ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Which data assimilation method to use and when: unlocking the potential of observations in shoreline modelling

M Alvarez-Cuesta[*] , A Toimil and I J Losada

IHCantabria—Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Isabel Torres 15, 39011 Santander, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: alvcuestam@unican.es

Abstract

Shoreline predictions are essential for coastal management. In this era of increasing amounts of data from different sources, it is imperative to use observations to ensure the reliability of shoreline forecasts. Data assimilation has emerged as a powerful tool to bridge the gap between episodic and imprecise spatiotemporal observations and the incomplete mathematical equations describing the physics of coastal dynamics. This research seeks to maximize this potential by assessing the effectiveness of different data assimilation algorithms considering different observational data characteristics and initial system knowledge to guide shoreline models towards delivering results as close as possible to the real world. Two statistical algorithms (stochastic ensemble and extended Kalman filters) and one variational algorithm (4D-Var) are incorporated into an equilibrium cross-shore model and a one-line longshore model. A twin experimental procedure is conducted to determine the observation requirements for these assimilation algorithms in terms of accuracy, length of the data collection campaign and sampling frequency. Similarly, the initial system knowledge needed and the ability of the assimilation methods to track the system nonstationarity are evaluated under synthetic scenarios. The results indicate that with noisy observations, the Kalman filter variants outperform 4D-Var. However, 4D-Var is less restrictive in terms of initial system knowledge and tracks nonstationary parametrizations more accurately for cross-shore processes. The findings are demonstrated at two real beaches governed by different processes with different data sources used for calibration. In this contribution, the coastal processes assimilated thus far in shoreline modelling are extended, the 4D-Var algorithm is applied for the first time in the field of shoreline modelling, and guidelines on which assimilation method can be most beneficial in terms of the available observational data and system knowledge are provided.

1. Introduction

Sandy shorelines are among the most dynamic coastal environments on Earth, cover one-third of ice-free coasts, host key socioeconomic activities and provide valuable flood protection for many coastal cities worldwide (Toimil *et al* 2023). Predicting how they react to natural and anthropogenic drivers is becoming one of the most important challenges in coastal science (Toimil *et al* 2020a, Splinter and Coco 2021). Physics-based modelling is the most viable alternative for efficiently generating various related predictions, from short-term cross-shore sediment transport (Miller and Dean 2004, Yates *et al* 2009, Splinter *et al* 2014) to long-term shoreline changes driven by longshore transport (Pelnard-Considère 1956) and water level changes (Bruun 1962, McCarroll *et al* 2021).

For shoreline models to provide reasonable predictions, their free parameters need to be tuned via calibration (Splinter *et al* 2013, Montaño *et al* 2020), which presents several challenges. One challenge is the computational cost associated with traditional calibration techniques, including Monte

Carlo variants such as general likelihood uncertainty estimation (Simmons *et al* 2017) or minimization of a cost function using gradient-free algorithms (Hewageegana and Canestrelli 2021).

Another challenge arises from the fact the validation of physics-based shoreline models is limited to a few monitored coastal sites (Turner *et al* 2016, Ludka *et al* 2019, Castelle *et al* 2020). In this sense, the release of satellite imagery and its use in coastal science (Luijendijk *et al* 2018, Vos *et al* 2019, Sabour *et al* 2020, Sánchez-García *et al* 2020), as well as the emergence of new forms of monitoring such as citizen initiatives (Harley and Kinsela 2022) and the application of the Internet of Things (IoT) technology (Tien *et al* 2023), have enabled important steps forwards. However, guidelines for determining the quantity (the length of the observation campaign and the sampling frequency) and quality (the accuracy of the measurements) of the data required for calibration to provide reliable estimates are still lacking.

Increased data availability has boosted the use of artificial intelligence (AI) and data assimilation (DA) in geophysical modelling in recent years (Kim *et al* 2021, Sonnewald *et al* 2021, Kelp *et al* 2023). While AI can be used to infer system dynamics directly from observational data (Kuglitsch *et al* 2023), DA blends a dynamic model with observations to better estimate the true state of a system and eventually tune its unknown parameters (Asch *et al* 2016). Unlike traditional calibration methods (Simmons *et al* 2017, Hewageegana and Castrelli 2021), DA allows for the joint tuning of parameters and predictions so that the true (unknown) state of the system can be efficiently reproduced considering the uncertainty of the model and the observations, through just one or a reduced number of model runs. There are two main families of DA methods: statistical and variational methods. The former requires the minimization of the uncertainty of the model error (Kalman 1960), while the latter seeks to minimize a cost function that represents the mismatch between the model and the observations (LeDimet and Talagrand 1986).

Variational methods, such as 4D-Var (the 4-dimensional variational method), and statistical methods, such as Kalman filter-based algorithms, have both been widely used for data assimilation in meteorology and oceanography (e.g. Houtekamer *et al* 2005, Neveu *et al* 2016, Stopa 2018). The choice between them boils down to a trade-off between accuracy, computational complexity, the characteristics of the observations and the characteristics of the type of analysis to be conducted (Kalnay *et al* 2007). Kalman filters are more flexible and easier to implement than 4D-Var in systems with real-time data updates (Lorenc 2003). Nevertheless, despite its complexity and the need for an adjoint model, 4D-Var also has advantages. It considers the entire time interval of interest as a whole to optimize the estimation of the initial state. This can be very useful in systems where the observations are irregularly distributed over time. Additionally, 4D-Var can be more accurate for nonlinear and dynamic systems (Caya *et al* 2005). In the field of coastal impacts, the latter advantage can translate into improvements in addressing changing trends and nonstationarities such as those arising from climate change.

To date, there have been few DA applications in shoreline modelling, and these applications have been limited to statistical methods (Kalman filters) combined with observations, whereas variational algorithms have remained unexplored. Shoreline prediction with DA was first introduced by Long and Plant (2012) by combining the extended Kalman filter (eKf) with a cross-shore model plus a linear term representing longshore processes using synthetic observations. Vitousek *et al* (2017) developed an integrated longshore and cross-shore model in combination with an eKf fed with LiDAR observations, while Alvarez-Cuesta *et al* (2021) used satellite data together with an eKf and a multiprocess shoreline model. Whereas the LiDAR observations were accurate (with a root mean square error, RMSE, of <1 m) and recorded once a year, the satellite observations were noisy (RMSE ~ 10 m) but sampled more frequently (biweekly). Ibaceta *et al* (2020) employed a cross-shore model and an Ensemble Kalman filter (EnKf) to track nonstationary parameters using observations from video cameras (RMSE ~ 0.5 m, sampled on a daily to weekly basis). Vitousek *et al* (2021) used an EnKf to assess the uncertainty of shoreline predictions due to wave forcing while additionally using camera-derived observations. Recently, Vitousek *et al* (2023) introduced a localized EnKf together with a previously developed multiprocess shoreline model that accounted for spatial dependencies in the littoral cells of California's coasts. They concluded that the use of satellite-derived shorelines yielded results similar to those obtained using monthly in situ GNSS data (RMSE ~ 1 m) from Ocean Beach.

The free model parameters to be tuned through DA should not change with variations in climate forcing (waves and water levels) because, by definition, they are independent of any other model input. In reality, however, this expectation may not hold (Ibaceta *et al* 2020, 2022), and climate variability may consequently compromise the quality of the calibration process, due to hidden dependencies between the wave climate and the free model parameters that should be disentangled by the assimilation algorithm. Thus, shoreline forecasts may benefit from the incorporation of climate-dependent parametrizations, as the world's coasts are expected to be exposed to unprecedented drivers due to climate change (Toimil *et al* 2020b, Lobeto *et al* 2021).

