Responses to the Reviewers' Comments and Suggestions

Journal: Geoscientific Model Development (GMD)
Manuscript number: gmd-2020-61
Manuscript title: Model-driven optimization of coastal sea observatories through data assimilation in a finite element hydrodynamic model (SHYFEM v.7_5_65)

We would like to thank the Reviewers for their valuable comments and effort to improve the manuscript. We have responded to all comments as can be seen in the following list. We believe that with these revisions the manuscript has been improved and we hope that it is now ready for publication.

The original Reviewers' comments and suggestions are shown in regular typeface, while our responses are shown in italics. The line and figures numbers we use refer to the revised document.

Response to Reviewer #1

R1.1 General comments: "Model-driven optimization of coastal sea observatories through data assimilation in a finite element hydrodynamic model (SHYFEM v.7_5_65)" is a well written manuscript concerned with using data assimilation to improve coastal modelling capabilities and optimise monitoring networks. The article contains a fair comparison of data interpolation (DI) and data assimilation (DA) methods, along with a further comparison of two different DA approaches. The Lagoon of Venice application involves a complex spatial domain with sensitive coastal dynamics. The objective of the numerical experimentation is clearly stated. Results are clearly presented in a number of attractive figures. In the conclusion, the objective is fulfilled and recommendations are made for modifying the monitoring network.

Response: We appreciate the comments and we improved the manuscript following all reviewer's suggestions.

R1.2 I particularly liked the introduction to DI and DA philosophy given in the paragraph starting on line 37.

Response: We thank the reviewer for this positive comment.

R1.3 Equations (6)-(7) could perhaps be introduced in a better way. There is a lot of notation all at once, some of which is not referred to in the text. For example, it would be good to elaborate on what is meant by the superscript "a" for "analysis" (although this does become clear at the end of Section 2).

Response: We concur with the reviewer that some of the mathematical passages were poorly explained. In the revised manuscript, we improved the explanation of the mentioned equations by adding more details on the different terms (lines 110-132).

R1.4 In the paragraph beginning on line 201, it would be beneficial to clarify how unforced boundary conditions are represented within the shallow water model. Forced boundaries

are mentioned, but the implementation of unforced boundary conditions (for example in urban areas) is unclear. Are free-slip conditions used?

Response: Unforced boundaries are solid boundaries that are implemented in the model with a free slip condition. The only condition that is enforced on these boundaries are the no-flux condition through these boundaries. No-slip conditions can also be implemented by the model, however, the resolution of the numerical grid is much too coarse for these kind of condition.

We have inserted the following sentence at lines 84-85: "At the boundaries, either water levels are prescribed at the open boundaries or the free-slip condition is implemented at solid (closed) boundaries".

R1.5 The statement on lines 274-275 claims that "results improved at all stations". However, there appears to be one exception at station 12, where the control simulation RMSE is 4.5 but the DA-EnSRF RMSE is 4.7. This should at least be mentioned and a sentence suggesting a reason for the anomaly would be beneficial.

Response: We thank the reviewer for highlighting this anomaly which was due to a typesetting error in the LaTeX file. The correct value of 3.7 is now reported in Table 1.

R1.6 There are a small number of typos and grammatical errors in the current manuscript.

Response: We corrected all reported typos and grammatical errors.

Response to Reviewer#2

R2.1 General comments: The paper analyzes optimization of observational grid via analyzing the impact that assimilation of station data has on the high resolution numerical model of the Venice lagoon. Several modes of assimilation are employed to introduce data into the model. I must say I really like the idea of how DA was used in the paper. The paper is interesting, contains new insight and is well written. The figures are clear. The abstract reflects the contents well.

I recommend publication after minor revision.

Response: We thank the reviewer for the positive comment and we improved the manuscript following all reviewer's suggestions.

R2.2 p3, L76: *pa* should be p_a ($_a$ denoting subscript)

Response: Corrected.

R2.3 p3, L77: ρ_q should be ρ_w

Response: Corrected.

R2.4 p4, L107: "the mean is:" should probably be "the ensemble mean is:"

Response: Corrected.

R2.5 p7 L185: I am not sure I understand this phrase "... at which degree the observations represent the state variable over the whole system." Can the authors please include a specific description and/or metrics by which this degree was measured?

Response: As explained in section 3.1, for both Data Interpolation and Data Assimilation experiments, the metric used to evaluate the representativeness of the method in describing the state variable over the system is the root mean square error. RMSE is evaluated in the station not considered in the DI or DA computations. The evaluation procedure was repeated for each monitoring station and the results are reported in Table 1.

R2.6 P8 L207: should sigma be a greek letter? Why did you set it to 2 km rather than something else?

Response: We corrected the sigma Greek letter. As specified at the beginning of the Results section (lines 247-249), all parameters (σ, τ, cut-off distance for the local analysis) were manually defined through trial and error calibration process and evaluating the goodness-of-fit of the water level RMSE in the DA-Nudging and DA-EnSRF base simulations.

R2.7 Perhaps I missed something but I still do not clearly understand how the boundary condition perturbations were generated. The paper states that 60 perturbations (gaussian, it seems?) were used as OBCs. Do I understand correctly that you used mean(A) as the open boundary conditions and then further added a constant (in space and time) perturbation to each ensemble member, where the amount of each member sea level perturbation was sampled from a gaussian N(mu, sigma)?

Response: We concur with the reviewer that the perturbation terms were not properly described. At each timestep (t), a random vector (r) of N perturbations is computed from a Gaussian distribution (with mean 0 and standard deviation of 30 cm) as:

$$r(t,n) = cos(2\pi r_2(t,n))\sqrt{-2log(r_1(t,n)+\epsilon)} \tag{1}$$

with n the number of the ensemble member $(1,N)$, r_1 and r_2 random vectors and ϵ a very small number.

The perturbation vector p at time t is computed using the random vector and the perturbation vector at the previous time (t_{-1}):

$$p(t,n) = \alpha p(t_{-1},n) + \sqrt{1-\alpha^2}r(t,n) \tag{2}$$

with $\alpha = 1-(t-t_{-1})/\tau$ and τ the decay time (2 days in our case). Then the new perturbation is stored for the next time step. This type of perturbations are classified as red noise.

We modified the manuscript to clarify the methodology adopted in this study. The text now reads (lines 220-225): "We used 60 perturbations for the sea-level boundary condition (member 0 is unperturbed) taken from a Gaussian distribution with a zero mean and a standard deviation set to 30 cm. This value was found empirically, in order to have a good spread at the boundary, which is then propagated to the variables computed by the model. As asserted, the perturbations are centred, having a null mean, and correlated in time. To do this, each perturbation at time t is obtained from a weighted average of a new perturbation and of the one at time $t-1$. This type of perturbations are classified as red noise and in the present case we used a decay time of two days."

R2.8 P9, L255: perhaps: "...towards the observations WHILE keeping the physical dynamics..."

Response: Corrected.

R2.9 p10, L298: I don't entirely see what is meant by "scalability". Can you please rephrase or clarify?

Response: We are referring to the computing scalability of the DA-EnSRF procedure on multiprocessor computers. The sentence has been rephrased as follow (lines 308-309): "So in this case the simulations are 26,535, but the computing scalability is high since the 61 simulations of the ensemble are independent and can be parallelized on multiple CPUs computers.

R2.10 P12, L351: These correlationS...

Response: Corrected.

Model-driven optimization of coastal sea observatories through data assimilation in a finite element hydrodynamic model (SHYFEM v. 7_5_65)

Christian Ferrarin[1], Marco Bajo[1], and Georg Umgiesser[1,2]

[1]CNR - National Research Council of Italy, ISMAR - Marine Sciences Institute, Venice, Italy
[2]Marine Research Institute, Klaipeda University, Klaipeda, Lithuania

Correspondence: Christian Ferrarin (c.ferrarin@ismar.cnr.it)

Abstract. Monitoring networks aims at capturing the spatial and temporal variability of one or several environmental variables in a specific environment. The optimal placement of sensors in an ocean or coastal observatory should maximize the amount of collected information and minimize the development and operational costs for the whole monitoring network. In this study, the problem of the design and optimization of ocean monitoring networks is tackled throughout the implementation of data assimilation techniques in the Shallow water HYdrodynamic Finite Element Model (SHYFEM). Two data assimilation methods - Nudging and Ensemble Square Root Filter - have been applied and tested in the Lagoon of Venice (Italy), where an extensive water level monitoring network exists. A total of 29 tide gauge stations were available and the assimilation of the observations result in an improvement of the performance of the SHYFEM model that went from an initial root mean square error (RMSE) on the water level of 5.8 cm to a final value of about 2.1 and 3.2 cm for the two data assimilation methods, respectively. In the monitoring network optimization procedure, by excluding just one tide gauge at a time, and always the station that contributes less to the improvement of the RMSE, a minimum number of tide gauges can be found that still allow for a successful description of the water level variability. Both data assimilation methods allow identifying the number of stations and their distribution that correctly represent the state variable in the investigated system. However, the more advanced Ensemble Square Root Filter has the benefit of keeping a physically and mass conservative solution of the governing equations, which results in a better reproduction of the hydrodynamics over the whole system. In the case of the Lagoon of Venice, we found that, with the help of a process-based and observation-driven numerical model, two-thirds of the monitoring network can be dismissed. In this way, if some of the stations must be decommissioned due to a lack of funding, an a-priori choice can be made, and the importance of the single monitoring site can be evaluated. The developed procedure may also be applied to the continuous monitoring of other ocean variables, like sea temperature and salinity.

1 Introduction

Ocean and coastal monitoring networks are fundamental for tracking contaminants in the water, assessing environmental change and water quality, observing sea level rise and developing strategies for managing resources in a changing climate (Stammer et al., 2019; Trowbridge et al., 2019). Coastal zones are dynamic and subject to changing environmental conditions

caused by natural and anthropogenic variations in climatic and oceanographic processes. The monitoring of the spatial and temporal complexity of the coastal ocean is challenging and a large number of observational sites are required to correctly describe the interactions at the land-sea transition, and coupled physical, chemical, and biological processes. However, the implementation and maintenance of such large monitoring networks are expensive and therefore their optimization is of crucial importance. In the last decades, satellite earth observation technologies have been widely used to integrate in-situ observatories for better understanding the current state of oceans and coastal seas (Levy et al., 2018).

Oceanographic models are increasingly used in coastal systems to describe sea dynamics induced by tide, atmospheric and terrestrial forcing, complementing thus the collected information retrieved by direct observations (Mey-Frémaux et al., 2019). Numerical models are also often used for predicting the ocean conditions, especially during storm events for endangered areas (Chaumillon et al., 2017). All models, however, need observations of the sea state to be calibrated and validated. Once the model is calibrated, new measurements can be used in a continuous validation of the model results. Observations can also be assimilated into the model, increasing its capacity to represent the dynamics of the investigated system (Edwards et al., 2015; Carrassi et al., 2018). In this case, we can speak of observations that improve the numerical model.

There is however another point of view. If only observations would be available, the best distribution of the monitored variable over the system could be given only by data interpolation (DI) of the observation points to the other areas. The direct observations of the sea conditions are considered to represent the true state at the monitoring point. However, the spatio-temporal interpolation of such true values is not meant to correctly describe the variability of the investigated state variable over the whole system. This is especially true in the coastal systems that are characterized by complex small scale and high-frequency dynamics. In this case the resulting picture of interpolated values may show non coherent features and inconsistency between data points. When an oceanographic model is available, the interpolation of these observations can be carried out by the model and much better representation of the environment can be achieved. In this contest, models are used to connect sparsely (in space and time) observations or synthesizing them through data assimilation (DA) techniques (Mey-Frémaux et al., 2019).

