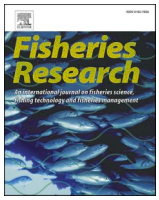
Fisheries Research 261 (2023) 106624

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



An ensemble modelling approach for spatiotemporally explicit estimation of fish distributions using data assimilation

Cian Kelly [a,*], Finn Are Michelsen [b], Morten Omholt Alver [a]

[a] Department of Engineering Cybernetics, NTNU, Trondheim, Norway
[b] Department of Fisheries and New Biomarine Industry, SINTEF Ocean, Trondheim, Norway

ARTICLE INFO

Handled by Jie Cao

Keywords:
Individual-based model
Ensemble
Data assimilation
Corrections
State estimation
Decision-support

ABSTRACT

This article presents a novel method for estimating large scale spatiotemporal distribution patterns of fish populations modelled at the individual level. A single realization of an individual-based model calibrated on historic data has weak predictive capacity, given the underlying uncertainties faced when modelling a relatively small cluster of individuals operating in a high dimensional spatial plane. By incorporating real-time data sources to update these models, we can improve their predictive capacity. When correcting estimates from a large population of individuals, we don't have access to information about individual histories, such as information derived from tagging data. We propose mapping individuals to derived density matrices, which can be corrected using conventional data sources which describe a mass of individuals e.g. catch data. An ensemble of derived states are used as forecast inputs to an assimilation procedure, that calculates an analysis state matrix of the same form. An individuals' position and biomass values are updated based on the analysis values. To assess the effect of corrections, we setup a simulation experiment to explore the impact the number of measurement points has on the updated spatiotemporal distribution. The measurement points were sampled from derived states of a twin model that resembles the original model. The output of the twin model serves as the true distribution. With an increasing number of measurement points the centre of mass of the modelled distribution converges on the true distribution and the two distributions increase in overlap. Additionally, the absolute error between model and true values decreases. This estimation method, applied to individual-based models and coupled with real-time fisheries data, can improve spatially explicit estimates of fish distributions.

1. Introduction

Individual-based models (IBMs) simulate interactions between a population of model individuals and their surrounding environment (Grimm and Railsback, 2005). IBMs capture large scale phenomena with simple interactions. Complexity arises from modelling bottom-up processes, rather than imposing population level parameters such as birth and death rates (DeAngelis and Grimm, 2014). It is the individuals local input information that produces unique responses. Infection transmission models in epidemiology demonstrate this. Contact rates and transmission probabilities vary in accordance with the unique behaviour of individuals. The social network of the individual matters too (Koopman and Lynch, 1999; Buchwald et al., 2020). For these reasons population-level features are not a simple sum of parts. The subtle differences between individuals alter system behaviour over time. Differences arise as individuals update state variables, such as position and velocity, at frequent time intervals (DeAngelis and Grimm, 2014). In this way, internal states represent the integration of past and present input over time. Incrementing states forward in time, in distinct simulation scenarios, can explain the evolution of higher level phenomena. For example, an individual fish's response to temperature and current explains variation in migratory routes (Barbaro et al., 2009; Tu et al., 2012).

These properties make IBMs attractive explanatory tools. However, IBMs have weak predictive capacity at a precise location and time, and are of limited operational use. As Baker et al. (2018) has pointed out, mechanistic models rely on oversimplified assumptions that are narrow in nature and limited in broad predictive power. Models are tuned once using historical data, validated on an independent dataset, before forecasting future estimates. This is useful for points trained on the historical data, but as the model progresses, states diverge from reality owing to uncertainties (Ward et al., 2016; Kieu et al., 2020). We consider model

* Corresponding author.
E-mail address: cian.kelly@ntnu.no (C. Kelly).

https://doi.org/10.1016/j.fishres.2023.106624
Received 28 April 2022; Received in revised form 5 October 2022; Accepted 17 January 2023
Available online 20 January 2023

simplifications, intialization values, mechanistic assumptions, parameters and inputs as main sources of uncertainty. Integrating real-time, real-world observations to correct model states is a way of controlling divergence. Data Assimilation updates model states based on real-time observations. It operates under the assumption models or observations alone cannot resolve the real system (Fu et al., 2011). It estimates the real system by applying a statistical correction term to model estimates (Alver and Michelsen, 2015. Data Assimilation has been applied successfully to applications in fishery models, predictive ecology, the terrestrial carbon cycle, traffic simulation, amongst other areas (Niu et al., 2014; Kieu et al., 2020).

The Ensemble Kalman Filter (EnKF) is a Data Assimilation method initially developed by Evensen (1994). It is used for state and parameter estimation of non-linear systems e.g. atmospheric and ocean systems (Houtekamer and Mitchell, 2001; Alver and Michelsen, 2015). The EnKF simulates separate instances of a model in a Monte Carlo setup where instances diverge over time due to random perturbations which represent the uncertainty in the model and its inputs. The divergence in model states is used to calculate error statistics. When observations are available, a correction term is applied to each instance of the model, based on these error statistics (Evensen, 2009). Although the EnKF implicitly assumes Gaussian distributions of prior states, it is effective in approximate estimation of states in non-linear systems that violate this assumption, which is often the case (Katzfuss et al., 2016).

Assimilating measurement data with IBM output can vastly improve predictions. This has been explored in the case of population-level estimates (Niu et al., 2014). Here we focus on high dimensional spatially explicit patterns of abundance in fish distributions. Real-time integration of available observations has the potential to facilitate the goal of time sensitive decision-support for stakeholders in the fishing industry. There are two main challenges to achieving this objective. Firstly, at large spatial scales, we currently don't have access to measurement data that describe individual fish with unique identities tracked through time to compare with IBM output directly. Secondly, the real spatially explicit distribution of fish stocks at any given time is highly uncertain, due to the sparsity of observations.

We propose a novel approach for correcting the IBM that is compatible with measurement sources that don't preserve an individual fishes identity, such as catch data. This method maps IBM output onto a two dimensional spatial grid, where derived density estimates are utilized as prior states in the EnKF. With minimal manipulation, the analysis estimate is remapped to the IBM individuals. The EnKF is advantageous for this purpose, as the IBM model mechanics are not altered directly, avoiding degeneracy of the model structure (Katzfuss et al., 2016). Additionally, the EnKF is suitable for assimilation when we don't fully understand the sources of errors.

To address the issue of the true underlying distribution, we use a twin model experiment to simulate observations. That is, we simulate an altered IBM and treat it as the true migration pattern. The IBM is based on the spawning migration of the Norwegian Spring Spawning Herring, as described in Kelly et al. (2022). The model IBM mechanics are extended from the single realization described, to an ensemble of estimates, through addition of stochastic perturbations to model components. In the true scenario, the deterministic realization is simulated alongside the ensemble of models, then sampled for measurement points. The measurement values are assimilated with each instance of the ensemble. We then investigate the impacts of measurements on the model distribution, given we have full knowledge of the true geographical distribution. The convergence of the ensemble on the true distribution indicates the capacity to correct the IBM. Spatial indices were used to measure this convergence and scenarios run in this study examined the influence of number of observations on spatial patterns.

With improvements in technology, observations will become less sparse and thus, our capacity to correct models shall improve (Fu et al., 2011). For example, acoustic technology today involves use of a multi-beam system that can resolve multiple fish at once (Chu, 2011). Additionally, studies have shown it is possible to classify fishing activity with high precision, from available vessel data at an individual level, such as position, speed and turning angle of boats (Bez et al., 2011; de Souza et al., 2016). Assimilating such sources of fisheries data with spatially explicit model predictions can improve our collective understanding of dynamics of large scale fish distribution patterns.

2. Materials and methods

The purpose of the model is to improve spatiotemporal estimates of fish distributions through integration of observations when they become available. The following description primarily focuses on two aspects: 1) Modifying the IBM to make it compatible with the EnKF procedure for assimilating data. 2) Setup of the twin model experiment to analyse the impact of measurements on the model (Fig. 1).

2.1. Ensemble of IBM trajectories

The IBM prediction model developed in Kelly et al. (2022) of a single model trajectory of herring is reproduced here for completeness, with the following set of difference equations at each time step k:

$$\mathbf{p}[k] = \mathbf{p}[k-1] + \Delta t\big(\mathbf{v}_f[k-1] + \mathbf{v}_c[k-1]\big) \tag{1}$$

$$\mathbf{v}_f[k-1] = -\Phi\mathbf{v}_c[k-1] + \mathbf{v}_b[k-1] \tag{2}$$

$$\mathbf{v}_b[k-1] = \mathbf{r}_b\left(\begin{bmatrix}\cos(\boldsymbol{\theta}[k-1])\\ \sin(\boldsymbol{\theta}[k-1])\end{bmatrix}\right) \tag{3}$$

$$\boldsymbol{\theta}[k-1] = f(\nabla T[k-1], \nabla D[k-1]) \tag{4}$$

where $\mathbf{p}$ is the vector of positions, $\mathbf{v}_c$ are the horizontal current components vector at the individuals position in m s^{-1}, Φ is a parameter that controls the response to the current, $\mathbf{r}_b$ is the swimming speed of the individual and $\boldsymbol{\theta}$ is the angle of orientation, which is a function temperature and bathymetry gradients (T and D). This configuration allows the individual to respond with a higher priority to the horizontal components of the prevailing current.