The main objective of this paper is to provide general guidelines for exploiting current shoreline observations to guide physics-based models of shoreline

evolution at different scales using the most appropriate assimilation algorithm. To achieve this goal, the coastal processes considered in data-assimilated shoreline modelling are extended; the 4D-Var assimilation algorithm is applied, for the first time to the authors' knowledge, to shoreline modelling; and the data and system knowledge requirements for the success of statistical and variational algorithms are analysed. Conclusions are drawn on the basis of synthetic exploratory cases and are subsequently verified at two real beaches subject to different coastal processes and with different data used for calibration.

2. Methods

2.1. Numerical models
In this work, shoreline evolution is regarded as a linear sum of cross-shore and longshore processes. The short-term coastal response to changes in wave action and water levels results in cross-shore sediment transport that moves sediment from the beach face to the offshore bar (erosion) and vice versa (accretion). Conversely, the obliqueness of the wave angle with respect to the coastline drives longshore sediment transport, whose gradients generate shoreline changes on a yearly time scale.

Using a physics-based modelling approach, cross-shore processes are modelled following Miller and Dean (2004), as shown in equation (1), but in addition, the astronomical tide is included, as in Toimil *et al* (2017). The model translates changes in waves and water levels to shoreline changes driven by the imbalance between the equilibrium shoreline Y_c^{eq} and the instantaneous shoreline Y_c

$$\frac{\partial Y_c}{\partial t} = K_c^{+/-} \left[Y_c^{\text{eq}} - Y_c \right] \tag{1}$$

where $K_c^{+/-}$ is the accretion/erosion cross-shore rate and Y_c^{eq} may be expressed as:

$$Y_c^{\text{eq}} = \Delta y_0 - W_{\text{b}}^* \left(\frac{0.106 H_{\text{sb}} + SS + AT}{B + h_{\text{b}}} \right). \tag{2}$$

Δy_0 is an empirical parameter that ensures that short-term fluctuations oscillate around a baseline position, W_{b}^* is the breaking line width, H_{sb} is the significant breaking wave height, SS is the storm surge, AT is the astronomical tide amplitude, B is the beach profile berm height and h_{b} is the breaking depth.

Longshore changes are modelled at fixed shore-normal transects, which define the cell boundaries where the longshore transport gradients are evaluated. Gradients in longshore sediment transport generate landward or seaward displacements of the active part of the beach profile:

$$\frac{\partial Y_l}{\partial t} = -\frac{1}{B + d_c} \frac{\partial Q}{\partial x}. \tag{3}$$

Y_l is the distance from the onshore point of the transect to the shoreline, d_c is the closure depth, ∂x is the transect spacing and Q is the longshore sediment transport calculated according to the Coastal Engineering Research Center (CERC) expression:

$$Q = K_l H_{\text{sb}}^{2.5} \sin(2\alpha) \tag{4}$$

where K_l is an empirical constant and α is the angle between the wave crests (θ_{b}) and the shoreline (θ_s).

2.2. Data assimilation algorithms
Given a set of observations $\mathbf{y}$ that are scattered in time and space, DA attempts to find the optimal state $\mathbf{x}$ defined by the variables evolved by the dynamic model. Shoreline observations $\mathbf{y}$ can be acquired through various means, such as video imagery, satellite imagery or in situ surveys conducted at sampling intervals spanning from hours to several months. These digitalized shorelines, whose spatial resolution depends on the acquisition method, are then interpolated onto a set of model transects. The state $\mathbf{x}$ is formed by the model predictions (Y_c and Y_l for cross-shore and longshore processes, respectively) and augmented by the free model parameters ($K^{+/-}$ and K_l for cross-shore and longshore processes, respectively). For cross-shore processes, a single transect is considered in the state vector. In contrast, the state vector for longshore transport is formed by Y_l and K_l at every transect used to discretize the shoreline.

Considering the state and the observations as random variables, the DA problem consists of solving the Bayes equation that gives the posterior distribution of the state ($\mathbf{x}$) given the observations based on the prior state distribution and the likelihood of the observations:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}) p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \tag{5}$$

where $p(\mathbf{y})$ is the marginal probability density function of the observations. It acts as a normalizing constant that ensures that $p(\mathbf{x}|\mathbf{y})$ is a valid probability distribution (integrates to one).

A common practise in statistical and variational methods is to assume that the distributions are normal so that they can be modelled in terms of the mean and the covariance matrix (**B**): $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}^b, \mathbf{B}^b)$, $p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(0, \mathbf{R})$, and $p(\mathbf{x}|\mathbf{y}) \sim \mathcal{N}(\mathbf{x}^a, \mathbf{B}^a)$, where the superscripts b and a stand for background and analysis, respectively. The initialization of $\mathbf{B}^b$ is key for the success of DA algorithms (Bannister 2008a); interested readers can refer to the supplementary material for further details. **R** is the observation error covariance matrix. It is common to assume unbiased and uncorrelated errors (i.e. a diagonal **R** matrix), where each term is the variance of the observation error distribution function, which depends on the accuracy of the data source (e.g. satellite images, LiDAR, GNSS,

video camera footage), plus the variance of the interpolation errors that arise when mapping the temporally and spatially scattered observations to the modelled state (negligible in shoreline modelling).

The covariance matrix for a Kalman filter is denoted by $\mathbf{P}$, and its initial value coincides with the initial background covariance matrix, $\mathbf{P}_0 = \mathbf{B}_0^b$. The initial background covariance matrix is constructed by assuming no correlations between the model state and the parameters (Evensen *et al* 2022). The background covariance matrix is formed from the errors of the state with respect to the background. The error in the initial shoreline position is assumed to be similar to the error in the observations, and the errors on the parameters are of the same order of magnitude as but smaller than the values of the initial constants. The interested reader is referred to the supplementary material for further details on the initialization of $\mathbf{P}_0$.

In sequential assimilation, there are two computational steps: forecasting, which involves a free model run, and correction, which involves an information flow from the observations to the model.

During the forecasting step (superscript *f*), the analysis state $\mathbf{x}_i^a$ is propagated in time considering the model's dynamic equation:

$$\mathbf{x}_{i+1}^f = \mathcal{M}_i\left(\mathbf{x}_i^a\right). \tag{6}$$

The eKf propagates the analysis covariance matrix $\mathbf{P}_i^a$ analytically using the tangent linear model $\mathbf{M_i}$ of $\mathcal{M}$ at instant *i* considering the covariance of the model's uncertainty or process noise $\mathbf{Q}$:

$$\mathbf{P}_{i+1}^f = \mathbf{M_i}\mathbf{P_i^a}\mathbf{M_i}^T + \mathbf{Q}. \tag{7}$$

During the correction step, the Kalman equations yield the mean $\mathbf{x}_{i+1}^a$ and the covariance $\mathbf{P}_{i+1}^a$ of the posterior distribution in a sequential fashion every time a new observation $\mathbf{y}_{i+1}$ is available (figure 1(a)):

$$\mathbf{x}_{i+1}^a = \mathbf{x}_{i+1}^f + \mathbf{K}_{i+1}\left(\mathbf{y}_{i+1} - \mathcal{H}\left(\mathbf{x}_{i+1}^f\right)\right)$$
$$\mathbf{K}_{i+1} = \mathbf{P}_{i+1}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_{i+1}^f\mathbf{H}^T + \mathbf{R}_{i+1}\right)^{-1}. \tag{8}$$
$$\mathbf{P}_{i+1}^a = \left(\mathbf{I} - \mathbf{K}_{i+1}\mathbf{H}\right)\mathbf{P}_{i+1}^f,$$

$\mathcal{H}$ is a nonlinear observation operator that maps the model state to the observation space, and $\mathbf{H}$ is its linearization. It is generally a temporal and spatial interpolation that enables the comparison of the model state (i.e. the shoreline) to the observations. $\mathbf{I}$ is the identity matrix, and $\mathbf{K}$ is the Kalman gain, which weights the model state with respect to the observations.

