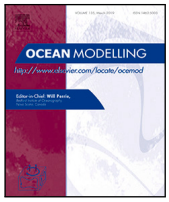
Ocean Modelling 186 (2023) 102278

Contents lists available at ScienceDirect

Ocean Modelling

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



Short communication

On a probabilistic evolutionary approach to ocean modelling: From Lorenz-63 to idealized ocean models

Igor Shevchenko [a,*], Pavel Berloff [a,b]

[a] Department of Mathematics, Imperial College London, Huxley Building, 180 Queen's Gate, London, SW7 2AZ, UK
[b] Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina 8, 119333, Moscow, Russia

ARTICLE INFO

Keywords:
Probabilistic evolutionary approach
Joint probability distribution
Probabilistic nudging
Eddy parameterization problem
Lorenz 63
Two-layer quasi-geostrophic model

ABSTRACT

In this study we develop an alternative way to model the ocean reflecting the chaotic nature of ocean flows and uncertainty of ocean models — instead of making use of classical deterministic or stochastic differential equations we offer a probabilistic evolutionary approach (PEA) that capitalizes on the use of probabilistic dynamics in phase space. The main feature of the data-driven version of PEA proposed in this work is that it does not require to know the physics behind the flow dynamics to model it. Within the PEA framework we develop two probabilistic evolutionary methods, which are based on probabilistic evolutionary models using quasi time-invariant structures in phase space.

The methods have been tested on complete and incomplete reference data sets generated by the Lorenz 63 system and by an idealized two-layer quasi-geostrophic model. The results show that both methods reproduce large- and small-scale features of the reference flow by keeping the probabilistic dynamics within the phase space of the reference flow. The proposed approach offers appealing benefits and a great flexibility to ocean modellers working with mathematical models and measurements. The most remarkable one is that it provides an alternative to the mainstream ocean parameterizations, requires no modification of existing ocean models, and is easy to implement. Moreover, it does not depend on the nature of input data, and therefore could work with both numerically-computed flows and real measurements from different sources (drifters, weather stations, etc.).

1. Introduction

The modern ocean modelling utilizes a wide spectrum of tools ranging from observations to using comprehensive ocean models (Ocean General Circulation Models). Most of the latter are based on deterministic or stochastic differential equations, and use both the physics- and data-driven paradigms, some rely on the statistical modelling (e.g., Storch and Zwiers, 2002; Vanem et al., 2022 and references there in). The majority operate in physical space (e.g., Marshall et al., 1997; Chassignet et al., 2007; Danilov et al., 2017; Madec and NEMO System Team, 2022), while some, umbrellaed under the recently proposed hyper-parameterization approach (e.g., Shevchenko and Berloff, 2021, 2022a,b, 2023), take advantage of working in phase space. In this study we develop an alternative way to model the ocean reflecting the chaotic nature of ocean flows and uncertainty of ocean models — instead of making use of classical deterministic or stochastic differential equations we offer a probabilistic evolutionary approach (PEA) that capitalizes on the use of probabilistic dynamics in phase space. In this study we develop a data-driven version of PEA, where the main feature is that it does not require to know the physics behind the flow dynamics to model it. It is achieved by its data-driven nature and by shifting the focus from the physical to the reference phase space (the phase space of the reference flow). The reference flow can be a numerical solution (generated by an ocean model), observational data from different sources (drifters, weather stations, etc.), or a combination of both. Within the PEA framework we develop two probabilistic evolutionary methods which are based on probabilistic evolutionary models.

The PEA offers appealing benefits and a great flexibility to ocean modellers working with mathematical models and measurements: (1) it requires no modification of existing ocean models, (2) is easy to implement, and (3) does not depend on the nature of input data. Most remarkably, the PEA provides an alternative to the mainstream ocean parameterizations. Namely, instead of running long high-resolution simulations of ocean models one can generate its relatively short coarse-grained version that retains flows that are nominally-resolved on the coarse grid (referred to as "nominally-resolved on the grid flows" below) and use it as input data for the PEA, thus replacing

* Corresponding author.
E-mail address: i.shevchenko@imperial.ac.uk (I. Shevchenko).

https://doi.org/10.1016/j.ocemod.2023.102278
Received 17 February 2023; Received in revised form 12 October 2023; Accepted 13 October 2023
Available online 20 October 2023

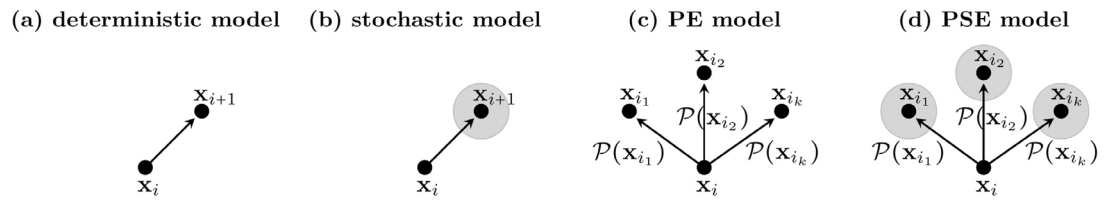


Fig. 1. Shown is a change of state in (a) deterministic models, (b) stochastic models, (c) probabilistic evolutionary (PE) models studied in this work, (d) probabilistic-stochastic evolutionary (PSE) models left for future studies; state and their neighbourhoods are denoted by black dots and grey discs, function $\mathcal{P}(\mathbf{x}_{i_k})$ is a transition probability function for state $\mathbf{x}_{i_k}$.

computationally-intensive ocean models with a way faster probabilistic evolutionary model. In other words, the PEA addresses the eddy-parameterization problem from a different angle. Namely, the PEA (1) shifts the focus from the physical space to the phase space of the model; and (2) considers the inability of the low-resolution model to reproduce the nominally-resolved flow structures as the persistent tendency of the phase space trajectory to escape the reference phase space. Note that the phase space trajectory represents the low-resolution solution and the reference phase space is the phase space of the reference solution.

First, we explain the probabilistic evolutionary approach and how to build probabilistic evolutionary models, and show how they work on the example of the Lorenz 63 system with complete and incomplete data sets. Then, we apply them to two-layer quasi-geostrophic (QG) flows with complete and incomplete reference data. It is worth mentioning that this work is intended as a proof of concept, therefore we deliberately reduce the technicalities beyond the PEA to a bare minimum, while focusing the attention on the key points of the approach.

2. The probabilistic evolutionary approach (PEA)

The probabilistic evolutionary approach is a new approach to ocean modelling that capitalizes on the chaotic nature of ocean dynamics by taking advantage of using the probability distribution of states in the reference phase space as opposed to making use of deterministic or stochastic differential equations. By construction, fluid dynamics models can be divided into two classes: deterministic and stochastic. In deterministic models, the dynamics is determined by a set of deterministic differential equations, i.e. the transition from a state $\mathbf{x}_i$ to a state $\mathbf{x}_{i+1}$ is unique (Fig. 1a). Stochastic models expands the class of deterministic models by introducing noise that does not lead to the unique state $\mathbf{x}_{i+1}$ but to a cloud of possible states neighbouring state $\mathbf{x}_{i+1}$. In other words, stochastic models describe the transit from a state $\mathbf{x}_i$ to a neighbourhood of state $\mathbf{x}_{i+1}$ (Fig. 1b), where the size of the neighbourhood depends on the noise statistics and the way it is included in the model (additively, multiplicatively, or otherwise).

The probabilistic evolutionary approach (proposed in this study) offers a different point of view: infinitely many states $(x_{i_1}, x_{i_2}, \ldots)$ can be reached from a state $\mathbf{x}_i$, and the transition to a particular state is defined by a transition probability function, $\mathcal{P}$, which assigns a probability to any possible transition (Fig. 1c). Note that probabilistic evolutionary models do not describe the evolution of a probability function, the evolution in these models is governed by a probability function. In the data-driven version of PEA proposed in this study, the transition probability function is calculated from available reference data as the current state changes (i.e., on the fly). In the physics-driven PEA, we envisage that the transition probability function can be defined from the physics of the studied phenomenon. Another class of models that naturally follows from the stochastic and probabilistic ones is the probabilistic-stochastic evolutionary models in which every possible state is replaced with a cloud of states neighbouring it (Fig. 1d). This gives rise to probabilistic evolutionary models with noise which are beyond the scope of this work.

In this study we are focused on the data-driven PEA, where the transition probability function is calculated locally from available reference data for every transition from one state to another, i.e. a new state of the probabilistic flow evolution is defined by the likelihood of reference states neighbouring to the current state of the probabilistic evolutionary model. Within the PEA framework, the probabilistic nature of the flow evolution implies that even very unlikely (rare) events are expected to occur once in a while thus echoing extreme weather and climate events. More importantly, it allows the probabilistic trajectory to cover regions of the reference phase space that are not presented in the reference data set, but can potentially happen. By an extreme event (state) we mean an event for which there exists a highly unlikely path in the phase space that moves the current state towards this extreme event. Extreme events per se are not supposed to be sampled. However, the highly unlikely path (or paths) that leads to extreme events needs to be sampled, otherwise extremes can probably not be found. More accurately, the reference data should include a set of states, sampling from which can give this path (or paths) with a non-zero probability.

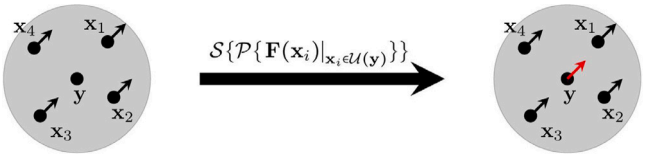


Fig. 2. Schematic of the data-driven PEA. A state $\mathbf{y}$ of the probabilistic evolutionary model (2) neighbouring to the reference states $\mathbf{x}_i$ equipped with directional vectors $\mathbf{F}(\mathbf{x}_i)$, $i = 1, 2, 3, 4$, denoted by black arrows. The neighbourhood $\mathcal{U}(\mathbf{y})$ of $\mathbf{y}$ is denoted by the grey disc, and $\mathcal{S}$ is an operator sampling from the transition probability function $\mathcal{P}$. The new vector $\mathbf{G}(\mathbf{y})$ is denoted by a red arrow.

We would like to remind the reader that a system of ordinary differential equations

$$\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n \tag{1}$$

can be geometrically interpreted as a vector field in the phase space of Eq. (1); here, the prime denotes a time derivative. The direction of the vector field at a given point $\mathbf{x}$ is determined by the vector $\mathbf{F}(\mathbf{x})$ for $\forall \mathbf{x} \in \mathbb{R}^n$. Once $\mathbf{F}(\mathbf{x})$ is known, it can be used to calculate a new position of point $\mathbf{x}$ in the phase space. This idea is used in the hyper-parameterization method "Advection of the image point" (Shevchenko and Berloff, 2021, 2023).

