Nonlinear Processes
in Geophysics

Open Access

EGU

Bridging classical data assimilation and optimal transport: the 3D-Var case

Marc Bocquet, Pierre J. Vanderbecken, Alban Farchi, Joffrey Dumont Le Brazidec, and Yelva Roustan

CEREA, École des Ponts and EDF R&D, Île-de-France, France

Correspondence: Marc Bocquet (marc.bocquet@enpc.fr)

Abstract. Because optimal transport (OT) acts as displacement interpolation in physical space rather than as interpolation in value space, it can avoid double-penalty errors generated by mislocations of geophysical fields. As such, it provides a very attractive metric for non-negative, sharp field comparison – the Wasserstein distance – which could further be used in data assimilation (DA) for the geosciences. However, the algorithmic and numerical implementations of such a distance are not straightforward. Moreover, its theoretical formulation within typical DA problems faces conceptual challenges, resulting in scarce contributions on the topic in the literature.

We formulate the problem in a way that offers a unified view with respect to both classical DA and OT. The resulting *OTDA* framework accounts for both the classical source of prior errors, background and observation, and a Wasserstein barycentre in between states which are pre-images of the background state and observation vector. We show that the hybrid OTDA analysis can be decomposed as a simpler OTDA problem involving a single Wasserstein distance, followed by a Wasserstein barycentre problem that ignores the prior errors and can be seen as a *McCann interpolant*. We also propose a less enlightening but straightforward solution to the full OTDA problem, which includes the derivation of its analysis error covariance matrix. Thanks to these theoretical developments, we are able to extend the classical 3D-Var/BLUE (best linear unbiased estimator) paradigm at the core of most classical DA schemes. The resulting formalism is very flexible and can account for sparse, noisy observations and non-Gaussian error statistics. It is illustrated by simple one- and two-dimensional examples that show the richness of the new types of analysis offered by this unification.

1 Introduction

1.1 Data assimilation and the double-penalty issue

Geophysical data assimilation (DA) is a set of methods and algorithms at the intersection of Earth sciences, mathematics, and computer science that are designed to enhance our understanding and predictive capability with respect to the complex systems that govern our planet (Carrassi et al., 2018). For example, these systems encompass the atmosphere, ocean, atmospheric chemistry and biogeochemistry, land surfaces, glaciology, hydrology, and climate system as a whole. DA is meant to optimally combine all sources of quantitative information, typically past and present observations, and numerical and statistical models of the system under consideration. DA is critical in forecasting chaotic geofluids by resetting the initial conditions of the flow, estimating physical and statistical parameters of the models, and providing a quantitative reanalysis of the past history of the climate system over decades. Because classical DA is applied to complex and high-dimensional dynamics, the DA algorithms often result from a compromise between the sophistication of the employed mathematical techniques and their numerical scalability and efficiency (Kalnay, 2003; Asch et al., 2016; Evensen et al., 2022). For instance, it is well-known that most DA methods are built around or from an update step – the analysis – where observations and background states are combined, an operation which often relies on Gaussian statistical assumptions.

Here, we would like to focus on one important issue that impacts classical DA, known as the *double-penalty error* in the geosciences. The double-penalty issue refers to the over-penalisation of errors in both the model and observational data (e.g. Amodei and Stein, 2009) and compromises the balance required for effective DA. It often stems from the mislocation of fields, which is caused by model error in either the forecasting or observation operator. A typical example is given by the slight mislocation of a plume of pollutant resulting in high predicted concentration values at positions where no pollutant is observed, while the model misses the observed concentration peaks (Farchi et al., 2016). This mismatch is heavily penalised due to the use, over the same discretised space, of the root-mean-square error (RMSE) for a point-by-point comparison. Figure 1 shows an exemplar of double-penalty error resulting in the inability to properly evaluate a model and learn from an analysis increment. This double-penalty error, a very common contribution to the *representation error* (Janjić et al., 2018), is ubiquitous in the geosciences, for example, in numerical weather prediction (in particular for water vapour), in atmospheric chemistry and air quality, in biogeochemistry, and in eddy-resolving ocean forecasting. This especially applies to sharp fields, whereas it may be of less relevance for smoother, larger-scale fields such as temperature.