As Evensen (2009) notes, the solution to a dynamical model is one of an infinite many realizations, and for meaningful solutions, we must consider the time series of the probability density function. The IBM modelled one realization of the herring migration pattern, optimized based on a narrow set of assumptions (Kelly et al., 2022). Numerous alternative realizations are possible, given uncertainties in model evolution over time. Here, we add random perturbations to the IBM state variables to generate a set of N divergent instances sequentially in time. This generates N trajectories of the original IBM, which are held in memory and updated independently at each time step.

Position $\mathbf{p}$ and velocity $\mathbf{v}$ of individuals were extended from the single IBM to N instances, notated by the state matrices $\mathbf{P}$ and $\mathbf{V}$, both with N columns. Additionally, biomass $\mathbf{B}$ of individuals in kg is added as another state here, where each individual was treated as a mass of fish (also referred to as a superindividual):

$$\mathbf{P}[k] = \mathbf{P}[k-1] + \Delta t(\mathbf{V}[k] + \widetilde{\mathbf{V}}[k]) \tag{5}$$

$$\mathbf{B}[k] = \mathbf{B}[k-1] - \Delta t(\widetilde{\mathbf{B}}[k] + \omega)\mathbf{B}[k-1] \tag{6}$$

where Δt was the time increment, reduction in biomass was controlled by the constant parameter ω, and divergence in states $\mathbf{V}$ and $\mathbf{B}$ were caused by the stochastic errors $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{B}}$. The expected value $\mathrm{E}[\widetilde{\mathbf{V}}] = \mathrm{E}[\widetilde{\mathbf{B}}] = 0$. These errors produce prediction uncertainty in the system, representing errors in individuals migration direction, speed and mass and were formulated as follows:

$$\widetilde{\mathbf{V}}[k] = \widetilde{\mathbf{R}}[k]\begin{bmatrix}\cos(\widetilde{\boldsymbol{\theta}}[k])\\ \sin(\widetilde{\boldsymbol{\theta}}[k])\end{bmatrix} \tag{7}$$

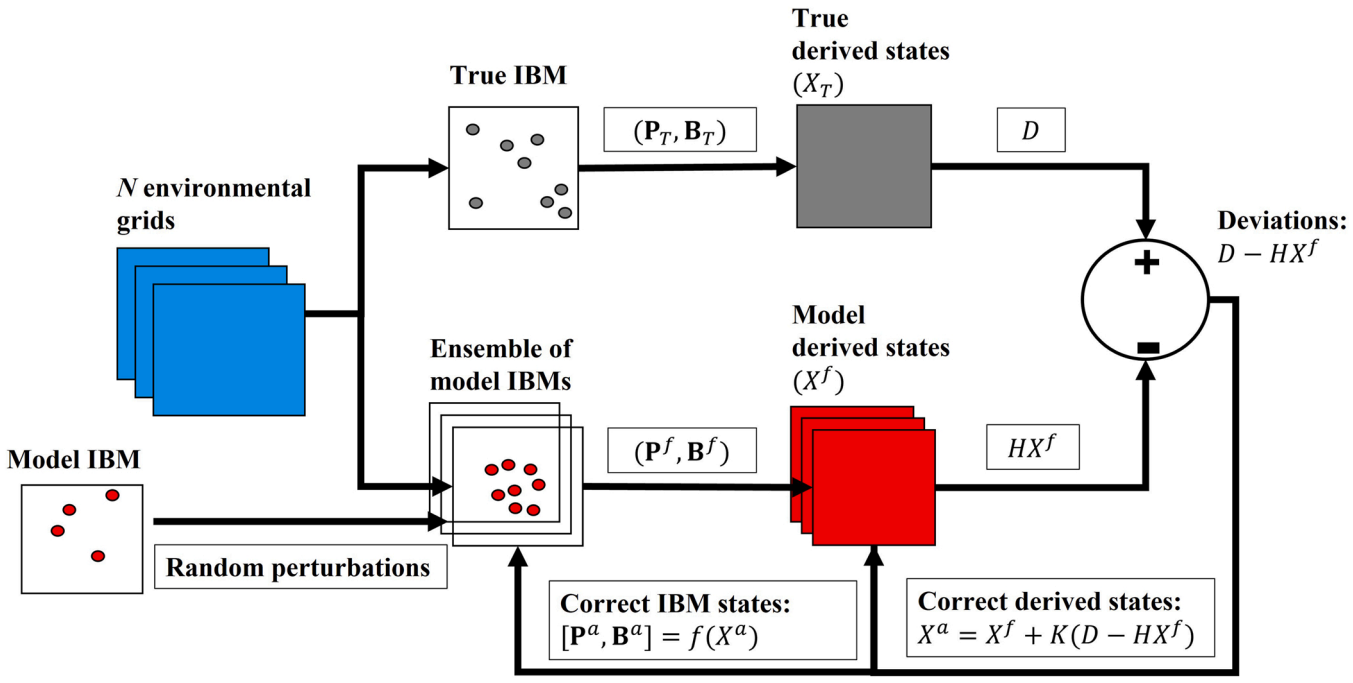


Fig. 1. A conceptualization of the assimilation of data using the twin model experiment.

$$\widetilde{\mathbf{R}}[k] = \alpha_1 \widetilde{\mathbf{R}}[k-1] + \alpha_2 \left(e_{sx1} \mathcal{N}\left(0, \epsilon_R^2\right)_{1xN} \right) + \mathcal{N}\left(0, \sigma_R^2\right)_{sxN} \tag{8}$$

$$\widetilde{\boldsymbol{\Theta}}[k] = \alpha_1 \widetilde{\boldsymbol{\Theta}}[k-1] + \alpha_2 \left(e_{sx1} \mathcal{N}\left(0, \epsilon_\Theta^2\right)_{1xN} \right) + \mathcal{N}\left(0, \sigma_\Theta^2\right)_{sxN} \tag{9}$$

$$\widetilde{\mathbf{B}}[k] = \alpha_1 \widetilde{\mathbf{B}}[k-1] + \alpha_2 \left(e_{sx1} \mathcal{N}\left(0, \epsilon_B^2\right)_{1xN} \right) + \mathcal{N}\left(0, \sigma_B^2\right)_{sxN} \tag{10}$$

where temporally correlated, slowly varying errors were controlled by the parameters α_1 and α_2, e_{sx1} is an $sx1$ vector of ones, where s is the number of individuals and ϵ_R, ϵ_Θ and ϵ_B represent standard deviation in speed, angle and biomass for each of the ensemble members. Applying these errors cause the N columns to diverge, creating an ensemble of random realizations.The standard deviations were calibrated to maintain spread between ensemble members and limit the severity of corrections when data was assimilated. This formulation is similar to system noise modelled in ocean models that use temporal autocorrelation of random noise to account for errors in representation of certain processes (Keppenne et al., 2008). In this case, we represent the errors in the individuals state matrix, resulting from uncertainties in the evolution of migration patterns. In addition to the temporally autocorrelated ensemble noise, spurious gaussian noise is added to each individual with mean of zero and standard deviations of σ_R, σ_Θ and σ_B. These individual noise components account for uncertainties in the migration of individual fish, regardless of ensemble member.

2.2. Data Assimilation framework

Before assimilation, the forecast IBM position $\mathbf{P}^f$ and biomass $\mathbf{B}^f$ from Equation (5) and (6) are converted to derived estimates:

$$X^f = f(\mathbf{P}^f, \mathbf{B}^f) \tag{11}$$

where X^f is an nxN grid of density values, with each cell representing the sum of the biomass of all individuals within that grid cell.