The PEA works in a different way. In the data-driven version of PEA the analytical form of Eq. (1) is not available. The only reference data available to the PEA is a numerical solution of (1), i.e. the reference solution, or observations if one works with data from weather stations, satellites, etc. Therefore, instead of directly calculating $\mathbf{F}(\mathbf{x})$ at a give point $\mathbf{x}$, the PEA computes a new vector, $\mathbf{G}(\mathbf{y})$, by sampling from the transition probability function, $\mathcal{P}$, based on the joint probability distribution of states neighbouring to the current state in the reference phase space (Fig. 2). Further below, $\mathbf{x}(t)$ is a reference state at time t, while $\mathbf{y}(t)$ is a state of the probabilistic model.

Thus, the probabilistic evolutionary model can be written as follows:

$$\mathbf{y}'(t) = \mathcal{S}\{\mathcal{P}\{\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t) \in \mathcal{U}(\mathbf{y}(t))}\}\}, \quad \mathbf{y}(t_0) = \mathbf{x}(t_0), \tag{2}$$

with $\mathcal{S}$ being an operator sampling from the joint probability distribution, i.e. it returns a point in phase space given a probability.

We have developed two methods within the PEA framework. These methods use different probability distributions to calculate the directional vector $\mathbf{G}(\mathbf{y})$ for the probabilistic evolutionary model. The first method calculates directional angles and the length of $\mathbf{G}(\mathbf{y})$ based on its joint probability distribution function computed from available reference data. Hence the probabilistic evolutionary model for the probabilistic solution $\mathbf{y}(t)$ is given by

$$\mathbf{y}'(t) = \mathbf{G}(\mathbf{y}) + \mathcal{N}(\mathbf{x}(t), \mathbf{y}(t)),$$
$$\mathbf{G}(\mathbf{y}) := C_S^{-1}\{\mathcal{S}\{\mathcal{P}_a\{C_S\{\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}\}\}\}\}, \quad \mathbf{y}(t_0) = \mathbf{x}(t_0), \tag{3}$$

where $\mathcal{U}(\mathbf{y}(t))$ is the neighbourhood of probabilistic solution $\mathbf{y}(t)$, C_S is the transformation from Cartesian to spherical coordinates (used to compute the angles and lengths of reference vectors), $\mathcal{P}_a$ is a transition probability function based on the joint probability distribution of directional angles and lengths of the reference vectors neighbouring to the current state $\mathbf{y}(t)$ in the phase space of (1), i.e. the vectors $\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}$; C_S^{-1} is the inverse of C_S used to compute $\mathbf{G}(\mathbf{y})$ in the Cartesian space. The second term on the right hand side of Eq. (3) is a nudging term:

$$\mathcal{N}(\mathbf{x}(t), \mathbf{y}(t)) := \eta\left(\frac{1}{M}\sum_{i\in\mathcal{U}(\mathbf{y}(t))}\mathbf{x}(t_i) - \mathbf{y}(t)\right), \tag{4}$$

where η is a nudging strength, M is the number of nearest (in l_2 norm) to the solution $\mathbf{y}(t)$ points over which the averaged reference solution $\mathbf{x}(t)$ is computed.

The second method does not use the joint probability distribution of the directional angles and lengths of reference vectors to compute $\mathbf{G}(\mathbf{y})$. Instead, it computes $\mathbf{G}(\mathbf{y})$ from the joint probability distribution of the coordinates of reference vectors neighbouring to the current state $\mathbf{y}(t)$ in the phase space of (1). Hence, the probabilistic evolutionary equation reads as follows:

$$\mathbf{y}'(t) = \mathbf{G}(\mathbf{y}) + \mathcal{N}(\mathbf{x}(t), \mathbf{y}(t)), \quad \mathbf{G}(\mathbf{y}) := \mathcal{S}\{\mathcal{P}_c\{\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}\}\},$$
$$\mathbf{y}(t_0) = \mathbf{x}(t_0), \tag{5}$$

where $\mathcal{P}_c$ is a transition probability function based on the joint probability distribution of the coordinates of the reference vectors neighbouring to the current state $\mathbf{y}(t)$ in the phase space of (1), i.e. the vectors $\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}$.

Calculation of the transition probability function. The transition probability function for the probabilistic models is based on the joint probability distribution (JPD) of states neighbouring to the current state in the reference phase space. The JPD is calculated with the histogram method from the "raw" reference data, i.e. only the solution and its tendencies are used; the tendencies are pre-computed with the central finite difference in time before the model run. We do not use model tendencies computed directly in the model, as they are not always available; moreover, they are hardly available when working with observations. The JPD is calculated at every integration step of the model, as the information is scooped from the neighbourhood of the probabilistic solution. Also note that the only difference between the models is in the reference coordinates used for the relevant neighbours for the joint probability distribution.

As the proposed approach is developed to work in high-dimensional phase spaces, computing a multidimensional JPD and keeping it in memory for further sampling is an unaffordable option. Instead, we calculate a coordinate-wise PDF of the tendencies. That gives us access to all necessary information for sampling.

The schematic of the JPD calculation for the second method is presented in Fig. 3. We deliberately exemplify it for the second method to avoid unnecessary and irrelevant (to this illustrative example) calculations of C_S and C_S^{-1}.

The classical approach to the JPD calculation (or more accurately its approximation) with the histogram method is to divide the domain (the neighbourhood $\mathcal{U}(\mathbf{y})$ in our case) into bins (boxes) and count the number of points (black dots in Fig. 3a) in every bin to compute the height of each column of the JPD (Fig. 3e). This approach is, however, only applicable when the number of dimensions is relatively low, as the space complexity (the total amount of memory used) grows as the number of bins to the power of dimensions. For example, the JPD in Fig. 3e has 20 bins in each dimension thus giving 400 in total. It does not look like much, but in spaces of tens of thousands of dimensions (which are typical for idealized ocean models not to mention those based on the primitive equations) this number rules the method out of use.

In order to use the histogram method in multi-dimensional spaces, we do not calculate the whole JPD, instead we compute the PDF for every component of the tendency $\mathbf{F}$ (Fig. 3f,g) and then use these PDFs to sample from the JPD. It is also worth noting that the histograms in Fig. 3e–g can be interpolated if there are not enough points in the neighbourhood to properly represent the JPD and individual PDFs.

Sampling from the JPD. In order to sample from the JPDs $\mathcal{P}_a$ and $\mathcal{P}_c$, we use the sampling operator $\mathcal{S}$ based on the inverse transform sampling method (Devroye, 1986); we also tried the rejection sampling but did not observe that much of a difference. However, we do not use the classical form of the inverse transform sampling method to sample directly from the JPD, as computing the JPD and its multivariate cumulative distribution function (CDF), which is needed for sampling, is too computationally intensive. Instead, we compute coordinate-wise PDFs and its CDFs for every component of the tendency $\mathbf{F}$ within the neighbourhood $\mathcal{U}(\mathbf{y})$.

To get more insights into the sampling procedure, we illustrate how it works on the two-dimensional case considered above; recall, we use the second method for this purpose. The sampling procedure starts with computing the CDF of F_1 (Fig. 4b), we use the PDF of F_1 for that (Fig. 4a). Then, we draw a random number, r_1, from a uniform distribution in the unit interval denoted as $\mathrm{U}[0,1]$ (it is the vertical coordinate in Fig. 4b) and find the corresponding horizontal coordinate $F_1(r_1) =: G_1$; the map $\mathcal{S}$ in Eq. (5) does that, i.e. $\mathcal{S} : r_1 \to F_1(r_1)$. It gives the first component of the new tendency G_1. In the next step we calculate the PDF of F_2 (Fig. 4c) given $F_1(r_1)$ and then compute its CDF (Fig. 4d). We draw another random number $r_2 \in \mathrm{U}[0,1]$ and compute $F_2(r_2) =: G_2$; it is the action of $\mathcal{S}$ onto r_2, i.e. $\mathcal{S} : r_2 \to F_2(r_2)$. It gives the second component of the new tendency G_2. Having two components of the new tendency $\mathbf{G}(\mathbf{y}) = (G_1, G_2)$, we plug it into Eq. (5) and integrate it over one time step. It completes the sampling procedure.

Probabilistic nudging. The form of the nudging term used in the probabilistic evolutionary models (3) and (5) is governed by our desire to keep the probabilistic evolutionary model as simple as possible. Different metrics and forms of the nudging term can be used instead (for example, adaptive Shevchenko and Berloff, 2022b or probabilistic nudging).

The idea behind probabilistic nudging is also based on using the probabilistic evolutionary machinery. However, instead of computing the joint probability distribution of the vectors $\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}$ as above, we compute the joint probability distribution of the states $\mathbf{x}(t)$ themselves. Thus, the probabilistic nudging term can be written as

$$\mathcal{N}(\mathbf{x}(t), \mathbf{y}(t)) := \eta\left(\mathcal{P}_c\{\mathbf{x}(t)\}|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))} - \mathbf{y}(t)\right), \tag{6}$$

where $\mathcal{P}_c$ is a probability function based on the joint probability distribution of the coordinates of the reference states $\mathbf{x}(t)$ neighbouring to the current state $\mathbf{y}(t)$ in the phase space of (1), i.e. the states $\mathbf{x}(t)|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y}(t))}$. We do not study how the probabilistic nudging performs in this work and leave it for future research.

