Ensemble Riemannian Data Assimilation over the Wasserstein Space

Sagar K. Tamang[1], Ardeshir Ebtehaj[1], Peter J. van Leeuwen[2], Dongmian Zou[3], and Gilad Lerman[4]

[1]Department of Civil, Environmental and Geo-Engineering and Saint Anthony Falls Laboratory, University of Minnesota-Twin Cities, Twin Cities, Minnesota, USA
[2]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA
[3]Duke Kunshan University, Kunshan, China
[4]School of Mathematics, University of Minnesota-Twin Cities, Twin Cities, Minnesota, USA

Correspondence: Sagar K. Tamang (taman011@umn.edu), Ardeshir Ebtehaj (ebtehaj@umn.edu)

Abstract. In this paper, we present an ensemble data assimilation paradigm over a Riemannian manifold equipped with the Wasserstein metric. Unlike the Euclidean distance used in classic data-assimilation methodologies, the Wasserstein metric can capture translation and difference between the shapes of square-integrable probability distributions of the background state and observations. This enables us to formally penalize geophysical biases in state-space with non-Gaussian distributions. The new approach is applied to dissipative and chaotic evolutionary dynamics and its potential advantages and limitations are highlighted compared to the classic ensemble data assimilation approaches under systematic errors.

1 Introduction

Extending the forecast skill of Earth System Models (ESM) relies on advancing the science of Data Assimilation (DA) (Tsuyuki and Miyoshi, 2007; Carrassi et al., 2018). A large body of current DA methodologies, either filtering (Kalman, 1960; Evensen, 1994; Reichle et al., 2002; Evensen, 2003) or variational approaches (Lorenc, 1986; Le Dimet and Talagrand, 1986; Talagrand and Courtier, 1987; Park and Županski, 2003; Trevisan et al., 2010; Carrassi and Vannitsem, 2010; Ebtehaj and Foufoula-Georgiou, 2013), are derived from basic principles of Bayesian inference under the assumption that the state-space is unbiased and can be represented well with Gaussian distributions, which are not often consistent with reality (Bocquet et al., 2010; Pires et al., 2010). It is well documented that this drawback often limits forecast skills of DA systems (Walker et al., 2001; Dee, 2005; Ebtehaj et al., 2014; Chen et al., 2019a) especially under the presence of systematic errors (Dee, 2003).

Apart from particle filters (Spiller et al., 2008; Van Leeuwen, 2010), which are intrinsically designed for state-space with a non-Gaussian distribution, numerous modifications to the variational DA (VDA) and ensemble-based filtering methods have been made to tackle non-Gaussianity of geophysical processes

(Pires et al., 1996; Han and Li, 2008; Mandel and Beezley, 2009; Anderson, 2010). As a few examples, in four-dimensional VDA, a quasi-static approach is proposed to ensure convergence by gradually increasing the assimilation intervals (Pires et al., 1996). To deal with multi-modal systems, Kim et al. (2003) proposed modifications to the ensemble Kalman filter (EnKF; Evensen, 1994; Li et al., 2009) using approximate implementation of Bayes' theorem in lieu of a linear interpolation via Kalman gain. For ensemble-based filters, Anderson (2010) proposed a new approach to account for non-Gaussian priors and posteriors by utilizing rank histograms (Anderson, 1996; Hamill, 2001). A hybrid ensemble approach was also suggested to combine advantages of both EnKF and particle filter (Mandel and Beezley, 2009).

Even though particle filters can handle non-Gaussian likelihood functions, when observations lie away from the support set of the particles, the ensemble variance tends to zero over time and can render the filter degenerate (Poterjoy and Anderson, 2016). In recent years, significant progress has been made to treat systematic errors through numerous ad hoc methods such as the field alignment technique (Ravela et al., 2007) and morphing EnKF (Beezley and Mandel, 2008) that can tackle position errors between observations and forecast. Dual state-parameter EnKF (Moradkhani et al., 2005) was also developed to resolve systematic errors originating from parameter uncertainties. Additionally, bias aware variants of the Kalman filter were designed (Drécourt et al., 2006; De Lannoy et al., 2007a, b; Kollat et al., 2008) to simultaneously update the state-space and an *a priori* estimate of the additive biases. In parallel, the cumulative distribution function matching (Reichle and Koster, 2004) has garnered widespread attention in land DA.

From a geometrical perspective, Gaussian statistical inference methods exhibit a flat geometry (Amari, 1985). In particular, it is proved that linear auto-regressive and moving average Markov stochastic models, which are driven by Gaussian noise, form dually flat manifolds (Amari, 2012). The notion of distance over such a geometrically flat space is defined over a straight line, which can be quantified by the Euclidean distance. Consequently, the Euclidean space has served as a major tool in explaining statistical inference techniques using linear Gaussian models and has been the cornerstone of DA techniques. It is important to note that the Euclidean distance remains (Ning et al., 2014) insensitive to the magnitude of translation between probability distributions with disjoint support sets – when used to interpolate between them.

Non-Gaussian statistical models often form geometrical manifolds, a topological space that is locally Euclidean. In the case of nonlinear regression, it is demonstrated that the statistical manifold exhibits a Riemannian geometry (Lauritzen, 1987) over which the notion of distance between probability distribu-

tions is geodesic that not only can capture translation but also the difference between the entire shape of probability distributions (Pennec, 2006). The question is – how can we equip DA with a Riemannian geometry? To answer this question, inspired by the theories of optimal mass transport (Villani, 2003), this paper presents the Ensemble Riemannian Data Assimilation (EnRDA) framework using the Wasserstein metric or distance, which is a distance function defined between probability distributions, as explained in detail in Section 2.3 .

In recent years, a few attempts have been made to utilize the Wasserstein metric, in geophysical DA. Reich (2013) introduced an ensemble transform particle filter, where the optimal transport framework was utilized to guide the resampling phase of the filter. Ning et al. (2014) used the Wasserstein distance to reduce forecast uncertainty due to parameter estimation errors in dissipative evolutionary equations. Feyeux et al. (2018) suggested a novel approach employing the Wasserstein distance in lieu of the Euclidean distance to penalize the position error between state and observations. More recently, Tamang et al. (2020) introduced a Wasserstein regularization in a variational setting to correct for geophysical biases under chaotic dynamics.

The EnRDA extends the previous work through the following main contributions: (a) EnRDA defines DA as a discrete barycenter problem over the Wasserstein space for assimilation in probability domain without any parametric or Gaussian assumption. The framework provides a continuum of non-parametric analysis probability histograms that naturally span between the distributions of the background state and observations through optimal transport of probability masses. (b) The presented methodology operates in an ensemble setting and utilizes a regularization technique for improved computational efficiency. (c) The paper studies advantages and limitations of DA over the Wasserstein space for dissipative advection-diffusion dynamics and nonlinear chaotic Lorenz-63 model in comparison with the well-known ensemble-based methodologies.

The organization of the paper is as follows: Section 2 provides a brief background on Bayesian DA formulations and optimal mass transport. The mathematical formalism of EnRDA is described in Section 3. Section 4 presents the results and compares them with their Euclidean counterparts. Section 5 discusses the findings and ideas for future research.

2 Background

2.1 Notations

Throughout, small bold letters represent m-element column vectors $\mathbf{x} = (x_1, \ldots, x_m)^{\mathrm{T}} \in \mathbb{R}^m$, where $(\cdot)^{\mathrm{T}}$ is the transposition operator. The m-by-n matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ are denoted by capital bold letters, whereas $\mathbb{R}^m_+ (\mathbb{R}^{m \times n}_+)$ denotes those vectors (matrices) only containing non-negative real numbers. The $\mathbb{1}_m$ refers to an m-element vector of ones and $\mathbf{I}_m$ is an $m \times m$ identity matrix. A diagonal matrix with entries given by $\mathbf{x} \in \mathbb{R}^m$ is represented by $\mathrm{diag}(\mathbf{x}) \in \mathbb{R}^{m \times m}$. Notation $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that the random vector $\mathbf{x}$ is drawn from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and $\mathbb{E}_X(\mathbf{x})$ is the expectation of $\mathbf{x}$. The ℓ_q-norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_q = \left(\sum_{i=1}^m |x_i|^q\right)^{1/q}$ with $q > 0$ and the square of the weighted ℓ_2-norm of $\mathbf{x}$ is represented as $\|\mathbf{x}\|^2_{\mathbf{B}^{-1}} = \mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}$, where $\mathbf{B}$ is a positive definite matrix. Notations of $\mathbf{x} \odot \mathbf{y}$ and $\mathbf{x} \oslash \mathbf{y}$ represent the element-wise Hadamard product and division between equal length vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. Notation $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ denotes the Frobenius inner product between matrices $\mathbf{A}$ and $\mathbf{B}$ and $\mathrm{tr}(\cdot)$ and $\det[\cdot]$ represent trace and determinant of a square matrix, respectively. Here, $p(\mathbf{x}) = \sum_{i=1}^M p_{\mathbf{x}_i} \delta_{\mathbf{x}_i}$ represents a discrete probability distribution with respective histogram $\{\mathbf{p}_x \in \mathbb{R}^M_+ : \sum_i p_{\mathbf{x}_i} = 1\}$ supported on $\mathbf{x}_i$, where $\delta_{\mathbf{x}_i}$ represents a Kronecker delta function at $\mathbf{x}_i$. Throughout, the dimension of the state or observations is denoted by small letters such as $\mathbf{x} \in \mathbb{R}^m$ while the number of ensembles or support points of their respective probability distribution is shown by capital letters such as $\mathbf{p}_x \in \mathbb{R}^M_+$.

2.2 Data Assimilation on Euclidean Space

In this section, we provide a brief review of the derivation of classic variational DA and particle filters based on Bayes' theorem to set the stage for the presented Ensemble Riemannian DA formalism.

2.2.1 Variational Formulation

Let us consider a discrete-time Markovian dynamics and its observations as follows:

$$\begin{aligned} \mathbf{x}^t &= \mathcal{M}(\mathbf{x}^{t-1}) + \boldsymbol{\omega}^t, & \boldsymbol{\omega}^t &\sim \mathcal{N}(\mathbf{0}, \mathbf{B}) \\ \mathbf{y}^t &= \mathcal{H}(\mathbf{x}^t) + \boldsymbol{v}^t, & \boldsymbol{v}^t &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \end{aligned} \tag{1}$$

where $\mathbf{x}^t \in \mathbb{R}^m$ and $\mathbf{y}^t \in \mathbb{R}^n$ represent the state variables and the observations at time t, $\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^m$ and $\mathcal{H} : \mathbb{R}^m \to \mathbb{R}^n$ are the deterministic forward model and observation operator, $\boldsymbol{\omega}^t \in \mathbb{R}^m$ and $\boldsymbol{v}^t \in \mathbb{R}^n$ are the independent and identically distributed model and observation errors, respectively.