It has been recognised that, although it can handle amplitude and smoothness mismatch, the weighted Euclidean (Mahalanobis) distance cannot cope with mislocation error; thus, it cannot account for the full *distortion* between mismatched fields (Hoffman et al., 1995). In the field of precipitation verification, one would alternatively speak of amplitude, structure, and location errors (e.g. Wernli et al., 2008). Hence, even though tuning covariances of Gaussian error distributions as in classical DA, such as increasing the correlation length, might help mitigate the double-penalty error, it is insufficient. In Fig. 1, one might replace the Euclidean norm with a weighted Euclidean one with a large correlation length. This would yield similar norm values for both cases. Unfortunately, it is not difficult to show that, in this limit, this (almost singular) norm can only distinguish between the spatial mean of both fields: it became blunt with no discriminating power. With respect to DA, Feyeux et al. (2018, their Fig. 1) also illustrate why the Euclidean distance cannot properly cope with mislocation error. Note that, if Feyeux et al. (2018) had used a weighted Euclidean distance instead, with the same covariance matrix for the two contributions of the cost function, the resulting analysis state would have been the same and, in particular, independent of the covariance matrix. A similar but two-dimensional illustration is given by Vanderbecken et al. (2023, their Fig. 3).

1.2 Nonlocal verification metrics

The issue can be attributed to the use of a *local* verification metric, meaning that it compares, through the RMSE, values at the same site, of the same grid cell. Thus, this issue goes beyond DA and pertains to the use of local metrics.

To avoid being impacted by the double-penalty issue stemming from the use of local verification metrics, smarter *non-local* or *multiscale* metrics have been proposed. A typical metric of this kind consists of the combination of a displacement map followed by the use of classical norm such as the RMSE (Hoffman et al., 1995; Keil and Craig, 2009). In this vein, effective verification metrics can be based on *optical flow-based warping* or on *deformed meshes*, prior to using classical norms (Gilleland et al., 2010a, b). These metrics can also be designed as *scale-dependent* and possibly *multiscale*, based on an empirical separation of scales, such as with *fuzzy* metrics (Ebert, 2008; Amodei and Stein, 2009) or *wavelets* (Briggs and Levine, 1997). They can be designed to grasp and quantify objects and features, such as lows and highs (Davis et al., 2006a, b; Lack et al., 2010). Metrics with similar capabilities (but not necessarily based on a displacement concept) have been introduced in computer vision, such as the *structural similarity index* (Zhou et al., 2004), or in the verification of precipitation (Wernli et al., 2008; Skok, 2023; Necker et al., 2023).

One of the most elegant approaches is based on the theory of *optimal transport* (OT) and the associated *Wasserstein distance*, which sits on solid mathematical foundations and significant developments; these are the main reasons why we will focus on OT in the following. Examples of the application of OT to the verification of tracer and greenhouse gases models are given in Farchi et al. (2016) and Vanderbecken et al. (2023).

## 1.3 Optimal transport and the Wasserstein distance

Before mentioning applications of the Wasserstein distance in the field of geoscience, let us first give a very brief introduction to the concept and mathematical formulation of OT.

The OT concept stems from an engineering, although rather universal, problem. Gaspard Monge (Monge, 1781) considered the *earth mover* problem, the goal of which is to efficiently move rubble to an embankment of about the same volume (see Fig. 2). Each displacement of a bit of earth has a known cost, so that the goal is to find the cheapest deterministic map that completely moves the rubble to the embankment. In mathematical terms, the goal is to find the map of minimal cost that transports the origin measure ρ_o to the target measure ρ_b; *measure* here means that both of them are non-negative and are integrable of integral 1. Note that the value 1 is arbitrary here and can be changed to $m > 0$, provided that this is the mass of both ρ_o and ρ_b. The cost is defined by a non-negative function $\mathcal{C}_{bo}$ of two variables (one for the origin space and the other for the target space). Let us assume a quadratic cost, defined for any couple of points (x, y) of a geometric domain Ω:

$$\mathcal{C}_{bo}(x, y) = \|x - y\|_2^2, \tag{1}$$
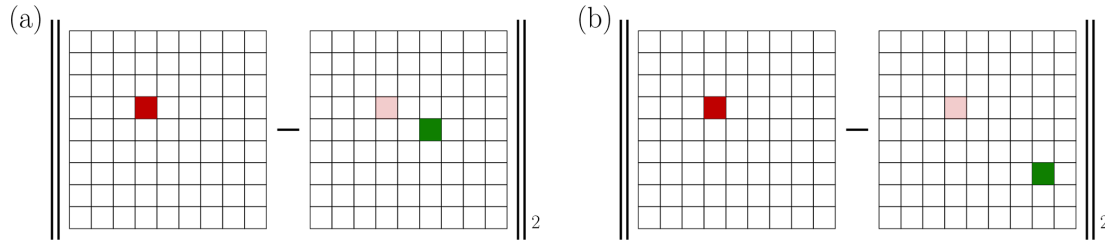
Figure 1. These two panels schematise the computation of the RMSE of two analysis increments. These increments are the difference between the truth (left mesh in between both pairs of the norm delimiters), concentrated here in the red grid cell, and the analysis, located in the green grid cells (right mesh in between both pairs of the norm delimiters). The increment in panel **(a)** is the outcome of a better analysis that is spatially closer to the truth, compared with that in panel **(b)**; however, both increments yield the same RMSE. Hence, this verification metric is impacted by the double-penalty error and does not help discriminate location errors.
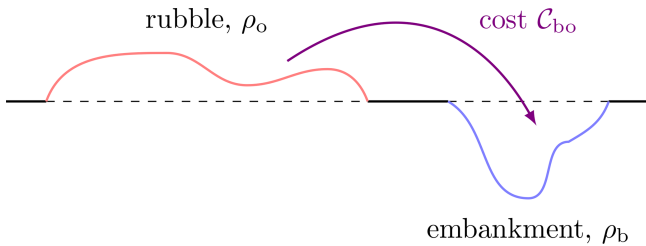
Figure 2. Illustration of the earth mover problem introduced by Monge in 1781 (see bulk of paper).

where $\|\cdot\|_2$ is the Euclidean norm. Let us define the set of all admissible differentiable maps T that transport ρ_o to ρ_b:

$$\mathcal{U}_{bo} = \{T : \Omega \longmapsto \Omega, \quad \rho_o = |\partial_x T| \, \rho_b \circ T\}, \tag{2}$$

where $|\partial_x T|$ is the absolute value of the determinant of the Jacobian of T, a factor which accounts for the deformation of the measure by the globally mass-conserving T. The *square* of the Wasserstein distance $\mathcal{W}_{\mathcal{C}_{bo}}$ is then defined by the following:

$$\mathcal{W}^2_{\mathcal{C}_{bo}} (\rho_o, \rho_b) = \min_{T \in \mathcal{U}_{bo}} \int_\Omega \mathcal{C}_{bo}(x, T(x)) \, \rho_o(x) \mathrm{d}x. \tag{3}$$

Here, the purpose is to minimise the total transport cost between ρ_o and ρ_b, and the optimal map T is often referred to as the *Monge map*. It can be shown that $\mathcal{W}_{\mathcal{C}_{bo}}$ is indeed a proper mathematical distance. The mathematical formulation is deceptively simple, as it is elegant, concise, and easy to grasp, but its theoretical and numerical solutions are far from trivial.

In the 20th century, a breakthrough was made by Leonid Kantorovich, who promoted the Monge problem to a probabilistic formulation. From his point of view, a bit of earth can be split and moved to many sites of the target measure support. Thus, the deterministic map T is replaced with a probabilistic measure π defined over $\Omega \times \Omega$. Hereafter, such a π is called a *transference plan*. An admissible transference plan

is integrable and has ρ_o and ρ_b as one-variable marginals; hence, the definition of the admissible set is as follows:

$$\mathcal{V}_{bo} = \left\{ \pi : \Omega \times \Omega \longmapsto \mathbb{R}_+, \quad \rho_o(x) = \int_\Omega \pi(x, y) \mathrm{d}y, \right.$$
$$\left. \rho_b(y) = \int_\Omega \pi(x, y) \mathrm{d}x \right\}. \tag{4}$$

As opposed to the deterministic Monge maps, the transference plans offer a symmetrical view of the origin and target space and their measures. An illustration of a discrete transference plan is given in Fig. 3. From this view, the squared Wasserstein distance can be reformulated as follows:

$$\mathcal{W}^2_{\mathcal{C}_{bo}} (\rho_o, \rho_b) = \min_{\pi \in \mathcal{V}_{bo}} \int_{\Omega \times \Omega} \mathcal{C}_{bo}(x, y) \pi(x, y) \mathrm{d}x \mathrm{d}y. \tag{5}$$

Equations (3) and (5) are the main continuous formulations of OT. In the rest of the paper, we will deal instead with *discrete* related formulations, which are more tangible and amenable to algorithmic and numerical implementations.