The EnKF uses the error covariance structure of the ensemble forecast X^f to calculate the correction term. However, the full covariance matrix is too large to be explicitly calculated here. We employ an equivalent implementation by Mandel (2006), which avoids the calculation of the full error covariance matrix and derives directly the prediction error covariance matrix in the observation space:

$$A = X^f - \frac{1}{N}\left(X^f e_{Nx1}\right) e_{1xN} \tag{12}$$

$$HA = HX^f - \frac{1}{N}\left(\left(HX^f\right) e_{Nx1}\right) e_{1xN} \tag{13}$$

$$P = \frac{1}{N-1} HA(HA)^T I_m + R \tag{14}$$

$$K = L \odot \left(\frac{1}{N-1} A(HA)^T P^{-1} \right) \tag{15}$$

where H is an mxn matrix that contains ones at m measured states, I_m is an mxm identity matrix, R is the mxm observation error covariance matrix, where each element on the diagonal is the variance of observation noise (σ_O^2), L is an mxN localization matrix and finally, K is the Kalman Gain, which is used to calculate the correction term. Localization adds a penalty to model covariances that are distant from the measurment point. For a small ensemble and high dimensional system, localization is necessary to limit the impact of observations (Houtekamer and Mitchell, 2005). The operator ($\odot$) is the Schur product, an elementwise operation acting on all covariance values. The full L matrix was calculated from a radial basis function:

$$L_{ij} = \begin{Bmatrix} 0, & \text{if } \| g_i - g_j \|^2 > c \\ \exp\left(\frac{- \| g_i - g_j \|^2}{2\rho^2} \right), & \text{otherwise} \end{Bmatrix} \tag{16}$$

where we calculate the euclidean distance between the xy grid coordinate for each model grid cell g_i and the measured grid cell g_j. The value is calculated for all model coordinates ($i = 1{:}\ n$) and measurement point coordinates ($j = 1{:}\ m$). When i equals j, the value of L equals one, and as i moves away from j there is exponential decline in the value of L. To controls spatial correlations around the measurement point, the constant parameter ρ is used. In addition, to avoid spurious correlations, a cut-off point c sets distant spatial covariances to zero.

The analysis estimate X^a is calculated as follows:

$$X^a = X^f + K(D - HX^f) \tag{17}$$

where D is the mxN measurement matrix. The standard EnKF adds $\mathcal{N}(0,R)$ realizations of observation errors to generate observational perturbations. However, in this study, where we are sampling from a non-negative distribution with mostly zero values, perturbation of observations led to inaccuracies in the posterior field. Instead, the columns of D are treated as replicates of the original measurement vector. Treating the observations as deterministic contracts the variance across the ensemble in the analysis estimate (Burgers et al., 1998). To compensate for this contraction in spread between columns of X^a, an inflation factor ψ was used to replace the analysis estimates, as mentioned in Evensen (2009):

$$X_z^a = \overline{X}^a + \psi\left(X_z^a - \overline{X}^a\right) \tag{18}$$

where z is the index for the ensemble member and $\overline{X}^a$ is the ensemble mean of the analysed derived states.

Following assimilation, the IBM is modified to reflect updated grid cell biomass values and this is achieved with minimum manipulation of the underlying model structure. The derived analysis estimate X^a is converted back to IBM states:

$$[\mathbf{P}^a, \mathbf{B}^a] = f(X^a) \tag{19}$$

where $\mathbf{B}^a$ and $\mathbf{P}^a$ are the analysis biomass and position values for individuals, calculated from X^a.

2.3. Data assimilation adapted for the IBM

In this section, we detail how the conversion between IBM and EnKF states was achieved in Equations (11) and (19). Mapping from individual representations to derived density states means aggregating information from individuals into a grid representation that describes geographical distribution and abundance. Cocucci et al. (2022) describes this as a transition between micro- and macro-states. To achieve this mapping, the forecasted states are derived individual by individual, as shown in Algorithm 1, until the X^f matrix is furnished with an ensemble of density fields.

Algorithm 1. Algorithm for mapping from IBM states to forecast states X^f in Equation (11).

```
for e ← 1 to N do
    for i ← 1 to s do
        1. Find cell coordinates xy of the individual from P^f(i, e)
        2. Find biomass b of the individual from B^f(i, e)
        3. Add derived value to forecast state matrix: X^f(x, y, e) = X^f(x, y, e) + b;
    end
end
```

Mapping from the high dimensional analysis field to relatively fewer individuals is more challenging, and Algorithm 2 was designed to maximize the retention of density values, while limiting adjustments to the IBM. Each cell is checked for the analysis estimate and if it is greater than zero, the value is assigned as individual biomass, divided evenly amongst individuals at that position. If the analysis estimate is greater than zero, but there are no individuals present, one individual positioned in a cell with a zero analysis estimate is randomly moved there (assuming there is an individual available to move).

Algorithm 2. Algorithm for mapping from analysis states X^a to IBM states in Equation (19).

(continued on next column)

(*continued*)

```
for e ← 1 to N do
    for c ← 1 to n do
        1. Find xy coordinates of cell from linear index c
        2. Find derived density d of analysis matrix: d = X^a(x, y, e)
        3. Check the number of individuals si at coordinates from P^f(:, e)
        4. if si > 0 & d > 0 then
            a) Each individual retains the prior position: P^a = P^f;
            b) Divide d evenly amongst si individuals in cell, so for each individual: B^a = d/si;
            end
        5. if si == 0 & d > 0 then
            a) Randomly move an individual from a cell where density is zero: P^a = xy;
            b) Set the biomass of moved individual to the density in this cell: B^a = d;
            end
    end
end
```

This method is similar to the randomized redistribution described in Cocucci et al. (2022), where individuals are moved between categories where needed and attributes are updated. In this case, $\mathbf{P}^a$ and $\mathbf{B}^a$ are estimated from the macro- to micro-state mapping. This mapping conserves density estimates with higher priority than individual histories, given real fisheries observations are of aggregated individuals.

2.4. Twin model development

The observations used in this study were synthetically generated using a twin model, which represents the true distribution here. Twin model design has been used to give insight into capacity to correct model components with few observed variables (Simon and Bertino, 2009. Specifically, we are testing the data assimilation procedure and observability of the system in our setup. Here, we observe a derived variable from the twin model (X_T), which is a density field with dimensions $nx1$. This was sampled in the assimilation procedure to furnish the D matrix in Equation (17). The samples were taken from a predefined grid along the Norwegian coast (Fig. 2). Like the model IBM, these values were derived from individual state variables:

$$X_T = f(\mathbf{P}_T, \mathbf{B}_T) \tag{20}$$

where $\mathbf{P}_T$ and $\mathbf{B}_T$ were position and biomass of twin model individuals. Unlike the model IBM, the twin individuals were stepped forward with no feedback from the assimilation procedure. The twin IBM was updated using the same dynamics as the main IBM, with the exception of the swimming speed $\mathbf{r}_b$ in Equation (3), which was reduced in the twin model. This hypothetical scenario represents a situation where the model overestimates the true migration speed of the herring.

2.5. Model Simulation

The environmental conditions were obtained from a run of the physical-biological ocean model SINMOD (Slagstad and McClimans, 2005 set up in a domain with 4 km horizontal resolution covering the Norwegian and Arctic Seas. The same grid resolution was used for the derived states, where $n = 941 \times 620$. The IBM modelled individuals in a 2D environment where position was updated on N continuous horizontal planes. The s individuals initialized in each ensemble member had their position $\mathbf{P}$ centred in an area in Northern Norway in mid-January. The biomass $\mathbf{B}$ states for each ensemble were initialized from a Gaussian distribution with mean μ_B and standard deviation Σ_B. These values were divided among individuals based on their proximity to the centre point of the starting position. The model was simulated for a period of 45 days during the herring spawning migration. The time increment Δt was 4 h, for a total of 270 time steps. The simulation period was split into assimilation and non-assimilation periods. The assimilation period operated from day 18–37. During the assimilation period, corrections were performed once per day. This left a period prior- and post-assimilation for the states to diverge from the ensemble mean. Model parameters were calibrated to stabilize the assimilation

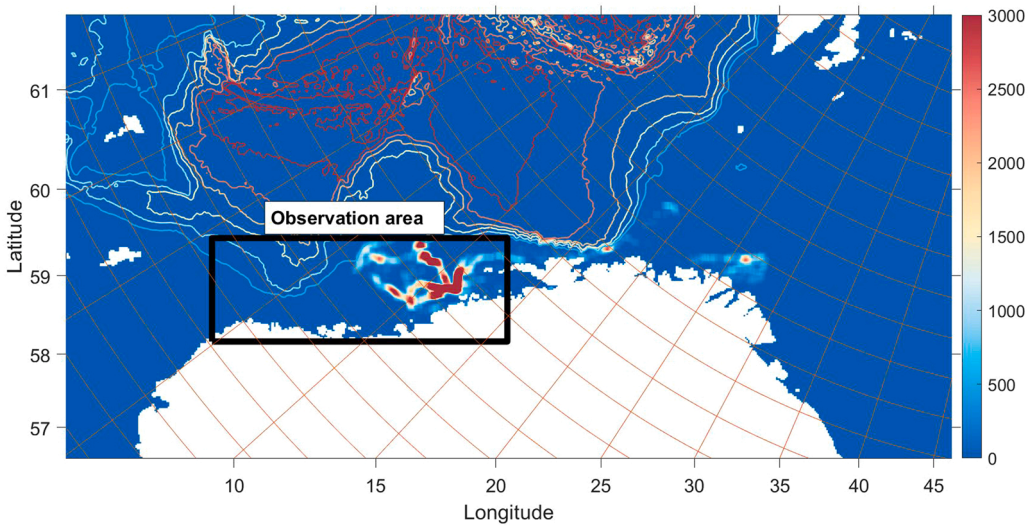


Fig. 2. The area sampled for measurement point during assimilation (black rectangle), where the colored cells represent the true distribution of derived density values (X_T), in kilograms, on day 25 of the simulation. The contours represent the depth in metres. For convenience, the colorbar represents the scale of both X_T and depth. Longitude and latitude ticks extend from the x and y axis, respectively, along the Norwegian coast.

procedure, specifically to avoid extreme correction terms (Table 1).