Recalling Bayes' theorem, dropping the time superscript, without loss of generality, the posterior probability density function (pdf) of the state given the observation can be obtained as $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x})/p(\mathbf{y})$, where $p(\mathbf{y}|\mathbf{x})$ is proportional to the likelihood function, $p(\mathbf{x})$ is the prior density and $p(\mathbf{y})$ denotes the distribution of observations. Letting $\mathbf{x}_b = \mathbb{E}_X(\mathbf{x}) \in \mathbb{R}^m$ represents the background state, ignoring the constant term $\log p(\mathbf{y})$ and assuming Gaussian distributions for the observation error and the prior, logarithm of the posterior density leads to the well-known three-dimensional variational (3D-Var) cost function (Lorenc, 1986; Kalnay, 2003):

$$\begin{aligned} -\log p(\mathbf{x}|\mathbf{y}) &\propto \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathcal{H}(\mathbf{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathcal{H}(\mathbf{x})) \\ &\propto \|\mathbf{x}-\mathbf{x}_b\|^2_{\mathbf{B}^{-1}} + \|\mathbf{y}-\mathcal{H}(\mathbf{x})\|^2_{\mathbf{R}^{-1}}\,. \end{aligned} \tag{2}$$

As a result, the analysis state obtained by minimization of the cost function in Eq. (2) is the mode of the posterior distribution that coincides with the posterior mean when errors are drawn from unbiased Gaussian densities and $\mathcal{H}(\cdot)$ is a linear operator. Using the Woodbury matrix inversion lemma (Woodbury, 1950), it can be easily demonstrated that for a linear observation operator, the analysis states in the 3D-Var and Kalman filter are equivalent (Tarantola, 1987). As is evident, zero-mean Gaussian assumptions lead to penalization of the error through the weighted Euclidean norm.

2.2.2 Particle Filters

Particle filters (Gordon et al., 1993; Doucet and Johansen, 2009; Van Leeuwen et al., 2019) in DA were introduced to address the issue of non-Gaussian distribution of the state by representing the prior and posterior distributions through a weighted ensemble of model outputs referred to as "particles". In its standard discrete setting, using Monte Carlo simulations, the prior distribution $p(\mathbf{x})$ is represented by a sum of equal-weight Kronecker delta functions as $p(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\delta_{\mathbf{x}_i}$, where $\mathbf{x}_i \in \mathbb{R}^m$ is the state variable represented by the i^{th} particle.

Each of these M particles are then evolved through the nonlinear model in Eq. (1). Assuming that the conditional distribution $p(\mathbf{y}|\mathbf{x}_i) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\left\{ -\frac{1}{2}[\mathbf{y} - \mathcal{H}(\mathbf{x}_i)]^{\mathrm{T}} \mathbf{R}^{-1} [\mathbf{y} - \mathcal{H}(\mathbf{x}_i)] \right\}$ is Gaussian, using Bayes' theorem, it can be shown that the posterior distribution $p(\mathbf{x}|\mathbf{y})$ can be approximated using a set of weighted particles as $p(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{M} w_i \, \delta_{\mathbf{x}_i}$, where $w_i = \frac{p(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{M} p(\mathbf{y}|\mathbf{x}_j)}$. The particles are then resampled from the posterior distribution $p(\mathbf{x}|\mathbf{y})$ based on their relative weights and propagated forward in time according to the model dynamics.

As is evident, in particle filters, weights of each particle are updated using the Gaussian likelihood function under a zero-mean error assumption. However, in the presence of systematic biases, when the support sets of particles and the observations are disjoint, only the weights of a few particles become significantly large and weights of other particles tend to zero. As the underlying dynamical system progresses in time, only those few particles, with relatively larger weights, are resampled and the filter can become degenerate gradually in time (Poterjoy and Anderson, 2016).

2.3 Optimal Mass Transport

The theory of optimal mass transport (OMT), coined by Gaspard Monge (Monge, 1781) and later extended by Kantorovich (Kantorovich, 1942), was developed to minimize transportation cost in resource allocation problems with purely practical motivations. Recent developments in mathematics discovered that OMT provides a rich ground to compare and morph probability distributions and uncovered new connections to partial differential equations (Jordan et al., 1998; Otto, 2001) and functional analysis (Brenier, 1987; Benamou and Brenier, 2000; Villani, 2003).

In a discrete setting, let us define two discrete probability distributions $p(\mathbf{x}) = \sum_{i=1}^{M} p_{\mathbf{x}_i} \delta_{\mathbf{x}_i}$ and $p(\mathbf{y}) = \sum_{j=1}^{N} p_{\mathbf{y}_j} \delta_{\mathbf{y}_j}$ with their respective histograms $\{\mathbf{p}_x \in \mathbb{R}_+^M : \sum_i p_{\mathbf{x}_i} = 1\}$ and $\{\mathbf{p}_y \in \mathbb{R}_+^N : \sum_j p_{\mathbf{y}_j} = 1\}$ supported on $\mathbf{x}_i$ and $\mathbf{y}_j$. A "ground" transportation cost matrix $\mathbf{C} \in \mathbb{R}_+^{M \times N}$ is defined such that its elements $c_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|_q^q$ represent the cost of transporting unit probability masses from location $\mathbf{x}_i$ to $\mathbf{y}_j$. The Kantorovich OMT problem determines an optimal "transportation plan" $\mathbf{U}^a \in \mathbb{R}_+^{M \times N}$ that can linearly map two probability histograms onto each other with minimum amount of total transportation cost as follows:

$$\mathbf{U}^a = \underset{\mathbf{U}}{\operatorname{argmin}} \, \langle \mathbf{C}, \mathbf{U} \rangle \qquad \text{s.t.} \qquad \mathbf{U} \geq 0, \quad \mathbf{U}\mathbb{1}_N = \mathbf{p}_x, \quad \mathbf{U}^{\mathrm{T}}\mathbb{1}_M = \mathbf{p}_y \,. \tag{3}$$

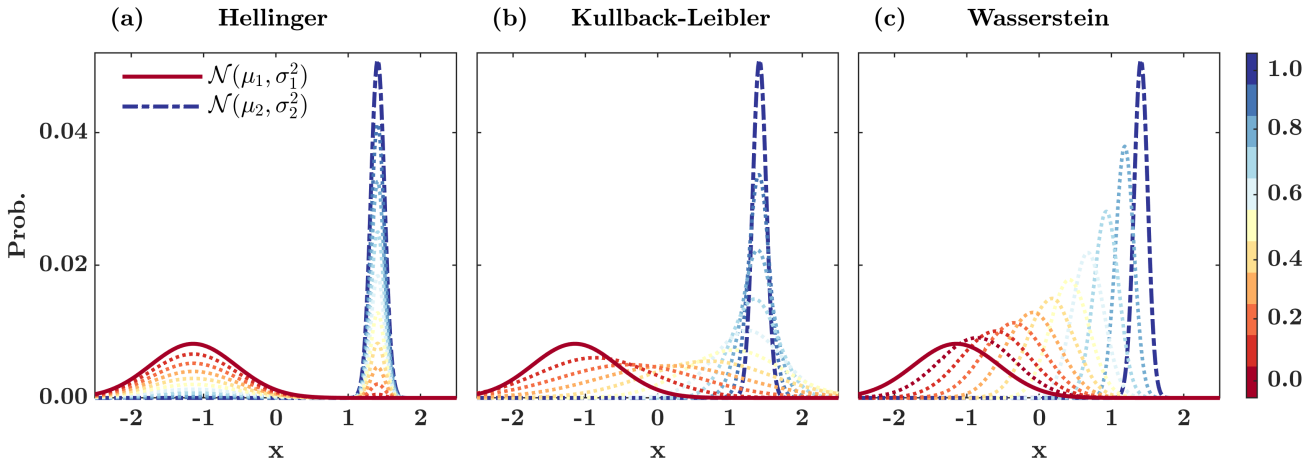

Figure 1. Interpolation between two Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ where, $\mu_1 = -1.1, \mu_2 = 1.4, \sigma_1^2 = 0.4$, and $\sigma_2^2 = 0.01$ as a function of the interpolation or displacement parameter $\eta \in [0, 1]$ for the (a) Hellinger distance, (b) Kullback-Leibler divergence, and (c) 2-Wasserstein distance (Peyré and Cuturi, 2019).

The transportation plan can be interpreted as a "joint distribution" that couples the marginals histograms $\mathbf{p}_x$ and $\mathbf{p}_y$. For the transportation cost with $q = 2$, the OMT problem in Eq. (3) is convex and defines the square of the 2-Wasserstein distance $d^2_{\mathcal{W}}(\mathbf{p}_x, \mathbf{p}_y) = \langle \mathbf{C}, \mathbf{U}^{\mathbf{a}} \rangle$ between the histograms.

Wasserstein distance has some advantages compared to other measures of proximity – such as the Hellinger distance (Hellinger, 1909) or the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). To elaborate on the advantages, we confine our consideration to the Gaussian densities over which the Wasserstein distance can be obtained in a closed form. In particular, interpolating over the 2-Wasserstein space between $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, using an interpolation parameter η, results in a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_\eta, \boldsymbol{\Sigma}_\eta)$, where $\boldsymbol{\mu}_\eta = \eta\, \boldsymbol{\mu}_x + (1 - \eta)\, \boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_x^{-1/2} \big(\eta\, \boldsymbol{\Sigma}_x + (1 - \eta)\, (\boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Sigma}_y \boldsymbol{\Sigma}_x^{1/2})^{1/2}\big)^2 \boldsymbol{\Sigma}_x^{-1/2}$ (Chen et al., 2019b).

Fig. 1 shows the spectrum of interpolated distributions between two Gaussian pdfs for a range of $\eta \in [0, 1]$. As shown, the interpolated densities using the Hellinger distance, which is Euclidean in the space of probability histograms, are bimodal. Although the Gaussian shape of the interpolated densities using the KL divergence is preserved, the variance of the interpolants is not necessarily bounded by the variances of the input Gaussian densities. Unlike these metrics, as shown, the Wasserstein distance moves the mean and preserves the shape of the interpolants through a natural morphing process.