Due to the size of $\mathbf{P}_{i+1}^f$, which scales with the size of the augmented state, this matrix is unfeasible to calculate in many environmental applications. Thus, it

can be approximated using a sample of evolved states, leading to the EnKf (Evensen 1994).

In the SEnKf, each ensemble member is analytically evolved following equation (6), and the covariance matrix $\mathbf{P}_{i+1}^f$ is directly obtained from the ensemble without analytical derivations (figure 1(b)):

$$\mathbf{P}_{i+1}^f = \frac{1}{N-1}\mathbf{X}_{i+1}^f{\mathbf{X}_{i+1}^f}^T \tag{9}$$

where $\mathbf{X}_{i+1}^f$ is the perturbation matrix defined by subtracting the ensemble mean from each ensemble member's state and *N* is the number of ensemble members.

In the SEnKf, the observations need to be perturbed for every ensemble member and updated so that the variance according to the SEnKf analysis correctly represents the variance according to the eKf analysis (i.e. $\left(\mathbf{I} - \mathbf{K}_{i+1}\mathbf{H}\right)\mathbf{P}_{i+1}^f$). Thus, equation (8) becomes:

$$\mathbf{x}_{i+1}^a = \mathbf{x}_{i+1}^f + \mathbf{K}_{i+1}\left(\mathbf{y}_{i+1} + \epsilon_y - \mathcal{H}\left(\mathbf{x}_{i+1}^f\right)\right),$$
$$\epsilon_{\mathbf{y}} = \mathcal{N}\left(0, \mathbf{R}_{i+1}\right). \tag{10}$$

Adding noise to the observations via random sampling from $\mathcal{N}\left(0, \mathbf{R}_{i+1}\right)$ avoids the divergence of the SEnKf by adjusting the analysis variance. As the magnitude of the sampling noise depends on the error of the observations via the covariance matrix $\mathbf{R}_{i+1}$, the larger the data source error is, the larger the sampling error introduced in the SEnKf. This fact limits the application of the SEnKf when the variations in the time of the state are relatively small compared to the data source error.

The magnitude of the sampling noise depends on the error of the observations via the covariance matrix $\mathbf{R}_{i+1}$.

In variational methods, the sought solution is the one that maximizes the posterior probability of $p\left(\mathbf{x}|\mathbf{y}\right)$ (i.e. the mode), and no uncertainty estimate of the analysis state is obtained. In 4D-Var, DA is performed in a time window that enables future observations to be considered in prediction at a previous time step (figure 1(c)).

Considering that inside the time window, the state trajectory is governed solely by the model dynamics (strong-constraint 4D-Var), the minimization of the cost function (equation (11)) yields the most probable state

$$J(\mathbf{x}_0) = J_b + \sum J_{obs} = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$$
$$+ \frac{1}{2}\sum_{i=0}^{K}\left(\mathcal{H}\left(\mathbf{x}_i\right) - \mathbf{y}_i\right)^T\mathbf{R}_i^{-1}\left(\mathcal{H}\left(\mathbf{x}_i\right) - \mathbf{y}_i\right) \tag{11}$$

where $\mathbf{x}_0 b$ is the a priori background state, $\mathbf{B}_0$ is the initial background error covariance matrix, $\mathbf{x}_0$ is the
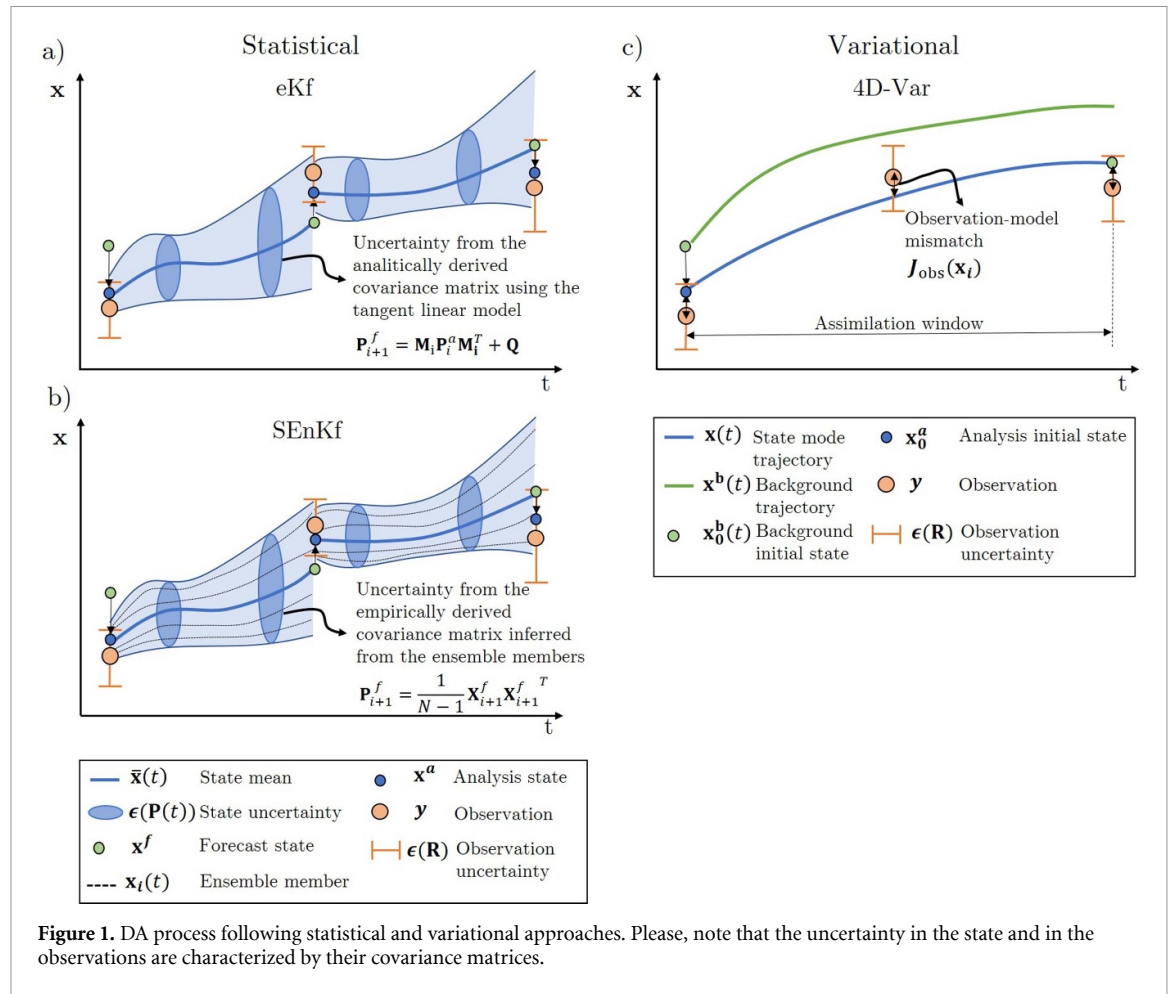
Figure 1. DA process following statistical and variational approaches. Please, note that the uncertainty in the state and in the observations are characterized by their covariance matrices.

initial state vector, $\mathbf{R}_i$ is the observation error covariance matrix, K is the number of steps in the time window and $\mathbf{x}_i$ is the state at time t_i of observation $\mathbf{y}_i$. Note that ultimately, J is a function of $\mathbf{x}_0$, as $\mathbf{x}_i = \mathcal{M}_{i-1}(\mathcal{M}_{i-2}\ldots\mathcal{M}_0(\mathbf{x}_0))$. Additionally, if there are no observations at a given step i, $\mathbf{y}_i = \mathcal{H}(\mathbf{x}_i) = \mathbf{0}$.