Validated ocean circulation model and DA can also assist the network design of a new observing system or optimizing existing observatory (Fujii et al., 2019). In the case of new monitoring networks, Observing System Simulation Experiments (OSSEs) are performed assimilating synthetic observation data (generated from a free-running model simulation that is intended to represent a virtual ~~"true"~~ "true" ocean) into other data-assimilative simulation runs in which different initial/forcing conditions are used (Raicich, 2006; Xue et al., 2011). The evaluation of the impact of the assimilated data in the OSSE simulations allows designing an optimal observing system. In order to evaluate existing monitoring networks, Observing System Experiments (OSEs) are performed by assimilating in several simulations a certain amount or type of observations and evaluating their impacts on the model against a reference dataset. Such an approach can be adopted in coastal regions to optimize existing observational arrays, with implications on sampling technology and networks (Frolov et al., 2008; Schulz-Stellenfleth and Stanev, 2010).

In this study, we show how data assimilation techniques are implemented in the Shallow water HYdrodynamic Finite Element Model (SHYFEM) for optimizing the tide gauge network of the Lagoon of Venice (Italy). Since one limitation of the

observing system evaluation procedure is that it depends on the properties of the DA employed for the evaluation (Fujii et al., 2019), here we adopted a multiple systems approach implementing the Nudging and the Ensemble Square Root Filter data assimilation methods.

2 Methods

2.1 SHYFEM model description

The numerical experiments consisted of simulating the circulation in the Lagoon of Venice using the open-source SHYFEM hydrodynamic model (Umgiesser et al., 2014). The model has ~~been already~~ already been applied to simulate hydrodynamics in the Mediterranean Sea (Ferrarin et al., 2018), in the Adriatic Sea (Bellafiore et al., 2018; Bajo et al., 2019) and in several coastal systems (Umgiesser et al., 2014, and references therein). The model solves the shallow-water equations in their formulations with levels and transports using a finite-element numerical method and semi-implicit time stepping. In the present work, a relatively simple two-dimensional configuration of the model has been used, solving the following equations:

$$\frac{dU}{dt} - fV = -H\left(g\frac{\partial \zeta}{\partial x} + \frac{1}{\rho_w}\frac{\partial p_a}{\partial x}\right) + A_H\nabla^2 U + \frac{1}{\rho_w}\left(\tau_{wx} - \tau_{bx}\right) \tag{1a}$$

$$\frac{dV}{dt} + fU = -H\left(g\frac{\partial \zeta}{\partial y} + \frac{1}{\rho_w}\frac{\partial p_a}{\partial y}\right) + A_H\nabla^2 V + \frac{1}{\rho_w}\left(\tau_{wy} - \tau_{by}\right) \tag{1b}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} = 0 \tag{1c}$$

where t is the time, x and y are the spatial Cartesian coordinates and $\eta = \eta(x,y,t)$ is the water level. $U = U(x,y,t)$ and $V = V(x,y,t)$ are the zonal and meridional water transport components, g is the acceleration due to gravity, ~~$pa = pa(x,y,t)$~~ $p_a = p_a(x,y,t)$ is the atmospheric pressure at mean sea level, ~~ρ_q~~ ρ_w the average density of sea water, $h = h(x,y)$ is the water depth at rest, while $H = h + \eta$ is the total water depth and $f = f(y)$ is the Coriolis parameter, varying with latitude. Smagorinsky's formulation (Smagorinsky, 1963; Blumberg and Mellor, 1987) is used to parameterize the horizontal eddy viscosity (A_h). τ_{wx} and τ_{wy} are the two components of the wind stress in the x and y directions and τ_{bx} and τ_{by} are the two components of the bottom stress.

The Coriolis term and pressure gradient in the momentum equation, and the divergence terms in the continuity equation are treated semi-implicitly. Bottom friction and vertical eddy viscosity are treated fully implicitly for stability reasons due to the shallow nature of the lagoon, while the remaining terms (advective and horizontal diffusion terms in the momentum equation) are treated explicitly. At the boundaries, either water levels are prescribed at the open boundaries or the free-slip condition is implemented at solid (closed) boundaries. A detailed description of the model equations is given in Umgiesser et al. (2014) and Bellafiore et al. (2018).

2.2 Data assimilation methods

2.2.1 Nudging

The nudging method is a flexible assimilation technique that is computationally more economical than other assimilation methods like variational data assimilation. First used in meteorology (Hoke and Anthes, 1976), the nudging method has been used with success in modelling the atmosphere (Stauffer and Seaman, 1990) and in oceanography (Verron, 1990; Blayo et al., 1994). Nudging is a simple assimilation technique where a new source term is added to the prognostic equations that drag the results versus the observed values. Therefore, it uses dynamical relaxation of the equations to tend to the observational points. The extra term to be introduced in the prognostic equation can be formulated as:

$$\partial S/\partial t = ... + (S_{obs} - S)/\tau \tag{2}$$

where S is the variable where nudging has to be applied, S_{obs} the observation value, and τ is the relaxation time scale. Depending on the value of τ, the relaxation is very strong (small τ) or weak (large τ). The value of τ can be different from point to point. It is worth mentioning that, by adding this extra term in the governing equations (e.g. the continuity Eq.1c for the water level), the numerical solution is no more mass conservative.

2.2.2 Ensemble Square Root Filter

The ensemble square root filter (hereinafter referred to as EnSRF) is a more complex assimilation method, widely used in environmental sciences (Evensen, 2004), and can be considered as an evolution of the Ensemble Kalman Filter (EnKF, Evensen, 2003). The assimilation code that allows one to use both these methods, has been recently implemented in SHYFEM ~~() and the code was~~ (Bajo, 2020) and used for the first time in a study on seiches and storm surges in the Adriatic Sea (Bajo et al., 2019).

The formulation of the EnSRF is slightly different from the EnKF and avoids the perturbation of the observations. Using the notation of Evensen (2004), if we define the model states as $\psi_i \in \mathbb{R}^n$ and the matrix holding them as:

$$\mathbf{A} = (\psi_1, \psi_2, ..., \psi_N) \in \mathbb{R}^{n \times N}, \tag{3}$$

with N the number of ensemble members and n the dimension of the states, the ensemble mean is:

$$\overline{\mathbf{A}} = \mathbf{A}\mathbf{1}_N, \tag{4}$$

where $\mathbf{1}_N \in \mathbb{R}^{N \times N}$ is a square matrix with each element equal to $1/N$. ~~The ensemble approximation of~~ If we define $\mathbf{P}$ as the background error covariance ~~, P, is:~~ matrix, which contains the covariance of the errors between all the model variables in the

whole computational domain, then its ensemble approximation is:

$$\mathbf{P}_e = \frac{\mathbf{A}'(\mathbf{A}')^T}{N-1}, \tag{5}$$

where $\mathbf{A}' = \mathbf{A} - \overline{\mathbf{A}}$, is the matrix containing the ensemble perturbations. ~~The covariance update in the Kalman filter is :~~

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f,$$

~~with the index a as analysis~~

In the traditional Kalman Filter formulation, the covariance matrix $\mathbf{P}$ is updated every time new observations are available. The matrix before the update is referred as *forecast*, f ~~as first guess,~~ , while after the update it is referred as *analysis*, a. The updating process is expressed by:

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f \tag{6}$$

where $\mathbf{H} \in \mathbb{R}^{m\times n}$ is the observation operator, with m the number of observations, and $\mathbf{R}$ the observation error covariance matrix.

In the ensemble methods, ~~this equation is written as:~~ using Eq.5 and the approximation $\mathbf{P} \approx \mathbf{P}_e$ in the Eq.6, we obtain:

$$\mathbf{A}^{a\prime}\mathbf{A}^{a\prime T} = \mathbf{A}'(\mathbf{I} - \mathbf{S}^T\mathbf{C}^{-1}\mathbf{S})\mathbf{A}'^T \tag{7}$$

where $\mathbf{S}$ and $\mathbf{C}$ are defined as:

$$\begin{aligned}\mathbf{S} &= \mathbf{H}\mathbf{A}' \\ \mathbf{C} &= \mathbf{S}\mathbf{S}^T + (N-1)\mathbf{R}\end{aligned} \tag{8}$$

After some eigenvalue and singular value decompositions (see the paper Evensen, 2004), the equation splits into two symmetrical parts:

$$\mathbf{A}^{a\prime}\mathbf{A}^{a\prime T} = \left(\mathbf{A}'\mathbf{V}_2\sqrt{\mathbf{I} - \mathbf{\Sigma}_2^T\mathbf{\Sigma}_2}\right)\left(\mathbf{A}'\mathbf{V}_2\sqrt{\mathbf{I} - \mathbf{\Sigma}_2^T\mathbf{\Sigma}_2}\right)^T, \tag{9}$$

where $\mathbf{V}_2 \in \mathbb{R}^{N\times N}$, and $\mathbf{\Sigma}_2 \in \mathbb{R}^{m\times N}$ are two matrices coming from the decomposition of $\mathbf{S}^T\mathbf{C}^{-1}\mathbf{S}$ and $\mathbf{I}$ is the identity matrix. The solutions are:

$$\mathbf{A}^{a\prime} = \mathbf{A}'\mathbf{V}_2\sqrt{\mathbf{I} - \mathbf{\Sigma}_2^T\mathbf{\Sigma}_2}\mathbf{\Theta}^T, \tag{10}$$

for any random orthogonal matrix $\mathbf{\Theta}^T$. This , which allows a random redistribution of the variance reduction among the ensemble members.

The approximation of the covariance matrix with the ensemble members perturbations (eq. 5) becomes perfect when N goes to infinity. However, with a finite number of ensemble members, model variables that are far from each other and not really correlated, can have have a variance different from zero. To avoid this issue, keeping a reasonable number of ensemble members, we apply a localisation scheme. Localisation is often used in ensemble data assimilation and can be done following different methods (Houtekamer and Mitchell, 2001; Hamill et al., 2001; Anderson, 2003). In the present case we used a local analysis method, which performs in a similar way of the covariance localisation method (Sakov and Bertino, 2011).

This method reduces the influence of the observations too far from the location of a model variable the model variable which is going to be modified. If the model has N variables, the distance of each of them from each observation is computed and a weighting factor, depending on such distance, is computed. We used a Gaspari-Cohn function (Gaspari and Cohn, 1999), with which the weight decreases in a way similar to a Gaussian, but vanishes for distances $r > 2d$, where d is a *cut-off* distance. Instead of making a global analysis, the analysis is made for each node of the grid near enough to some observations and the matrices are reduced to a local dimension. Then, the total analysis is the sum of all the local contributions (Carrassi et al., 2018).

2.3 The optimization procedure

Starting from the DA run with the assimilation of all stations (N), the monitoring network evaluation procedure was designed as an iterative process in which several numerical simulations are carried out excluding one tide gauge from the assimilation at a time. In this study, we consider the root mean square error (RMSE) of the simulated values respect to the observations as the cost function to be minimized in the optimization process. Similar to the approach described in the previous section, for each run the RMSE is evaluated for all data points. After doing this for all remaining stations, the observation site that contributes less to the improvement of the RMSE (the one having the lowest RMSE value) is excluded in the next optimization step (assimilation of $N-1$ stations). The iterative process continues ($N-2, N-3, N-4$, ...) until only one station is assimilated. At each optimization step, the mean RMSE over the whole monitoring network is evaluated. The whole optimization procedure requires $N \times (N+1)/2$ numerical simulations. In the case of the DA-EnSRF, the computational effort is much higher and depends on the number of members of the ensemble.

The optimization procedure is easily and efficiently parallelized since all simulations within each iteration step are independent of each other. Similarly, all members of each DA-EnSRF process are independent and can be carried out simultaneously on different processors.