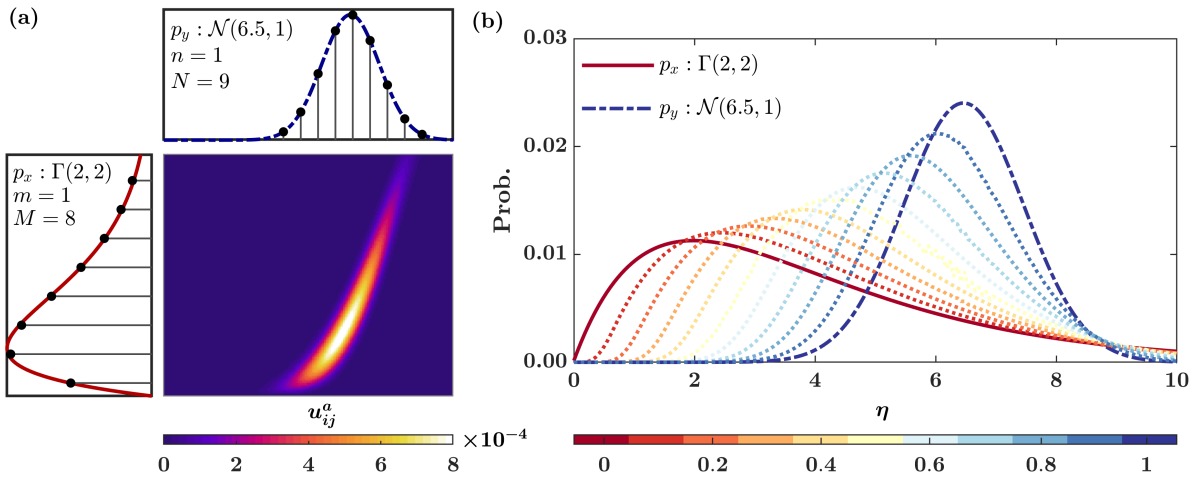

Figure 2. (a) Optimal transportation plan or the joint distribution $\mathbf{U}^a$ between a gamma $\Gamma(2,2)$ and a Gaussian marginal distribution $\mathcal{N}(6.5,1)$ as well as (b) the 2-Wasserstein interpolation between them for different values of the displacement parameter $\eta \in [0,1]$.

As is previously noted, this metric is not limited to any Gaussian assumption. Fig. 2 shows the 2-Wasserstein interpolation between a gamma and a Gaussian distribution. The results show that the Wasserstein metric penalizes the translation and mismatch between the shapes of the pdfs. It can be shown that $d^2_{\mathcal{W}}(\mathbf{p}_x, \mathbf{p}_y) = d^2_{\mathcal{W}}(\overline{\mathbf{p}}_x, \overline{\mathbf{p}}_y) + \left\|\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right\|_2^2$, where $\overline{\mathbf{p}}_x$ and $\overline{\mathbf{p}}_y$ are the centered zero-mean probability masses and $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are the respective mean values (Peyré and Cuturi, 2019).

3 Ensemble Riemannian Data Assimilation

3.1 Problem Formulation

First, let us recall that the weighted mean of a cloud of points $\{\mathbf{x}_i\}_{i=1}^M \in \mathbb{R}^m$ in the Euclidean space is $\boldsymbol{\mu}_x = \sum_{i=1}^M \eta_i \mathbf{x}_i$ for a given family of non-negative weights $\sum_i \eta_i = 1$. This expected value is equivalent to solving the following variational problem:

$$\boldsymbol{\mu}_x = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{M} \eta_i \left\|\mathbf{x}_i - \mathbf{x}\right\|_2^2 . \tag{4}$$

The 3D-Var problem in Eq. (2) reduces to the above barycenter problem when the model and observation error covariances are diagonal with uniform error variances across multiple dimensions of the state-space. For non-diagonal error covariances, it can be shown that the weight of the background and observation

are $(\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{B}^{-1}$ and $(\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}$ respectively, where $\mathbf{H}$ is the linear approximation of the observation operator. Therefore, the 3D-Var DA can be interpreted as a "barycenter problem" in the Euclidean space, where the analysis state is the weighted mean of the background state and observation.

By changing the distance metric from Euclidean to the Wasserstein (Agueh and Carlier, 2011), a Riemannian barycenter can be defined as the Fréchet mean (Fréchet, 1948) of N_p probability histograms with finite second-order moments as follows:

$$\mathbf{p}_\eta = \underset{\mathbf{p}}{\operatorname{argmin}} \sum_{k=1}^{N_p} \eta_k \, d^2_{\mathcal{W}}(\mathbf{p}, \mathbf{p}_k)\,. \tag{5}$$

Inspired by (Feyeux et al., 2018), the EnRDA defines the probability distribution of the analysis state $p(\mathbf{x}_a) \in \mathbb{R}^M$ as the Fréchet barycenter over the Wasserstein space as follows:

$$p(\mathbf{x}_a) = \underset{\mathbf{p}_x}{\operatorname{argmin}} \left\{ \eta \, d^2_{\mathcal{W}}(\mathbf{p}_x, \mathbf{p}_{x_b}) + (1-\eta)\, d^2_{\mathcal{W}}(\mathbf{p}_x, |\det[\mathcal{H}'(\mathbf{x})]\,|\,\mathbf{p}_y) \right\}, \tag{6}$$

where the displacement parameter $\eta > 0$ assigns the relative weights to the observation and background term to capture their respective geodesic distances from the true state. Here $\mathcal{H}'(\cdot)$ is the Jacobian of the observation operator assuming that $\mathcal{H} : \mathbf{x} \to \mathbf{y}$ is a smooth and a square (i.e., $m = n$) bijective map. The displacement parameter η is a hyperparameter and its optimal value should be determined empirically using some reference data through cross-validation studies. For example, in practice, one may compute the mean squared error as a function of η by comparing the analysis state and some ground-based observations and use the minimum point of that function statically over a window of multiple assimilation cycles. It is also important to note that due to the bijective assumption for the observation operator, the above formalism currently lacks the ability to propagate the information content of observed dimensions to unobserved ones. This limitation is discussed further in the Section 5.

The solution of the above DA formalism involves finding the optimal analysis transportation plan or the joint distribution $\mathbf{U}^a \in \mathbb{R}^{M\times N}$, using Eq. (3), that couples the background and observation marginal histograms. We use the McCann's method (McCann, 1997; Peyré et al., 2019) to obtain the analysis

probability histogram:

$$p(\mathbf{x}_a) = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij}^{a} \, \delta_{\mathbf{z}_{ij}} \,, \tag{7}$$

where the analysis support points are $\mathbf{z}_{ij} = \eta\, \mathbf{x}_i + (1-\eta)\, \mathbf{y}_j$. To solve Eq. (3), the widely used interior-point methods (Altman and Gondzio, 1999) and the Orlin's (Orlin, 1993) algorithm have super-cubic run time with a computational complexity of $O(M^3 \log M)$, where $M = N$. Therefore, the use of original OMT framework in EnRDA is a limitation in high-dimensional geophysical DA problems. To alleviate this computational cost, in the next subsection 3.2, we discuss the use of entropic regularization.

To solve Eq. (6) in an ensemble setting, let us assume that in the absence of any a priori information, initially the background probability distribution is represented by $i = 1 \ldots M$ ensemble members of the state variable $\mathbf{x}_i \in \mathbb{R}^m$ as $p(\mathbf{x}_b) = \frac{1}{M} \sum_{i=1}^{M} \delta_{\mathbf{x}_i}$. An a priori assumption is needed to reconstruct the observation distribution $p(\mathbf{y}) = \sum_{i=1}^{N} \mathbf{p}_{y_j} \delta_{\mathbf{y}_j}$ at $j = 1 \ldots N$ supporting points. To that end, one may choose a parametric or a non-parametric model based on the past climatological information. Here, for simplicity, we assume a zero-mean Gaussian representation with covariance $\mathbf{R} \in \mathbb{R}^{n \times n}$ (Burgers et al., 1998) to perturb the observation at each assimilation cycle. After each cycle, the probability histogram of the analysis state $p(\mathbf{x}_a)$ is estimated using Eq. (7) over $\mathbf{z}_{ij}$ at $M \times N$ support points. Then $p(\mathbf{x}_a)$ is resampled at M points using the multinomial sampling scheme (Li et al., 2015) to initialize the next time step forecast.

3.2 Entropic Regularization of EnRDA

In order to speed up the computation of the coupling, the problem in Eq. (3) can be regularized (Cuturi, 2013):

$$\mathbf{U}^a = \underset{\mathbf{U}}{\operatorname{argmin}} \, \langle \mathbf{C}, \mathbf{U} \rangle - \gamma\, H(\mathbf{U}) \qquad \text{s.t. } \mathbf{U} \geq 0, \;\; \mathbf{U}\mathbb{1}_N = \mathbf{p}_{x_b}, \;\; \mathbf{U}^{\mathrm{T}}\mathbb{1}_M = \mathbf{p}_y \,, \tag{8}$$

where $\gamma > 0$ is the regularization parameter and $H(\mathbf{U}) := \langle \mathbf{U}, \log \mathbf{U} - \mathbb{1}_M \mathbb{1}_N^{\mathrm{T}} \rangle$ represents the Gibbs-Boltzmann relative entropy function. Note that the relative entropy is a concave function and thus its negative value is convex.

The Lagrangian function ($\mathcal{L}$) of Eq. (8) can be obtained by adding two dual variables or Lagrangian multipliers $\mathbf{q} \in \mathbb{R}^M$ and $\mathbf{r} \in \mathbb{R}^N$:

$$\mathcal{L}(\mathbf{U}, \mathbf{q}, \mathbf{r}) = \langle \mathbf{C}, \mathbf{U} \rangle - \gamma\, H(\mathbf{U}) - \langle \mathbf{q}, \mathbf{U}\mathbb{1}_N - \mathbf{p}_{x_b} \rangle - \langle \mathbf{r}, \mathbf{U}^{\mathrm{T}}\mathbb{1}_M - \mathbf{p}_y \rangle \,. \tag{9}$$

Setting the derivative of the Lagrangian function to zero, we have

$$\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{q}, \mathbf{r})}{\partial u_{ij}} = c_{ij} + \gamma \log(u_{ij}) - q_i - r_j = 0 \,, \qquad \forall i, j \,. \tag{10}$$

The entropic regularization keeps the problem in Eq. (8) strongly convex and it can be shown (Peyré et al., 2019) that Eq. (10) leads to a unique optimal joint density with the following form:

$$\mathbf{U}^a = \mathrm{diag}(\mathbf{v})\, \mathbf{K}\, \mathrm{diag}(\mathbf{w}) \,, \tag{11}$$

where $\mathbf{v} = \exp(\mathbf{q}) \oslash (\gamma \mathbb{1}_M) \in \mathbb{R}^M$ and $\mathbf{w} = \exp(\mathbf{r}) \oslash (\gamma \mathbb{1}_N) \in \mathbb{R}^N$ are the unknown scaling variables and $\mathbf{K} \in \mathbb{R}_+^{M \times N}$ is the Gibbs kernel with elements $k_{ij} = \exp\big(-\frac{c_{ij}}{\gamma}\big)$, where c_{ij} are the elements of the cost matrix $\mathbf{C}$.