Minimizing J requires the term-by-term calculation of the gradient of the cost function to guide a descent algorithm:

$$\nabla J(\mathbf{x}_0) = \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$$
$$+ \sum_{i=0}^{K} M_0^T \ldots M_{i-1}^T \mathrm{H}_i^T \mathbf{R}_i^{-1}\left(\mathcal{H}(\mathbf{x_i}) - \mathbf{y_i}\right). \tag{12}$$

Solving for $\nabla J(\mathbf{x}_0)$ implies the backpropagation of the innovation vector $\mathbf{y_i} - \mathcal{H}(\mathbf{x_i})$ using the transpose (M^T) or adjoint of the model operator dynamics. Deriving the adjoint of the forward model enables the computation of the gradient of the cost function, as shown in the supplementary material.

The 4D-Var algorithm consists of iteratively running the forward model (equation (6)), calculating the cost function (equation (11)) and its gradient (equation (12)), and applying a descent algorithm to

find the optimal initial state vector $\mathbf{x}_0$. $J(\mathbf{x}_0)$ is usually nonquadratic, preventing the descent algorithm from reaching the global minimum, which corresponds to the usually unknown ground-truth (reference) solution. Hence, incremental 4D-Var and control variable transformations are usually required in practical applications. These techniques are comprehensively described in the supplementary material.

2.3. Description of the test cases

The parameter estimation skill of the different DA algorithms is tested considering the observation data requirements in terms of accuracy, length and sampling frequency and also the required system knowledge in terms of initial conditions and parameter nonstationarity. Cross-shore and longshore processes are studied in isolation following the twin experimental procedure. This method consists of generating a ground-truth simulation (reference) using a set of parameters by freely running the dynamical model. This simulation is then used to generate the observations for other simulations, in which the initial parameter set is altered but the dynamical model is enhanced by the DA algorithms.

Cross-shore processes are modelled at a single transect. The reference cross-shore shoreline position is $Y_{st,0} = 0$, and the accretion and erosion constants are $K^+ = 0.002$ and $K^+ = 0.04$, respectively. Waves and water levels correspond to real time series obtained from Alvarez-Cuesta *et al* (2023). The equilibrium profile for the cross-shore model is characterized by Dean's parameter (Dean 1991) with a value of $A = 0.25$. Longshore processes are modelled by considering a 3 km embayment discretized with 15 shore-normal transects. The reference initial shoreline is a parabola calculated such that the cross-shore length of the embayment is 100 m. Null flow conditions are imposed at the boundaries to simulate a closed system, and the wave dynamics are constant, with $H_{sb} = 2$ m and $\theta_b = -30°$. A uniform longshore constant $K_l = 50\,\mathrm{m}^{1/2}/\mathrm{day}$ is considered at every transect. Further details about the reference state of the system and the initialization of the perturbed simulations for the twin experiments are provided in the supplementary material.

The length of the simulations is 30 years, divided into a calibration phase (DA-guided run) and a validation phase (free run without DA). When analysing the observations and the initial system knowledge requirements, the validation phase covers the last 10 years. When analysing the nonstationary evolution of the system, because the ability of the DA algorithm to track changes in the model parameters from the observations is measured, DA is never switched off, and the calibration period covers the whole time span. The main characteristics of the terms influencing the physics-based process equations (equations (2)–(4)) in the test cases and the specific characteristics of the different DA methods (i.e. the number of ensemble members for the SEnKf and its initialization, the temporal window for 4D-Var, and the definitions of the state covariance matrix **B** and the process noise matrix **Q**) are described in the supplementary material.

Different error metrics are used to test the capabilities of the DA algorithms throughout the analysis during the validation phase:

- DA benefit (DAB). This metric measures the improvement of a DA algorithm with respect to the background simulation (without DA) in terms of the root mean squared error (RMSE) between the altered simulation and the reference simulation during the validation period:

$$\mathrm{DAB}\,[\%] = \left(1 - \frac{\mathrm{RMSE}\,(Y_{\mathrm{alt,\ DA}})}{RMSE\,(Y_{\mathrm{alt,}\ b})}\right) * 100 \quad (13)$$

where the subindex *alt* stands for 'altered' and *b* for 'background'.

- Parameter error (eK). This metric measures the difference between the DA value at the end of the calibration period and the reference value

$$eK\,[\%] = \left(\frac{K_{\mathrm{alt,\ DA}} - K_r}{K_r}\right) * 100. \quad (14)$$

The subindex *r* refers to the reference simulation. For longshore processes, *eK* is calculated based on the mean DA parameter across all transects.

- Computational burden (CPU). This quantity is measured in seconds.
- Percentage of time within acceptable limits (PTWL). This metric is used to evaluate whether a DA algorithm is capable of tracking nonstationary variations in the model parameters. The acceptable limits are defined as a 20% range of a parameter around its reference value.

3. Results

3.1. The role of the observations: quantity and quality

The performance of the DA algorithms is tested by using observations with different levels of added noise ($\mathcal{N}\,(0, \epsilon^2)$), where ϵ includes the error on the observations, as well as variations due to different observation lengths and different sampling frequencies. In figure 2(a), the performance metrics are analysed for different observation lengths and added noise levels considering a fixed sampling frequency of 15 days. In figure 2(b), the sampling frequency is analysed together with the length of the observations for an observation error of 1 m.

This analysis shows that the Kalman filters perform better than 4D-Var for both cross-shore and longshore processes when the reference parameters do not vary in time. For cross-shore processes, the eKf slightly outperforms the SEnKf, as it achieves a DAB larger than 80% in almost every scenario at a smaller computational expense. In the case of longshore processes, the SEnKf outperforms the other algorithms in terms of DAB and eK when the error on the observations is greater than 7 m and when the length of the observations is shorter than 3 years. Conversely, 4D-Var performs well when the error on the observations is less than 4 m for both cross-shore and longshore processes. For greater observation errors, increasing the length of the observations is not clearly correlated with an improvement in DAB. This fact is explained by the random nature of the observations together with the lack of flow dependency of the state covariance matrix between assimilation windows (interested readers are referred to the supplementary material for a detailed explanation). For longshore processes, the length of the observations should exceed 5 years to yield a relative error on the model constant of less than 20% of the reference value
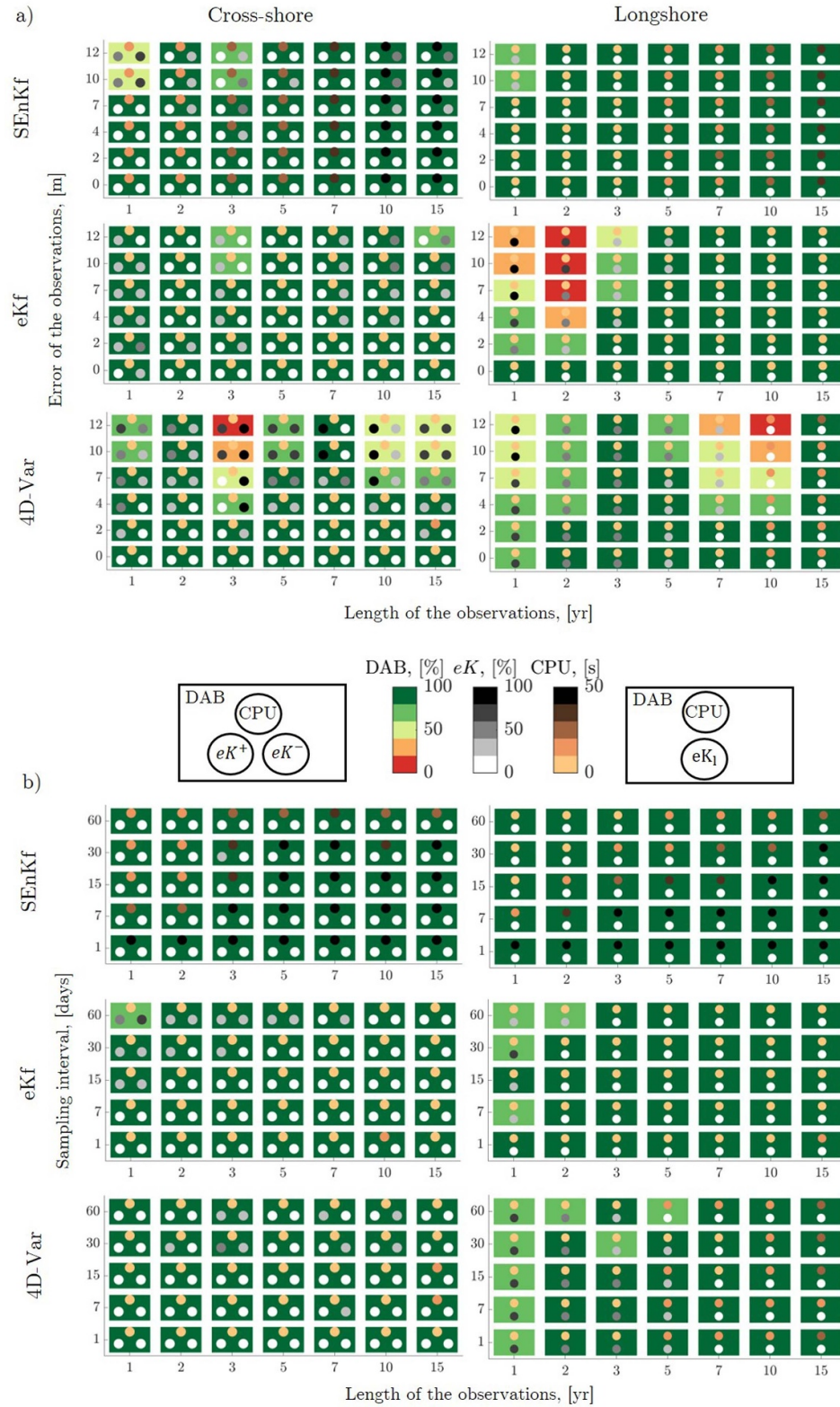
Figure 2. Analysis of the role of observations on the performance of the DA algorithms for cross-shore and longshore processes. Panel (a): influence of the length of the observations campaign and the error of the observations. Panel (b): influence of the length of the observations campaign and the sampling interval. The DAB is shown in coloured squares and the inner dots represent the eK and CPU cost.