2.4 Application to the Lagoon of Venice

The Lagoon of Venice (Fig. 1) is situated in the Northern Adriatic Sea and is the largest Mediterranean lagoon (area of 550 km^2). The principal hydraulic forcings of the Lagoon of Venice are the tide and the wind Umgiesser et al. (2004b). Even if the lagoon is a micro-tidal system (tidal range of about 80 cm), tides are a major factor in shaping landforms and driving

ecological gradients and biological communities. The lagoon is separated from the open sea by barrier islands, and three inlets (Lido, Malamocco, and Chioggia) ensure an active renewal of the lagoon waters (Ferrarin et al., 2017). The lagoon is characterized by a complex system of tidal channels. The density of the drainage network increases landward as main tidal collectors departing from the inlets branch in progressively smaller-size channels, ranging in depth from a more than 15 m of main reaches to few decimetres of salt marsh creeks (Madricardo et al., 2017). Such a drainage network cuts across a large extent of shallow water areas, which have an average depth of 1 m and include mudflats and salt marshes.

The city of Venice is located in the centre of the lagoon and is composed of more than a hundred islands linked by bridges. The elevation of these islands is extremely low, subjecting them to flooding during storms, which in turn threatens the unique cultural heritage of this city and affects its everyday life. The northern Adriatic Sea is frequently affected by storm surge events, mainly triggered by strong south-easterly wind (Orlić et al., 1994). It is therefore of crucial importance for the management of this environment to monitor water level variations outside and inside the lagoon.

2.4.1 The tide gauge network

The Lagoon of Venice has two tide gauge networks for supporting the local real-time storm surge prediction and warning system. They are managed by the Institute for Environmental Protection and Research - National Centre for Coastal Zone and Characterization Marine Climatology and for Operational Oceanography (ISPRA, Unit for Tides and Lagoons, http://www.venezia.isprambiente.it/, last access 10 January 2020) and the Tide Forecast and Early Warning Center of the City of Venice (CPSM, https://www.comune.venezia.it/it/content/centro-previsioni-e-segnalazioni-maree, last access 10 January 2020). ISPRA manages a network of 45 tide gauge stations equipped for the systematic measurement of water level and other related parameters, such as wind direction, wind speed, atmospheric pressure, precipitation, and wave-height inside the Lagoon of Venice and in the north-western Adriatic coastline. The monitoring network of CPSM consists of 17 hydro-meteorological stations distributed within the lagoon and along the Venetian littoral for the real-time monitoring of the water levels, waves and meteorological parameters. Some locations with high valuable relevance are monitored by both institutions.

In this study, we collected all the available data from both the ISPRA and CPSM monitoring networks over a one-month period (November 2013) with the highest number of stations without missing data. The selected dataset consists of quality-controlled 10-minute values of sea level measured at the 29 tide gauge stations marked with red dots in Fig. 1. As shown in the figure, all tide gauges are installed within navigational channels in order to allow their installation and maintenance. Most of the tide gauges are located in the central and northern parts of the lagoons, where most of the urban settlements are placed (Venice, Murano and Burano), at the inlets and in the southern end of the lagoon near Chioggia. The selected period of investigation comprises both calm weather conditions as well as significant wind events.

In order to investigate at which degree the observations represent the state variable over the whole system, a field approximation through optimal interpolation (OI) of the data have been performed. OI is a commonly used and fairly simple method to perform interpolation of sparse data and also in data assimilation. OI was first described in Gandin (1965) and other references and implementations can be found also in Daley (1991). It is also often referred to as statistical interpolation. In OI, starting

from a background grid, observation points are used to correct the background grid. Points that lie close to each other are given less weight. The interpolation of the water levels was carried out on a 0.5×0.5 km regular grid.

2.4.2 Simulation set-up

The water circulation in the Lagoon of Venice, induced by tide and wind was simulated by the unstructured model SHYFEM applied over a spatial domain that represents the entire Lagoon and its adjacent shore. The model adequately reproduces the complex geometry and bathymetry of the Lagoon of Venice using unstructured numerical meshes composed of triangular elements of variable form and size, going down to a few meters in the channels (Fig. 1). The model bathymetry was obtained from the data collected in 2002 by Magistrato alle Acque di Venezia - merged with later surveys - and the 2014 MBES bathymetry acquired by CNR-ISMAR in the main channels of the lagoon (Madricardo et al., 2017).

The application of the SHYFEM model to the Lagoon of Venice has been validated in previous works reproducing correctly tidal propagation, storm surge, water flows at the lagoons' inlets and water temperature and salinity variability (Umgiesser et al., 2004a; Ferrarin et al., 2008, 2010; Ghezzo et al., 2011).

In this study, hydrodynamics in the lagoon was simulated using 10-minute observed forcing and boundary conditions (i.e., wind stress and open sea level). The initial condition is always a calm state. This is certainly no problem for the current velocity and the water level since these quantities approach a dynamic state very fast (less than a day). The numerical simulations were performed over the period covered by the selected dataset (November 2013).

In order to apply the Nudging DA method, a value for the relaxation parameter τ has to be determined. In our case, it was supposed that every observation point would only influence the grid points up to a certain distance. For every observation, a Gaussian bell curve was constructed. The standard deviation of the curve (~~*sigma*~~σ) was set to 2 km, and all points further than 3 standard deviations are excluded from the computations (Fig. 2a). Overlapping areas of influence are considered by summing the value of the Gaussian curve in these points. The τ value at the peak point of the Gaussian curve was set to 100 seconds, and this value then increases smoothly to infinity in order to simulate an influence which becomes lower when moving away from the observation point.

The EnSRF needs an ensemble of model states that should ideally represent the error of the simulation. In the present case the ensemble of the model states is created varying the boundary condition. We used 60 perturbations for the sea-level boundary condition ~~,~~ (member 0 is unperturbed) taken from a Gaussian distribution with a zero mean and a standard deviation set to 30 cm. This value was found empirically, in order to have a good spread at the boundary, which is then propagated to the variables computed by the model. As asserted, the perturbations are centred, having a null mean, and correlated in time~~with~~. To do this, each perturbation at time t is obtained from a weighted average of a new perturbation and of the one at time $t-1$. This type of perturbations are classified as red noise and in the present case we used a decay time of two days~~(red noise)~~. We made also perturbations for the wind, with the same method, but using a standard deviation proportional to 40% of the wind speed. Due to the small study area, we considered the wind constant in space so that the perturbations can vary only in time, as the boundary conditions. However, because of the smallness of our system, the perturbations on the wind are not very effective,

as well as perturbations on the initial state. Therefore, the perturbations at the boundary condition are necessary both to create the initial ensemble of states and to keep the spread of the ensemble during the whole time of the simulation.

~~We run also several preliminary tests to empirically found~~ After several preliminary numerical tests, the best cut-off distance for the local analysis ~~, which was fixed~~ was set to 0.1 geographical degrees (about 10 km). In order to illustrate the important effect of the localisation, in Fig. 2b we show the correlation values between each observation station and each model level in each node of the model grid, at a specific time-step. The correlation is weighted with the Gaspari-Cohn function, which vanishes the values too far from the station. This quantity is not used directly by the local analysis routine, but it is useful to understand its effect. Note also that this is the correlation with the water levels, but the EnSRF considers also the cross-correlations with the water velocities and corrects them as well. The strong difference with the relaxation time used by the nudging to weight the observations (Fig. 2a), is that the use of the real correlations between the model variables produces an anisotropic distribution of the observation correction, which respects the water dynamics forced by the channels, by the tidal flats and by the basin morphology. Moreover, as the dynamics varies at each time-step, so does the correlation between model variables and the weight of the assimilation increments.

The EnSRF assimilates water level from the selected stations considering them independent (the R error covariance matrix is diagonal) and the error of each station is set to 1 cm. The model evolves forward in time the ensemble members, each one with different boundary condition and wind forcing, and an analysis step is done every hour. The results considered in this work are extracted by the analysis states, which are saved every hour.

3 Results

In the exposition of the results, we defined the model run without data assimilation as the control simulation, while, for both the DA schemes, the base run accounts for the assimilation of all the 29 monitoring stations. All mentioned parameters (~~*tau*, *sigma*~~τ, σ, cut-off distance for the local analysis) were manually defined through trial and error calibration process and evaluating the goodness-of-fit of the water level RMSE in the DA-Nudging and DA-EnSRF base simulations.

3.1 Data interpolation vs. data assimilation

When entering a shallow basin, as the Venice lagoon, the tidal wave is deformed, either damped or amplified, according to a relationship between local flow resistance and inertia, and the characteristics of the incoming tidal wave (Ferrarin et al., 2010). In the data interpolation method, the distribution of the water levels is given by a spatial interpolation of the observations. Fig. 3a reports a snapshot of the interpolated water levels over the lagoon during a flood tide. The map shows, for this particular time frame, a patchy non coherent distribution with the lowest values in the nearshore area close to the inlets, while the highest are in proximity of the central and northern lagoon's margins.

Does the interpolation of the observations provide a realistic spatial representation of the water level variability over the lagoon domain? To answer to this question, we show in Fig. 3b the water level computed by the model, without any data assimilation (Control sim.). The nudging run (DA-Nudging) is shown in Fig. 3c and the EnSRF run (DA-EnSRF) in Fig. 3d.

The control simulation has a completely different distribution of the water levels with respect to the data interpolation. The mode simulation shows the lowest water level in the open sea, which gradually increases going from the inlets to the inner lagoon, describing the propagation of the tidal wave. The three inlets lead the water circulation in three sub-basins, divided by narrow areas with little water exchange (these zones are identified as dynamical watersheds). The modelled maps (control, DA-Nudging and DA-EnSRF) clearly account for islands and marsh boundaries. DA-Nudging shows a similar representation of the control simulation, but with slightly higher values of the water levels on the central and southern tidal flats (Fig. 3c). Similarly, the DA-EnSRF adjusts the water levels towards the observations while keeping the physical dynamics of the flow (Fig. 3d). It is worth mentioning that the water level distributions at different tidal phases would lead to similar DI and DA considerations.

In order to establish which method better represents the water level variability over the lagoon, we need to evaluate the capacity of each approach to describe the parameter at locations not included in the computation. Thanks to a large number of available tide gauges in the Lagoon of Venice, the model skill assessment (in terms of the root mean square error, RMSE) is determined by re-running DI and DA experiments removing one station from the assimilation and comparing the water level in this station with the modelled one. The evaluation procedure was repeated for each monitoring station and the results are reported in Table 1. When using the optimal interpolation approach, the average RMSE is 3.9 cm, with values ranging from 0.8 to 8.5 cm. The highest RMSE is found at stations located at the lagoon margins (9, 14, 25 and 27) and the Chioggia and Malamocco inlets (4 and 12). The control SHYFEM simulation, the one without data assimilation, has a mean RMSE of 5.8 cm, with the highest errors found at the stations located near the lagoon margins (1, 9, 14, 24 and 29). The correlation coefficient (not reported in the table) is everywhere higher than 0.97, except for station 24 where it is 0.47. Therefore, from the statistics we deduce that the control simulation has a worse performance with respect to the direct interpolation of the data and that it slightly fails in reproducing correctly the water dynamics in border areas, especially in the small creeks surrounded by mashes (e.g. station 24). However, even if data interpolation is statistically better, looking at Fig. 3a the spatial distribution of the water level is clearly unphysical.