From the mass conservation constraints in Eq. (8) and scaling form of the optimal joint density in Eq. (11), we can derive:

$$\mathrm{diag}(\mathbf{v})\, \mathbf{K}\, \mathrm{diag}(\mathbf{w}) \mathbb{1}_N = \mathbf{p}_{x_b} \qquad \text{and} \qquad \mathrm{diag}(\mathbf{w})\, \mathbf{K}^{\mathrm{T}}\, \mathrm{diag}(\mathbf{v}) \mathbb{1}_M = \mathbf{p}_y \,. \tag{12}$$

The two unknown scaling variables $\mathbf{v}$ and $\mathbf{w}$ can then be iteratively solved using Sinkhorn's matrix scaling algorithm (Cuturi, 2013) as follows:

$$\mathbf{v}^{l+1} = \mathbf{p}_{x_b} \oslash (\mathbf{K}\mathbf{w}^l) \qquad \text{and} \qquad \mathbf{w}^{l+1} = \mathbf{p}_y \oslash (\mathbf{K}^{\mathrm{T}}\mathbf{v}^l) \,. \tag{13}$$

The Sinkhorn algorithm is initialized using $\mathbf{w}^0 = \mathbb{1}_N$ and since the marginal densities $\mathbf{p}_{x_b}$ and $\mathbf{p}_y$ are not zero vectors, the Hadamard division in Equation 13 remains valid throughout the iterations. A summary of the EnRDA implementation is demonstrated in Algorithm 1.

Algorithm 1 Ensemble Riemannian Data Assimilation

Inputs: Ensemble size M, number of perturbed observations N from a chosen observation pdf, displacement parameter η, entropic regularization parameter γ and total number of time steps T.
Initialize: $\mathbf{x}_i^0 \sim p(\mathbf{x}^0),\ i = 1, \ldots, M.$
for $t = 1, \ldots, T$ **do**
  $\mathbf{x}_i^t = \mathcal{M}(\mathbf{x}_i^{t-1}) + \boldsymbol{\omega}_i^t,\ i = 1, \ldots, M.$
  Generating ensemble of observations $\mathbf{y}_j^t,\ j = 1, \ldots, N.$
  At initial time obtain probability histogram of the background state and observations:

$$p(\mathbf{x}_b) = \frac{1}{M} \sum_{i=1}^{M} \delta_{\mathbf{x}_i^t}\,, \quad p(\mathbf{y}) = \sum_{j=1}^{N} \mathbf{p}_{y_j} \delta_{\mathbf{y}_j^t}.$$

  Compute the joint histogram as follows:

$$\mathbf{U}^a = \underset{\mathbf{U}}{\operatorname{argmin}} \sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij}\, c_{ij} - \gamma \langle \mathbf{U}, \log \mathbf{U} - \mathbb{1}_M \mathbb{1}_N^{\mathrm{T}} \rangle \quad \text{s.t. } \mathbf{U} \geq 0,\ \mathbf{U}\mathbb{1}_N = \mathbf{p}_{x_b},\ \mathbf{U}^{\mathrm{T}}\mathbb{1}_M = \mathbf{p}_y\,,$$

$$\text{where } c_{ij} = \left\| \mathbf{x}_i^t - \mathbf{y}_j^t \right\|_2^2 .$$

  Obtain analysis probability distribution $p(\mathbf{x}_a) = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij}^a\, \delta_{\mathbf{z}_{ij}}$ where $\mathbf{z}_{ij} = \eta\, \mathbf{x}_i^t + (1 - \eta)\, \mathbf{y}_j^t$.
  Obtain M analysis ensemble members $\mathbf{x}_{ai} \in \mathbb{R}^m$ by multinomial sampling from $p(\mathbf{x}_a)$.
  Set $\mathbf{x}_i^t := \mathbf{x}_{ai}$.
end for

The entropic regularization parameter plays an important role in characterization of the joint density; however, there exists no closed-form solution for its optimal selection. Generally speaking, increasing the value of γ will increase convexity of the cost function and thus computational efficiency; however, at the expense of reduced coupling between the marginal histograms, consistent with the second law of thermodynamics.

As an example, the effects of γ on the coupling between two Gaussian mixture models $\mathbf{p}_{x_b}$ and $\mathbf{p}_y$ are demonstrated in Fig. 3. It can be seen that at smaller values of $\gamma = 0.001$, the probability masses of the joint distribution are sparse and lie compactly along the main diagonal – capturing a strong coupling between the background state and observations. However, as the value of γ increases, the probability masses of the joint distribution spread out – reflecting less degree of dependencies between the marginals. It is important to note that in limiting cases, as $\gamma \to 0$, the solution of Eq. (8) converges to the true optimal joint histogram, while as $\gamma \to \infty$, the entropy of the analysis state increases and tends to $\mathbf{p}_{x_b}\mathbf{p}_y^{\mathrm{T}}$. The regularization parameter is dependent on the elements of the transportation cost matrix $\mathbf{C}$ and varies

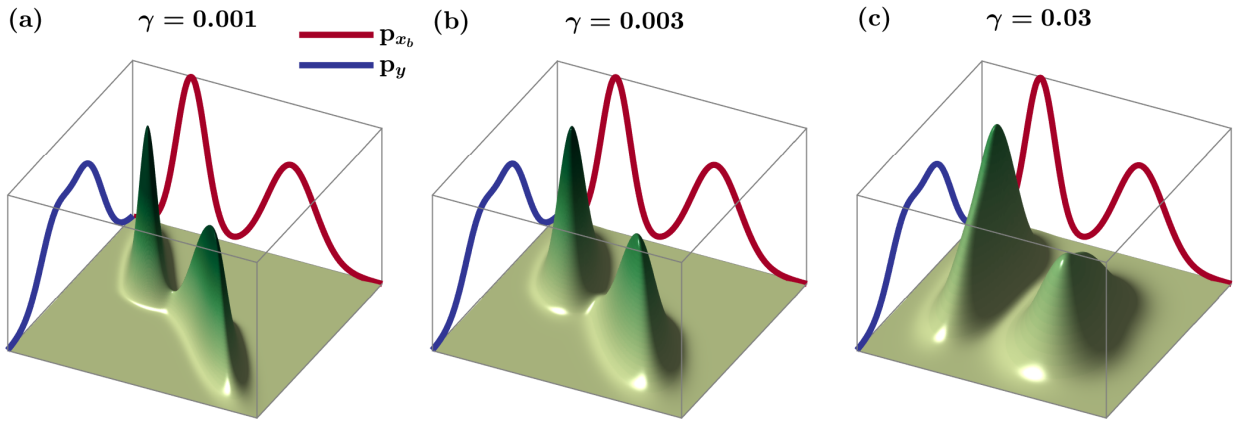

Figure 3. The effect of the entropic regularization parameter γ on the optimal joint histogram coupling two Gaussian mixture models $\mathbf{p}_{x_b} : 0.5\mathcal{N}(-12, 0.4) + 0.5\mathcal{N}(-8, 0.8)$ and $\mathbf{p}_y : 0.55\mathcal{N}(5, 4) + 0.45\mathcal{N}(9.5, 4)$.

according to the experimental settings. In practice, one can begin with the value of γ set as the largest element of the cost matrix $\mathbf{C}$ and gradually reduce it to find the minimum value of γ that provides a stable solution.

4 Numerical Experiments and Results

In order to demonstrate the performance of EnRDA and quantify its effectiveness, we focus on the linear advection-diffusion equation and the chaotic Lorenz-63 model (Lorenz, 1963). The advection-diffusion model explains a wide range of heat, mass, and momentum transport across the land, vegetation, and atmospheric continuum, and has been utilized to evaluate the performance of DA methodologies (Zhang et al., 1997; Hurkmans et al., 2006; Ning et al., 2014; Ebtehaj et al., 2014; Berardi et al., 2016). Similarly, the Lorenz-63 model, as a chaotic model of atmospheric convection, has been widely used in testing the performance of DA methodologies (Miller et al., 1994; Nakano et al., 2007; Van Leeuwen, 2010; Goodliff et al., 2015; Tandeo et al., 2015; Tamang et al., 2020). Throughout, under controlled experimental settings with foreknown model and observation errors, we run the forward models under systematic errors and compare the results of EnRDA with a particle filter (PF) and an EnKF.

4.1 Advection-Diffusion Equation

4.1.1 State-space Characterization

The advection-diffusion is a special case of the Navier-Stokes partial differential equation. In its linear form, with constant diffusivity in an incompressible fluid flow, it is expressed for a mass conserved physical quantity $\mathbf{x}(\mathbf{s},t)$ as follows:

$$\frac{\partial \mathbf{x}(\mathbf{s},t)}{\partial t} + \mathbf{a} \odot \nabla \mathbf{x}(\mathbf{s},t) = \mathbf{D}\,\nabla^2 \mathbf{x}(\mathbf{s},t)\,, \tag{14}$$

where $\mathbf{s} \in \mathbb{R}^n$ represents a $n-$dimensional spatial domain at time t. In the above expression, $\mathbf{a} = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ is the advection velocity vector and $\mathbf{D} = \text{diag}(D_1, \ldots, D_n) \in \mathbb{R}^{n \times n}$ represents the diffusivity matrix. Given initial condition $\mathbf{x}(\mathbf{s}, t = 0)$, owing to its linearity, the solution at time t can be obtained by convolving the initial condition with a Kronecker delta function $\delta(\mathbf{s} - \mathbf{a}\,t)$ followed by a convolution with the fundamental Gaussian kernel $\mathcal{G}(\mathbf{s},t) = \dfrac{1}{(2\pi)^{n/2}\,|\mathbf{\Sigma}|^{1/2}} \exp\left(-\dfrac{1}{2}\mathbf{s}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{s}\right)$, where $\mathbf{\Sigma} = 2\,\mathbf{D}\,t$.

