugem, T. Verbrugghe and B. Decrop, "A continental shelf model in TELEMAC 2D". TELEMAC User Conference, Proceedings. Presented at the TELEMAC User Conference, 2014.

[10] K. Chu, W.A. Breugem and B. Decrop, "Improvement of a Continental Shelf Model of the North Sea". Telemac user conference 2020, Antwerp, Belgium, 2020.

[11] M. I. Apecechea, M. Verlaan, F. Zijl, C.L. Coz and H. Kernkamp, "Effects of self-attraction and loading at a regional scale: a test case for the Northwest European Shelf". Ocean Dynamics, 67, 2017, pp. 729–749.

[12] K. Chu, W.A. Breugem, L. Wang, and B. Decrop, "Automatic calibration of a continental shelf model of the North Sea using data assimilation algorithm". Proc. 39th IAHR World Congr. Int. Assoc. Hydro-Environ. Eng. Res., Granada, Spain, 2022.

[13] Q.H. Zhang, S. Doorme, J. Figard, W.A. Breugem and K. Bakhtiari, "Application of TOMAWAC for wave propagation and wave energy assessment: a reliable 20-year database for North Sea metocean condition, with a focus near the Belgian coast". *Submitted to*: Telemac user conference, Karlsruhe, Germany, 2023.