Differently, both DA methods strongly improved the model skills in all parts of the lagoon. The average RMSE resulted in 2.1 cm and 3.2 cm for DA-Nudging and DA-EnSRF, respectively. The results reported in Table 1 show that results improved at all stations, even those affected by the highest errors in the control simulation. The capacity of the different methods in reproducing the temporal evolution of the water level is shown in Fig. 4 for the station n. 12 (in this case removed from the assimilation/interpolation). It is evident that in this case, the interpolation does not represent correctly the water level variability, being influenced by values recorded outside the lagoon domain, which do not take into account the correct tidal propagation dynamics. On the other hand, the data assimilation results adjust the water levels towards the observations keeping the physical dynamics of the flow. Therefore, the model simulation with a DA scheme is the approach that better represents the variability of the water levels in the lagoon.

Additionally, in a multivariate analysis approach we tested the capability of the applied DA-driven simulations in reproducing the current velocities recorded by an acoustic Doppler current profiler (ADCP) mounted on the bottom of the Lido inlet, close to the station n. 15 shown in Fig. 1. Time series of observed and simulated vertically integrated velocities are illustrated in

Fig. 5, while the statistical results are summarized in Table 2. Since the DA-Nudging does not adjust the velocities according to the correction of the water level, the model computes spurious velocities according to the pressure gradients generated by the water level increments. The ~~simulated~~ DA-Nudging current velocities - and therefore the water exchange through the inlets - ~~using DA-Nudging~~ resulted to be ~~higher~~ overestimated and slightly out of ~~phase than the observations. Interesting~~tidal phase. Interestingly, the DA-Nudging performances on the current velocity are even worse than those of the control simulation. On the other hand, since DA-EnSRF uses cross-correlation to propagate the observation correction to the other model variables, the velocities are corrected according to the modification of the levels, towards a better agreement with the ADCP currents. This is a demonstration of the potentiality of a complex DA method, where a correct specification of the cross-correlations in the background covariance matrix allows a correction of model variables even if they are not directly correlated with the assimilated quantities.

3.2 Monitoring network optimization

The next step is to use DA methods to find the minimum number of stations - and their distribution - that correctly represent the state variable in the investigated system. The optimization procedure of this tide gauge network, composed ~~by~~ of $N = 29$ stations, requires 435 $(N \times (N+1)/2)$ numerical simulations. However, the computational cost of the DA-EnSRF is much higher, since the ensemble is composed by 61 members. So in this case the simulations are 26,535, but the computing scalability is high since the 61 simulations ~~are independent~~of the ensemble are independent and can be parallelized on multiple CPUs computers. The results of the water level observatory evaluation are reported in Fig. 6 in terms of the model RMSE as a function of the number of stations considered in the assimilation. For comparison, the same procedure was applied to the data interpolation.

The evaluation procedure allows finding the minimum number of tide gauges for a successful description of the water level in the lagoon. However, the optimization criterion (the RMSE threshold) is arbitrary and may differ for different environments, state variables and monitoring networks. In the present case, we can see that using both DA-Nudging and DA-EnSRF, the RMSE does not change too much passing from 29 to 10-12 assimilated stations. Even if the EnSRF has an average RMSE higher than the DA-Nudging, the RMSE of the EnSRF has a slower increase with the reduction of the stations. The initial decrease of the RMSE is probably due to the fact that observations have errors, and tide gauges close to each other can provide slightly different data. The EnSRF considers the observation error in the observation covariance matrix, but it is difficult to find the right value and normally the nominal instrument error is used.

Considering the spatial interpolation method, the use of 10 stations has a RMSE comparable to the error of the control simulation. But we have to stress that in this case the spatial representation of the water level is clearly wrong. We should also mention that the model with the assimilation of only three stations ~~give~~ gives a lower RMSE than DI with all 29 stations, apart from the fact that results are physically more coherent and consistent.

The resulting optimal distributions of the 10 tide gauge stations determined by DA-Nudging and DA-EnSRF are shown in Fig. 7. In both cases, the optimization procedure selected tide gauges located near the inlets (one each, avoiding redundancy of nearby stations), in some of the islands in the northern part of the lagoon, and stations along the lagoon margins. We therefore

can consider that, with the help of DA methods, only 10 of the considered 29 tide gauges are necessary for properly describing the spatial and temporal variability of the water level in the Lagoon of Venice. Considering that the average annual maintenance cost of a tide gauge in the Lagoon of Venice is approximately 3,500 € (Alvise Papa, CPSM, personal communication), the optimization of the monitoring network could allow saving about 66,000 € per year.

However, the choice of which stations to keep in the monitoring network depends also on many practical factors. As an example, the monitoring authority would decide to keep some stations because of their strategic relevance, maintenance costs, distance from the laboratory or for continuing long-term time series. The optimization method can be easily customized based on predetermined specific constraints. As a realistic exercise, we fixed the stations at the inlets (4, 12, 11) and in the main urban settlements (2, 6, 17, 19) in the monitoring network. The evaluation procedure is then repeated using the DA-Nudging method, keeping these 7 stations and the results are presented in Fig. 8. In this customized optimization exercise, the results show that 15 stations are necessary to guarantee a proper description of the water level variability in the lagoon.

4 Discussion and Conclusions

The methodology presented in this study allows for the evaluation of existing coastal observatories. Using a DA system, which is an observation-driven and process-based method, the iterative optimization procedure establishes the relevance of each single monitoring station on the description of the considered environment. The example reported in this study describes the optimization of an existing observatory with defined monitoring points. However, the methodology could be applied also to design new monitoring networks. As described by Raicich (2006) and Xue et al. (2011), in an observing system simulation experiment, synthetic observations are generated by a model run in some locations and then they are assimilated as real observations. The procedure is similar to a *twin* experiment, a method used to assess the quality of a data assimilation system.

As indicated by Fujii et al. (2019), the goodness of the results of such methods strongly depends on the numerical model applied, on the DA scheme implemented and on the optimization procedure. This effect is evident in the results presented above, where the numerical model performances differ when using a different methodology for assimilating the observations. Moreover, the optimization procedure selected some stations at the lagoon edges, where the RMSE of the control simulation was the highest. The DA scheme should be selected not only considering the computational cost, but also considering the capacity in reproducing a correct multivariate dynamics of the system. This can be done using observations not assimilated of the same type of the assimilated ones and also observations of other variables of the model (as the ADCP data in our case). In semi-enclosed basins such as lagoon environments, the fluxes through the inlets control the water and the sediment and the nutrient exchanges between the sea and the lagoon, influencing the whole dynamic of the system (Ferrarin et al., 2010). Indeed, the more advanced EnSRF method improved not only the assimilated water level but also the current velocity, and therefore the fluxes, at the inlet. Therefore, as also outlined by many authors (e.g. Jones et al., 2012; Edwards et al., 2015; Bajo et al., 2019), the description of the coastal sea environment can be improved with the use of a modelling, process-based approach and the use of observations in a complex data assimilation system.

Additionally, as specified at section 2.4.2, the perturbation method implemented in the ensemble data assimilation system allows the creation of ensemble members that are dynamically consistent and generates realistic correlations in the background-error covariance matrix. These ~~correlation~~correlations, as well as the covariance matrix, are not constant in time, but vary accordingly with the dynamics induced by the periodic tide and by the non-periodic stress of the wind. In designing or optimizing a monitoring network, such correlation matrix represents a precious source of information which can be used to investigate the area of representativeness of each selected station. To better understand the potentiality of the ensemble data assimilation methods, we show in Fig. 9 the correlation between the sea level at each station locations with the other nodes of the grid, weighted by the Gaspari-Cohn function. The figure is similar to Fig. 2b but in this case the correlations are averaged over the duration of the whole simulation and considers only the stations selected by the optimization procedure. Even at a first glance, this maps give information about the influence area of each station. These areas do not spread isotropically from the station locations, but they are constrained by the morphology and by the water dynamics, which is considered in the model. This is true not only for the water level but, as asserted before, also the other variables should benefit by the assimilation of water level observations. Maps similar to that in Fig. 9 can be obtained considering the cross-correlation of the sea level with the water current or with other variables like temperature or salinity, in case of a baroclinic model.

The combination of observations and numerical models is particularly important in coastal regions with scarce monitoring resources. However, to reduce the model error, the applied numerical models must correctly reproduce the complex morphology of the coastline and the exchange processes between the shelf and the open seas. The processes in such complex systems at the land-sea transition are extremely dynamic and require a holistic approach in which all the hydrological entities (river mouth, salt marshes, lagoons, swamps, coastal sea) should be considered as integral parts of the entire domain of computation. Moreover, due to the complex geometry and morphology of the coastal regions, the numerical models need to be able to represent hydrodynamic conditions with very high resolution, on the horizontal, vertical and temporal dimensions. With respect to the above-cited requirement, unstructured models - as the one applied in this study - realise a seamless transition between different spatial scales for reproducing the coastal-sea interactions, adopting a variable resolution of the mesh elements (Ferrarin et al., 2018; Kärnä et al., 2018; Maicu et al., 2018; Stanev et al., 2018; Androsov et al., 2019). The applied numerical models need to be continuously evaluated and upgraded to maintain the highest accuracy.

The model-driven optimization procedure was here applied using hindcast simulations, but it can be also used in an forecasting modelling for evaluating the effect of the assimilated data on the predictions (Cummings and Smedstad, 2014; Bajo et al., 2017). An observation assessment is particularly important when the assimilated data come from different data sources (e.g., fixed monitoring stations, satellite, radar, gliders), or for a priori estimation of new data sources in an already existing DA system (Bonaduce et al., 2018). It is crucial in operational oceanography to have a DA scheme keeping the correct physical description of the dynamics in the investigated environment, without introducing errors that can propagate in time. As indicated by Fujii et al. (2019), in an operational framework a DA system can also be used as an automatic control system for the quality of observations.

In the case of the Lagoon of Venice tide gauge network, we demonstrated how numerical models with data assimilation can play a valuable role in optimizing and designing coastal observatories. The iterative optimization process was based on the

evaluation of the RMSE at the stations not assimilated. It is worth noting that the existing monitoring network can be reduced by a factor of 2/3 using the tide gauge system in conjunction with a high-resolution numerical model, by means of DA. The applied methodology is easily exportable to other coastal environments and can be extended to other physical variables.

Code and data availability. The SHYFEM hydrodynamic model is open source (GNU General Public License as published by the Free Software Foundation) and freely available through GitHub at https://github.com/SHYFEM-model (last access: 10 January 2020). The SHYFEM code version v. 7_5_65 can be accessed from Zenodo (Umgiesser, 2019). The SHYFEM model v. 7_5_65 with with the data assimilation code (version ens2.1) is available on Zenodo (Bajo, 2020). The data assimilation code is based on the Geir Evensen's routines, available at the web-page https://github.com/geirev/EnKF_analysis (last access: 20 January 2020). Configuration files, data and scripts used to run the models and analyse the results presented in this work are available on Zenodo (Ferrarin et al., 2020).

Author contributions. GU conceived the idea of the study with the support of CF. GU developed the optimization procedure and the nudging data assimilation routines, and MB developed the Ensemble Square Root Filter data assimilation software. CF and MB performed the numerical simulations. All authors discussed, reviewed and edited the different versions of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Venezia2021 research program funded by the Provveditorato for the Public Works of Veneto, Trentino Alto Adige and Friuli Venezia Giulia, provided through the concessionary of State Consorzio Venezia Nuova and coordinated by CORILA. The authors wish to thank the Tide Forecast and Early Warning Center of the City of Venice and the Italian Institute for Environmental Protection and Research (ISPRA) for providing tide gauge and current velocity data.

References

Anderson, J. L.: A Local Least Squares Framework for Ensemble Filtering, Mon. Weather Rev., 131, 634–642, https://doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2, 2003.

Androsov, A., Fofonova, V., Kuznetsov, I., Danilov, S., Rakowsky, N., Harig, S., Brix, H., and Wiltshire, K. H.: FESOM-C v.2: coastal dynamics on hybrid unstructured meshes, Geosci. Model Dev., 12, 1009–1028, https://doi.org/10.5194/gmd-12-1009-2019, 2019.