On the optimal choice of parameters. Note that the neighbourhood $\mathcal{U}(\mathbf{y}(t))$ in Eqs. (3) and (5) is computed as N (and M for the nudging term) nearest (in l_2 norm) to the solution $\mathbf{y}(t)$ points. The neighbourhood can be computed differently, and the way it is computed affects the solution. The optimal value of N and M (as well as η)

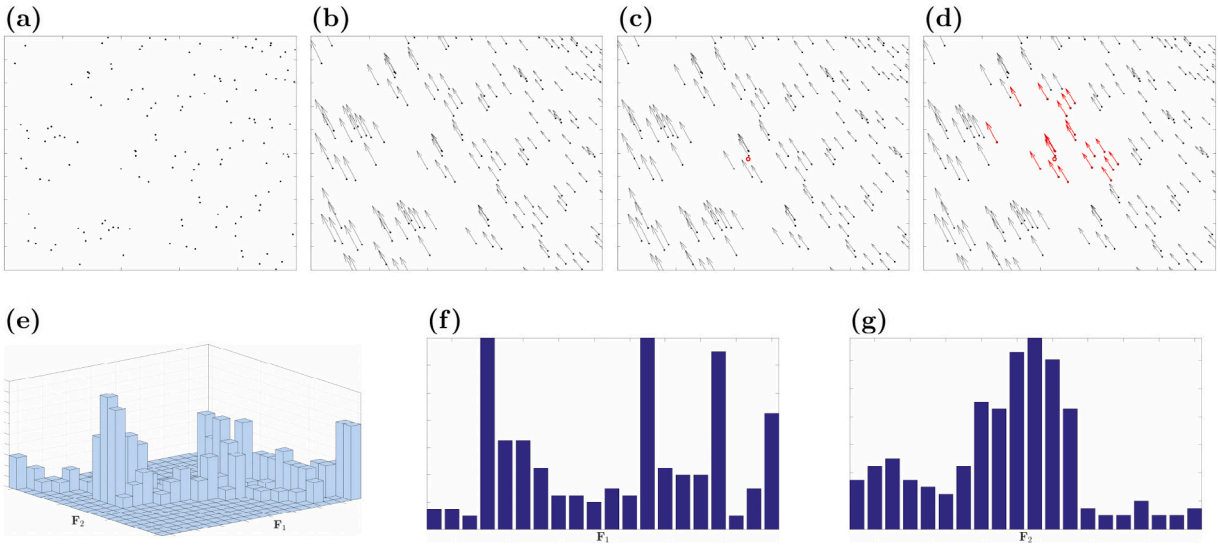


Fig. 3. Schematic of the JPD calculation. Given the reference data $\mathbf{x}_i$, $i = 1, \dots, n$ (black dots in (a)), we calculate its tendencies with the central finite difference in time (black vectors in (b)); the reference data can be a numerical solution (generated by an ocean model), observational data from different sources (drifters, weather stations, satellites, etc.), or a combination of both. Given a point, say $\mathbf{y}$, (red circle in (c)) we find N nearest (in l_2 norm) points $\mathbf{x}_j$ to $\mathbf{y}$ and their tendencies $\mathbf{F}(\mathbf{x}_j)$, $j = j_1, \dots, j_N$ (red vectors in (d)), it gives the neighbourhood $\mathbf{F}(\mathbf{x}_j)\big|_{\mathbf{x}_j \in \mathcal{U}(\mathbf{y})}$. Given the tendencies $\mathbf{F}(\mathbf{x}_j)$, $j = j_1, \dots, j_N$ we calculate the JPD (or more accurately, an approximation to the JPD) as a histogram (e); this JPD represents the term $\mathcal{P}_c\{\,\mathbf{F}(\mathbf{x}(t))|_{\mathbf{x}(t)\in\mathcal{U}(\mathbf{y})}\}$ in Eq. (5). We compute the JPD in the coordinate-wise manner by using PDFs for every component of the tendency $\mathbf{F} = (F_1, F_2)$ (see (f) and (g)). Note that we denote axes only when it is relevant for the schematic.

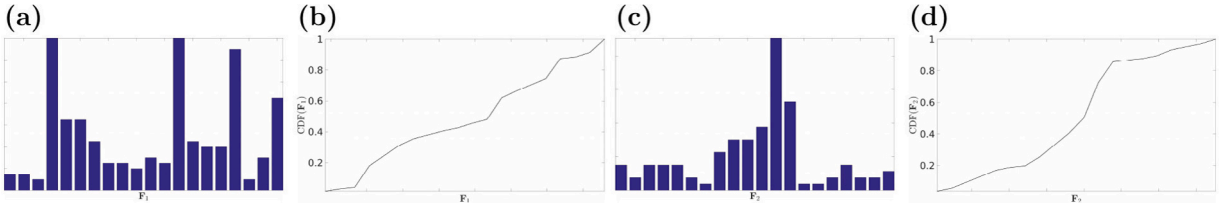


Fig. 4. Schematic of sampling from the JPD. Given the PDF of $\mathbf{F}_1$ (a), we compute its CDF (b), draw a random number $r_1 \in \mathrm{U}[0,1]$, and find $\mathbf{F}_1(r_1)$ — the first component of the new tendency G_1. Then, we compute the PDF of $\mathbf{F}_2$ given $\mathbf{F}_1(r_1)$ (c) and its CDF (d), draw another random number $r_2 \in \mathrm{U}[0,1]$ and find $\mathbf{F}_2(r_2)$ — the second component of the new tendency G_2.

for a given reference solution can be computed by solving the following optimization problem

$$\min_{N,M,\eta} \mathcal{F}(\mathbf{x}(t), \mathbf{y}(t)), \quad t \in [0, T] \tag{7}$$

where $\mathcal{F}$ is a problem-specific function, and T is the length of the reference solution $\mathbf{x}(t)$. For example, $\mathcal{F}$ can be defined as a norm of the difference between the reference and probabilistic solutions. Our choice of N, M, and η is driven by our measure of goodness (to keep the probabilistic solution $\mathbf{y}(t)$ within the reference phase space). This measure is used because it allows the probabilistic solution to evolve in the neighbourhood of the reference phase space, since the failure to do so results in a wrong flow dynamics typically shown by low-resolution ocean models. Studying optimal strategies of computing the neighbourhood and its size as well as the nudging strength η is a topic beyond the scope of the present paper.

Conservation laws. In order to address conservation laws within the PEA context, let us consider a quantity ϕ in a domain Ω. The conservation of ϕ in Ω is given by

$$(I_\phi := \int_\Omega \phi \, d\Omega = const\,. \tag{8}$$

To check whether ϕ is conserved along the probabilistic trajectory let us focus on the right hand side of the probabilistic model. It consists of two terms: the directional vector $\mathbf{G}$ and the nudging term $\mathcal{N}$. First we note that nudging defined by a linear operator, as in (4), does not affect the conservation of ϕ given the integration starts from a ϕ_0 for which $I_{\phi_0} = 0$. On the other note, the probabilistic nudging (6) and the directional vector $\mathbf{G}$ can potentially compromise the conservation law, and a special treatment might be needed in this case, depending on how $\mathbf{G}$ and $\mathcal{P}_c$ (in the nudging term) are calculated; brute forcing (8) is always an option. The coordinate-wise calculation of the transition probability function in $\mathbf{G}$, as in our case, is likely to lower the accuracy (compared to the reference ϕ) with which (8) holds. The vector-wise calculation of the transition probability function is less error-prone and therefore can be an alternative to the brute force approach when a higher accuracy is needed. We will return to the conservation laws in the context of quasi-geostrophic dynamics discussed later. We would also like to note that the hyper-parameterization method "Advection of the image point" (Shevchenko and Berloff, 2021, 2023) preserves ϕ with the reference accuracy, as the right-hand side of the equation used to compute $\mathbf{y}$ is defined by an average of directional vectors and the nudging term is given by (4).

As an example, we consider the Lorenz 63 system (Lorenz, 1963):

$$\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t)), \quad \mathbf{F} := \begin{pmatrix} \sigma(y - x) \\ x(\rho - z) - y \\ xy - \beta z \end{pmatrix}, \tag{9}$$

with $\mathbf{x}(t) = (x(t), y(t), z(t))$, and $\sigma = 10$, $\beta = 8/3$, $\rho = 28$. As an initial condition, we take $\mathbf{x}(t_0) = (-8.6, -12.4, 21.0)$ to make sure the solution is close to the Lorenz attractor (Fig. 5a). Along with the solution of the Lorenz system, we compute the probabilistic solutions to Eqs. (3) and (5) with $N = 10$ (Fig. 5b,c); we have also tested the probabilistic evolutionary methods for $N = 5$ and the results are qualitatively the same (not shown). Note that all probabilistic solutions are computed

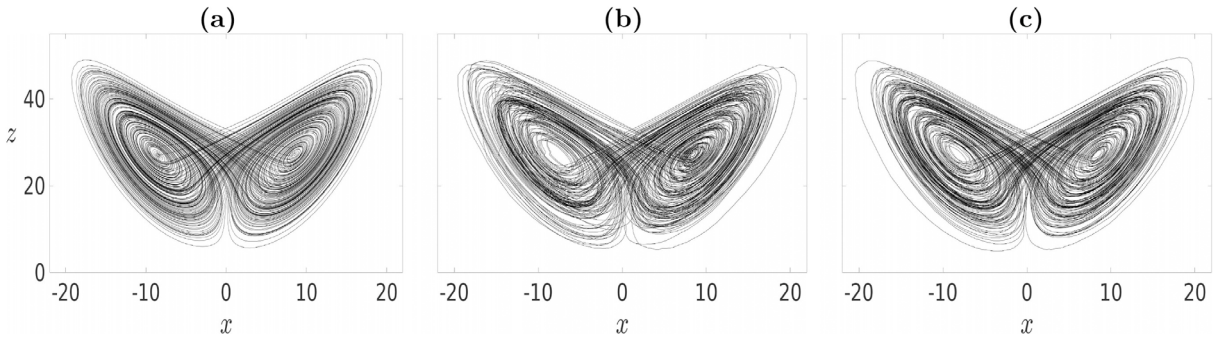


Fig. 5. Shown is (a) the solution of the Lorenz system (9) for the time interval $t \in [0, 200]$, (b) and (c) the probabilistic solutions of (3) and (5), respectively. The probabilistic solutions use only the first half of the reference data (i.e., $t \in [0, 100]$), and over the second half the probabilistic evolutionary methods work out of the sample. Both probabilistic solutions stay in the reference phase space, and reproduce the Lorenz attractor.

without nudging (i.e., for $\eta = 0$), as it is not required to reproduce the Lorenz attractor; however, nudging will play an essential role in simulations of the QG model discussed in Section 3.

As seen in Fig. 5b,c, the probabilistic solution stays within the same region of the phase space as the reference solution of the Lorenz system, despite that only the first half of the reference solution is available (i.e., $t \in [0, 100]$), and the method runs out of the sample over the second half, i.e. for $t \in [100, 200]$. This is important, as for the probabilistic evolutionary approach the measure of goodness is how close the probabilistic solution is to the reference phase space.

Incomplete reference data. As the proposed probabilistic evolutionary approach is intended to work with ocean models, it is instructive to test its methods on incomplete reference data sets which are not uncommon in ocean modelling, when using observational data for reanalysis in comprehensive ocean models. By incomplete data we mean incomplete in time, i.e. some tendencies and prognostic variables in the phase space are removed from the reference data set. We consider three test cases: (1) gappy dynamics, (2) holey attractor, and (3) disjoint wings.

Gappy dynamics. For this test we generate two data sets that are used as reference data for both methods. Namely, we take the Lorenz solution over the time period $[0, 100]$ and keep every second and every fourth point of the solution, thus retaining only 50% and 25% of the original reference data, respectively. We set $N = 10$ and present the results in Fig. 6; the results for $N = 5$ are qualitatively the same (not shown). As we see in Fig. 6, both methods keep the probabilistic solution on the Lorenz attractor, despite the reference solution missing a substantial portion of data.