4.1.2 Experimental Setup and Results

In this subsection, we first highlight the difference between Euclidean and Wasserstein barycenters using a 2-D advection-diffusion model and then compare the results of EnRDA with the PF and EnKF on a 1-D advection-diffusion equation.

Fig. 4 shows the results of an assimilation experiment using the 2-D advection-diffusion equation where the underlying state is bimodal. This experiment is designed to demonstrate the differences between the Euclidean and Wasserstein barycenters in the presences of bias in a non-Gaussian state-space. In particular, the state-space is characterized over a spatial domain $s_1 = (0, 10]$ and $s_2 = (0, 10]$ with a discretization of $\Delta s_1 = \Delta s_2 = 0.1$. The advection-diffusion is considered to be an isotropic process with true model parameters set as $a_1 = a_2 = 0.08$ [L/T], and $D_1 = D_2 = 0.02$ [L^2/T]. The shown state variable is obtained after evolving two Kronecker delta functions $\mathbf{x}(\mathbf{s},t) = 1000\,\delta(s_1, s_2)$ and $\mathbf{x}(\mathbf{s},t) = 4000\,\delta(s_1, s_2)$ for T =0–25 and $T = 0$–35 [t], respectively.

To resemble a model with systematic errors, background state is obtained by increasing the advective velocity to 0.12 [L/T] while diffusivity is reduced to 0.01 [L^2/T] (Fig. 4b). Observations are not considered to have position biases; however, a systematic representative error is imposed assuming that the sensing

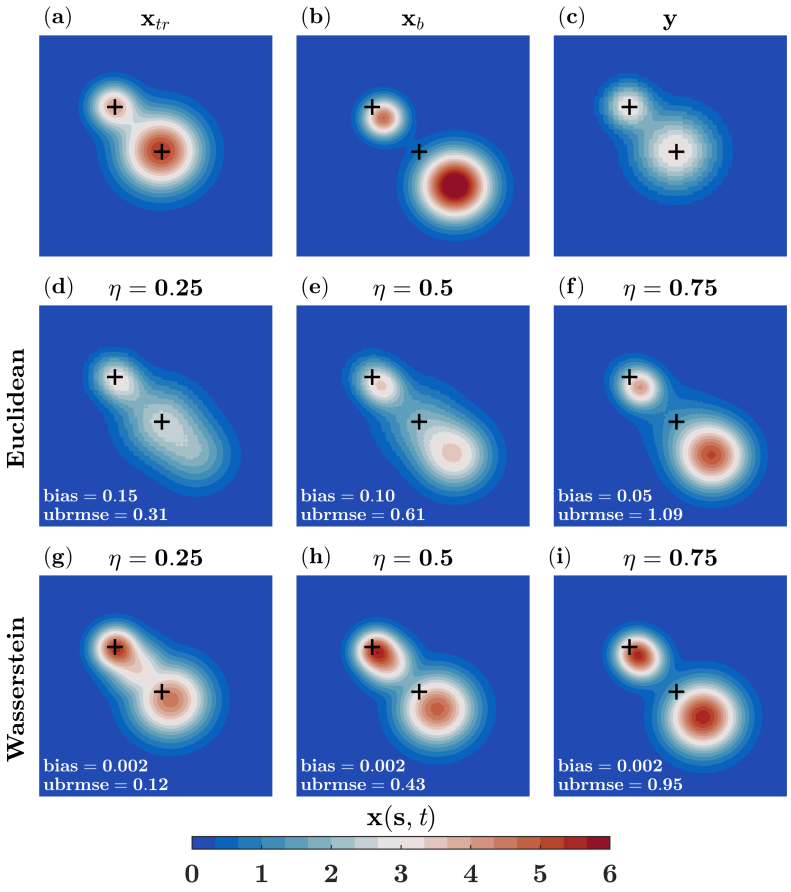

Figure 4. The true state $\mathbf{x}_{tr}$, the background state $\mathbf{x}_b$ and observation $\mathbf{y}$ (a–c) with systematic errors under a 2-D advection-diffusion dynamics as well as the Euclidean (d–f) and the Wasserstein barycenters (g–i) between $\mathbf{x}_b$ and $\mathbf{y}$ for different displacement parameters η. The entropic regularization parameter is set to $\gamma = 0.003$ and the black plus signs show the location of the modes for the true state.

system has a lower resolution than the model. To that end, we evolve two Kronecker delta functions, $\mathbf{x}(\mathbf{s}, t) = 800\,\delta(s_1, s_2)$ and $\mathbf{x}(\mathbf{s}, t) = 2400\,\delta(s_1, s_2)$, with mass less than the true state for same time period of $T = 0$–25 and $T = 0$–35 [t] and then up-scaled the field by a factor of two through box averaging (Fig. 4c).

As shown in Fig. 4(g–i), the Wasserstein barycenter preserves the shape of the state variable well and gradually moves the mass towards the background state as the value of η increases, while the bias remains almost constant and the ubrmse increases from 0.12 to 0.95. The error quality metrics are constantly below the Euclidean counterpart. As shown in Fig. 4(d–f), the shape of the Euclidean barycenter for smaller values of η is not well recovered due to the position error. As η increases from 0.25 to 0.75, the Euclidean barycenter is nudged towards the background state and begins to recover the shape. The bias is

reduced by more than 30%, from 0.15 to 0.05; however, this occurs at the expense of almost three folds increase in unbiased root mean squared error (ubrmse), from 0.3 to 1.1. The reason for reduction of the bias is that the positive differences between the Euclidean barycenter and true state are compensated by their negative differences. However, ubrmse is quadratic and thus measures the average magnitude of the error irrespective of its signs.

For the 1-D case, the state-space is characterized over a spatial domain $s \in (0, 60]$ with a discretization of $\Delta s = 0.1$. The model parameters are chosen to be $a = 0.8$ [L/T] and $D = 0.25$ [L^2/T]. The initial state resembles a bimodal mixture of Gaussian distributions obtained by superposition of two Kronecker delta functions $x(s, t = 0) = 300\,\delta(s)$ – evolved for time 15 and 25 [t], respectively. The ground truth of the trajectory is then obtained by evolving the initial state at a time step of $\Delta t = 0.5$ over a window of $T = 0$–30 [t] in the absence of any model error.

The observations are obtained at assimilation intervals $10\Delta t$, assuming an identity observation operator, through corrupting the ground truth by a heteroscedastic Gaussian noise with a variance $\epsilon_y = 5\%$ of the squared values of the ground truth state. We introduce both systematic and random errors in model simulations. For the systematic error, model velocity and diffusivity coefficient are set to $a' = 0.12$ [L/T] and $D' = 0.4$ [L^2/T] respectively. To impose the random error, a heteroscedastic Gaussian noise with variance $\epsilon_b = 2\%$ is added at every Δt to model simulations. One hundred ensemble members are used in EnRDA and the regularization and displacement parameters are set to $\gamma = 3$ and $\eta = 0.2$, through the previously outlined trial and error procedures. To obtain a robust comparison of EnRDA with PF and EnKF, each with 100 particles (ensemble members), the experiment is repeated for 50 independent simulations.

The evolution of the initial state over a time period $T = 0 - 30$ [t] and the results comparing EnRDA with PF and EnKF at 5, 15, and 25 [t] are shown in Fig. 5. As demonstrated, during all time steps, EnRDA reduces the analysis uncertainty, in terms of both bias and ubrmse. The shape of the entire state-space is also properly preserved and remains closer to the ground truth. The EnKF performs comparable to EnRDA during the initial time steps, however, the performance degrades and the error statistics gradually increases as the system evolves over time. Among the two traditional ensemble-based methodologies, the PF acquires the highest error statistics owing to the well-known problem of filter degeneracy (Poterjoy and Anderson, 2016), which is exacerbated by the presence of systematic errors.

We should emphasize that the presented results do not imply that EnRDA always performs better than PF and EnKF. The EnKF at the limiting case $M \to \infty$, in the absence of bias, is a minimum mean squared

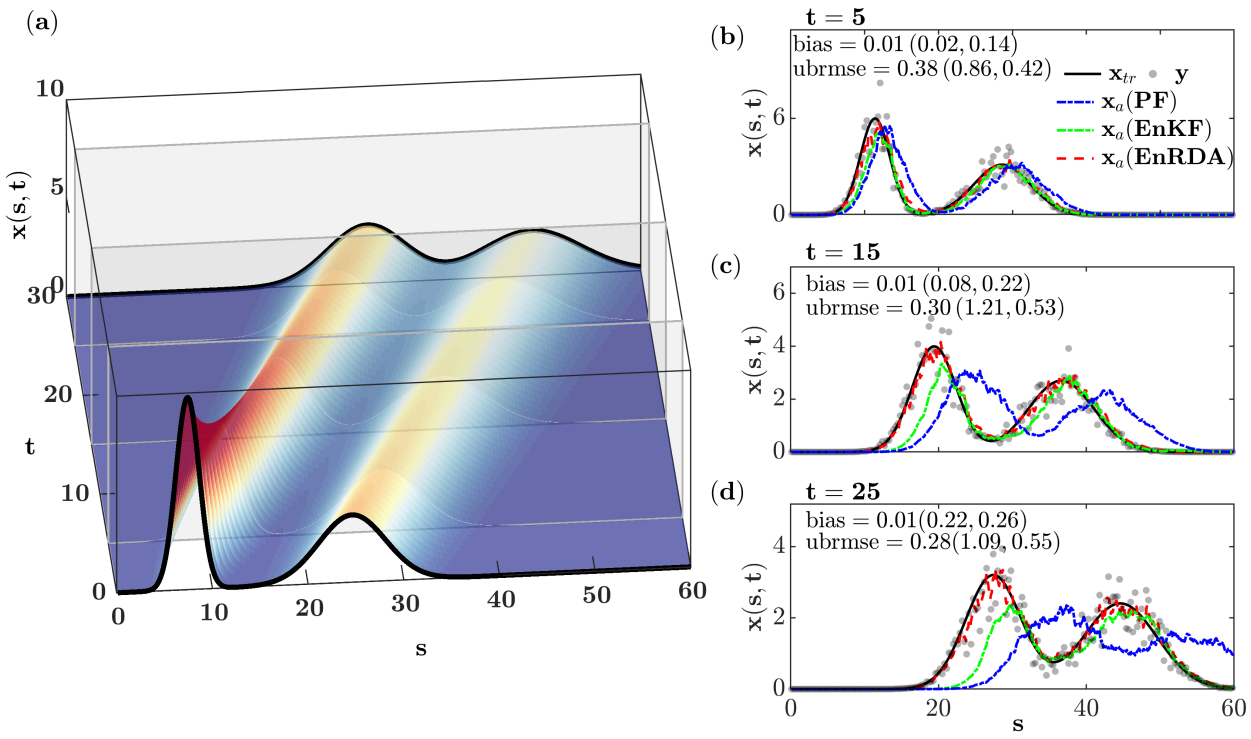

Figure 5. (a) Temporal evolution of a bimodal initial state under a linear 1-D advection-diffusion equation (b–d) by particle filter (PF), ensemble Kalman filter (EnKF) and ensemble Riemannian data assimilation for three time snapshots at 5, 15 and 25 [t]. The bias and ubrmse of the analysis states are also reported in the legends.

error estimator and attains the lowest possible posterior variance for linear systems, also referred to as the Cramer-Rao lower bound (Cramér, 1999; Rao et al., 1973). Thus, when the errors are drawn from zero-mean Gaussian distributions with a linear observation operator, EnKF can outperform EnRDA in terms of the mean squared error.