Bajo, M.: SHYFEM v. 7_5_65 with the data assimilation code version ens2.1, https://doi.org/10.5281/zenodo.3757843, 2020.

Bajo, M., De Biasio, F., Umgiesser, G., Vignudelli, S., and Zecchetto, S.: Impact of using scatterometer and altimeter data on storm surge forecasting, Ocean Model., 113, 85 – 94, https://doi.org/10.1016/j.ocemod.2017.03.014, 2017.

Bajo, M., Medugorac, I., Umgiesser, G., and Orlić, M.: Storm surge and seiche modelling in the Adriatic Sea and the impact of data assimilation, Quart. J. Roy. Meteor. Soc., 145, 2070–2084, https://doi.org/10.1002/qj.3544, 2019.

Bellafiore, D., Mc Kiver, W., Ferrarin, C., and Umgiesser, G.: The importance of modeling nonhydrostatic processes for dense water reproduction in the southern Adriatic Sea, Ocean Model., 125, 22–28, https://doi.org/10.1016/j.ocemod.2018.03.001, 2018.

Blayo, E., Verron, J., and Molines, J. M.: Assimilation of TOPEX/POSEIDON altimeter data into a circulation model of the North Atlantic, J. Geophys. Res., 99, 24 691, https://doi.org/10.1029/94jc01644, 1994.

Blumberg, A. and Mellor, G. L.: A description of a three-dimensional coastal ocean circulation model, in: Three-Dimensional Coastal Ocean Models, edited by N. S. Heaps, pp. 1–16, AGU, Washington, DC, 1987.

Bonaduce, A., Benkiran, M., Remy, E., Traon, P. Y. L., and Garric, G.: Contribution of future wide-swath altimetry missions to ocean analysis and forecasting, Ocean Sci., 14, 1405–1421, https://doi.org/10.5194/os-14-1405-2018, 2018.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Chaumillon, E., Bertin, X., Fortunato, A., Bajo, M., Schneider, J.-L., Dezileau, L., Walsh, J. P., Michelot, A., Chauveau, E., Créach, A., Hénaff, A., Sauzeau, T., Waeles, B., Gervais, B., Jan, G., Baumann, J., Breilh, J.-F., and Pedreros, R.: Storm-induced marine flooding: lessons from a multidisciplinary approach, Earth-Sci. Rev., 165, 151 – 184, https://doi.org/10.1016/j.earscirev.2016.12.005, 2017.

Cummings, J. A. and Smedstad, O. M.: Ocean Data Impacts in Global HYCOM, J. Atmos. Ocean. Technol., 31, 1771–1791, https://doi.org/10.1175/jtech-d-14-00011.1, 2014.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, Cambridge, UK, pp. 457, 1991.

Edwards, C. A., Moore, A. M., Hoteit, I., and Cornuelle, B. D.: Regional Ocean Data Assimilation, Annu. Rev. Mar. Sci., 7, 21–42, https://doi.org/10.1146/annurev-marine-010814-015821, 2015.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dyn., 53, 343–367, https://doi.org/10.1007/s10236-003-0036-9, 2003.

Evensen, G.: Sampling strategies and square root analysis schemes for the EnKF, Ocean Dyn., 54, 539–560, https://doi.org/10.1007/s10236-004-0099-2, 2004.

Ferrarin, C., Umgiesser, G., Cucco, A., Hsu, T.-W., Roland, A., and Amos, C. L.: Development and validation of a finite element morphological model for shallow water basins, Coast. Eng., 55, 716–731, https://doi.org/10.1016/j.coastaleng.2008.02.016, 2008.

Ferrarin, C., Cucco, A., Umgiesser, G., Bellafiore, D., and Amos, C. L.: Modelling fluxes of water and sediment between the Venice Lagoon and the sea, Cont. Shelf Res., 30, 904–914, https://doi.org/10.1016/j.csr.2009.08.014, 2010.

Ferrarin, C., Maicu, F., and Umgiesser, G.: The effect of lagoons on Adriatic Sea tidal dynamics, Ocean Model., 119, 57–71, https://doi.org/10.1016/j.ocemod.2017.09.009, 2017.

Ferrarin, C., Bellafiore, D., Sannino, G., Bajo, M., and Umgiesser, G.: Tidal dynamics in the inter-connected Mediterranean, Marmara, Black and Azov seas, Prog. Oceanogr., 161, 102–115, https://doi.org/10.1016/j.pocean.2018.02.006, 2018.

Ferrarin, C., Bajo, M., and Umgiesser, G.: Dataset: SHYFEM set-up for model driven optimization of the tide gauge monitoring network in the Lagoon of Venice, https://doi.org/10.5281/zenodo.3770173, 2020.

Frolov, S., Baptista, A., and Wilkin, M.: Optimizing fixed observational assets in a coastal observatory, Cont. Shelf Res., 28, 2644–2658, https://doi.org/10.1016/j.csr.2008.08.009, 2008.

Fujii, Y., Rémy, E., Zuo, H., Oke, P., Halliwell, G., Gasparin, F., Benkiran, M., Loose, N., Cummings, J., Xie, J., Xue, Y., Masuda, S., Smith, G. C., Balmaseda, M., Germineaud, C., Lea, D. J., Larnicol, G., Bertino, L., Bonaduce, A., Brasseur, P., Donlon, C., Heimbach, P., Kim, Y., Kourafalou, V., Traon, P.-Y. L., Martin, M., Paturi, S., Tranchant, B., and Usui, N.: Observing System Evaluation Based on Ocean Data Assimilation and Prediction Systems: On-Going Challenges and a Future Vision for Designing and Supporting Ocean Observational Networks, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00417, 2019.

Gandin, L. S.: Objective analysis of meteorological fields, Israel Program for Scientific Translation, Jerusalem, pp. 242, 1965.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Q. J. R. Meteorol. Soc., 125, 723–757, https://doi.org/10.1002/qj.49712555417, 1999.

Ghezzo, M., Sarretta, A., Sigovini, M., Guerzoni, S., Tagliapietra, D., and Umgiesser, G.: Modeling the inter-annual variability of salinity in the lagoon of Venice in relation to the water framework directive typologies, Ocean Coast. Manage., 54, 706 – 719, https://doi.org/10.1016/j.ocecoaman.2011.06.007, 2011.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter, Mon. Weather Rev., 129, 2776–2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2, 2001.

Hoke, J. E. and Anthes, R. A.: The Initialization of Numerical Models by a Dynamic-Initialization Technique, Mon. Weather Rev., 104, 1551–1556, https://doi.org/10.1175/1520-0493(1976)104<1551:tionmb>2.0.co;2, 1976.

Houtekamer, P. L. and Mitchell, H. L.: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation, Mon. Weather Rev., 129, 123–137, https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2, 2001.

Jones, E. M., Oke, P. R., Rizwi, F., and Murray, L. M.: Assimilation of glider and mooring data into a coastal ocean model, Ocean Modelling, 47, 1 – 13, https://doi.org/10.1016/j.ocemod.2011.12.009, 2012.

Kärnä, T., Kramer, S. C., Mitchell, L., Ham, D. A., Piggott, M. D., and Baptista, A. M.: Thetis coastal ocean model: discontinuous Galerkin discretization for the three-dimensional hydrostatic equations, Geosci. Model Dev., 11, 4359–4382, https://doi.org/10.5194/gmd-11-4359-2018, 2018.

Levy, G., Vignudelli, S., and Gower, J.: Enabling earth observations in support of global, coastal, ocean, and climate change research and monitoring, Int. J. Remote Sens., 39, 4287–4292, https://doi.org/10.1080/01431161.2018.1464101, 2018.

Madricardo, F., Foglini, F., Kruss, A., Ferrarin, C., Pizzeghello, N. M., Murri, C., Rossi, M., Bajo, M., Bellafiore, D., Campiani, E., Fogarin, S., Grande, V., Janowski, L., Keppel, E., Leidi, E., Lorenzetti, G., Maicu, F., Maselli, V., Mercorella, A., Gavazzi, G. M., Minuzzo, T., Pellegrini, C., Petrizzo, A., Prampolini, M., Remia, A., Rizzetto, F., Rovere, M., Sarretta, A., Sigovini, M., Sinapi, L., Umgiesser, G., and Trincardi, F.: High-resolution multibeam and hydrodynamic datasets of tidal channels and inlets of the Lagoon of Venice, Sci. Data, 4, https://doi.org/10.1038/sdata.2017.121, 2017.

Maicu, F., De Pascalis, F., Ferrarin, C., and Umgiesser, G.: Hydrodynamics of the Po River-Delta-Sea system, J. Geophys. Res. Oceans, 123, 6349–6372, https://doi.org/10.1029/2017JC013601, 2018.

Mey-Frémaux, P. D., Ayoub, N., Barth, A., Brewin, R., Charria, G., Campuzano, F., Ciavatta, S., Cirano, M., Edwards, C. A., Federico, I., Gao, S., Hermosa, I. G., Sotillo, M. G., Hewitt, H., Hole, L. R., Holt, J., King, R., Kourafalou, V., Lu, Y., Mourre, B., Pascual, A., Staneva, J., Stanev, E. V., Wang, H., and Zhu, X.: Model-Observations Synergy in the Coastal Ocean, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00436, 2019.

Orlić, M., Kuzmić, M., and Pasarić, Z.: Response of the Adriatic Sea to the Bora and Sirocco forcing, Cont. Shelf Res., 14, 91 – 116, https://doi.org/10.1016/0278-4343(94)90007-8, 1994.

Raicich, F.: The assessment of temperature and salinity sampling strategies in the Mediterranean Sea: idealized and real cases, Ocean Sci., 2, 97–112, https://doi.org/10.5194/os-2-97-2006, 2006.

Sakov, P. and Bertino, L.: Relation between two common localisation methods for the EnKF, Computational Geosciences, 15, 225–237, https://doi.org/10.1007/s10596-010-9202-6, 2011.

Schulz-Stellenfleth, J. and Stanev, E.: Statistical assessment of ocean observing networks: A study of water level measurements in the German Bight, Ocean Model., 33, 270–282, https://doi.org/10.1016/j.ocemod.2010.03.001, 2010.

Smagorinsky, J.: General circulation experiments with the primitive equations, I. The basic experiment, Mon. Weather Rev., 91, 99–152, 1963.

Stammer, D., Bracco, A., AchutaRao, K., Beal, L., Bindoff, N. L., Braconnot, P., Cai, W., Chen, D., Collins, M., Danabasoglu, G., Dewitte, B., Farneti, R., Fox-Kemper, B., Fyfe, J., Griffies, S. M., Jayne, S. R., Lazar, A., Lengaigne, M., Lin, X., Marsland, S., Minobe, S., Monteiro, P. M. S., Robinson, W., Roxy, M. K., Rykaczewski, R. R., Speich, S., Smith, I. J., Solomon, A., Storto, A., Takahashi, K., Toniazzo, T., and Vialard, J.: Ocean Climate Observing Requirements in Support of Climate Research and Climate Information, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00444, 2019.

Stanev, E., Pein, J., Grashorn, S., Zhang, Y., and Schrum, C.: Dynamics of the Baltic Sea straits via numerical simulation of exchange flows, Ocean Model., 131, 40–58, https://doi.org/10.1016/j.ocemod.2018.08.009, 2018.

Stauffer, D. R. and Seaman, N. L.: Use of Four-Dimensional Data Assimilation in a Limited-Area Mesoscale Model. Part I: Experiments with Synoptic-Scale Data, Mon. Weather Rev., 118, 1250–1277, https://doi.org/10.1175/1520-0493(1990)118<1250:uofdda>2.0.co;2, 1990.