Holey attractor. In this test case we cut out some regions of the reference dynamics by making three holes of radius 4 in the attractor itself (Fig. 7a). As in the first test, both probabilistic evolutionary methods (Fig. 7b,c) keep the solution on the attractor. More importantly, the methods restore the dynamics on the attractor as if there are no holes.

Disjoint wings. In this test we cut the attractor into two disjoint sets (Fig. 8a) thus simulating a substantially corrupted data set; the cut width is 2, which corresponds to a 13% loss of reference data. As seen in Fig. 8b,c, both probabilistic evolutionary methods not only recover the attractor but also restore the dynamics in between the wings where the reference solution is unavailable.

On the detrimental role of nudging. We did not use nudging in the probabilistic evolutionary methods to model the Lorenz system, as it is not necessary to reproduce the reference dynamics, i.e. the attractor. But, within the context of the QG model discussed below we will see the beneficial effect of nudging on the flow dynamics. However, it should be noted in advance that nudging can also play a detrimental role when using out of place. As an example, we take the Lorenz system and demonstrate how improper use of nudging can affect the solution. As seen in Fig. 9, the nudging strength has a substantial effect on the solution. Namely, for a relatively strong nudging the probabilistic solution cannot properly develop on the attractor, and the whole dynamics is confined in a narrow strip. It is a simple but illustrative example of how one should use caution to properly adjust the nudging strength to get a good solution.

Summing up our findings for the Lorenz 63 system, we conclude that the proposed probabilistic evolutionary methods have strong potential for modelling geophysical flows. It may seem that the first method is somewhat inferior to the second one, as the former gives the trajectories which do stay in the reference phase space, but can experience rapid changes of the direction that may lead to undesirable effects in geophysical flows; in fact, these changes can be mitigated by using a shorter integration step (not shown) for the reference and probabilistic solutions. Despite this seeming disadvantage we do not disregard the first method and test it too on the QG model, as this effect might be of minor or no influence within the context of QG dynamics.

3. Multilayer quasi-geostrophic equations

In this section we apply the probabilistic evolutionary methods to the 2-layer quasi-geostrophic (QG) model describing the evolution of potential vorticity (PV) anomaly $\mathbf{q} = (q_1, q_2)$ in a domain Ω (Pedlosky, 1987):

$$\begin{aligned} &\partial_t q_1 + \mathbf{u}_1 \cdot \nabla q_1 = \nu \nabla^4 \psi_1 - \beta \partial_x \psi_1, \\ &\partial_t q_2 + \mathbf{u}_2 \cdot \nabla q_2 = \nu \nabla^4 \psi_2 - \mu \nabla^2 \psi_2 - \beta \partial_x \psi_2, \end{aligned} \tag{10}$$

where $\mathbf{u} = (u, v)$ is a horizontal velocity vector, $\boldsymbol{\psi} = (\psi_1, \psi_2)$ is the stream function in the top and bottom layers, $\nu = 3.125\,\mathrm{m^2\ s^{-1}}$ is the lateral eddy viscosity, $\beta = 2 \times 10^{-11}\,\mathrm{m^{-1}\ s^{-1}}$ is the planetary vorticity gradient, and $\mu = 4 \times 10^{-8}\,\mathrm{s^{-1}}$ is the bottom friction parameter. The computational domain $\Omega = [0, L_x] \times [0, L_y] \times [0, H]$ is a horizontally periodic flat-bottom channel of depth $H = H_1 + H_2$ given by two stacked isopycnal fluid layers of depth $H_1 = 1.0\,\mathrm{km}$, $H_2 = 3.0\,\mathrm{km}$, and $L_x = 3840\,\mathrm{km}$, $L_y = L_x/2$.

Forcing in (10) is introduced via a vertically sheared, baroclinically unstable background flow (e.g., Cotter et al., 2020):

$$\psi_i \to -U_i\, y + \psi_i, \quad i = 1, 2, \tag{11}$$

with the background-flow zonal velocities $U_1 = 6.0$, $U_2 = 0.0\ \mathrm{cm\,s^{-1}}$.

The PV anomaly and stream function are related through the system of elliptic equations:

$$q_1 = \nabla^2 \psi_1 + s_1(\psi_2 - \psi_1), \tag{12a}$$

$$q_2 = \nabla^2 \psi_2 + s_2(\psi_1 - \psi_2), \tag{12b}$$

with the stratification parameters $s_1 = 4.22 \cdot 10^{-3}\,\mathrm{km^{-2}}$, $s_2 = 1.41 \cdot 10^{-3}\,\mathrm{km^{-2}}$; chosen so that the first Rossby deformation radius is $Rd_1 = 25\,\mathrm{km}$.

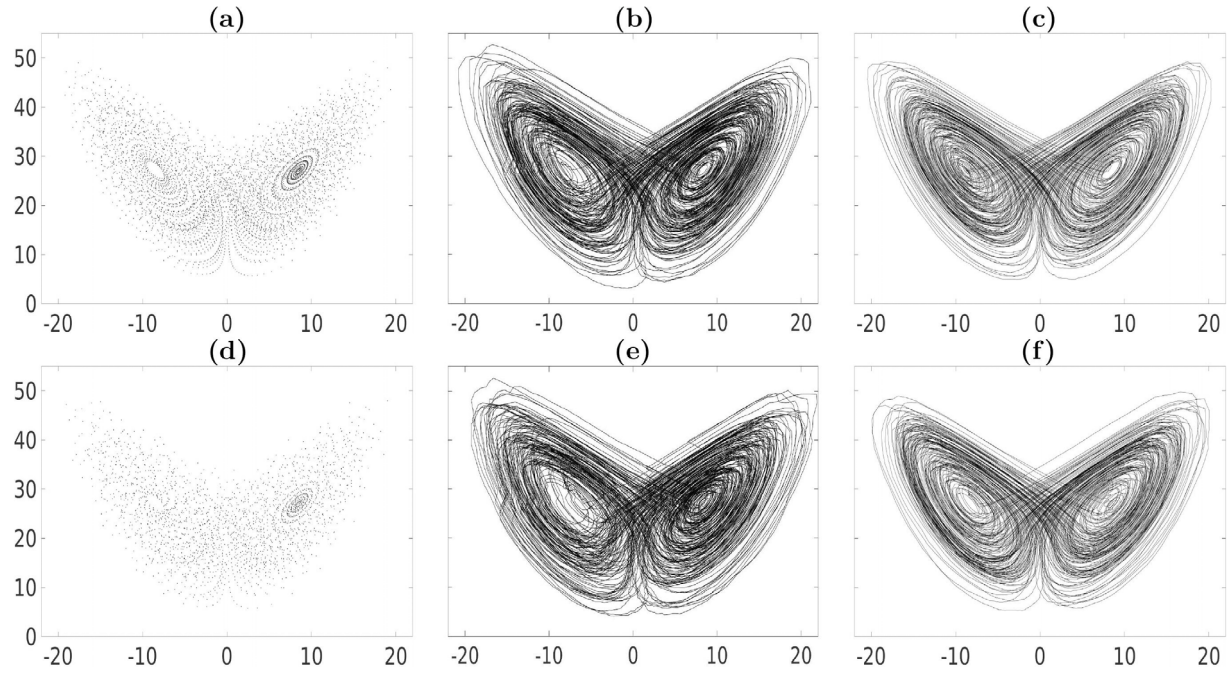


Fig. 6. Shown is (a) the reference solution (over $t \in [0, 100]$) with every second point retained, (b)/(c) the first/second probabilistic solution (computed with the first/second method) over $t \in [0, 200]$; (d)–(f) are the same as (a)–(c) but for the reference solution with every forth point retained. The axes are the same as in Fig. 5.

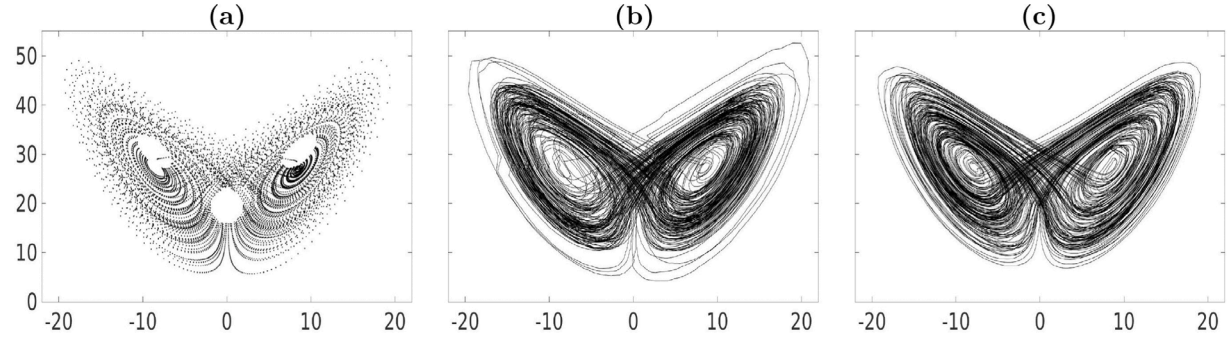


Fig. 7. Shown is (a) the reference solution (over $t \in [0, 100]$) with three holes of radius 4, (b)/(c) the first/second probabilistic solution over $t \in [0, 200]$. The axes are the same as in Fig. 5.

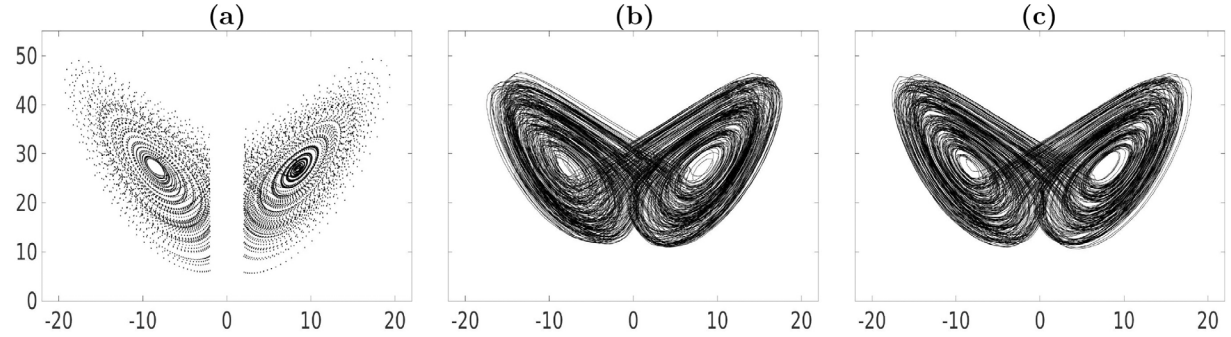


Fig. 8. Shown is (a) the reference solution (over $t \in [0, 100]$) with disjoint wings, (b)/(c) the first/second probabilistic solution over $t \in [0, 200]$. The axes are the same as in Fig. 5.