4.2 Lorenz-63

4.2.1 State-space Characterization

The Lorenz system (Lorenz-63, Lorenz, 1963) is derived through truncation of the Fourier series of the Rayleigh-Bénard convection model. This model can be interpreted as a simplistic local weather system only involving the effect of local shear stress and buoyancy forces. The system is expressed using coupled ordinary differential equations that describe the temporal evolution of three coordinates x, y, and z

representing the rate of convective overturn, horizontal, and vertical temperature variations:

$$\begin{aligned}\frac{dx}{dt} &= -\sigma(x-y)\\ \frac{dy}{dt} &= \rho x - y - xz\\ \frac{dz}{dt} &= xy - \beta z\,,\end{aligned} \tag{15}$$

where σ represents the Prandtl number, ρ is a normalized Rayleigh number proportional to the difference in temperature gradient through the depth of the fluid and β denotes a horizontal wave number of the convective motion. It is well established that for parameter values of $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, the system exhibits chaotic behavior with the phase space revolving around two unstable stationary points located at $(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1)$ and $(-\sqrt{\beta(\rho-1)}, -\sqrt{\beta(\rho-1)}, \rho-1)$.

4.2.2 Experimental Setup and Results

Throughout, we use the classic multinomial resampling for implementation of EnRDA and particle filter. Apart from the systematic error component, we utilize the standard experimental setting used in numerous DA studies (Miller et al., 1994; Furtado et al., 2008; Van Leeuwen, 2010; Amezcua et al., 2014). In order to obtain the ground truth of the model trajectory, the system is initialized at $\mathbf{x}_0 = (1.508870, -1.531271, 25.46091)$ and integrated with a time step of $\Delta t = 0.01$ over a time period of $T = 0$–20 [t] using the fourth-order Runge-Kutta approximation (Runge, 1895; Kutta, 1901). The observations are obtained at every assimilation interval $40\Delta t$ by assuming identity observation operator and perturbing the ground truth with Gaussian noise $\boldsymbol{v}_t \sim \mathcal{N}(0, \sigma^2_{obs}\,\boldsymbol{\Sigma}_\rho)$, where $\sigma^2_{obs} = 2$ and the correlation matrix $\boldsymbol{\Sigma}_\rho \in \mathbb{R}^{3\times 3}$ is populated with 1 on the diagonal entries, 0.5 on the first sub and super diagonals, and 0.25 on the second sub and super diagonals.

In order to characterize the distribution of the background state, 100 particles (ensemble members) are used, among all DA methods, by adding to the ground truth a zero-mean Gaussian noise $\boldsymbol{\omega}_0 \sim \mathcal{N}(0, \sigma_0^2\,\mathbf{I}_3)$ with $\sigma_0^2 = 2$, at the initial time. For introducing systematic errors, the model parameters are set to $\sigma' = 10.5$, $\rho' = 27$, and $\beta' = 10/3$. The random errors are also introduced in time by adding a Gaussian noise $\boldsymbol{\omega}_t \sim \mathcal{N}(0, \sigma_b^2\,\mathbf{I}_3)$ at every Δt, with $\sigma_b^2 = 0.02$. Throughout, to draw a robust statistical inference about the error statistics, the DA experiments are repeated for 50 independent simulations. As described

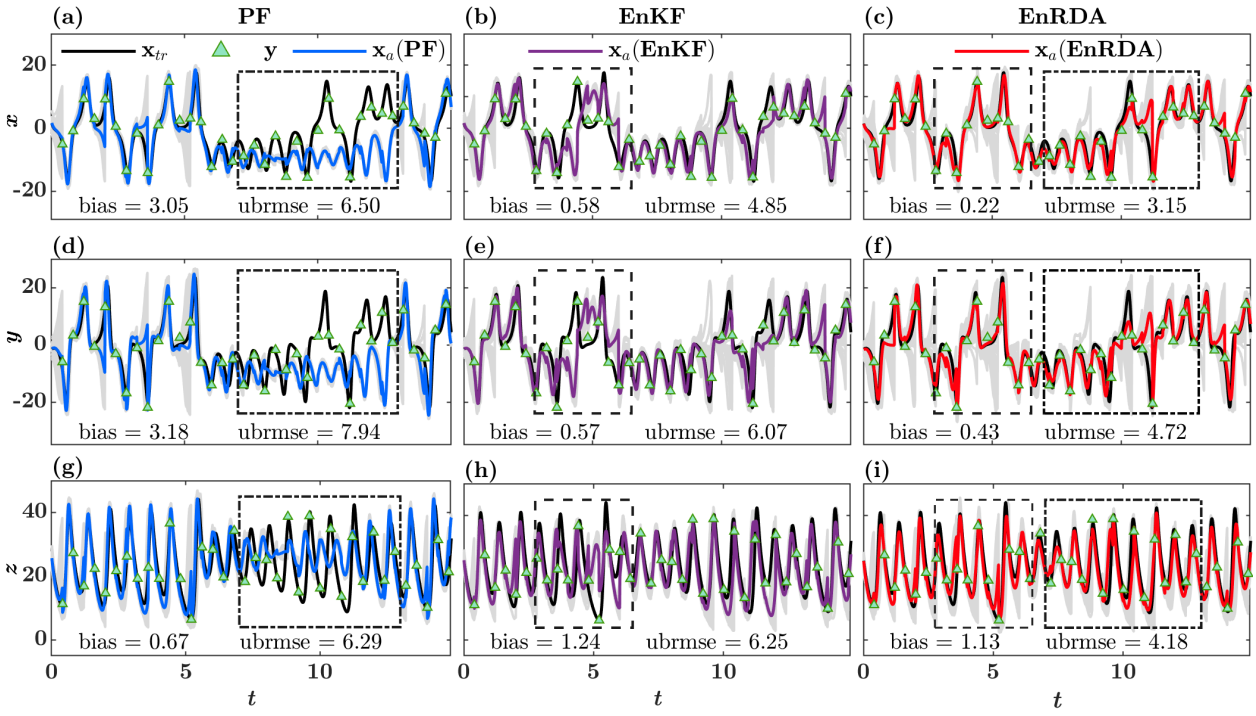

Figure 6. Temporal evolution of the true state $\mathbf{x}_{tr}$ in Lorenz-63, observations $\mathbf{y}$ as well as the analysis state $\mathbf{x}_a$ for the particle filter (PF) (first column), ensemble Kalman filter (EnKF) (second column) and ensemble Riemannian data assimilation (EnRDA) (third column) with 100 particles (ensemble members). The temporal evolution of the particles and ensemble members are shown with solid gray lines. Also shown within dashed rectangles are the windows of time over which the DA methods exhibit large errors.

previously, to properly account for the effects of both bias and ubrmse, the optimal value of the displacement parameter η in EnRDA can be selected based on an offline analysis of the minimum mean squared error. However, to provide a fair comparison between EnRDA and other filtering methods, at each assimilation cycle, we set $\eta = \mathrm{tr}(\mathbf{R})/\mathrm{tr}(\mathbf{R}+\mathbf{B})$, when observation operator is an identity matrix. Note that while the observation error covariance remains constant in time, the background error covariance is obtained from simulated ensembles and changes in time dynamically. This selection assures that the relative weights assigned to the background state and observations remain at the same order of magnitude among different methods.

Fig. 6 shows the temporal evolution of the ground truth and the analysis state by the PF (first column), EnKF (second column), and EnRDA (third column) over a time period of $T = 0$–15 [t]. As is evident, the PF is well capable of capturing the ground truth when observations lie within the particle spread. However, when the observations lie far apart from the support set of particles (Fig. 6, dashed box) and

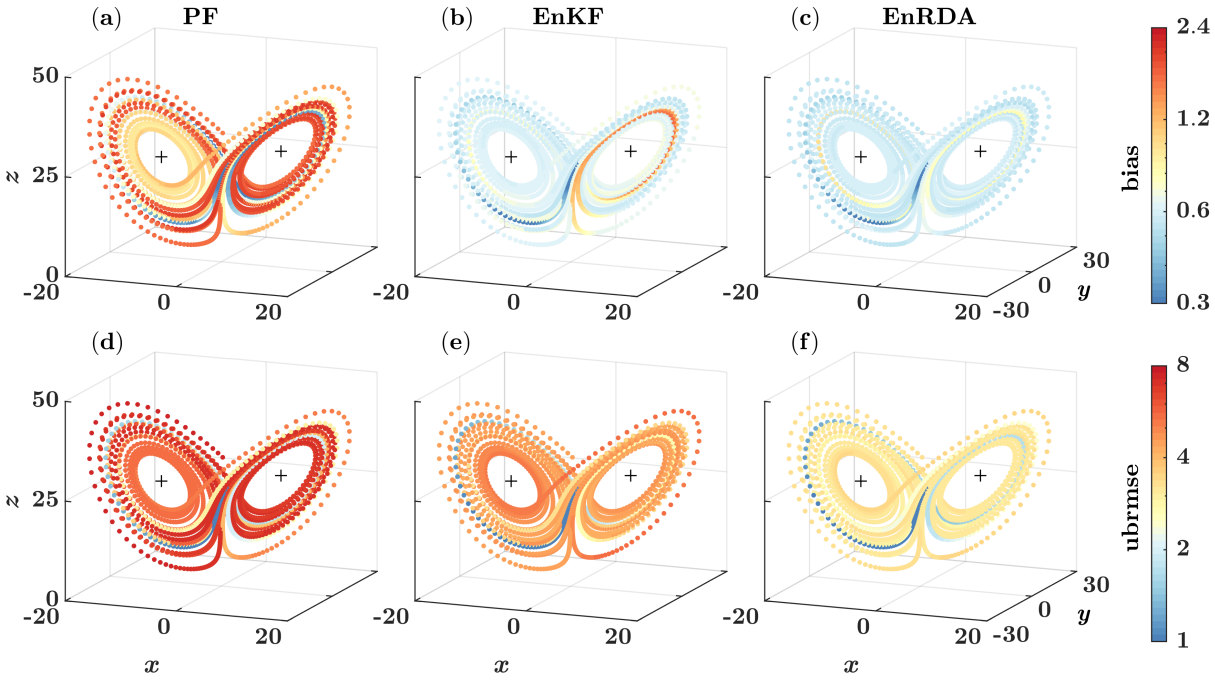

Figure 7. Temporal evolution of bias and ubrmse along three dimensions of the Lorenz-63 for (a, d) particle filter (PF), (b, e) ensemble Kalman filter (EnKF), and (c, f) ensemble Riemannian data assimilation (EnRDA), each with 100 particles (ensemble members). The mean values are computed over 50 independent simulations.