Trowbridge, J., Weller, R., Kelley, D., Dever, E., Plueddemann, A., Barth, J. A., and Kawka, O.: The Ocean Observatories Initiative, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00074, 2019.

Umgiesser, G.: SHYFEM v. 7_5_65, https://doi.org/10.5281/zenodo.3757785, 2019.

Umgiesser, G., Melaku Canu, D., Cucco, A., and Solidoro, C.: A finite element model for the Venice Lagoon. Development, set up, calibration and validation, J. Mar. Syst., 51, 123–145, https://doi.org/10.1016/j.jmarsys.2004.05.009, 2004a.

Umgiesser, G., Sclavo, M., Carniel, S., and Bergamasco, A.: Exploring the bottom shear stress variability in the Venice Lagoon, J. Mar. Syst., 51, 161–178, 2004b.

Umgiesser, G., Ferrarin, C., Cucco, A., De Pascalis, F., Bellafiore, D., Ghezzo, M., and Bajo, M.: Comparative hydrodynamics of 10 Mediterranean lagoons by means of numerical modeling, J. Geophys. Res. Oceans, 119, 2212–2226, https://doi.org/10.1002/2013JC009512, 2014.

Verron, J.: Altimeter data assimilation into an ocean circulation model: Sensitivity to orbital parameters, J. Geophys. Res., 95, 11 443, https://doi.org/10.1029/jc095ic07p11443, 1990.

Xue, P., Chen, C., Beardsley, R. C., and Limeburner, R.: Observing system simulation experiments with ensemble Kalman filters in Nantucket Sound, Massachusetts, J. Geophys. Res., 116, https://doi.org/10.1029/2010jc006428, 2011.

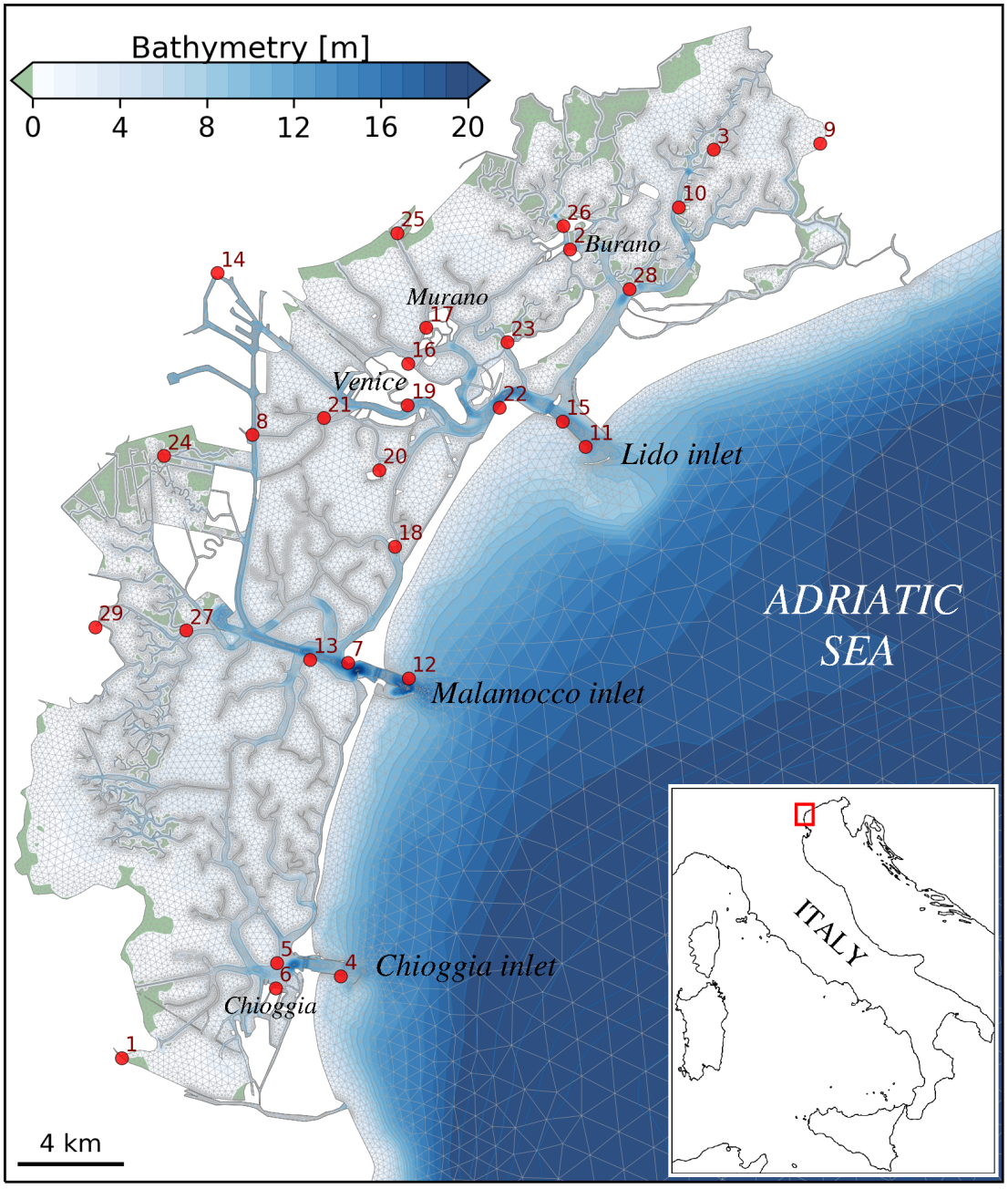

Figure 1. Bathymetry and unstructured mesh of the Lagoon of Venice. The red dots mark the tide gauge monitoring station.

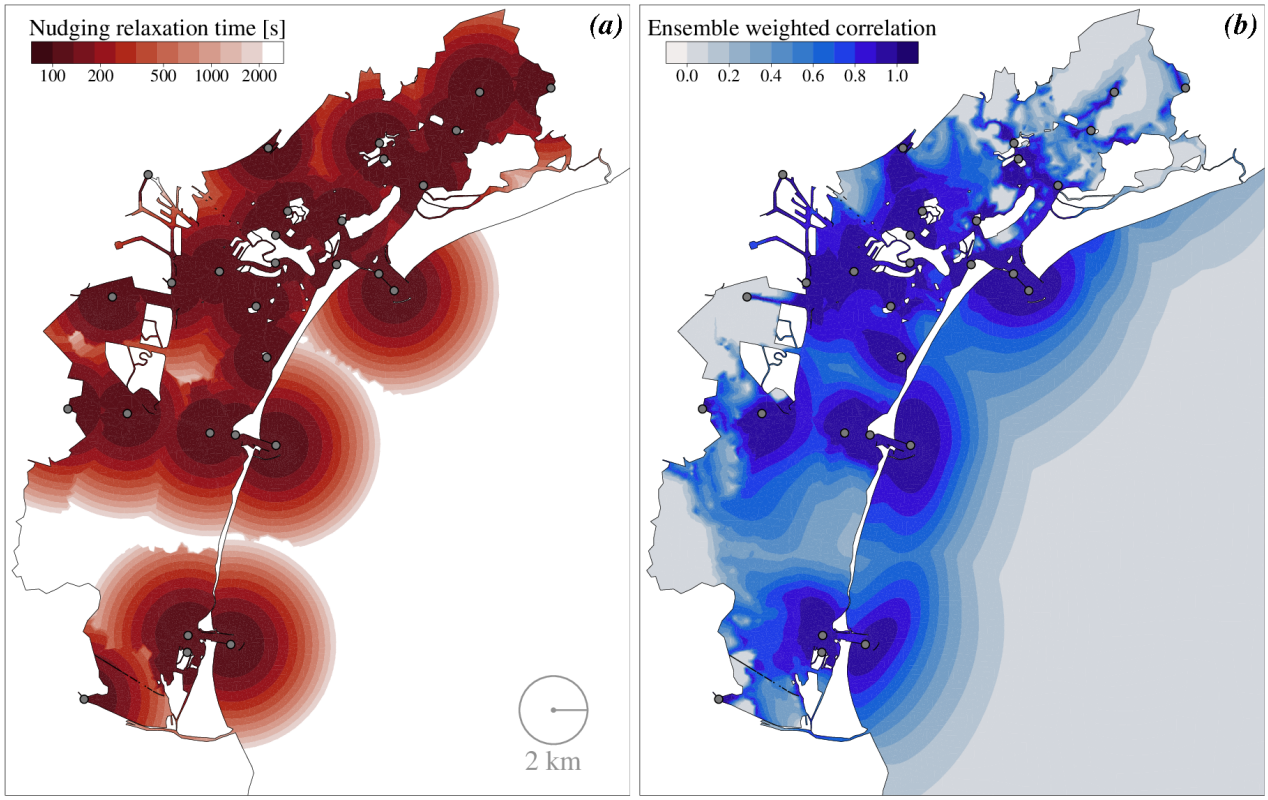

Figure 2. Spatial distribution of (a) the relaxation time adopted in the Nudging and (b) the weighted correlation of the ensemble considered in the EnSRF method.

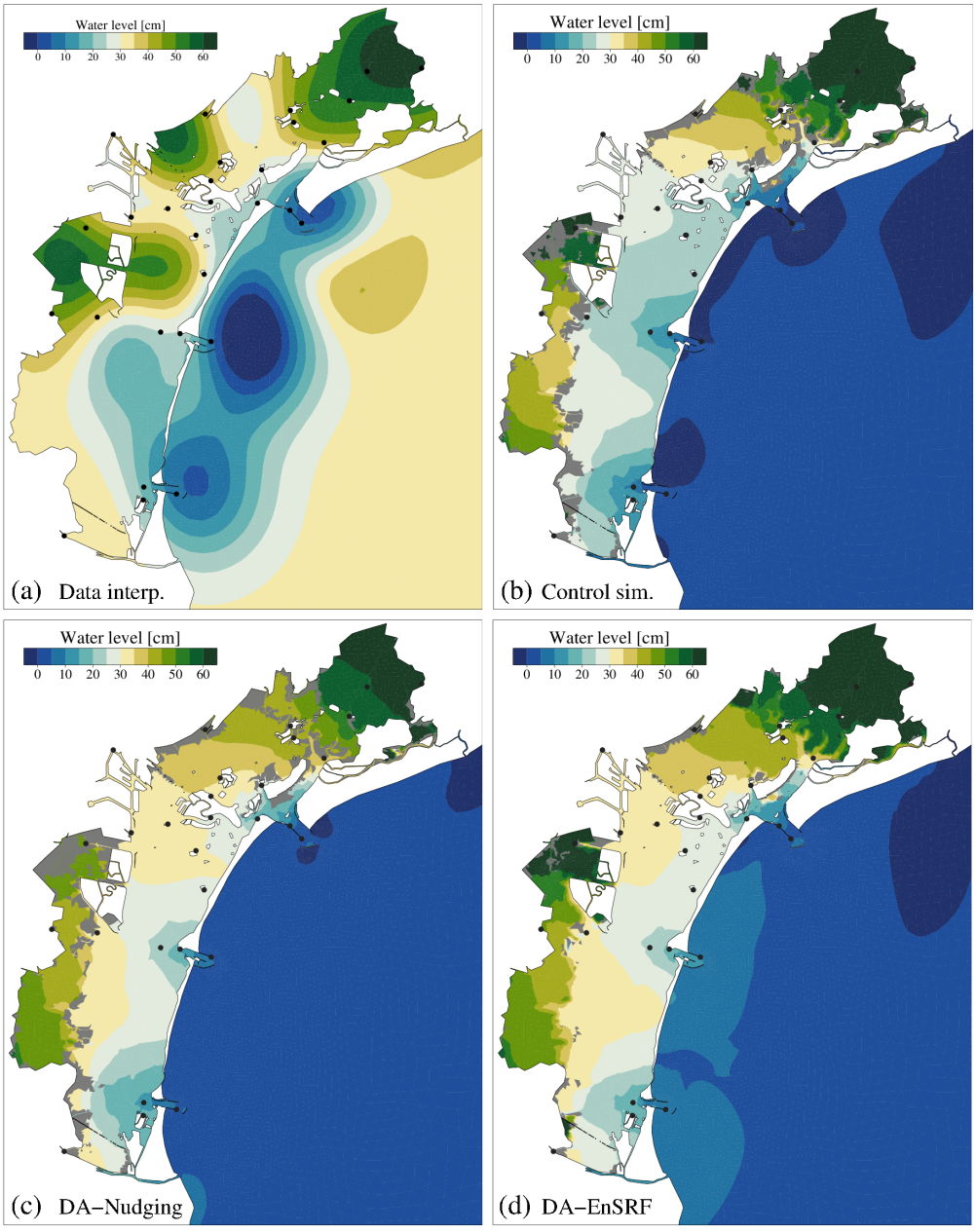

Figure 3. A snapshot on 2013-11-04 at 14:00 UTC of the water level distribution in the Lagoon of Venice as obtained by the optimal interpolation (a), the control simulation without assimilation (b), the DA-Nudging base run (c) and the DA-EnSRF base run (d). The gray colour indicates dry salt marshes.