of the parameter. From a computational perspective, the SEnKf is the most demanding algorithm because of the use of an ensemble, and its cost increases with the number of observations for simulations exceeding

50 s. The most efficient algorithm in the case of cross-shore processes is the eKf, with an average simulation cost of less than 10 s, while 4D-Var is also computationally efficient but becomes more expensive as the

length of the observations increases, as the minimization algorithm needs to be executed more times.

Regarding the sampling interval of the observations, the DA algorithms do not show great differences in the 1–60 days range. The SEnKf yields a DAB greater than 80% for all sampling intervals from 1 to 60 days, and the relative parameter error is always smaller than 20% for both longshore and cross-shore processes. The eKf yields a DAB of more than 80% in every case, except when the sampling interval is 60 days and the length of the observations is one year for cross-shore processes or two years for longshore processes. However, the longer the sampling interval is, the more observations are required to reduce the relative error of the parameters from the 20%–40% range to the 0%–20% range. Similarly, 4D-Var yields a DAB of more than 80% in every case for cross-shore processes, but its parameter convergence is compromised when the sampling interval of the observations is greater than 30 days for campaign durations of less than four years. In the case of longshore processes, to guarantee good convergence of the algorithm, observations over more than 3 years are required if the sampling interval is greater than one month.

3.2. The role of system knowledge: prior knowledge and temporal evolution (parameter nonstationarity)

The initial state of the perturbed simulations is modified by taking different quantiles from the upper and lower tails of the CDF of the reference simulation, as detailed in the supplementary material. Observations with no added noise are assimilated every 15 days.

In figure 3(a), the performance metrics are analysed for different campaign durations and initialization values. The Kalman filters require more precise initialization than does 4D-Var for cross-shore processes, which is especially noticeable for extreme quantiles (0.999). In this case, the SEnKf requires more than 10 years of observations; the eKf, 5 years; and 4D-Var, 2 years to achieve a DAB greater than 80%. For longshore processes, the Kalman filters converge to the reference value even for extreme quantiles, but 4D-Var requires more than 5 years to successfully converge in the case of the 0.999 quantile. The different time scales associated with longshore and cross-shore processes may explain the different convergence rates.

The skill of the DA algorithms in tracking the variations in time of the free model parameters (parameter nonstationarity) is measured by forcing the cross-shore and longshore numerical models using square waves. A square wave is formed by pulses such that the value of the varying parameter alternates between two discrete levels in a regular and repetitive manner. The pulse period, defined as the time it

takes for one complete cycle of the waveform to occur, is used in combination with the error of the observations to rank the different DA algorithms. The DA performance indicators are calculated for the whole simulation period, as the calibration and validation phases coincide, and biweekly observations are used for assimilation.

In figure 3(b), the performance of the DA methods is analysed for different pulse periods of the square wave and for different observational errors. In this case, 4D-Var clearly outperforms the Kalman filters for cross-shore processes. All DA algorithms perform better for longer pulse periods, but the error on the observations is found to be a critical parameter when analysing parameter nonstationary. For cross-shore processes, the SEnKf never reaches a DAB greater than 80%, even for clean observations. The eKf performs slightly better, but clean observations are required for this algorithm to show an improvement with respect to the background simulation. In comparison, 4D-Var outperforms the Kalman filters, achieving a DAB greater than 50% for observations with an error of 3 m. For longshore processes, due to its nature, the DAB is always greater than 80% because the shoreline position associated with longshore changes does not oscillate around an equilibrium position, as in the case of cross-shore processes, meaning that the background simulation completely diverges if no DA is applied. However, the PTWL is compromised in the cases of shorter pulse periods and noisy observations. In these cases, the best performing algorithm is the eKf, which achieves a PTWL greater than 80% for periods of 12 years or more with clean observations. For shorter periods or noisier observations, the parameter estimates from the eKf oscillate around the reference value but may fall outside the acceptable limits for the PTWL, as shown in figure 3(c).

3.3. Application to real cases

The different DA algorithms were tested at two coastal sites, Tairua and Castellón. Tairua is a cross-shore-dominated 1.2 km long pocket beach located on the eastern coast of the Coromandel Peninsula, New Zealand. Shoreline, wave and water parameters were obtained from Montaño *et al* (2020). Shoreline observations were obtained from video cameras on a daily basis, with an acquisition error of 0.5 m. Castellón is a longshore-dominated beach located on the Mediterranean coast of Spain. Data were obtained from Álvarez-Cuesta *et al* (2021), and shoreline observations were extracted using the CoastSat algorithm on a bimonthly basis, with an acquisition error of 10 m (Vos *et al* 2019).