the distribution of the background state and observations become disjoint, the filter becomes degenerate and the analysis state (particle mean) deviates from the ground truth (Fig. 6g, dashed box). As a result, the bias of PF along the z-dimension is markedly lower than that of the EnKF and EnRDA while ubrmse is significantly higher. Whereas, EnRDA is capable of capturing the true state well even when ensemble spread and observations are far apart from each other. Although EnKF does not suffer from the same problem of filter degeneracy as the particle filter, in earlier time steps from 2.5 to 7.5 [t], it struggles to adequately nudge the analysis state towards the ground truth when ensemble members are far from the observations due to the imposed systematic bias. EnRDA seems to be robust to the propagation of systematic errors in this region and follows the true trajectory well.

The time evolution of the bias and ubrmse for 50 independent simulations, with the same error structure, is color coded over the phase space in Fig. 7. As shown, the error quality metrics are relatively lower for EnRDA than other DA methodologies. Nevertheless, we can see that the improvement compared to EnKF is modest. In particular, across all dimensions of the problem, the mean bias and ubrmse are decreased in

Table 1. Expected values of the bias and ubrmse for the particle filter (PF), ensemble Kalman filter (EnKF) and ensemble Riemannian data assimilation (EnRDA) for 50 independent simulations of Lorenz-63 across all problem dimensions.

Methods	bias				ubrmse			
	x	y	z	$x-z$	x	y	z	$x-z$
Particle Filter	2.24	2.41	0.59	1.75	6.25	7.95	7.88	7.36
EnKF	0.33	0.35	1.23	0.64	3.80	5.41	5.02	4.74
EnRDA	0.17	0.24	1.25	0.56	2.63	4.0	3.78	3.47

EnRDA by 68 (13)% and 53 (27)% compared to the particle filter (EnKF). More details about the expected values of the bias and ubrmse are reported in Table 1. We emphasize that the presented results shall be interpreted in light of the presence of systematic errors. In fact, EnRDA cannot reduce the analysis error variance beyond a minimum mean squared estimator such as EnKF in the absence of bias.

5 Discussion and Concluding Remarks

In this study, we introduced an ensemble data assimilation (DA) methodology over a Riemannian manifold, namely Ensemble Riemannian DA (EnRDA), and illustrated its performance in comparison with other ensemble-based DA techniques for dissipative and chaotic dynamics. We demonstrated that the presented methodology is capable of assimilating information in probability domain – characterized by the families of distributions with finite second-order moments. The key message is that when the probability distribution of the forecast and observations exhibit non-Gaussian structure and their support sets are disjoint, due to the presence of systematic errors; the Wasserstein metric can be leveraged to potentially extend geophysical forecast skills. Even though, future research for a comprehensive comparison with existing filtering and bias correction methodologies is needed to completely characterize relative pros and cons of the proposed approach – especially when it comes to the ensemble size and optimal selection of the displacement parameter η.

We explained the role of static regularization and displacement parameter in EnRDA and empirically examined their effects on the optimal joint histogram, coupling the background state and observations, and consequently on the analysis state. Nevertheless, future studies are required to characterize closed-form or heuristic expressions to expand our understating of their impacts on the forecast uncertainty dynamically. As was explained earlier, unlike the Euclidean DA methodologies that assimilate available information using different relative weights across multiple dimensions through the error covariance matrices; a scalar

displacement parameter is utilized in EnRDA that interpolates uniformly between all dimensions of the problem. Future research can be devoted to developing a framework that utilizes a dynamic vector representation of the displacement parameters to effectively tackle possible heterogeneity of uncertainty across multiple dimensions.

In it's current form, EnRDA requires the observation operator to be smooth and bijectve. This is a limitation when observations of all problem dimensions are not available and propagation of observations to non-observed dimensions is desired. Extending the EnRDA methodology to include partially observed systems seems to be an important future research area. This could include performing a rough inversion for unobserved components of the system offline or extending the methodology in the direction of particle flows (Van Leeuwen et al., 2019). Another promising area is to use EnRDA only over the observed dimensions of the state-space and, similar to the EnKF, use the ensemble covariance to update the unobserved part of state-space through a hybrid approach. Furthermore, it is important to note that several bias correction methodologies are available that explicitly add a bias term to the control vector in variational and filtering DA techniques (Dee, 2003; Reichle and Koster, 2004; De Lannoy et al., 2007b). Future research is required to compare the performance of EnRDA with other bias correction methodologies to fully characterize its relative advantages and disadvantages.

We should mention that EnRDA is computationally expensive as it involves estimation of the coupling through the Wasserstein distance. On a desktop machine with a 3.4 GHz CPU clock rate, it took around 1600 s to complete 50 independent simulations on Lorenz-63 for EnRDA compared to 651 (590) s for the PF (EnKF) with 100 particles (ensemble members). Since the computational cost is nonlinearly related to the problem dimension, it is expected that it grows significantly for large-scale geophysical DA and becomes a limiting factor. Furthermore, in high-dimensional geophysical problems, the computational cost of determining optimal displacement parameter η through cross-validation can be high. Although the entropic regularization works well for the presented low dimensional problems, future research is needed to test its efficiency in high-dimensional problems. Constraining the solution of the coupling on a submanifold of probability distributions with a Gaussian mixture structure (Chen et al., 2019b) can also be a future research direction for lowering the computational cost. Furthermore, recent advances in approximation of the Wasserstein distance using a combination of 1-D Radon projections and dimensionality reduction (Meng et al., 2019), can significantly reduce the computational cost to make EnRDA a viable methodology for tackling high-dimensional geophysical DA problems.

Lastly, in the presented formalism, we define the analysis state distribution through an optimal coupling between forecast and observation distributions. Future line of research can be devoted to coupling the forecast distribution with a normalized version of the likelihood function towards establishing connections with Bayesian data assimilation.

Code availability. A demo code for EnRDA in the MATLAB programming language can be downloaded at https://github.com/tamangsk/EnRDA

Author contributions. S.K.T. and A.E. designed the study. S.K.T. implemented the formulation and analyzed the results. P.J.L. and G.L. provided conceptual advice and all authors contributed to the writing.

Competing interests. The authors declare no competing interests.

Acknowledgements. The first and second author acknowledge the grant from the National Aeronautics and Space Administration (NASA) Terrestrial Hydrology Program (THP, 80NSSC18K1528) through Dr. J. Entin and the New (Early Career) Investigator Program (NIP, 80NSSC18K0742) through Dr. T. Lee and Dr. A. Leidner. The third author acknowledges support from the European Research Council for funding via the Horizon2020 CUNDA project under number 694509. The fifth author also acknowledges support from National Science Foundation (NSF, DMS1830418).

References

Agueh, M. and Carlier, G.: Barycenters in the wasserstein space, SIAM Journal on Mathematical Analysis, https://doi.org/10.1137/100805741, 2011.

Altman, A. and Gondzio, J.: Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization, Optimization Methods and Software, 11, 275–302, 1999.

Amari, S.-i.: Differential-Geometrical Methods in Statistics, 1985.

Amari, S.-i.: Differential-geometrical methods in statistics, vol. 28, Springer Science & Business Media, 2012.

Amezcua, J., Ide, K., Kalnay, E., and Reich, S.: Ensemble transform Kalman–Bucy filters, Quarterly Journal of the Royal Meteorological Society, 140, 995–1004, 2014.

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, Journal of climate, 9, 1518–1530, 1996.

Anderson, J. L.: A non-Gaussian ensemble filter update for data assimilation, Monthly Weather Review, 138, 4186–4198, 2010.

Beezley, J. D. and Mandel, J.: Morphing ensemble Kalman filters, Tellus A: Dynamic Meteorology and Oceanography, 60, 131–140, 2008.

Benamou, J.-D. and Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, Numerische Mathematik, 84, 375–393, 2000.

Berardi, M., Andrisani, A., Lopez, L., and Vurro, M.: A new data assimilation technique based on ensemble Kalman filter and Brownian bridges: an application to Richards' equation, Computer Physics Communications, 208, 43–53, 2016.

Bocquet, M., Pires, C. A., and Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation, Monthly Weather Review, 138, 2997–3023, 2010.

Brenier, Y.: Décomposition polaire et réarrangement monotone des champs de vecteurs, CR Acad. Sci. Paris Sér. I Math., 305, 805–808, 1987.

Burgers, G., Van Leeuwen, P. J., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Monthly Weather Review, https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2, 1998.

Carrassi, A. and Vannitsem, S.: Accounting for model error in variational data assimilation: A deterministic formulation, Monthly Weather Review, 138, 3369–3386, 2010.

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

Chen, B., Dang, L., Gu, Y., Zheng, N., and Principe, J. C.: Minimum Error Entropy Kalman Filter, IEEE Transactions on Systems, Man, and Cybernetics: Systems, https://doi.org/10.1109/tsmc.2019.2957269, 2019a.

Chen, Y., Georgiou, T. T., and Tannenbaum, A.: Optimal transport for Gaussian mixture models, IEEE Access, 7, 6269–6278, 2019b.

Cramér, H.: Mathematical methods of statistics, vol. 9, Princeton university press, 1999.

Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport, in: Advances in neural information processing systems, pp. 2292–2300, 2013.

De Lannoy, G. J., Houser, P. R., Pauwels, V. R., and Verhoest, N. E.: State and bias estimation for soil moisture profiles by an ensemble Kalman filter: Effect of assimilation depth and frequency, Water resources research, 43, 2007a.