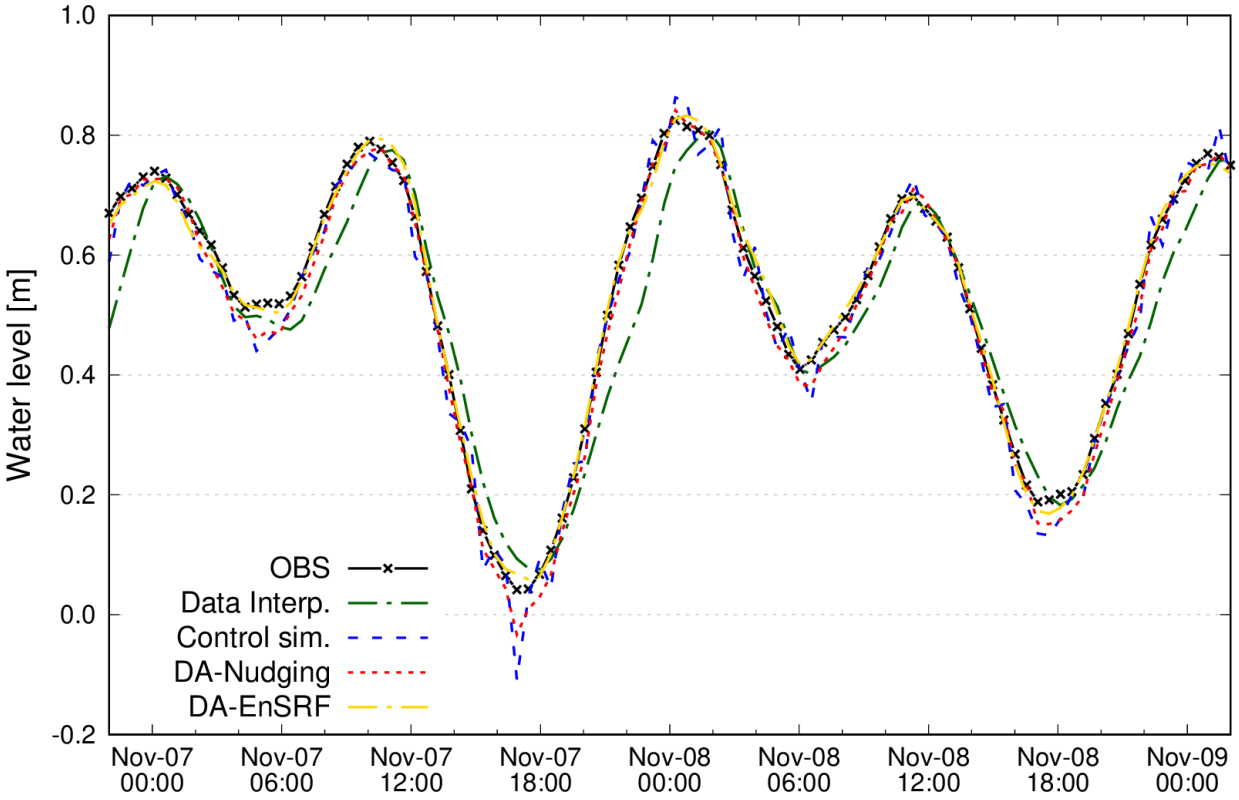

Figure 4. Observed, interpolated and simulated water levels at station 12. In this computation, station 12 was not included in the Di and DA algorithms.

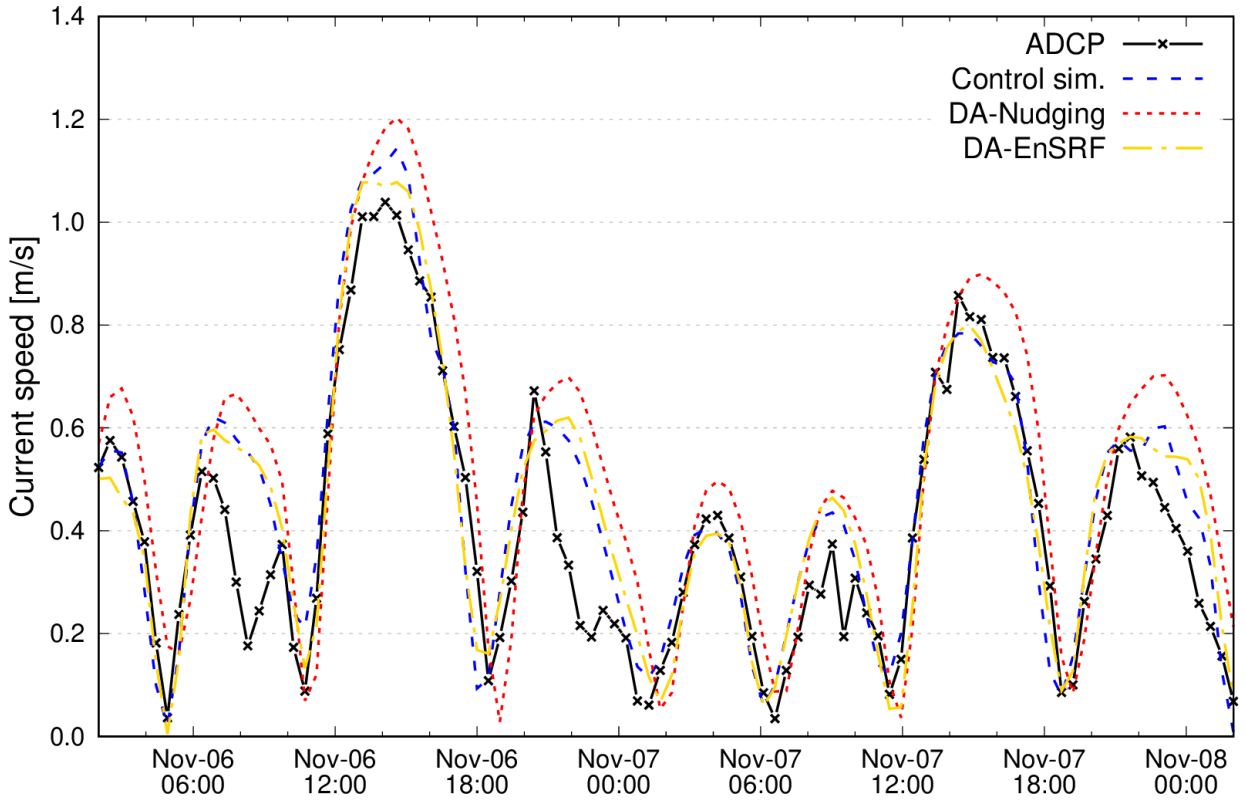

Figure 5. Observed and simulated vertically integrated current velocity at the Lido inlet. The ADCP was located close to tide gauge number 15.

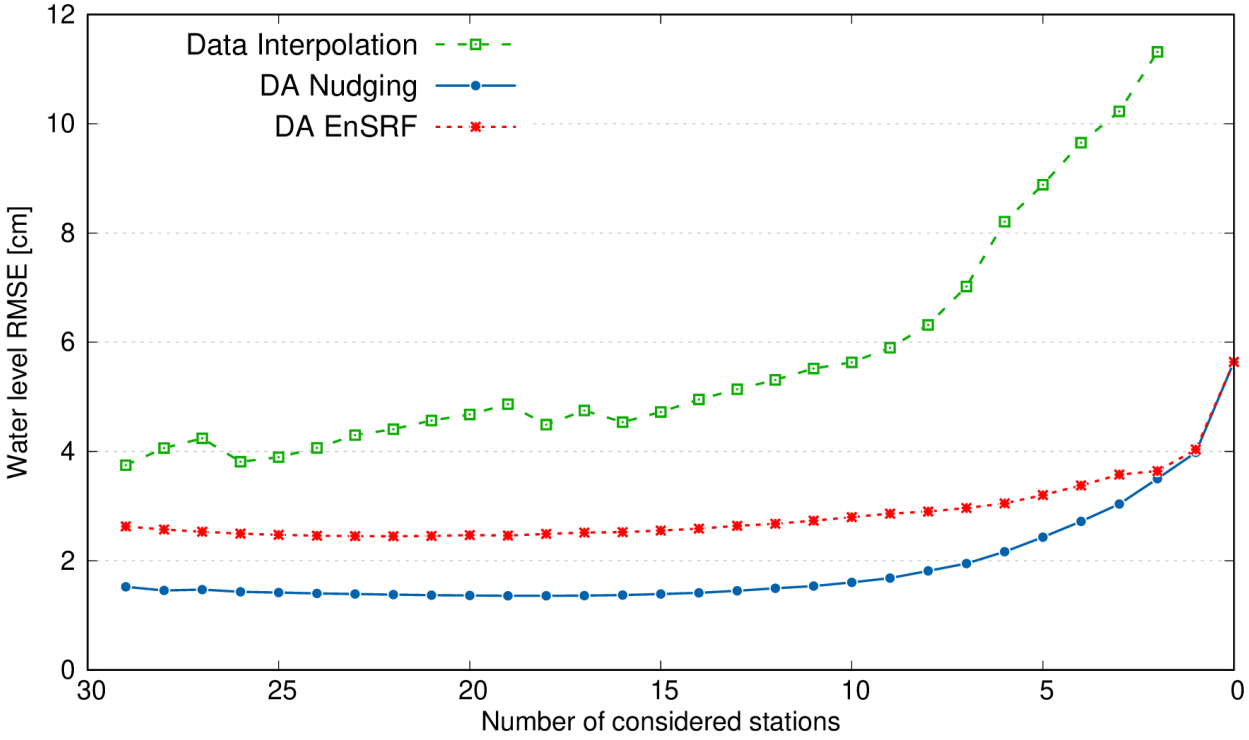

Figure 6. Root mean square error of the water levels as a function of the number of tide gauge stations interpolated or assimilated. The RMSE value with zero considered stations for DA is also indicating the error of the base simulation when no DA methods are applied.