In this case, the no DA (background) and DA simulations were initialized using typical values for the model parameters following USACE (1984) and
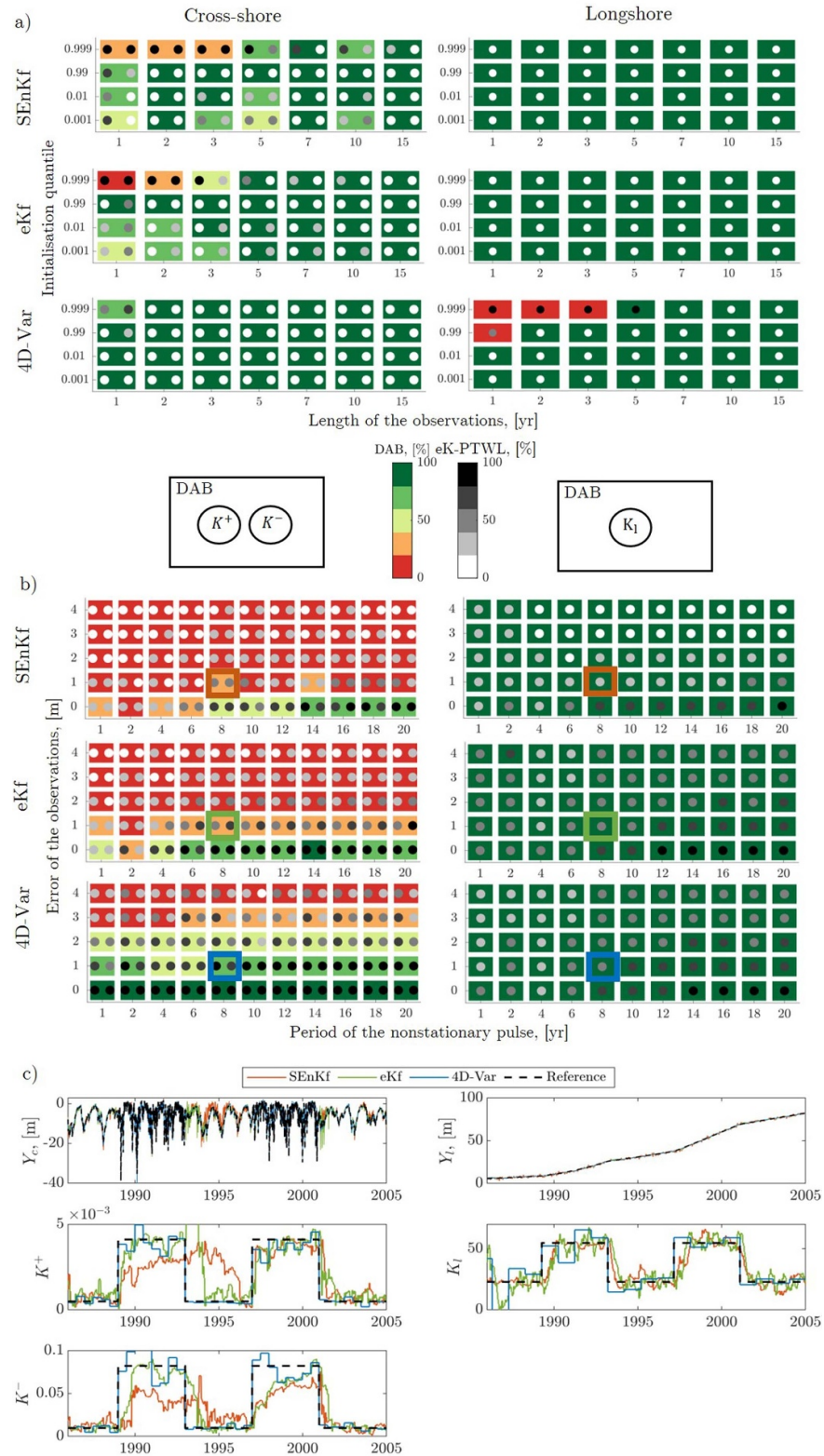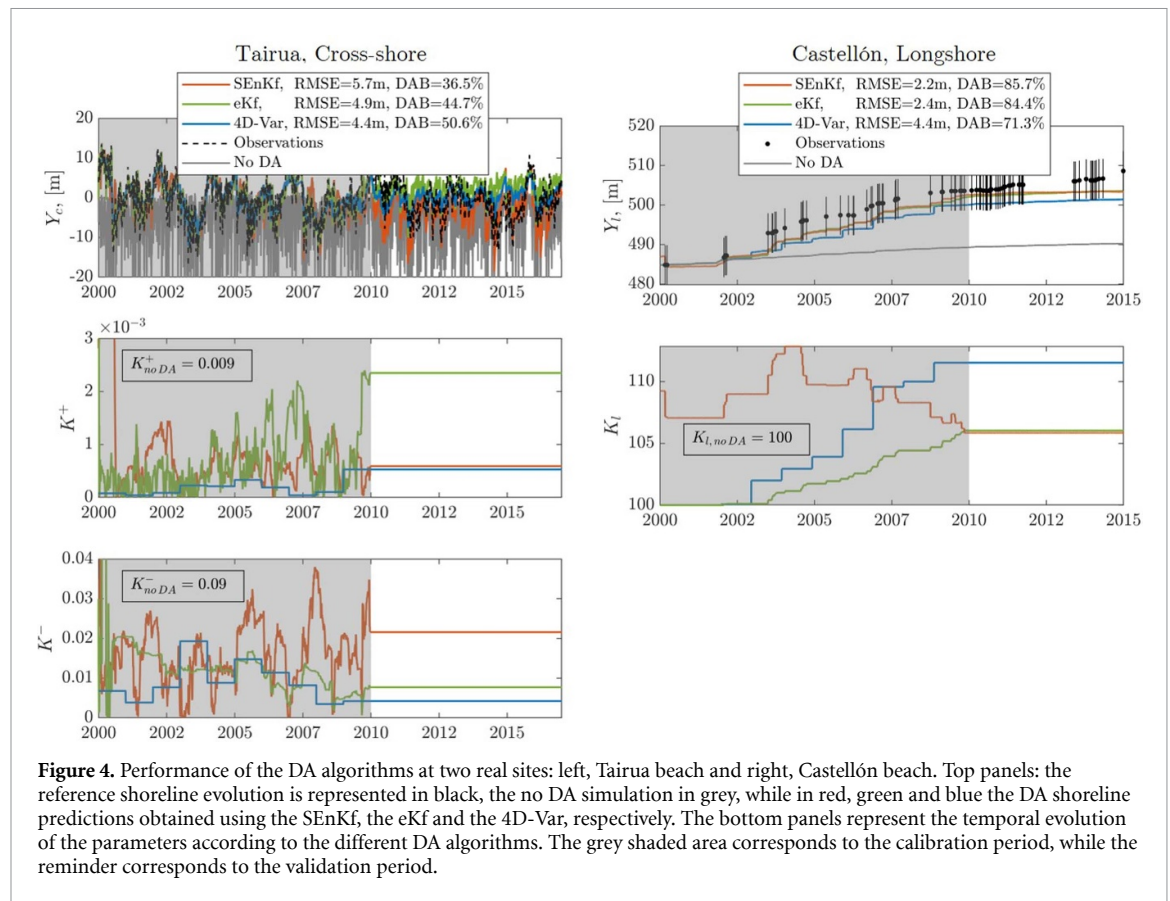
Figure 3. Analysis of the role of the system knowledge on the performance of the DA algorithms for cross-shore and longshore processes. Panel (a): influence of the prior knowledge of the system, measured as the length of the observations campaign and the initialisation quantile. Panel (b): influence of the error of the observations and the period of the non-stationary pulse of the true parameter's variation. Panel (c): time evolution of the shoreline position and the non-stationary evolution of the free parameters is displayed for cross-shore and longshore processes. The DAB is shown in coloured squares and the inner dots represent the eK in panel a and the PTWL in panel b.