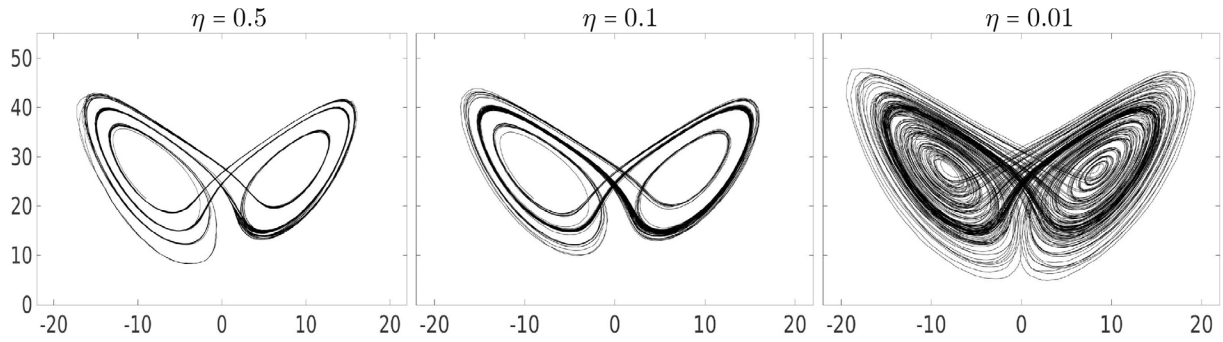


Fig. 9. Shown is the dependence of the Lorenz solution (computed with the second probabilistic evolutionary method) on the nudging strength η for $t \in [0, 200]$. The results for the first method are qualitatively similar (not shown).

System (10)–(12) is augmented by the integral mass conservation constraint (McWilliams, 1977):

$$\partial_t \iint_{\Omega} (\psi_1 - \psi_2)\, dydx = 0, \tag{13}$$

by the periodic horizontal boundary conditions set at eastern, Γ_2, and western, Γ_4, boundaries

$$\boldsymbol{\psi}\big|_{\Gamma_2} = \boldsymbol{\psi}\big|_{\Gamma_4}, \tag{14}$$

and no-slip boundary conditions

$$\boldsymbol{u}\big|_{\Gamma_1} = \boldsymbol{u}\big|_{\Gamma_3} = 0\,. \tag{15}$$

set at northern, Γ_1, and southern, Γ_3, boundaries of the domain Ω.

The QG equations (10) can be recast in the form of Eq. (1) as follows:

$$\mathbf{q}'(t) = \mathbf{F}(\mathbf{q}, \boldsymbol{\psi}, \mathbf{u}), \tag{16}$$

where the right hand side $\mathbf{F}$ defines the vector field used to evolve $\mathbf{q}$; $\mathbf{F}$ is computed with the central finite difference in time from the available reference data. The only difference with the Lorenz system (9) is the vector $\mathbf{F}$ and the advected quantity $\mathbf{q}$; note that the dimension of phase space is defined by the number of degrees of freedom used to discretize the equation in space. Thus, the analogue of Eqs. (3) and (5) for the QG equations reads:

$$\mathbf{y}'(t) = \mathbf{G}(\mathbf{y}) + \mathcal{N}(\mathbf{q}(t), \mathbf{y}(t)), \quad \mathbf{y}(t_0) = \mathbf{q}(t_0). \tag{17}$$

For the purpose of this study both high- and low-resolution solutions are needed. We compute these solutions on uniform grids of size 513×257 and 129×65 over a period of 4 years after a 10-year initial spin up; the resolution of these grids is 7.5 km and 30 km, respectively. In order to test the probabilistic evolutionary methods in different regimes, we use both the high-resolution solution and its point-to-point projection onto the coarse grid 129×65; the coarse-graining is of little importance to the probabilistic evolutionary approach, and any other method (e.g., interpolation schemes, spatial averaging, or filters) can be used. The high-resolution solution is needed to study to what extent the probabilistic evolutionary methods can be used as an alternative to high-resolution ocean simulations, whereas its low-resolution projection is to compare the performance of the methods with the low-resolution solution computed on the coarse grid with the QG model. We should also note that these solutions are denoted as $\mathbf{x}(t)$ in Eq. (1), while their corresponding probabilistic solutions are denoted as $\mathbf{y}(t)$ in Eqs. (3) and (5). It is an important comparison which will show whether the probabilistic evolutionary methods can reproduce flow features that are presented in the low-resolution projection but are missing in the low-resolution QG solution. For the purpose of this study, it is enough to consider the first layer PV anomaly, as it is much more energetic than the second layer and full of both large- and small-scale flow features.

In order to demonstrate the ability of the probabilistic evolutionary methods to reproduce nominally-resolved on the coarse grid flow features, we take only the first two years of the 4-year long high-resolution solution, coarse-grain it onto the grid of size 129×65 and then use it as a reference solution (Fig. 10a). As for the Lorenz 63, we firstly apply the methods without nudging.

The build-up effect. As seen in Fig. 10b,c, at the very beginning the probabilistic solution reproduces both large-scale flow structures (two zonally-elongated jets) as well as small-scale vortices and meanders along the jets of the reference solution (Fig. 10a). It is because the probabilistic solution remains in the reference phase space. But, after a short period of time the build-up effect takes over and makes the probabilistic solution drift away from the reference phase space, and eventually to settle to a virtually constant in time direction (i.e., the amplitude of the solution changes much faster compared to its structure); "virtually constant" comes from the fact that sampling gives some variance of the directional vector, but (depending on the JPD) this variance can be too small thus leading to negligible changes in the structure of the solution compared to its amplitude. After 40- and 60-day solutions already show almost no change in the structure. It takes only 60 days for the solution to "freeze" down (almost no structural changes) to a point of no use.

The lack of reference data for the PEA and/or bad choice of parameters (N, M, η) can steer the probabilistic trajectory away from the reference phase space. We refer to this as the build-up effect, meaning that after a period of time, say T, the neighbourhood of the nearest points stalls (the points in the neighbourhood become the closest ones to the point $\mathbf{y}(t)$ for $\forall t > T$); therefore, the same points are used again and again during the integration thus driving the probabilistic trajectory away from the reference phase space. In principle, building up numerical errors may terminate this drift, and the solution can return back to the reference phase space, but this is case dependent. However, if the return time is relatively long (longer than the characteristic time of the reference solution) then the flow dynamics can be seriously distorted over the period of the trajectory injection. For more details on the build-up effect we refer the reader to Shevchenko and Berloff (2023).

In order to avoid the build-up effect we use the nudging methodology. We set the nudging parameter as $\eta = 0.1$ (it is not the only choice) in Eqs. (3) and (5), and present the results in Fig. 11. As seen in Fig. 11b,c, both methods reproduce the large-scale reference flow structures (two zonally-elongated jets) as well as small-scale vortices and meanders along the jets of the reference solution (Fig. 11a). However, the coarse-grid QG model cannot reproduce the large-scale flow structures not to mention the small-scale structures (Fig. 11d).

It might seem counterintuitive that for the Lorenz model nudging leads to a collapse of the attractor, while for the QG model without nudging it results in a "frozen" state. The reason for this behaviour can be rooted in the structure of the phase space itself. If the phase space is dense and the trajectory wanders in a kind of cyclic way

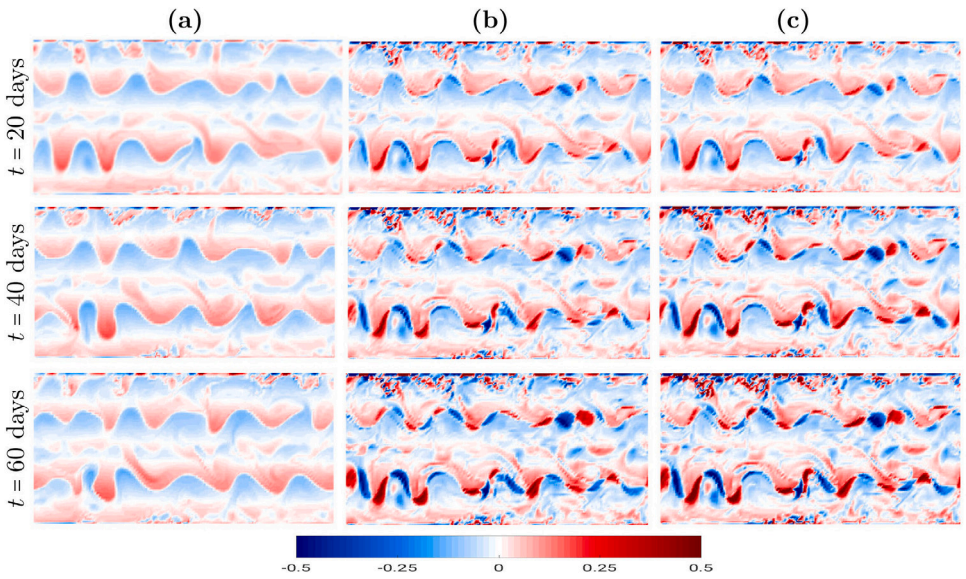


Fig. 10. Shown are snapshots of the top layer PV anomaly: (a) the reference solution (computed on the 513 × 257 grid and then projected onto the 129 × 65 grid), (b)/(c) probabilistic solution computed with the first/second method. All solutions are given in units of $[s^{-1} f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\,s^{-1}$ is the Coriolis parameter. Both methods reproduce nominally-resolved on the coarse grid flow features but over a short period the build-up effect kicks in and arrests the dynamics.

(like in the Lorenz case) then it is likely that a strong nudging will force the solution to evolve in a narrow band and therefore will not allow it to properly develop on the attractor. It happens because the neighbourhood is always formed of the points that lay in this narrow band. On the other note, if the vector field in the phase space can form directions leading the trajectory outside of the field (it happens because of lack of data) then after a period of time the neighbourhood of the nearest points can stall (i.e. the points in the neighbourhood become the closest ones to the image point for all time). Therefore, the same points are used again and again during the integration thus leading to a stagnation of the flow dynamics. It is what happens in the QG model. Stagnations can be detected and the phase space can be deformed to prevent them but this technique is currently under development.

Note that in Fig. 10b,c (where the solutions are computed without nudging) and in Fig. 11b,c (where the solutions are computed with nudging) the solutions computed with the two methods stay close to each other (even after 4 years of integration in the second case), while some form of chaotic behaviour is introduced by the sampling of the probability distributions, so that, even for the same method, after a certain time, two simulations seeded with different random selections of the probability distributions should diverge substantially. In the first case, this similarity comes from the fact that there is not enough time for the solutions to diverge (they "freeze" very quickly). In the second case, it happens because the nudging might be too strong for this case thus substantially affecting the probabilistic evolution. If the nudging strength is weaker (Fig. 12) then solutions computed with different probabilistic methods diverge quicker.