To investigate the number of observation points needed to make the ensemble converge towards the true state, four separate scenarios were setup to test varying number of observation points, where $S_1 = 100$ points, $S_2 = 200$ points, $S_3 = 400$ points and $S_4 = 800$ points. The points were sampled at equally spaced intervals in the observation area (Fig. 2). To implement a fixed virtual observer system, these same points were sampled on each day of the assimilation period. The output from these scenarios was compared to a control model, which was run in parallel with no assimilation of data.

2.6. Analysis

Quantitative and qualitative measures of performance investigated the capacity to correct the four model scenarios with samples of measurement points and thus, represent the spatiotemporal patterns of the true fish distribution. This is important in the geographical mapping of fish stocks. The quantitative measures used were based on equations from Woillez et al. (2007). The Centre of Gravity (CG) measures the weighted position of the density estimates at a given time. We investigated how this diverged from the control model and converged towards the true CG. Global index of Collocation (GIC) is a measure of the overlap of two separate distributions. It takes into account the CG of the two distributions and the variance around the CG. A value of one is perfect overlap between the two and a value of zero indicates distinct populations. Both CG and GIC were described in terms of latitude and longitude coordinates. They were calculated from the densities of the derived states $\overline{X}$ and X_T, where $\overline{X}$ is the ensemble mean of the model. The derived states were saved once per day during the model simulation, and after the assimilation step.

In addition to spatial estimates, the raw error between the density values of the model and true model were analyzed. This ground truth error was taken as the difference root-mean squared difference between density estimates from the model and the true derived density estimates:

$$e_T = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(\overline{X}_i - X_{T_i})^2} \tag{21}$$

where i is the model coordinate and l is the number of indices within the observation area (Fig. 2).

Table 1
List of model variables and parameters for ensemble simulations.

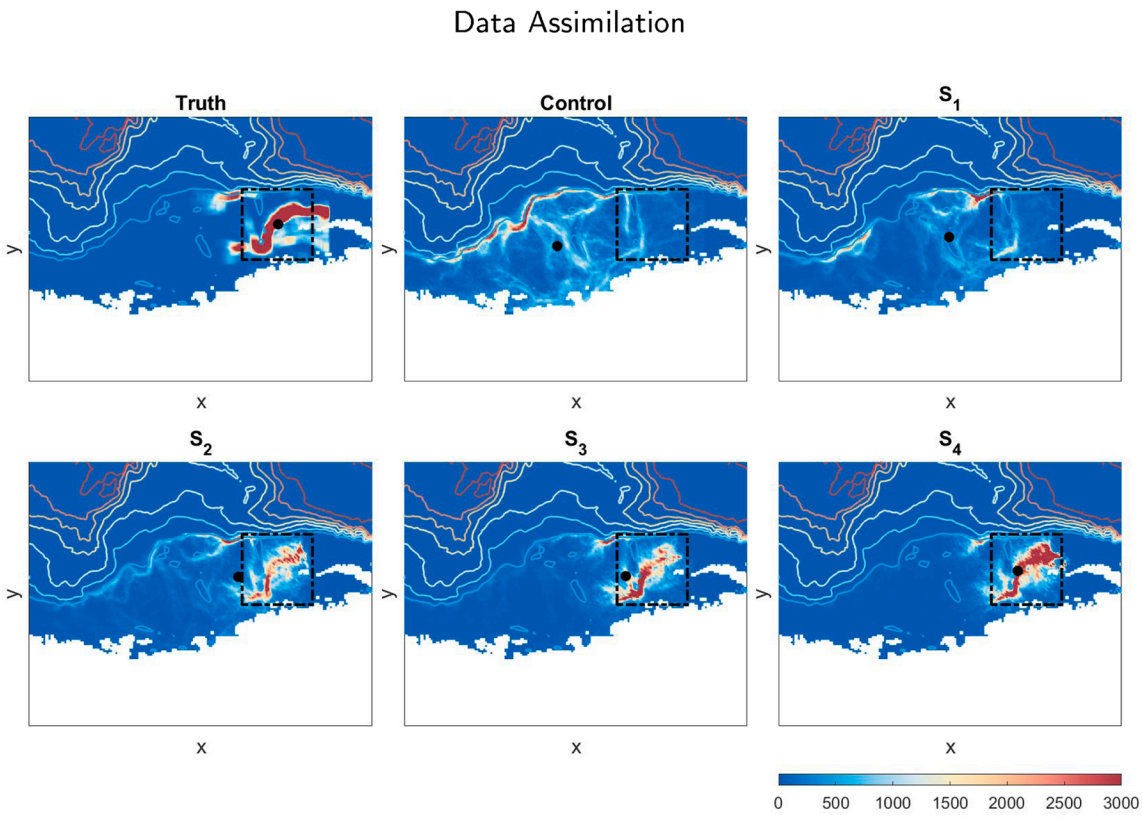
Name	Description	Unit	Value
State variables			
$\mathbf{P}$	Model position of individuals		
$\mathbf{B}$	Model biomass of individuals		
X	Model derived density states		
$\mathbf{P}_T$	True position of individuals		
$\mathbf{B}_T$	True biomass of individuals		
X_T	True derived density states		
Parameters			
N	Number of ensemble members	–	100
s	Number of individuals	–	10000
m	Number of observations	–	–
n	Number of derived states	–	583420
Δt	Time step	h	4
ϵ_R	Standard deviation in ensemble magnitude	m s^{-1}	0.1
ϵ_Θ	Standard deviation in ensemble angle	$^\circ$	45
ϵ_B	Standard deviation in ensemble biomass	–	0.002
ω	Biomass reduction for ensemble	–	0.005
μ_B	Mean total initial biomass	kg	5e06
σ_R	Standard deviation in individual magnitude	m s^{-1}	0.01
σ_Θ	Standard deviation in individual angle	$^\circ$	4.5
σ_B	Standard deviation in individual biomass	–	0.002
σ_O	Observation noise	kg	250
Σ_B	Standard deviation in initial biomass	kg	1e06
α_1	Temporal correlation parameter	–	0.984
α_2	Temporal correlation parameter	–	0.129
ρ	Localization parameter	–	6
c	Localization cut-off	–	15
ψ	Inflation factor	–	1.01