Miller and Dean (2004) without assuming any prior system knowledge. The results are shown in figure 4. At Tairua, the DAB, calculated with respect to the background simulation during the validation period, exceeds 50% for 4D-Var and 44% and 36% for the eKf and SEnKf, respectively. All three algorithms

Figure 4. Performance of the DA algorithms at two real sites: left, Tairua beach and right, Castellón beach. Top panels: the reference shoreline evolution is represented in black, the no DA simulation in grey, while in red, green and blue the DA shoreline predictions obtained using the SEnKf, the eKf and the 4D-Var, respectively. The bottom panels represent the temporal evolution of the parameters according to the different DA algorithms. The grey shaded area corresponds to the calibration period, while the reminder corresponds to the validation period.

provide satisfactory results, but 4D-Var outperforms the Kalman filters, yielding an RMSE of 4.4 m. At Castellón, after calibration with noisy satellite observations for longshore processes, 4D-Var is the worst performing algorithm, achieving a DAB of 71%, whereas the SEnKf reaches a DAB of 85%. The results are consistent with the conclusions derived from the synthetic cases for cross-shore and longshore processes. For cross-shore processes and almost clean observations, all the methods yield similar results, but 4D-Var slightly outperforms the Kalman filters. When modelling longshore processes using noisy observations, the Kalman filters perform better than 4D-Var, as the latter requires more accurate observations to achieve optimal performance. However, the DAB values obtained in the real cases are smaller than those obtained for the equivalent theoretical case shown in figure 2. This fact can be attributed to cascading forcing conditions and model errors arising from our incomplete knowledge of the physical system.

4. Discussion and conclusions

The results of the different DA methods are summarized in the form of a scorecard in table 1. For low-quality (noisy) data, the Kalman filters outperform 4D-Var for both cross-shore and longshore processes.

The strong constraint assumption of 4D-Var, which imposes the requirement that the system trajectory is determined by the model equations within the assimilation window, may need to be relaxed to properly integrate inaccurate data, as in weak-constraint 4D-Var (Kalnay *et al* 2007). However, 4D-Var and the eKf both provide better results than the SEnKf when the system is initialized far from the reference state and when it evolves nonlinearly for both cross-shore and longshore processes. In these cases, both 4D-Var and the eKf explicitly incorporate the tangent linear and adjoint operators of the system dynamics, while the SEnKf derives the corresponding magnitudes by propagating an ensemble of states. In such cases, the ensemble may not properly sample the true system covariance, resulting in filter divergence even if additive covariance inflation is applied to the ensemble members by considering the process noise covariance matrix **Q**. In contrast to the eKf and 4D-Var, the SEnKf is relatively easy to implement because this DA algorithm derives the required information from the physical system by propagating an ensemble of states without the need for further analytical derivations. This fact may explain why Kalman filters have thus far been favoured over variational methods in shoreline modelling. For simulations with reduced state vectors ($\mathcal{O} \sim 10$–100 elements), the eKf is more efficient than the SEnKf because it requires only one

Table 1. Score card of the different DA algorithms for cross-shore and longshore processes. The more stars, the better the performance of the algorithm.

	DA Method	Performance with low quality data	Performance with limited observations	Performance with reduced prior system knowledge.	Performance in tracking system non-stationarity	CPU efficiency	Ease of Implementation
Cross-shore	SEnKF	***	***	**	*	*	***
	eKF	***	**	**	**	***	**
	4D-Var	*	***	***	***	**	*
Longshore	SEnKF	***	***	***	**	*	***
	eKF	**	*	***	***	**	**
	4D-Var	**	**	**	***	**	*

forward simulation. For larger states, however, the propagation of the full analytically derived covariance matrix of the eKf is prohibitive, and its representation in the ensemble space of the SEnKf becomes a more affordable solution.

The results presented in this work seek to provide general guidelines to help decide which DA method to use on the basis of the main physical processes of interest as well as the available observations and system knowledge. While we have extensively examined conventional and widely used formulations for shoreline modelling, the adoption of alternative process-based or data-guided formulations would necessitate a comparable sensitivity analysis akin to the one presented here. This step is crucial for determining the most appropriate DA method for a given modelling approach. Additionally, similar conditions and simple rules have been used here to initialize all the methods in terms of the background error covariance **B** or process noise covariance **Q**. Although substantial work has been done regarding the initialization of **B** for state estimation problems (Bannister 2008a, 2008b), its definition for parameter estimation problems remains challenging (Smith *et al* 2013), as does the initialization of an ensemble consistent with prescribed **B** and **Q** matrices. The results may be substantially improved through careful fine tuning of those magnitudes, for instance, with a case-specific derivation of the matrix **Q**, as in Ibaceta *et al* (2020). Importantly, the use of DA does not guarantee a significant improvement in the results in real cases if the background covariance and process noise are not properly set. This is because the perfect model hypothesis is not satisfied and errors from wave dynamics and model parametrizations may cascade, hampering the convergence of the DA algorithms. Thus, further research is required to define the **B** and **Q** matrices and ensemble initialization for DA problems involving combined state and parameter estimation in shoreline modelling.

As remote sensing data are revolutionizing coastal science, this study aims to optimize the way such data are employed in combination with traditional physics-based equations to produce more accurate shoreline projections. Such optimization will lead to improved hazard forecasts for risk management and climate change adaptation. DA can overcome some of the main limitations of purely data-driven methods while also complementing them. The data dependence of data-driven methods can be compensated for by the consideration of the physics of the relevant processes in DA methods, while data-driven models can efficiently process raw observational data to feed into DA algorithms. DA is also an excellent tool for data handling, as it allows new observations to be directly integrated into real-time modelling systems, ultimately contributing to the development of digital twin models (Hoffmann *et al* 2023).

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: 10.5281/zenodo.10078543.

Acknowledgments

ORCID iDs

M Alvarez-Cuesta ● https://orcid.org/0000-0002-1180-0746
A Toimil ● https://orcid.org/0000-0002-2067-872X
I J Losada ● https://orcid.org/0000-0002-9651-9709

References

Álvarez-Cuesta M, Losada I J and Toimil A 2023 A nearshore evolution model for sandy coasts: IH-LANSloc *Environ. Modelling Softw.* **169** 105827

Alvarez-Cuesta M, Toimil A and Losada I J 2021 Modelling long-term shoreline evolution in highly anthropized coastal areas. Part 1: model description and validation *Coast. Eng.* **169** 103960

Asch M, Bocquet M and Nodet M 2016 *Data Assimilation Methods, Algorithms, and Applications* (Society for Industrial and Applied Mathematics(SIAM))

Bannister R N 2008a A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances *Q. J. R. Meteorol. Soc.* **134** 1951–70

Bannister R N 2008b A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics *Q. J. R. Meteorol. Soc.* **134** 1971–96

Bischof C H, Bucker H M, Lang B, Rasch A and Vehreschild A 2002 Combining source transformation and operator overloading techniques to compute derivatives for MATLAB programs *Proc. 2nd IEEE Int. Workshop on Source Code Analysis and Manipulation* (*October*) (IEEE) pp 65–72

Bruun P 1962 Sea-level rise as a cause of shore erosion *J. Waterw. Harb. Div.* **88** 117–30

Castelle B, Bujan S, Marieu V and Ferreira S 2020 16 years of topographic surveys of rip-channelled high-energy meso-macrotidal sandy beach *Sci. Data* **7** 410

Caya A, Sun J and Snyder C 2005 A comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation *Mon. Weather Rev.* **133** 3081–94

Coastal Engineering Research Center (US) 1984 *Shore Protection Manual* vol 1 (Department of the Army, Waterways Experiment Station, Corps of Engineers, Coastal Engineering Research Center)

Dean R G 1991 Equilibrium beach profiles: characteristics and applications *J. Coast. Res.* **7** 53–84

Evensen G 1994 Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics *J. Geophys. Res. Oceans* **99** 10143–62

Evensen G, Vossepoel F C and van Leeuwen P J 2022 *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem* (Springer Nature)

Harley M D and Kinsela M A 2022 CoastSnap: a global citizen science program to monitor changing coastlines *Cont. Shelf Res.* **245** 104796

Hewageegana V H and Canestrelli A 2021 On the predictive skill of morphodynamic models for onshore sandbar migration *Earth Surf. Process. Landf.* **46** 1692–712

Hoffmann J, Bauer P, Sandu I, Wedi N, Geenen T and Thiemert D 2023 Destination Earth–a digital twin in support of climate services *Clim. Serv.* **30** 100394