De Lannoy, G. J., Reichle, R. H., Houser, P. R., Pauwels, V., and Verhoest, N. E.: Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter, Water Resources Research, 43, 2007b.

Dee, D. P.: Detection and correction of model bias during data assimilation, Meteorological Training Course Lecture Series (ECMWF), 2003.

Dee, D. P.: Bias and data assimilation, Quarterly Journal of the Royal Meteorological Society, 131, 3323–3343, 2005.

Doucet, A. and Johansen, A. M.: A tutorial on particle filtering and smoothing: Fifteen years later, Handbook of nonlinear filtering, 12, 3, 2009.

Drécourt, J.-P., Madsen, H., and Rosbjerg, D.: Bias aware Kalman filters: Comparison and improvements, Advances in Water Resources, 29, 707–718, 2006.

Ebtehaj, A. M. and Foufoula-Georgiou, E.: On variational downscaling, fusion, and assimilation of hydrometeorological states: A unified framework via regularization, Water Resources Research, 49, 5944–5963, https://doi.org/10.1002/wrcr.20424, 2013.

Ebtehaj, A. M., Zupanski, M., Lerman, G., and Foufoula-Georgiou, E.: Variational data assimilation via sparse regularisation, Tellus A: Dynamic Meteorology and Oceanography, 66, 21 789, 2014.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, 1994.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.

Feyeux, N., Vidard, A., and Nodet, M.: Optimal transport for variational data assimilation, Nonlinear Processes in Geophysics, 25, 55–66, 2018.

Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié, in: Annales de l'institut Henri Poincaré, vol. 10, pp. 215–310, 1948.

Furtado, H. C. M., de Campos Velho, H. F., and Macau, E. E. N.: Data assimilation: Particle filter and artificial neural networks, in: Journal of Physics: Conference Series, vol. 135, p. 012073, IOP Publishing, 2008.

Goodliff, M., Amezcua, J., and Van Leeuwen, P. J.: Comparing hybrid data assimilation methods on the Lorenz 1963 model with increasing non-linearity, Tellus A: Dynamic Meteorology and Oceanography, 67, 26 928, 2015.

Gordon, N. J., Salmond, D. J., and Smith, A. F.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation, in: IEE proceedings F (radar and signal processing), vol. 140, pp. 107–113, IET, 1993.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550–560, 2001.

Han, X. and Li, X.: An evaluation of the nonlinear/non-Gaussian filters for the sequential data assimilation, Remote Sensing of Environment, 112, 1434–1449, 2008.

Hellinger, E.: Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen., Journal für die reine und angewandte Mathematik (Crelles Journal), 1909, 210–271, 1909.

Hurkmans, R., Paniconi, C., and Troch, P. A.: Numerical assessment of a dynamical relaxation data assimilation scheme for a catchment hydrological model, Hydrological Processes: An International Journal, 20, 549–563, 2006.

Jordan, R., Kinderlehrer, D., and Otto, F.: The variational formulation of the Fokker-Planck equation, SIAM Journal on Mathematical Analysis, https://doi.org/10.1137/S0036141096303359, 1998.

Kalman, R. E.: A new approach to linear filtering and prediction problems, Journal of basic Engineering, 82, 35–45, 1960.

Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge university press, 2003.

Kantorovich, L. V.: On the translocation of masses, in: Dokl. Akad. Nauk. USSR (NS), vol. 37, pp. 199–201, 1942.

Kim, S., Eyink, G. L., Restrepo, J. M., Alexander, F. J., and Johnson, G.: Ensemble filtering for nonlinear dynamics, Monthly Weather Review, 131, 2586–2594, 2003.

Kollat, J., Reed, P., and Rizzo, D.: Addressing model bias and uncertainty in three dimensional groundwater transport forecasts for a physical aquifer experiment, Geophysical research letters, 35, 2008.

Kullback, S. and Leibler, R. A.: On information and sufficiency, The annals of mathematical statistics, 22, 79–86, 1951.

Kutta, W.: Beitrag zur naherungsweisen Integration totaler Differentialgleichungen, Z. Math. Phys., 46, 435–453, 1901.

Lauritzen, S. L.: Statistical manifolds, Differential geometry in statistical inference, 10, 163–216, 1987.

Le Dimet, F.-X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, Tellus A: Dynamic Meteorology and Oceanography, 38, 97–110, 1986.

Li, H., Kalnay, E., and Miyoshi, T.: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 135, 523–533, 2009.

Li, T., Bolic, M., and Djuric, P. M.: Resampling methods for particle filtering: classification, implementation, and strategies, IEEE Signal processing magazine, 32, 70–86, 2015.

Lorenc, A. C.: Analysis methods for numerical weather prediction, Quarterly Journal of the Royal Meteorological Society, 112, 1177–1194, 1986.

Lorenz, E. N.: Deterministic nonperiodic flow, Journal of the atmospheric sciences, 20, 130–141, 1963.

Mandel, J. and Beezley, J. D.: An ensemble Kalman-particle predictor-corrector filter for non-Gaussian data assimilation, in: International Conference on Computational Science, pp. 470–478, Springer, 2009.

McCann, R. J.: A convexity principle for interacting gases, Advances in mathematics, 128, 153–179, 1997.

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., and Ma, P.: Large-scale optimal transport map estimation using projection pursuit, in: Advances in Neural Information Processing Systems, edited by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., vol. 32, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2019/file/4bbbe6cb5982b9110413c40f3cce680b-Paper.pdf, 2019.

Miller, R. N., Ghil, M., and Gauthiez, F.: Advanced data assimilation in strongly nonlinear dynamical systems, Journal of the atmospheric sciences, 51, 1037–1056, 1994.

Monge, G.: Mémoire sur la théorie des déblais et des remblais, Histoire de l'Académie Royale des Sciences de Paris, 1781.

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R.: Dual state–parameter estimation of hydrological models using ensemble Kalman filter, Advances in water resources, 28, 135–147, 2005.

Nakano, S., Ueno, G., and Higuchi, T.: Merging particle filter for sequential data assimilation, Nonlinear Processes in Geophysics, https://doi.org/10.5194/npg-14-395-2007, 2007.

Ning, L., Carli, F. P., Ebtehaj, A. M., Foufoula-Georgiou, E., and Georgiou, T. T.: Coping with model error in variational data assimilation using optimal mass transport, Water Resources Research, 50, 5817–5830, 2014.

Orlin, J. B.: A faster strongly polynomial minimum cost flow algorithm, Operations research, 41, 338–350, 1993.

Otto, F.: The geometry of dissipative evolution equations: The porous medium equation, Communications in Partial Differential Equations, https://doi.org/10.1081/PDE-100002243, 2001.

Park, S. K. and Županski, D.: Four-dimensional variational data assimilation for mesoscale and storm-scale applications, Meteorology and Atmospheric Physics, 82, 173–208, 2003.

Pennec, X.: Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements, Journal of Mathematical Imaging and Vision, https://doi.org/10.1007/s10851-006-6228-4, 2006.

Peyré, G. and Cuturi, M.: Computational optimal transport, Foundations and Trends in Machine Learning, https://doi.org/10.1561/2200000073, 2019.

Peyré, G., Cuturi, M., et al.: Computational optimal transport, Foundations and Trends® in Machine Learning, 11, 355–607, 2019.

Pires, C., Vautard, R., and Talagrand, O.: On extending the limits of variational assimilation in nonlinear chaotic systems, Tellus A, 48, 96–121, 1996.

Pires, C. A., Talagrand, O., and Bocquet, M.: Diagnosis and impacts of non-Gaussianity of innovations in data assimilation, Physica D: Nonlinear Phenomena, 239, 1701–1717, 2010.

Poterjoy, J. and Anderson, J. L.: Efficient assimilation of simulated observations in a high-dimensional geophysical system using a localized particle filter, Monthly Weather Review, 144, 2007–2020, 2016.

Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., and Rao, C. R.: Linear statistical inference and its applications, vol. 2, Wiley New York, 1973.

Ravela, S., Emanuel, K., and McLaughlin, D.: Data assimilation by field alignment, Physica D: Nonlinear Phenomena, 230, 127–145, 2007.

Reich, S.: A nonparametric ensemble transform method for Bayesian inference, SIAM Journal on Scientific Computing, 35, A2013–A2024, 2013.

Reichle, R. H. and Koster, R. D.: Bias reduction in short records of satellite soil moisture, Geophysical Research Letters, 31, 2004.

Reichle, R. H., McLaughlin, D. B., and Entekhabi, D.: Hydrologic data assimilation with the ensemble Kalman filter, Monthly Weather Review, 130, 103–114, 2002.

Runge, C.: Über die numerische Auflösung von Differentialgleichungen, Mathematische Annalen, 46, 167–178, 1895.

Spiller, E. T., Budhiraja, A., Ide, K., and Jones, C. K.: Modified particle filter methods for assimilating Lagrangian data into a point-vortex model, Physica D: Nonlinear Phenomena, 237, 1498–1506, 2008.

Talagrand, O. and Courtier, P.: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory, Quarterly Journal of the Royal Meteorological Society, 113, 1311–1328, 1987.

Tamang, S. K., Ebtehaj, A., Zou, D., and Lerman, G.: Regularized Variational Data Assimilation for Bias Treatment using the Wasserstein Metric, Quarterly Journal of the Royal Meteorological Society, 146, 2332–2346, 2020.

Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., and Fablet, R.: Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system, in: Machine learning and data mining approaches to climate science, pp. 3–12, Springer, 2015.

Tarantola, A.: Inverse problem theory: methods for data fitting and model parameter estimation., Inverse problem theory: methods for data fitting and model parameter estimation., https://doi.org/10.1016/0031-9201(89)90124-6, 1987.

Trevisan, A., D'Isidoro, M., and Talagrand, O.: Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 136, 487–496, 2010.

Tsuyuki, T. and Miyoshi, T.: Recent progress of data assimilation methods in meteorology, Journal of the Meteorological Society of Japan. Ser. II, 85, 331–361, 2007.

Van Leeuwen, P. J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Quarterly Journal of the Royal Meteorological Society, 136, 1991–1999, 2010.

Van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, 2019.

Villani, C.: Topics in optimal transportation, 58, American Mathematical Soc., 2003.

Walker, J. P., Willgoose, G. R., and Kalma, J. D.: One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application, Journal of Hydrometeorology, 2, 356–373, 2001.

Woodbury, M. A.: Inverting modified matrices, Statistical Research Group, 1950.

Zhang, X., Heemink, A., and Van Eijkeren, J.: Data assimilation in transport models, Applied mathematical modelling, 21, 2–14, 1997.