3. Results

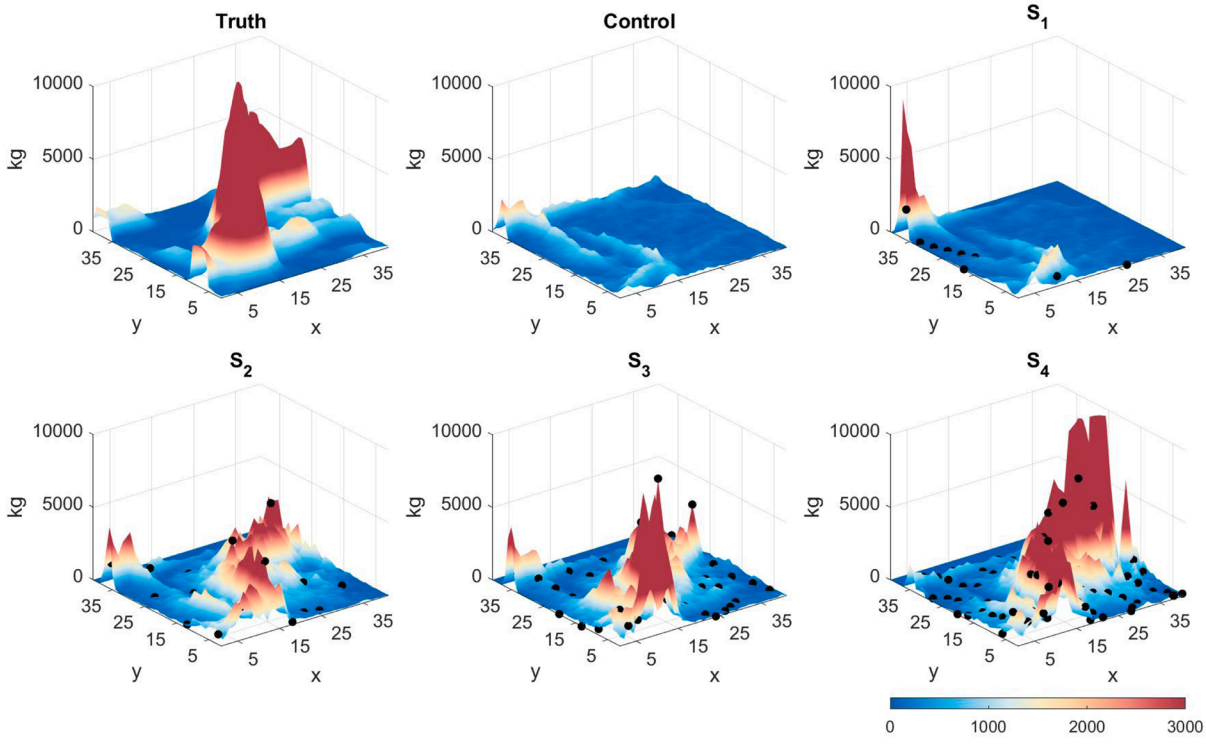
3.1. Qualitative analysis

The ensemble mean of the model ($\overline{X}$) is the best estimate of the model at each time step, and thus in the following results, $\overline{X}$ is the focus for analysis. The modelled migration moves south, with an offshore distribution prior to assimilation. During assimilation, the control simulation continues with this development, while the corrected scenarios develop a more coastal distribution, reflecting the true distribution. Following assimilation, all corrected scenarios deviate from the true distribution, but to a lesser degree than the control distribution.

To visualize the impact of corrections, we plotted derived density maps from two time stamps during the assimilation period, one at day 25 (Fig. 3) and another at day 35 (Fig. 4). The visual comparisons show

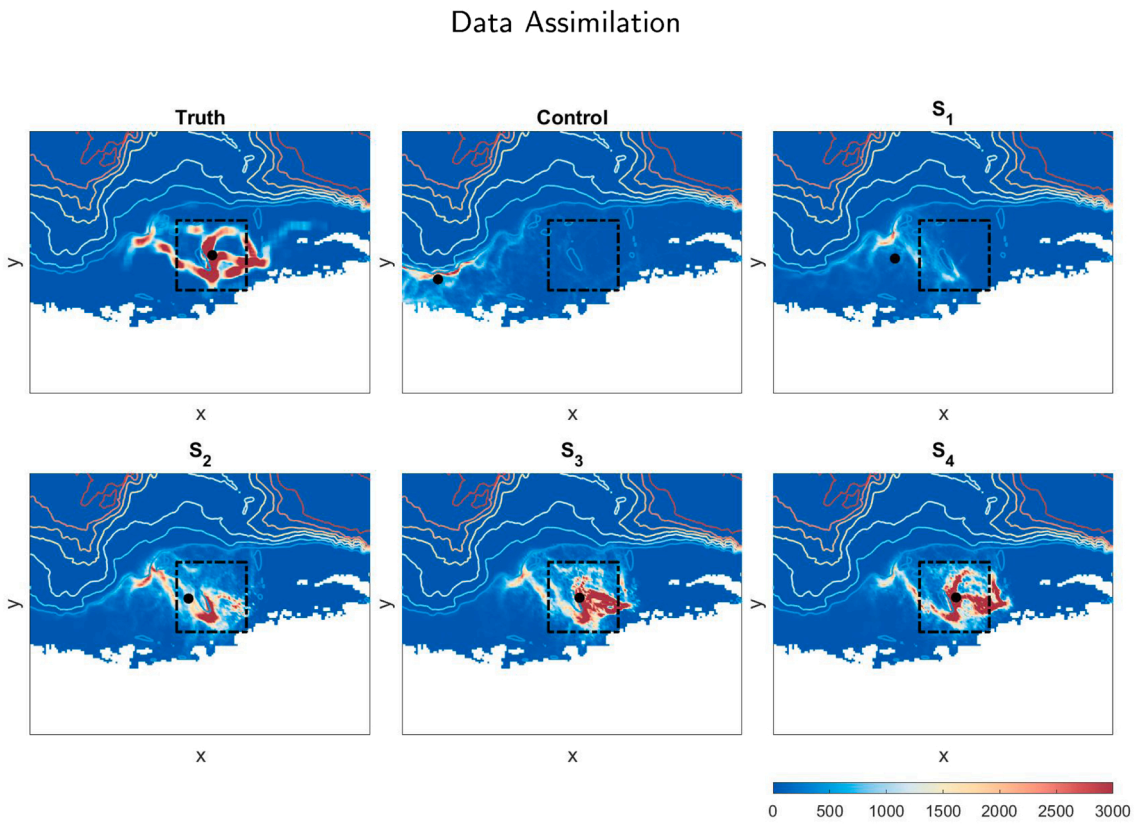


(a) Large scale distribution on day 20.

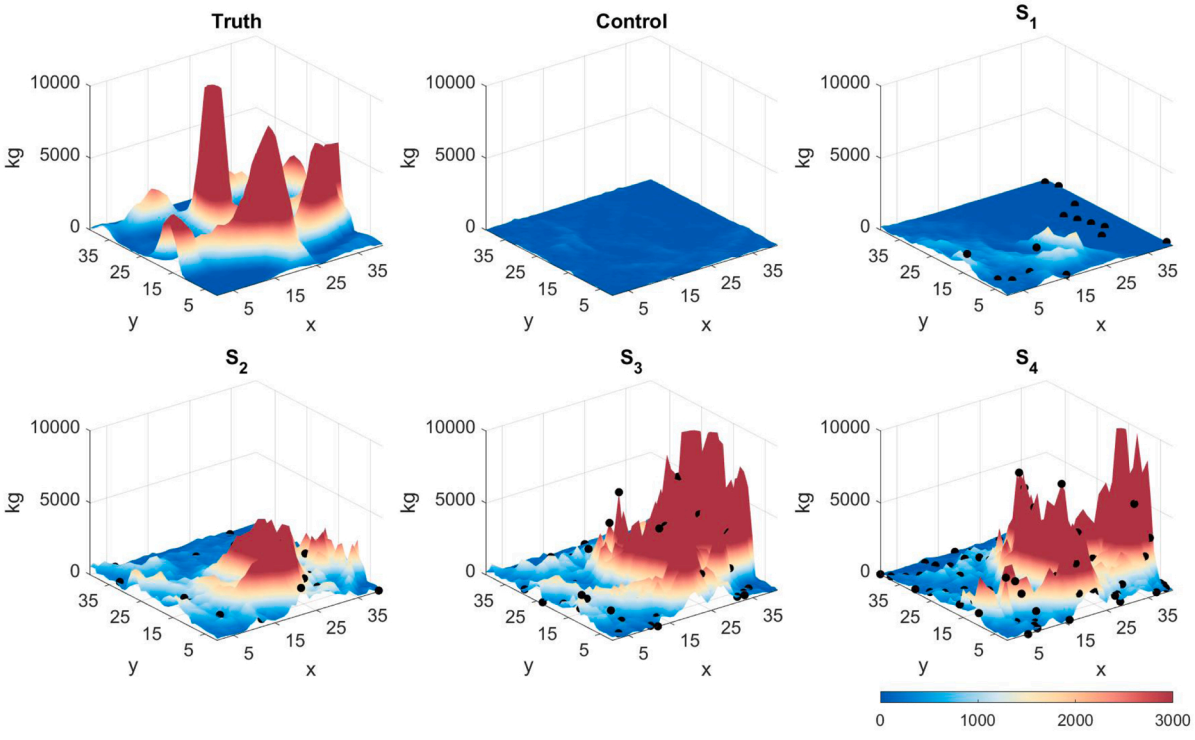


(b) Local distribution on day 20.

Fig. 3. (a) Large scale 2D plot of derived density states of model ($\overline{X}$) and true distribution (X_T) over a selected area of the Norwegian coast on day 20 of the simulation. The density colormap shows values in kg, while the contour lines show depth in metres. The black point shows the centre point (CG). The same colorbar scale is used for both. (b) The local 3D representations of derived states taken from the squared area in (a). Black dots show the location of measurement points. No measurement points were sampled for the control scenario.



(a) Large scale distribution on day 35.



(b) Local distribution on day 35.

Fig. 4. (a) Large scale 2D plot of derived density states of model ($\overline{X}$) and true distribution (X_T) over a selected area of the Norwegian coast on day 35 of the simulation. The density colormap shows values in kg, while the contour lines show depth in metres. The same colorbar scale is used for both. The black point shows the centre point (CG). (b) The local 3D representations of derived states taken from the squared area in (a). Black dots show the location of measurement points. No measurement points were sampled for the control scenario.

large scale distributions (2D plot) concatenated vertically with local distributions (3D surface plot). The large scale model distributions become more similar to the true distribution in the assimilated scenarios. The model CG converges on the true value also. The true distribution tends towards the coast and is concentrated more northerly. Further north (higher on the x axis) there is a clear increase in density values in assimilated scenarios, where measurement values from the true distribution are higher. Further south (lower on the x axis), the density values decrease, as a result of adjustments using zero measurement values.