It is also important to remark that the rapid change of the trajectory computed with the first method (which we observed for the Lorenz system) does not seem to reveal itself in the QG dynamics, and both methods give qualitatively the same results. Therefore we further use the second method as it is somewhat faster than the first one.

Incomplete reference data. Incomplete observational data are typical in ocean modelling when using observational data for reanalysis in comprehensive ocean models. Obviously, such data cannot be directly used for numerical modelling, as they may include undefined values, missing parts of data records, or a combination of both. There are different interpolation methods and reanalysis data to overcome the problem. As the probabilistic evolutionary approach is intended to work with comprehensive ocean models, it would be instructive to study its performance on incomplete, raw reference flows, i.e. without engaging interpolation or reanalysis. For doing so, we take the second method and consider similar to the Lorenz system test cases (Fig. 13): (1) gappy trajectory, (2) holey dynamics, and (3) disjoint space. We use the same 2-year long reference solution as above, while running the probabilistic evolutionary model for four years.

In the *gappy trajectory* test we remove every second (Fig. 13a) and every fourth (Fig. 13b) point from the original 2-year long reference solution thus retaining only 50% and 25% of the reference data. As seen in Fig. 14a,b, the probabilistic evolutionary solution reproduces the nominally-resolved reference flow structures way beyond the time period over which the reference data is available.

In the *holey dynamics* test we remove a vast region of the reference dynamics. Namely, the reference solution contained in the sphere of radius $r = 0.979$ centred at its time mean has been excluded from the reference solution thus making voids in different parts of the reference trajectory; note that only half of the reference solution remained after this resection. As with the previous test case, the probabilistic evolutionary method restores (Fig. 14c) the nominally-resolved flow structures of the reference solution.

The *disjoint space* is a test where we cut the phase space into two disjoint regions (divided by a gap of width 0.02). Despite that, the probabilistic evolutionary solution is still able to evolve in the reference phase space and reproduce nominally-resolved reference flow features (Fig. 14d).

Although the probabilistic solution reproduces both large- and small-scale reference flow structures for substantially corrupted data sets, the results in Fig. 14 clearly show that it is overheated, i.e. substantially larger in the absolute value than the reference solution. It happens because the incomplete reference data works as a repeller thus pressing the probabilistic trajectory out of the reference space. On the one hand, it might be considered a weakness of the proposed approach, while, on the other hand, it might be an option to explore vaster regions of the reference phase space, and thus study reference solutions that can potentially be simulated with the reference model. In order to keep the probabilistic evolutionary solution in the reference phase space (and thus make its amplitude closer to the reference one), we crank up the nudging strength (Fig. 15). It can be done either manually (as

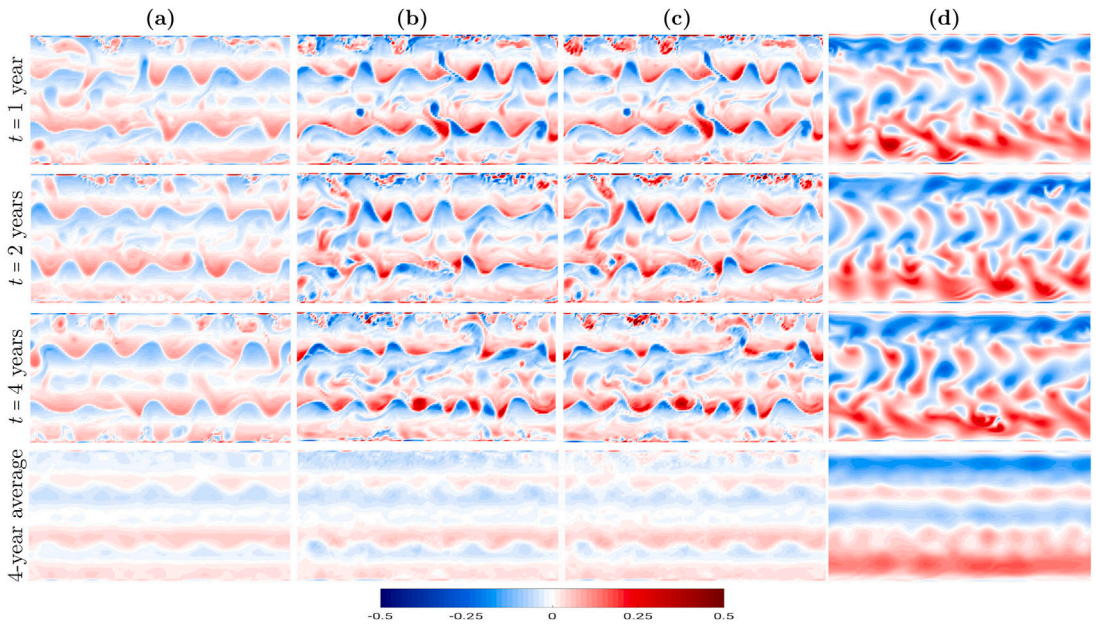


Fig. 11. Shown are snapshots of the top layer PV anomaly: (a) the reference solution (computed on the 513 × 257 grid and then projected onto the 129 × 65 grid), (b)/(c) probabilistic solution computed with the first/second method, (d) modelled solution computed with the QG Eqs. (10) on the coarse grid 129 × 65, and a 4-year time-average (last row); the nudging strength is $\eta = 0.1$ for both probabilistic solutions. All solutions are given in units of $[s^{-1}f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\ s^{-1}$ is the Coriolis parameter. Note that the probabilistic evolutionary methods use only the first 2 years out of the 4-year long reference solution. Both methods reproduce nominally-resolved on the coarse grid flow features, while the coarse-grid QG solution results in complete failure to reproduce even large-scale jets not to mention nominally-resolved features.

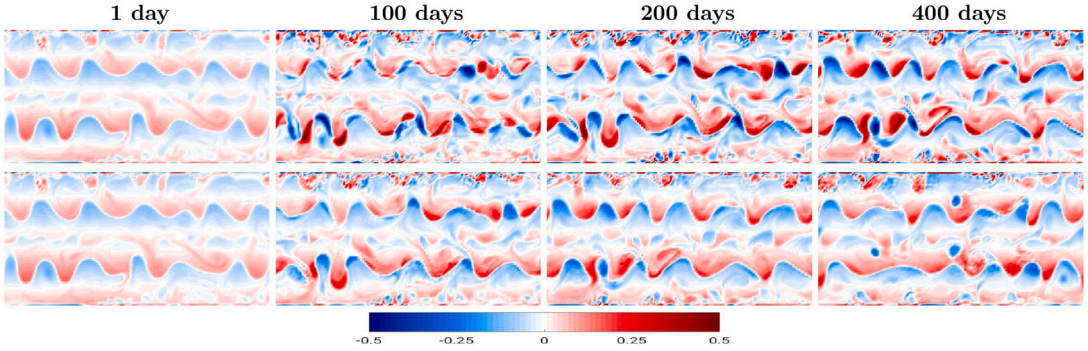


Fig. 12. Shown are snapshots of the top layer PV anomaly computed with the first (top row) and second (bottom row) probabilistic methods; the nudging strength is $\eta = 0.075$ (it is weaker than that in Fig. 11). All solutions are given in units of $[s^{-1}f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\ s^{-1}$ is the Coriolis parameter. The results show that the two probabilistic solutions diverge quicker when the nudging is weaker.

in this study) or automatically with the adaptive nudging (Shevchenko and Berloff, 2022b). The stronger nudging makes the amplitude of all probabilistic solutions smaller by keeping them closer to the reference phase space, although we adjusted the nudging strength individually for each case. Based on the PEA performance for the incomplete reference solutions, we conclude that the PEA works well even with substantially corrupted reference data. This is an appealing feature not only for ocean modellers working with models but also for those working with measurements.

Long simulations. It might seem from the results above that simulations with the PEA can only be twice as long as the reference solution, thus setting the upper time limit for PEA simulations. In what follows, we demonstrate how the PEA works for longer periods. We take the second method and run it for 8 years, while using the same 2-year long reference solution (Fig. 16). As seen in the figure, the method reproduces nominally-resolved flow features (large-scale jets, small-scale vortices, and meanders along the jets) of the reference solution. This, once again, ensures that the PEA can model ocean flows far beyond the reference data set.

Further insights into the dynamics can be provided with more quantitative diagnostics (Fig. 17). The energy spectral density (ESD) of the 8-year averaged probabilistic solution (Fig. 17c) is two orders of magnitude closer to that of the reference solution (Fig. 17a) than the ESD of the modelled solution (Fig. 17b). The root-mean-square error is lower for the probabilistic solution (Fig. 17e) compared to the modelled solution (Fig. 17d). Besides, the integral of the PV anomaly (conserved quantity in the QG model) over the domain is of order 10^{-4} which indicates that the PEA respects conservation laws (Fig. 17f). The integral of the PV anomaly for the modelled solution (not shown) is of order 10^{-12}. The accuracy with which conservation laws hold in the PEA can be improved by (1) cranking up the accuracy of the sampling procedure, or using (2) the vector-wise calculation of the transition probability function (i.e. sampling whole directional vectors rather than their coordinates) or (3) the brute-force approach (imposing conservation laws in every time step).