Houtekamer P L, Mitchell H L, Pellerin G, Buehner M, Charron M, Spacek L and Hansen B 2005 Atmospheric data assimilation with an ensemble Kalman filter: results with real observations *Mon. Weather Rev.* **133** 604–20

Ibaceta R, Splinter K D, Harley M D and Turner I L 2020 Enhanced coastal shoreline modeling using an ensemble Kalman filter to include nonstationarity in future wave climates *Geophys. Res. Lett.* **47** e2020GL090724

Ibaceta R, Splinter K D, Harley M D and Turner I L 2022 Improving multi-decadal coastal shoreline change predictions by including model parameter non-stationarity *Front. Mar. Sci.* **9** 1012041

Kalman R E 1960 A new approach to linear filtering and prediction problems *ASME J. Basic Eng.* **82** 35–45

Kalnay E, Li H, Miyoshi T, Yang S-C and Ballabrera-Poy J 2007 4-D-Var or Ensemble Kalman filter? *Tellus* A **59** 758–73

Kelp M M, Keller C A, Wargan K, Karpowicz B M and Jacob D J 2023 Tropospheric ozone data assimilation in the NASA GEOS composition forecast modeling system (GEOS-CF v2.0) using satellite data for ozone vertical profiles (MLS), total ozone columns (OMI), and thermal infrared radiances (AIRS, IASI) *Environ. Res. Lett.* **18** 094036

Kim H, Lakshmi V, Kwon Y and Kumar S V 2021 First attempt of global-scale assimilation of subdaily scale soil moisture estimates from CYGNSS and SMAP into a land surface model *Environ. Res. Lett.* **16** 074041

Kuglitsch M M *et al* 2023 When it comes to Earth observations in AI for disaster risk reduction, is it feast or famine? A topical review *Environ. Res. Lett.* **18** 093004

LeDimet F-X and Talagrand O 1986 Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects *Tellus* A **38** 97–110

Lobeto H, Menendez M and Losada I J 2021 Future behavior of wind wave extremes due to climate change *Sci. Rep.* **11** 7869

Long J W and Plant N G 2012 Extended Kalman filter framework for forecasting shoreline evolution *Geophys. Res. Lett.* **39** L13603

Lorenc A C 2003 The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var *Q. J. R. Meteorol. Soc.* **129** 3183–203

Ludka B C *et al* 2019 Sixteen years of bathymetry and waves at San Diego beaches *Sci. Data* **6** 161

Luijendijk A, Hagenaars G, Ranasinghe R, Baart F, Donchyts G and Aarninkhof S 2018 The state of the world's beaches *Sci. Rep.* **8** 6641

McCarroll R J, Masselink G, Valiente N G, Scott T, Wiggins M, Kirby J-A and Davidson M 2021 A rules-based shoreface translation and sediment budgeting tool for estimating coastal change: ShoreTrans *Mar. Geol.* **435** 106466

Miller J K and Dean R G 2004 A simple new shoreline change model *Coast. Eng.* **51** 531–56

Montaño J *et al* 2020 Blind testing of shoreline evolution models *Sci. Rep.* **10**

Neveu E, Moore A M, Edwards C A, Fiechter J, Drake P, Crawford W J, Jacox M G and Nuss E 2016 An historical analysis of the California current circulation using ROMS 4D-Var: system configuration and diagnostics *Ocean Modelling* **99** 133–51

Pelnard-Considère R 1956 Essai de théorie de l'évolution des formes de rivage en plages de sable et de galets *4th Journées de l'Hydraulique, Les Energies de La Mer, Question III* pp 289–298

Sabour S, Brown S, Nicholls R J, Haigh I D and Luijendijk A P 2020 Multi-decadal shoreline change in coastal natural world heritage sites–a global assessment *Environ. Res. Lett.* **15** 104047

Sánchez-García E, Palomar-Vázquez J M, Pardo-Pascual J E, Almonacid-Caballer J, Cabezas-Rabadán C and Gómez-Pujol L 2020 An efficient protocol for accurate and massive shoreline definition from mid-resolution satellite imagery *Coast. Eng.* **160** 103732

Simmons J A, Harley M D, Marshall L A, Turner I L, Splinter K D and Cox R J 2017 Calibrating and assessing uncertainty in coastal numerical models *Coast. Eng.* **125** 28–41

Smith P J, Thornhill G D, Dance S L, Lawless A S, Mason D C and Nichols N K 2013 Data assimilation for state and parameter estimation: application to morphodynamic modelling *Q. J. R. Meteorol. Soc.* **139** 314–27

Sonnewald M, Lguensat R, Jones D C, Dueben P D, Brajard J and Balaji V 2021 Bridging observations, theory and numerical

simulation of the ocean using machine learning *Environ. Res. Lett.* **16** 073008

Splinter K D and Coco G 2021 Challenges and opportunities in coastal shoreline prediction *Front. Mar. Sci.* **8** 788657

Splinter K D, Turner I L and Davidson M A 2013 How much data is enough? The importance of morphological sampling interval and duration for calibration of empirical shoreline models *Coast. Eng.* **77** 14–27

Splinter K D, Turner I L, Davidson M A, Barnard P, Castelle B and Oltman-Shay J 2014 A generalized equilibrium model for predicting daily to interannual shoreline response *J. Geophys. Res.* **119** 1936–58

Stopa J E 2018 Wind forcing calibration and wave hindcast comparison using multiple reanalysis and merged satellite wind datasets *Ocean Modelling* **127** 55–69

Tien I, Lozano J-M and Chavan A 2023 Locating real-time water level sensors in coastal communities to assess flood risk by optimizing across multiple objectives *Commun. Earth Environ.* **4** 96

Toimil A, Camus P, Losada I J, Le Cozannet G, Nicholls R J, Idier D and Maspataud A 2020b Climate change-driven coastal erosion modelling in temperate sandy beaches: methods and uncertainty treatment *Earth Sci. Rev.* **202** 103110

Toimil A, Losada I J, Álvarez-Cuesta M and Le Cozannet G 2023 Demonstrating the value of beaches for adaptation to future coastal flood risk *Nat. Commun.* **14** 3474

Toimil A, Losada I J, Camus P and Díaz-Simal P 2017 Managing coastal erosion under climate change at the regional scale *Coast. Eng.* **128** 106–22

Toimil A, Losada I J, Nicholls R J, Dalrymple R A and Stive M J F 2020a Addressing the challenges of climate change risks and adaptation in coastal areas: a review *Coastal Engineering* vol 156 (Elsevier B.V.)

Turner I L, Harley M D, Short A D, Simmons J A, Bracs M A, Phillips M S and Splinter K D 2016 A multi-decade dataset of monthly beach profile surveys and inshore wave forcing at Narrabeen, Australia *Sci. Data* **3** 160024

Vitousek S, Barnard P L, Limber P, Erikson L and Cole B 2017 A model integrating longshore and cross-shore processes for predicting long-term shoreline response to climate change *J. Geophys. Res.* **122** 782–806

Vitousek S, Cagigal L, Montaño J, Rueda A, Mendez F, Coco G and Barnard P L 2021 The application of ensemble wave forcing to quantify uncertainty of shoreline change predictions *J. Geophys. Res.* **126** e2019JF005506

Vitousek S, Vos K, Splinter K D, Erikson L and Barnard P L 2023 A model integrating satellite-derived shoreline observations for predicting fine-scale shoreline response to waves and sea-level rise across large coastal regions *J. Geophys. Res.* **128** e2022JF006936

Vos K, Splinter K D, Harley M D, Simmons J A and Turner I L 2019 CoastSat: a Google Earth Engine-enabled Python toolkit to extract shorelines from publicly available satellite imagery *Environ. Modelling Softw.* **122** 104528

Yates M L, Guza R T and O'Reilly W C 2009 Equilibrium shoreline response: observations and modeling *J. Geophys. Res. Oceans* **114** C09014