In any given cell, local densities vary from true values, but on average the ensemble mean approaches the topography of X_T. Location and density of measurement points impact the scale of corrections. The peaks and valleys in the local densities of Fig. 3b and 4b are concentrated in varying locations, related to the position of measurement points. In the case of the control model, the derived density topology is distinct from the true derived topology. With an increasing number of measurement points, the density map starts to resemble the true map. For example, the ridge in S_4 resembles the surface features of the true model (Fig. 4b). In any given cell, the density estimates from $\overline{X}$ may not reflect those from X_T, but on average with increased observation numbers recreates a similar topography.

3.2. Quantitative analysis

The time series of CG of the ensemble mean for three scenarios was compared to the true CG. The CG is calculated in both latitude and longitude axes (Fig. 5). During the assimilation period there is convergence of CG towards the true point. The standard deviation across the ensemble reduces during the assimilation period and is sharply reduced with a higher density of observations. This sharp reduction is pronounced on the first call to the assimilation function. A large number of instances of the model are heavily penalized at this point. The standard deviation increases rapidly post-assimilation. There is faster convergence on latitude, reflecting the greater difference in latitude points, which was the main axis of variation for the simulation period. The inflation factor (ψ) is partly responsible for maintaining the standard deviation across the ensemble. The CG and standard deviation identical in all scenarios prior to assimilation. With a low density of observations, there are relatively weak corrections and convergence on the truth. In all cases there is divergence from the true CG post-assimilation. However, with a higher density of observations, there is less divergence. This is clear when we compare S_3 to S_1.

In Fig. 6, we compare each scenario to the true and control CG (Fig. 6a) and overlap (Fig. 6b). Before the assimilation period, the model and true distribution diverge and there is less overlap. The non-assimilated control model continues to diverge from the truth during the assimilation period. There is immediate divergence from the control on day 18 and convergence to the true CG for all scenarios. This is reflected in the overlap, which approaches a value of one over time.

The ground truth error (e_T) evaluates the raw error in the observation area between the model derived density values and the true derived density estimates from Equation (21). The error increases initially as the initial distribution of the truth and ensemble diverge in spatial characteristics. This pattern continues for the control model, until it eventually plateaus. The e_T is generally reduced from S_1 to S_4, with an initial sharp reduction, followed by a gradual decline in errors, with some irregularities. The e_T remains lower than the control for some days post-assimilation, until it eventually converges to a similar value at the end of the simulation.

4. Discussion

In this article, we have presented a novel general method for

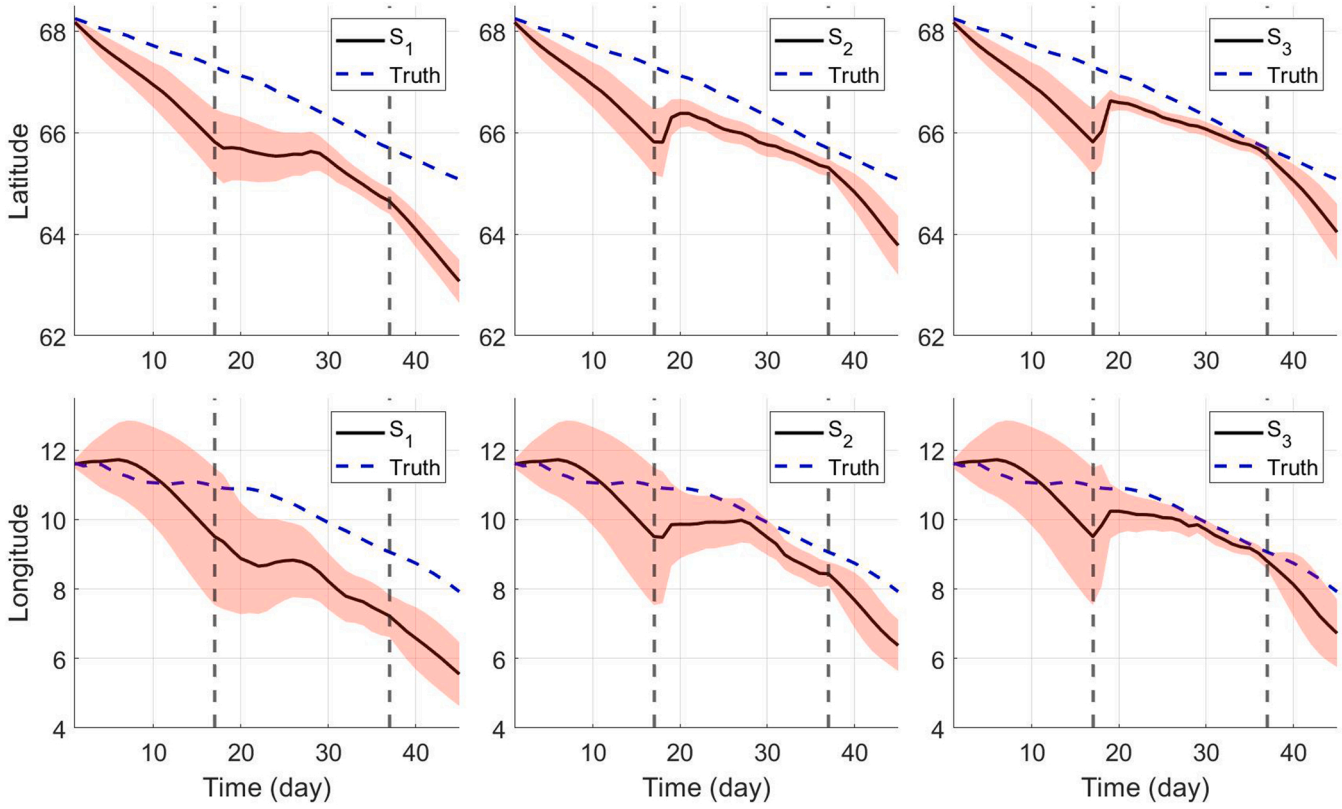
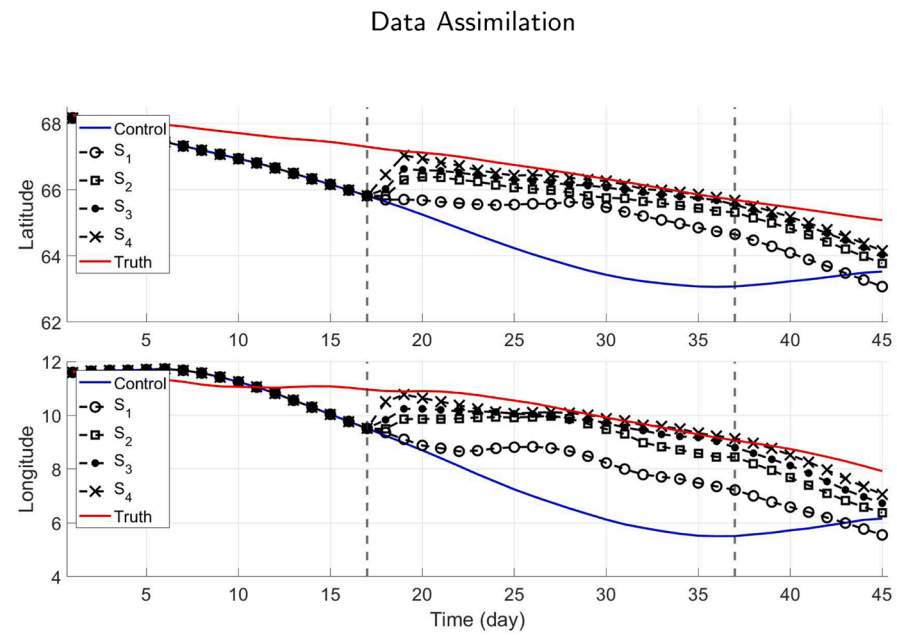
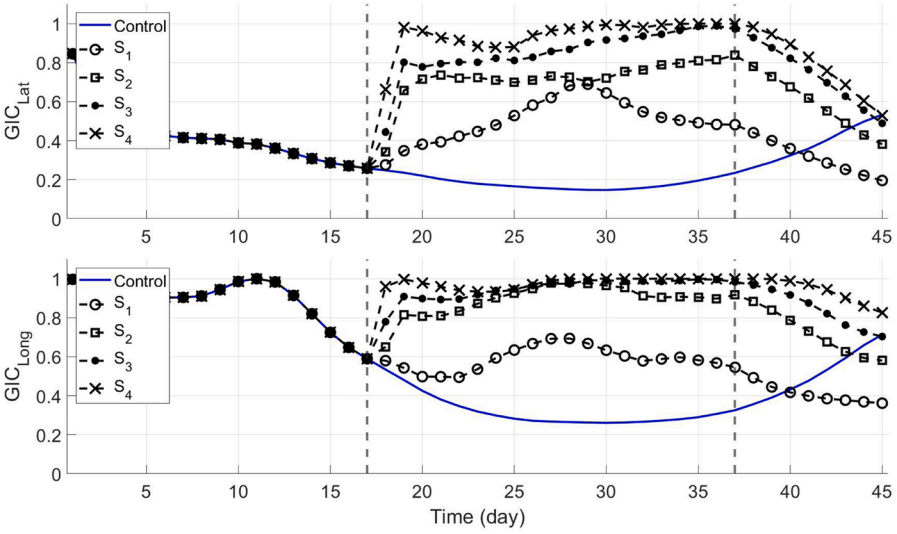