The high-resolution simulation. As the PEA is proposed as an alternative to modelling the ocean, it would be instructive to assess it on high-resolution data as well. In order to do it, we take a 2-year long high-resolution QG solution (computed on the grid 513 × 257) as a reference solution, and study how the second method performs (Fig. 18). As with the low-resolution reference solution, the method

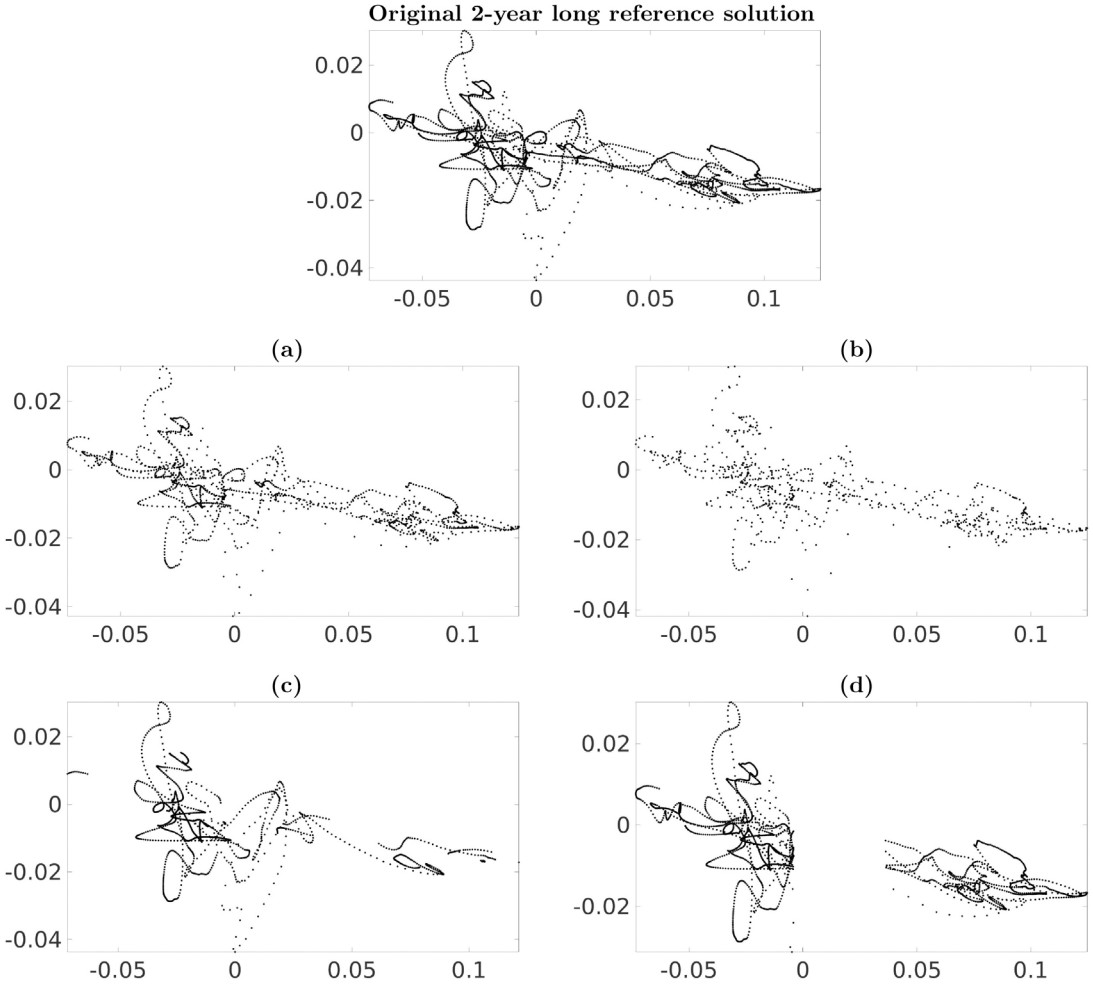


Fig. 13. Shown is the trajectory of the top-layer PV anomaly $(q_1(x_i, y_i, t), q_1(x_j, y_j, t))$ for two randomly-chosen points (x_i, y_i) and (x_j, y_j) for the original 2-year long reference solution used in the QG simulations above (the axes denote the PV anomaly value at the corresponding point), (a)/(b) gappy trajectory with every second/fourth point retained, (c) holey dynamics (the reference solution contained in a sphere of radius $r = 0.979$ centred at its time mean is removed), (d) disjoint space (the reference phase space is cut into two disjoint regions; the cut width is 0.02).

performs equally well and reproduces the nominally-resolved reference flow features.

4. Conclusions and discussion

In this work we have proposed a probabilistic evolutionary approach (PEA) to ocean modelling that capitalizes on the chaotic nature of ocean dynamics by taking advantage of using the probability distribution of neighbourhood states in the reference phase space as opposed to making use of deterministic or stochastic differential equations. A new state of the model is determined by the likelihood of the states neighbouring to the current state. The probabilistic nature of the flow evolution implies that even very unlikely (rare) events are expected to occur once in a while thus echoing observations of extreme weather and climate events.

Within the PEA framework we have developed two probabilistic evolutionary methods. These methods have been tested on the Lorenz 63 system and showed that both methods reproduce the Lorenz attractor even for substantially corrupted reference data sets. In addition, we have shown how the nudging strength can influence the probabilistic solution. Being assured in the PEA potential for modelling geophysical flows, we have considered an idealized ocean model (two-layer quasi-geostrophic model configured for a horizontally periodic flat-bottom channel) and showed that a non-eddy-resolving solution can be substantially improved towards the reference eddy-resolving solution compared to the low-resolution simulation. Within the context of QG dynamics we have demonstrated the build-up effect and its detrimental consequences on the probabilistic solution, and how to avoid it with the nudging methodology. We have also studied how the probabilistic evolutionary models work on incomplete reference data and demonstrated that they reproduce nominally-resolved on the coarse grid reference flow features even for substantially corrupted reference data sets. In addition, we have demonstrated that the PEA works very well over long time periods (8 years in our case) even for short reference solutions (2-year long). Our results show that the probabilistic evolutionary approach performs equally well for both low- and high-resolution reference solutions.

The appealing advantages of the probabilistic evolutionary approach are: (1) it requires no modification of the ocean model; (2) easy to implement; (3) it can take not only the reference solution as input data but also real measurements from different sources (drifters, weather stations, etc.), or a combination of both; note that points in the phase space represent states of the system and are irrelevant to how these states are computed or observed (for example, if the Argo fleet provides irregular in space data then it should be preprocessed to properly cover the area of interest); (4) it is ready out of the box for generating ensembles of solutions, (5) copes with substantially corrupted data sets, (6) performs equally well for both low- and high-resolution reference data, and reproduces nominally-resolved reference dynamics; (7) works over long periods without degradation of the

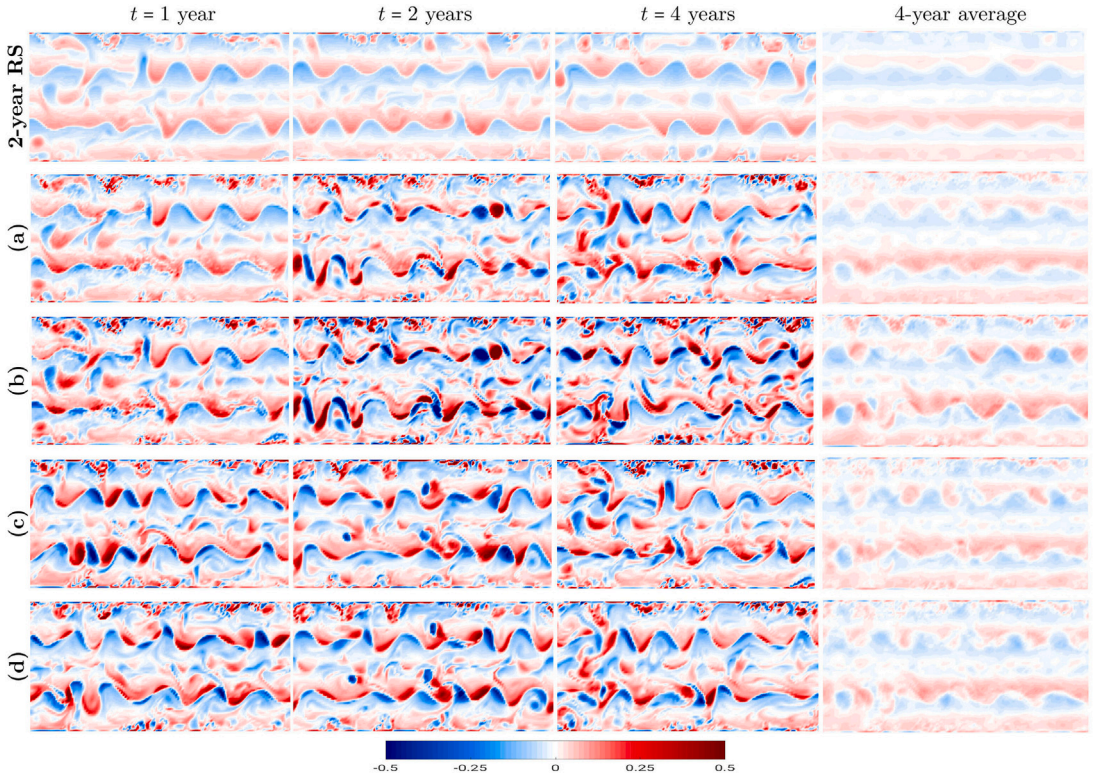


Fig. 14. Shown are snapshots of the top layer PV anomaly for the incomplete reference solution (Fig. 13): **(top row)** the original 2-year long complete reference solution (as in Fig. 11a), (a)/(b) probabilistic evolutionary dynamics for the gappy reference solution with every second/fourth point removed (Fig. 13a,b), (c) probabilistic evolutionary dynamics for the holey reference solution (Fig. 13c), (d) probabilistic evolutionary dynamics for the disjoint reference solution (Fig. 13d). and a 4-year time-average (last column); the nudging strength is $\eta = 0.1$. All solutions are given in units of $[s^{-1} f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\ s^{-1}$ is the Coriolis parameter. Note that the probabilistic evolutionary method reproduces both large- and small-scale flow features even for substantially incomplete reference data sets.

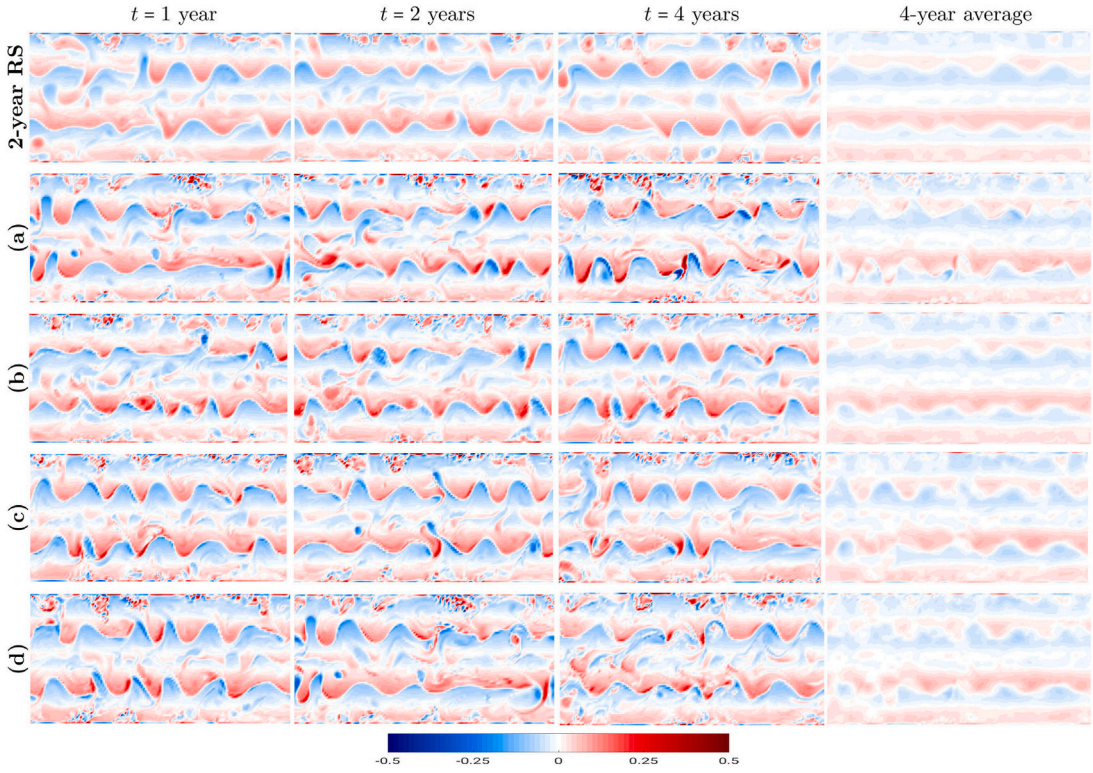


Fig. 15. The same as Fig. 14 but with the nudging strength $\eta = 0.2$ for (a)–(b), $\eta = 0.6$ for (c), and $\eta = 0.4$ for (d).