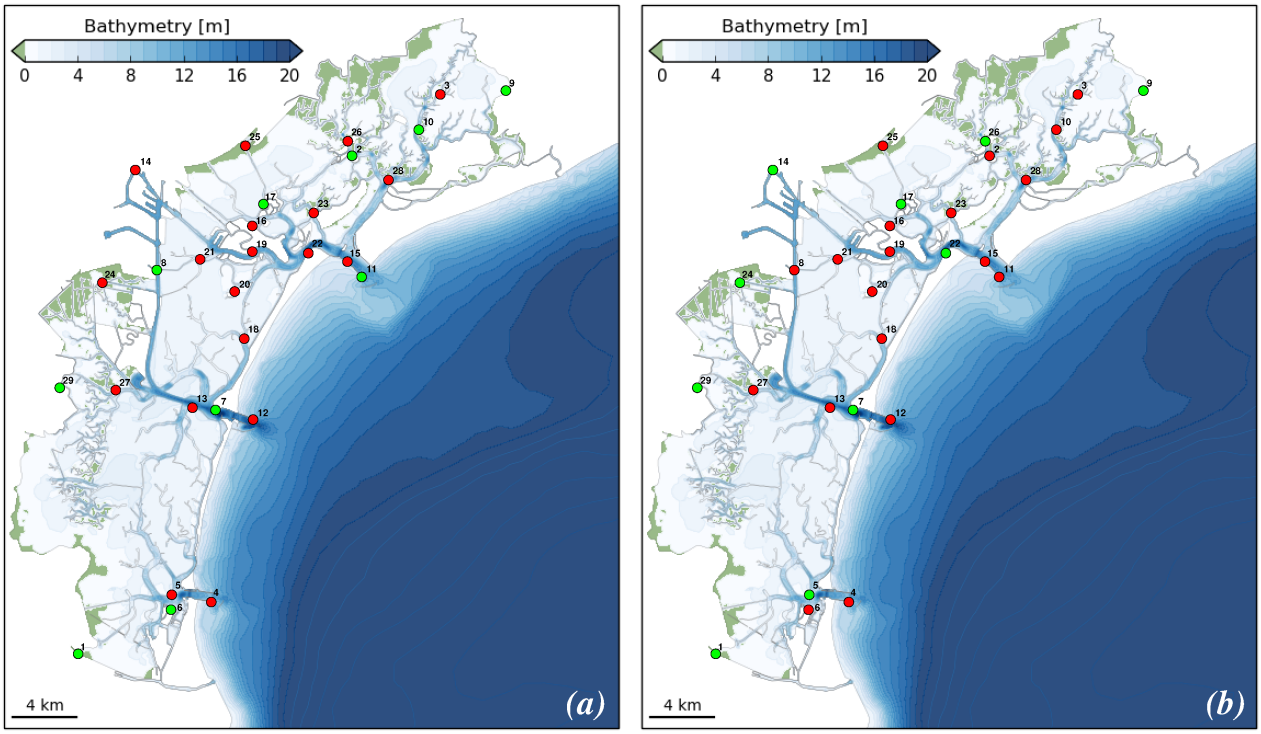

Figure 7. The optimal distribution of 10 tide gauge stations (marked with green dots) according to DA-Nudging (a) and DA-EnSRF (b).

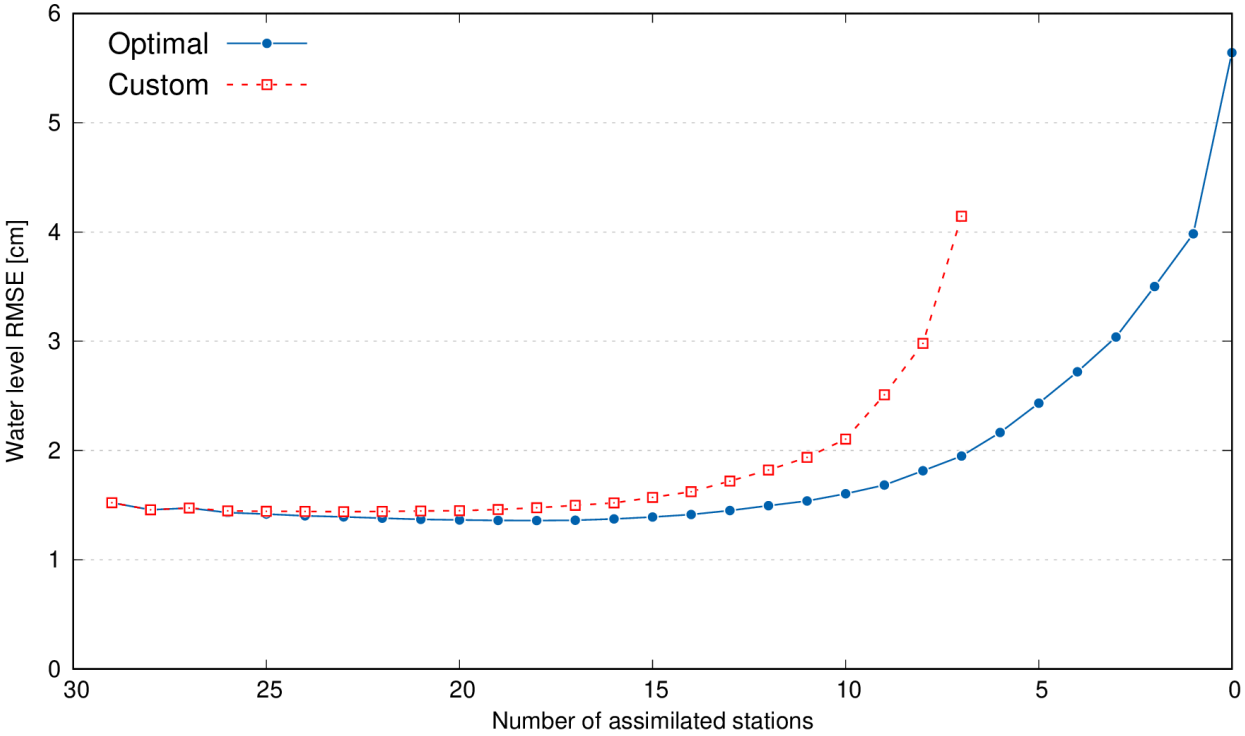

Figure 8. Same as Fig. 6, but for the optimal and the custom network experiment using Nudging data assimilation.

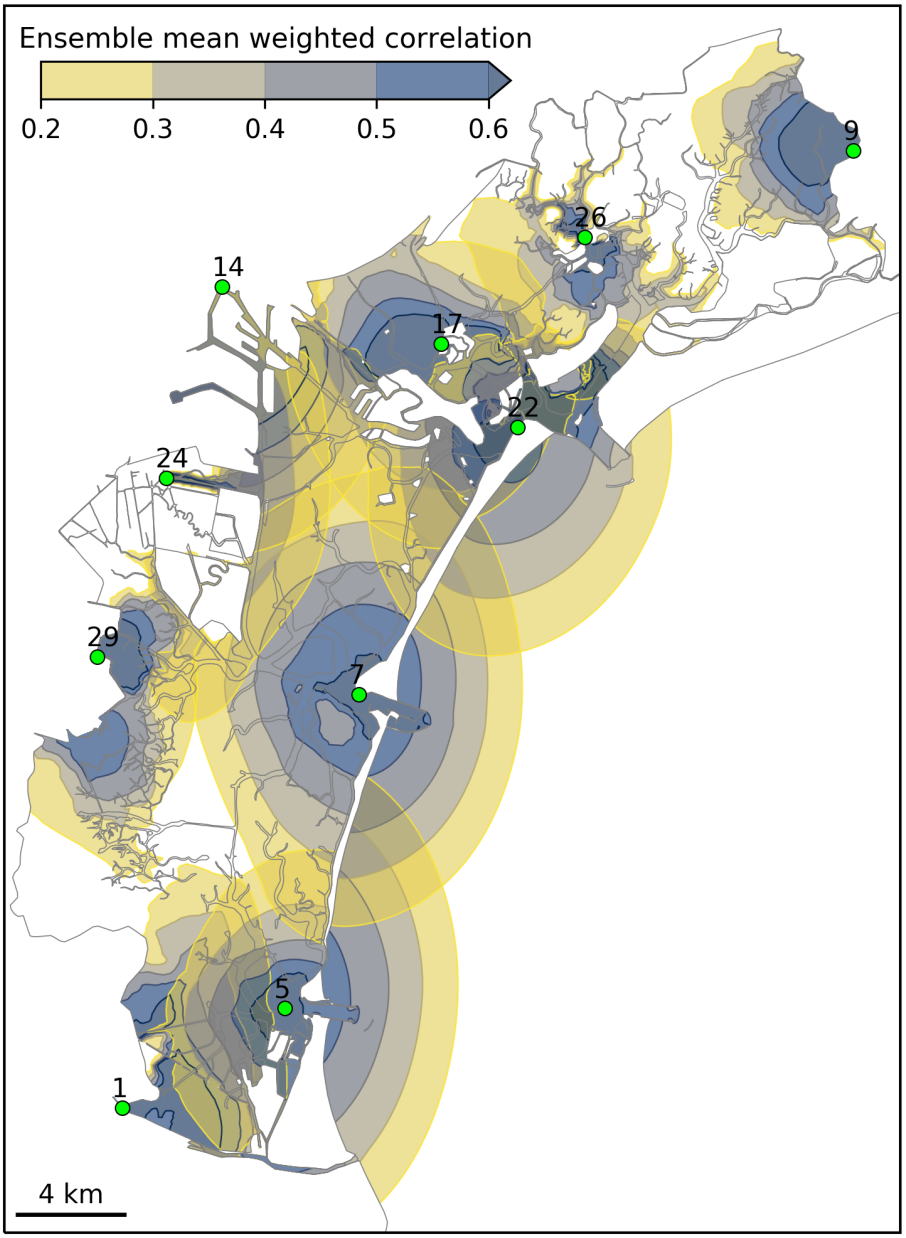

Figure 9. Ensemble weighted correlation (averaged over the simulation period) of the 10 monitoring stations selected using the EnSRF method.

Table 1. Root mean square errors (in cm) of DI and DA considering all other stations except the one for which the index is computed. The RMSEs of the control simulation are also reported.

Station ID	Data interp.	Control sim.	DA-Nudging	DA-EnSRF
1	4.5	8.3	5.8	7.5
2	0.8	4.1	0.9	2.4
3	3.0	5.2	2.8	3.4
4	6.5	4.1	2.5	3.5
5	1.4	4.1	1.8	2.1
6	1.5	4.2	1.4	2.3
7	3.1	5.2	1.9	3.1
8	2.4	5.3	1.4	2.3
9	7.3	9.5	3.3	9.4
10	2.5	4.0	0.9	2.4
11	3.8	4.3	2.2	3.0
12	8.5	4.5	2.5	~~4.7~~ 3.7
13	3.6	4.6	0.9	2.3
14	6.0	6.6	3.9	2.5
15	2.8	3.7	2.7	2.6
16	1.9	4.1	1.0	1.8
17	2.6	4.5	1.4	2.1
18	7.3	4.3	1.1	2.3
19	2.6	3.6	0.9	1.5
20	4.2	3.8	0.7	1.6
21	2.5	4.1	0.8	1.4
22	4.3	3.7	0.9	2.0
23	2.3	3.4	2.5	1.9
24	5.2	28.3	4.4	7.0
25	8.2	4.9	2.1	2.8
26	1.1	4.2	1.3	2.5
27	6.4	6.6	1.9	4.0
28	2.8	5.2	1.7	3.6
29	3.5	8.5	5.4	6.2
MEAN	3.9	5.8	2.1	3.2

Table 2. Statistical analysis of simulated current velocity at the Lido inlet. Results are given as RMSE (root mean square error, cm s^{-1}), BIAS (difference between the mean of simulation results and observations, cm s^{-1}) and R (correlation coefficient between model results and observations).

Simulation	RMSE	BIAS	R
Control	14.1	0.6	0.84
DA-Nudging	15.7	5.3	0.83
DA-EnSRF	13.6	0.4	0.85