Fig. 5. Time series of Centre of Gravity (CG) in terms of latitude (first row) and longitude (second row) values during the simulation period. The CG of the true derived states (X_T) is shown with the dotted blue line in each panel, while the CG of the ensemble mean of the model derived states ($\overline{X}$) is shown with the black line, with each column representing a separate scenario. The vertical dotted grey lines represent the boundaries of the assimilation period.



(a) Centre points for model and true distribution.



(b) Overlap between model and true distribution.

Fig. 6. (a) Time series of centre points (CG) for the four scenarios, true distribution and control model in terms of latitude (top panel) and longitude (bottom panel). The vertical dotted grey lines represent the boundaries of the assimilation period. (b) Time series of overlap (GIC) between the model and true distributions for the four scenarios and control model in terms of latitude (top panel) and longitude (bottom panel). The vertical dotted grey lines represent the boundaries of the assimilation period.

assimilating data with an IBM operating in a high dimensional system. Assimilating data sources with a population of unique, discrete individuals is challenging, given observation sources like catch data, which do not preserve individual identity. Our suggestion is the use of derived states, which map individuals onto a discrete grid, with each grid cell expressing total densities of individuals. These derived states can then be assimilated with observation data, using an ensemble approach, to calculate a posterior density grid. Derived states can be remapped to the IBM states, without excessive manipulation of the model structure. Such a method is particularly useful for spatially and temporally explicit predictions of fish distributions. In the setup tested here, we compared scenarios for a bounded time period, where observations were available at frequent discrete intervals. The prior- and post-assimilation periods assumed no access to observations. In scenarios with access to many measurement points, the large scale and local density field converge on the true distribution. Importantly, we have shown how the assimilated scenarios outperform the non-assimilated control scenario in spatiotemporal predictions during the assimilation period. Performance is also superior for the time stamps directly succeeding assimilation. Towards the end of the time series, the model estimates eventually diverge from the true distribution and converge on the control case. Future work on incorporating fisheries dependent data can improve predictions and validate this method with real data.

4.1. *Making the IBM compatible with the EnKF*

The EnKF was chosen given the highly non-linear nature of the system modelled. Additionally, the EnKF shifts values in the model, rather than reinitializing model components. This prevents degeneracy of the model structure since each IBM instance is altered with minimal manipulation during assimilation. The IBM states were perturbed with Gaussian errors, but upon simulation the distribution of the ensemble of

derived states becomes non-Gaussian. However, while the EnKF implicitly assumes a Gaussian state-space, it provides good approximate solutions in cases where systems violate this assumption (Katzfuss et al., 2016).

In the real system, the observations will be sampled from an underlying non-negative concentration field (fish per unit area), and furthermore the field will have a bias towards values of zero in locations outside of the distribution of the migrating fish at any time. This has two consequences. First, the assumption of gaussian measurement noise is a poor fit to real world observations outside of the area covered by the migrating fish. Second, perturbation of observations using gaussian distributed random values will lead to a high number of negative values in those same areas. For these reasons, observations were treated as deterministic in this study. To compensate for the lower ensemble spread resulting from this choice, an ensemble inflation factor was applied (Evensen, 2009). To relax the need for adding observation errors, a square root EnKF variant could be considered (for example: Bishop et al., 2001). In future studies using real world observations, the statistics of the sampling process should be investigated in detail for the actual observations made, and the assimilation process customized accordingly. One approach could be to use approximate Bayesian inference along the lines proposed by Eidsvik et al. (2008).

An innovation of our method is the use of derived states that convert from particles to a field of density values. This allows corrections of density values rather than unique individual values, for which we don't have measurement data to describe. This would require, for example, large-scale tagging studies or time-sensitive acoustic back-scattering data, which are not fully developed as of now. Also, a spatial density field is easier to interpret and compare with observation data sources, relative to a cloud of particles at large spatial scales. When mapping the posterior state back to individual states, there were two manipulations. Firstly, the negative X^a values were removed to omit negative biomass values. Secondly, individuals that had zero biomass values (post-assimilation), were moved into positions with positive X^a values, until either none remained to be moved or all positive X^a values were assigned, in a process similar to the randomized redistribution described in Cocucci et al. (2022). This prevented loss of information during assimilation, without heavily intruding on the mechanics of the IBM directly.

The parameter values in assimilation were calibrated to ensure corrections were applied without extreme effects. The inflation factor kept spread around the ensemble, preventing excessive convergence of model on the observations, given observations were treated as deterministic. Localization was used to limit impacts of observations spatially and the choice of localization distance affects the corrections of cells between measurement points. Random perturbations on model states generated variance in the evolution of the migration scenarios. Balance between observation noise and model perturbations determined the overall scale of the corrections. One must note that assimilation is an approximate method of estimation, and operates under the assumption of uncertainty in model states and parameters. More persistent effects of the data assimilation can be achieved by also estimating model parameters in the data assimilation process, and for the present system the average swimming speed is a natural choice. Using parameter estimation, one would not only update the model state, but also attempt to tune the model to better match the real system at a fundamental level.

4.2. Impacts of measurements on the fish distribution

We used the twin model experiment (Fig. 1) to generate virtual observations, gauging the impact of corrections on the model IBM. The twin model was designed to configure a hypothetical shift in the spatial distribution of the fish relative to the prior assumption of our model. Inter-annual shifts in distribution are common in many migratory fish species, as captured often in surveys. For example, the herring spawning migration usually ends with masses of individuals spawning around Møre, but often, spawning occurs further north (Slotte and Fiksen, 2000). Our intention was not to explicate those reasons, but to gain insight into how assimilation of real-time data may modify the distribution to reflect a hypothetical disagreement between modelled and true distributions. In reality, fish distributions are highly uncertain in real-time as we have access to sparse observations, such as catch data. The twin model experiment design is useful as we are omniscient of the underlying true distribution and can easily analyse the impact of measurements.

Qualitatively, the large scale and local spatiotemporal distributions increasingly resemble the true distribution with denser clusters of observations (Fig. 3 and 4). Quantitatively, the centre points and overlap of the model converge on the true indices during assimilation to an increasing degree with more observations (Fig. 6a and b. Additionally, the deviations between ensemble instances is reduced with measurements, meaning the estimates are of higher certainty (Fig. 5). Finally, the ground truth error between the model and true derived density states is reduced with observations, showing, with access to more measurements, the model becomes more predictive in an absolute sense (Fig. 7).

At any one location, corrections are highly sensitive to placement of observation points. For example, at the coordinate (5,38) in (Fig. 3b) there are high density values in scenarios S_1, S_2 and S_3, while this peak is absent in S_4. This is related to the position of measurement points at this step of the analysis and the previous position of measurement points. However, on average, the denser the observations, the more the features reflect the true spatial distribution. The overall topography of S_4 resembles the true distribution more closely at the large and local scale (Fig. 3a and 3b).

4.3. Opportunities for model implementation

Today, there is much interest in utilizing fisher's knowledge, as it is considered part of the best available information for research studies. This is complementary to research survey data, which much work has relied on until now. Utilizing spatially explicit data, such as position and speed from vessel monitoring systems, we can improve our understanding of the state of the fishery in real-time. The estimation approach presented in this article is intended to be coupled with such data sources and thus, facilitate real-time monitoring of fish stocks. This has potential applications in fisheries management, marine planning and tracking of migrations. We note that this method is suggested to support decisions in these areas alongside complementary sources of information. Explicit decisions in fisheries systems are complex and require human deliberation and intervention. Thus, our model offers increased situational awareness without explicitly directing the decision-making process. Decision-making is the responsibility of the end user.