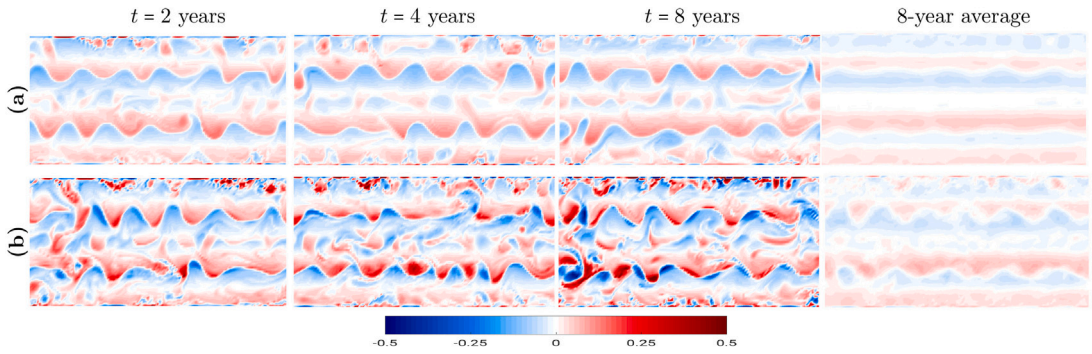


Fig. 16. Shown are snapshots of the top layer PV anomaly: (a) the reference solution, (b) probabilistic solution computed with the second method, and a 8-year time-average (last column); the nudging strength is $\eta = 0.1$. All solutions are given in units of $[s^{-1} f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\,\mathrm{s}^{-1}$ is the Coriolis parameter. As with shorter runs, the probabilistic evolutionary method reproduces the nominally-resolved reference flow features (both large and small scales) over the period of 8 years.

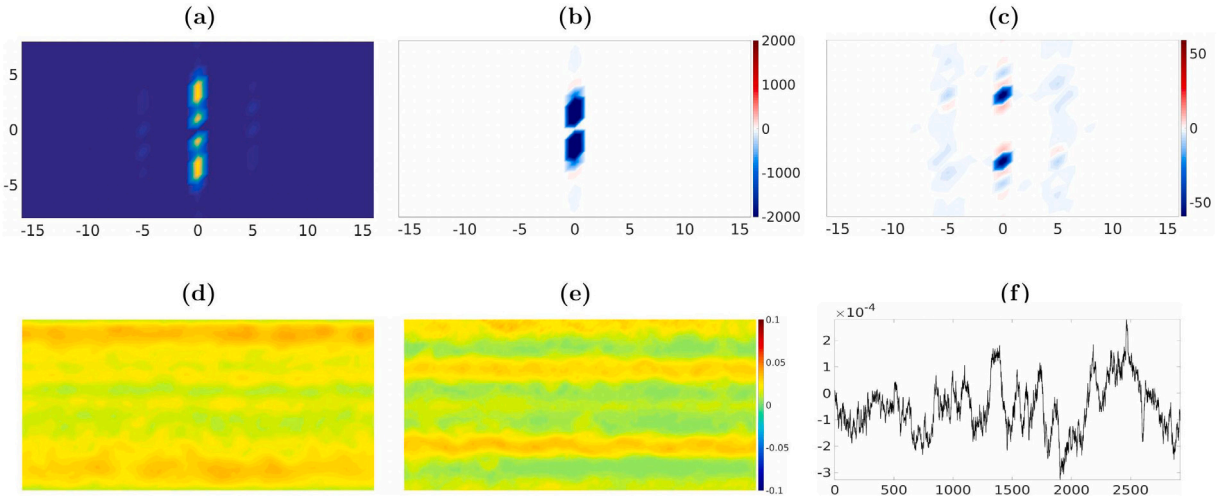


Fig. 17. Shown are different diagnostics of the top layer PV anomaly, q_1, computed over 8 years: the energy spectral density (ESD) for (a) 8-year averaged reference solution (last column in Fig. 16a), (b) difference between (a) and ESD for 8-year averaged modelled solution, (c) difference between (a) and ESD for 8-year averaged probabilistic solution computed with the second method (last column in Fig. 16b) (ESD is in units of $[s^{-1} f_0^{-1} m^2]^2$, and the axes represent wavenumbers), (d) root mean square error between the reference solution and modelled solution, (e) root mean square error between the reference solution and solution computed with the second method, (f) evolution of the integral I_{q_1} from the second method (the horizontal axis shows time in days, and the vertical axis represents the integral value); (d)–(f) are in units of $[s^{-1} f_0^{-1}]$, with $f_0 = 0.83 \times 10^{-4}\,\mathrm{s}^{-1}$ being the Coriolis parameter.

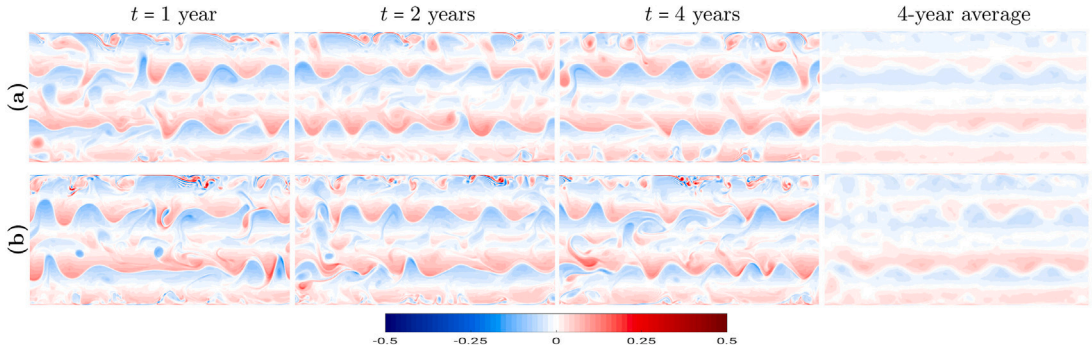


Fig. 18. Shown are snapshots of the top layer PV anomaly: (a) the reference solution (computed on the 513 × 257 grid), (b) probabilistic solution (computed with the second method on the grid 513 × 257), and a 4-year time-average (last column); the nudging strength is $\eta = 0.1$. All solutions are given in units of $[s^{-1} f_0^{-1}]$, where $f_0 = 0.83 \times 10^{-4}\,\mathrm{s}^{-1}$ is the Coriolis parameter. Note that the probabilistic evolutionary method reproduces the nominally-resolved reference flow features (large-scale jets and small-scale vortices).

probabilistic solution (even for short reference records) thus operating well beyond the reference data range.

All this offers a great flexibility to ocean modellers working with comprehensive ocean models and measurements, and allows us to expect that the proposed approach has strong potential for the use in the context of primitive equations which we plan to approach in future research.

CRediT authorship contribution statement

Igor Shevchenko: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing. **Pavel Berloff:** Writing – review & editing, Funding acquisition.

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

Data will be made available on request.

Acknowledgements

The authors thank The Leverhulme Trust for the support of this work through the grant RPG-2019-024 and the anonymous referees for their constructive comments, suggestions, and efforts which helped us improve the paper. Pavel Berloff was supported by the Natural Environment Research Council grant NE/T002220/1 and by the Moscow Centre for Fundamental and Applied Mathematics (supported by the Agreement 075-15-2019-1624 with the Ministry of Education and Science of the Russian Federation, Russia).

References

Chassignet, E., Hurlburt, H., Smedstad, O., Halliwell, G., Hogan, P., Wallcraft, A., Baraille, R., Bleck, R., 2007. The HYCOM (HYbrid Coordinate Ocean Model) data assimilative system. J. Mar. Syst. 65, 60–83.

Cotter, C., Crisan, D., Holm, D., Pan, W., Shevchenko, I., 2020. Modelling uncertainty using stochastic transport noise in a 2-layer quasi-geostrophic model. Found. Data Sci. 2, 173–205.

Danilov, S., Sidorenko, D., Wang, Q., Jung, T., 2017. The finite-volumE sea ice–ocean model (FESOM2). Geosci. Model Dev. 10, 765–789.

Devroye, L., 1986. Non-Uniform Random Variate Generation. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo.

Lorenz, E., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20, 130–141.

Madec, G., NEMO System Team, 2022. NEMO Ocean Engine. In: Scientific Notes of Climate Modelling Center, vol. 27, Institut Pierre-Simon Laplace.

Marshall, J., Adcroft, A., Hill, C., Perelman, L., Heisey, C., 1997. A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. J. Geophys. Res. 102, 5753–5766.

McWilliams, J., 1977. A note on a consistent quasigeostrophic model in a multiply connected domain. Dynam. Atmos. Ocean 5, 427–441.

Pedlosky, J., 1987. Geophysical Fluid Dynamics. Springer-Verlag, New York.

Shevchenko, I., Berloff, P., 2021. A method for preserving large-scale flow patterns in low-resolution ocean simulations. Ocean Model. 161, 101795.

Shevchenko, I., Berloff, P., 2022a. A method for preserving nominally-resolved flow patterns in low-resolution ocean simulations: Constrained dynamics. Ocean Model. 178, 102098.

Shevchenko, I., Berloff, P., 2022b. A method for preserving nominally-resolved flow patterns in low-resolution ocean simulations: Dynamical system reconstruction. Ocean Model. 170, 101939.

Shevchenko, I., Berloff, P., 2023. A hyper-parameterization method for comprehensive ocean models: Advection of the image point. Ocean Model. 184, 102208.

Storch, H.V., Zwiers, F., 2002. Statistical Analysis in Climate Research. Cambridge Univ. Press, Cambridge.

Vanem, E., Zhu, T., Babanin, A., 2022. Statistical modelling of the ocean environment – A review of recent developments in theory and applications. Mar. Struct. 86.