The method also has theoretical value for tuning parameters and

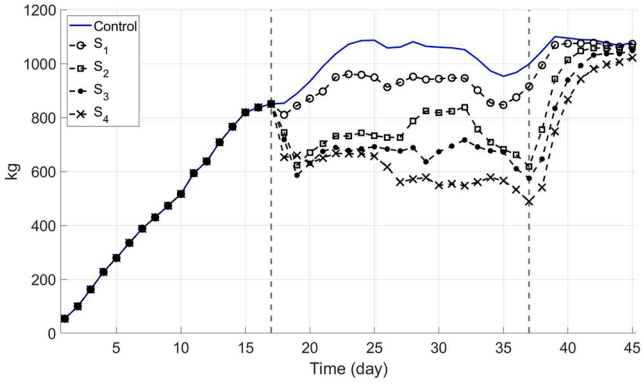


Fig. 7. The ground truth error (e_T) between the model and true distribution for the simulation period in kg. The vertical dotted grey lines represent the boundaries of the assimilation period.

improving models of fish dynamics. We have shown that applying corrections to model estimates improve prior predictions and with enough coverage, model estimates converge on the true spatial distribution. Further work will attempt to validate this method with real fisheries observation data. Furthermore, we wish to improve predictions when observations are not available, for example during time windows with little access to measurements.

CRediT authorship contribution statement

Cian Kelly: Conceptualization, Methodology, Writing, Simulation. **Finn Are Michelsen**: Writing – review & editing. **Morten Omholt Alver**: Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No data was used for the research described in the article.

Acknowledgements

We greatly acknowledge the support, input and feedback from the project participants: NTNU, SINTEF, NAIS and UiB. The work is part of the FishGuider project, which is funded by the project participants and the Norwegian Research Council (project number 296321).

References

Alver, M.O., Michelsen, F.A., 2015.Data assimilation with SINMOD. Technical Report. SINTEF Fisheries and Aquaculture.

Baker, R.E., Peña, J.M., Jayamohan, J., Jérusalem, A., 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol. Lett. 14, 20170660 https://doi.org/10.1098/rsbl.2017.0660.

Barbaro, A., Einarsson, B., Birnir, B., Sigurðsson, S., Valdimarsson, H., Pálsson, O., Sveinbjörnsson, S., Sigurðsson, T., 2009. Modelling and simulations of the migration of pelagic fish. ICES J. Mar. Sci. 66, 826–838. https://doi.org/10.1093/icesjms/fsp067.

Bez, N., Walker, E., Gaertner, D., Rivoirard, J., Gaspar, P., 2011. Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data. Can. J. Fish. Aquat. Sci. 68, 1998–2010. https://doi.org/10.1139/f2011-114.

Bishop, C.H., Etherton, B.J., Majumdar, S.J., 2001. Adaptive sampling with the ensemble transform Kalman Filter. Part I: theoretical aspects. Mon. Weather Rev. 129, 420–436. https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Buchwald, A.G., Adams, J., Bortz, D.M., Carlton, E.J., 2020. Infectious disease transmission models to predict, evaluate, and improve understanding of COVID-19 trajectory and interventions. Ann. ATS 17, 1204–1206. https://doi.org/10.1513/AnnalsATS.202005-501PS.

Burgers, G., Jan van Leeuwen, P., Evensen, G., 1998. Analysis scheme in the ensemble Kalman filter. Mon. Wea. Rev. 126, 1719–1724. https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Chu, D., 2011. Technology evolution and advances in fisheries acoustics. J. Mar. Sci. Technol. 19 https://doi.org/10.51400/2709-6998.2188.

Cocucci, T.J., Pulido, M., Aparicio, J.P., Ruíz, J., Simoy, M.I., Rosa, S., 2022. Inference in epidemiological agent-based models using ensemble-based data assimilation. PLOS ONE 17, e0264892. https://doi.org/10.1371/journal.pone.0264892.

DeAngelis, D.L., Grimm, V., 2014. Individual-based models in ecology after four decades. F1000Prime Rep. 6 https://doi.org/10.12703/P6-39. ⟨https://facultyopinions.com/prime/reports/b/6/39/⟩.

Eidsvik, J., Martino, S., Rue, H., 2008. Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models. Scand. J. Stat. https://doi.org/10.1111/j.1467-9469.2008.00621.x.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572.

Evensen, G., 2009. Data Assimilation: The Ensemble Kalman Filter. Springer.

Fu, W., She, J., Zhuang, S., 2011. Application of an ensemble optimal interpolation in a north/baltic sea model: Assimilating temperature and salinity profiles. Ocean Model. 40, 227–245. https://doi.org/10.1016/j.ocemod.2011.09.004.

Grimm, V., Railsback, S.F., 2005. Individual-based Modeling and Ecology. stu - student edition ed. Princeton University Press, pp. 3–21 (pp.). ⟨http://www.jstor.org/stable/j.ctt5hhnk8.5⟩ (pp.).

Houtekamer, P., Mitchell, H.L., 2005. Ensemble Kalman filtering. Q. J. R. Meteorol. Soc. 131, 3269–3289. https://doi.org/10.1256/qj.05.135.

Houtekamer, P.L., Mitchell, H.L., 2001. 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.

Katzfuss, M., Stroud, J.R., Wikle, C.K., 2016. Understanding the ensemble Kalman Filter. Am. Stat. 70, 350–357. https://doi.org/10.1080/00031305.2016.1141709.

Kelly, C., Michelsen, F.A., Kolding, J., Alver, M.O., 2022. Tuning and development of an individual-based Model of the herring spawning migration. Front. Mar. Sci. 8, 754476 https://doi.org/10.3389/fmars.2021.754476.

Keppenne, C.L., Rienecker, M.M., Jacob, J.P., Kovach, R., 2008. Error covariance modeling in the GMAO ocean ensemble Kalman filter. Mon. Weather Rev. 136, 2964–2982. https://doi.org/10.1175/2007MWR2243.1.

Kieu, L.M., Malleson, N., Heppenstall, A., 2020. Dealing with uncertainty in agent-based models for short-term predictions. R. Soc. Open Sci. 7, 191074 https://doi.org/10.1098/rsos.191074.

Koopman, J.S., Lynch, J.W., 1999. Individual causal models and population system models in epidemiology. Am. J. Public Health 89, 1170–1174. https://doi.org/10.2105/AJPH.89.8.1170.

Mandel, J., 2006.Efficient implementation of the ensemble kalman filter. Center for Computational Mathematics Reports.

Niu, S., Luo, Y., Dietze, M.C., Keenan, T.F., Shi, Z., Li, J., Iii, F.S.C., 2014. The role of data assimilation in predictive ecology. art65 Ecosphere 5. https://doi.org/10.1890/ES13-00273.1.

Simon, E., Bertino, L., 2009. Application of the gaussian anamorphosis to assimilation in a 3-d coupled physical-ecosystem model of the north atlantic with the enkf: a twin experiment. Ocean Sci. 5, 495–510. https://doi.org/10.5194/os-5-495-2009. ⟨https://os.copernicus.org/articles/5/495/2009/⟩.

Slagstad, D., McClimans, T.A., 2005. Modeling the ecosystem dynamics of the barents sea including the marginal ice zone: I. physical and chemical oceanography. J. Mar. Syst. 58, 1–18. https://doi.org/10.1016/j.jmarsys.2005.05.005. ⟨https://www.sciencedirect.com/science/article/pii/S0924796305001296⟩.

Slotte, A., Fiksen, Ø., 2000. State-dependent spawning migration in Norwegian spring-spawning herring. J. Fish. Biol. 56, 138–162. https://doi.org/10.1111/j.1095-8649.2000.tb02091.x.

de Souza, E.N., Boerder, K., Matwin, S., Worm, B., 2016. Improving fishing pattern detection from satellite AIS using data mining and machine learning. PLoS ONE 11, e0158248. https://doi.org/10.1371/journal.pone.0158248.

Tu, C.Y., Tseng, Y.H., Chiu, T.S., Shen, M.L., Hsieh, C.H., 2012. Using coupled fish behavior-hydrodynamic model to investigate spawning migration of Japanese anchovy, Engraulis japonicus, from the East China Sea to Taiwan: Spawning migration model of Japanese anchovy. Fish. Oceanogr. 21, 255–268. https://doi.org/10.1111/j.1365-2419.2012.00619.x.

Ward, J.A., Evans, A.J., Malleson, N.S., 2016. Dynamic calibration of agent-based models using data assimilation. R. Soc. Open Sci. 3, 150703 https://doi.org/10.1098/rsos.150703.

Woillez, M., Poulard, J.C., Rivoirard, J., Petitgas, P., Bez, N., 2007. Indices for capturing spatial patterns and their evolution in time, with application to European hake (Merluccius merluccius) in the Bay of Biscay. ICES J. Mar. Sci. 64, 537–550. https://doi.org/10.1093/icesjms/fsm025.