The field has attracted a lot of attention from pure and applied mathematicians as well as computer scientists. A complete introduction to the topic by experts can be found in the stimulating text books by Vilani (2003, 2009) and Peyré and Cuturi (2019). Peyré and Cuturi (2019) provide concrete examples, numerical methods, and a broad coverage of the topic from the perspective of applied mathematicians and computer scientists; hence, their work will be referred to quite often in the rest of the paper.

1.4 Nonlocal, multiscale metrics and data assimilation

Let us now go back to DA and narrow our focus to the use of advanced metrics in DA. Accounting for displacement error in DA, and hence relying on nonlocal verification metrics, has been advocated by Hoffman and Grassotti (1996), Ravela et al. (2007), and Plu (2013). Metrics built on a multiscale
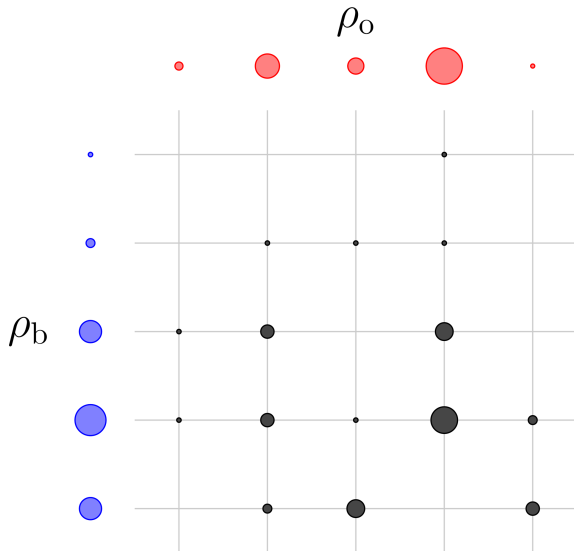
Figure 3. A representation of a discrete transport plan between two discrete origin (blue) and target (red) measures. The black dots represent the value of the transference plan. The radius of the dots is proportional to the values of these measures. This transference plan is checked to be admissible but is not necessarily optimal.

analysis of the fields to achieve a similar goal have been proposed by Ying (2019) and Ying et al. (2023).

The Wasserstein distance and closely related formulations have been advocated in the flow formulation of the analysis (DA update) to seamlessly transport the prior to the posterior (El Moselhy and Marzouk, 2012; Oliver, 2014; Marzouk et al., 2017; Farchi and Bocquet, 2018; Tamang et al., 2020). It can, for instance, be used to adjust the posterior discrete probability density functions (pdfs) in the particle filter. It has similarly been used to assist ensemble DA (Tamang et al., 2021, 2022). Finally, it has also, very recently, been used to compare forecast ensembles for sub-seasonal prediction (Le Coz et al., 2023; Lledó et al., 2023) or precipitation (Duc and Sawada, 2024).

In the context of this paper, it is *critical* to be aware that the use of OT in practical DA has, thus far, focused on applying OT independently to the pdf of each single scalar variable. Quite often, OT is applied to the pdf of a single random variable for the following two reasons:

- OT in one dimension (the space of the values taken by this random variable), with a quadratic cost, has a very simple solution that only relies on the cumulative distribution functions of the origin and target measures (see e.g. Remark 2.30 in Peyré and Cuturi, 2019), a technique known in statistics as *quantile matching*.

- Increasing the number of random variables is subject to the curse of dimensionality, necessitating an exponential increase in computational resources when increasing the resolution of the discretised fields.

This is very different from our context and objective where the objects dealt with by OT are (non-negative) physical field states, not the pdf of one of their scalar variables.

1.5　Feyeux et al. (2016) proposal

The present paper stands more in the wake of the seminal proposals of Ning et al. (2014), Feyeux (2016), and Feyeux et al. (2018). Their idea is to replace the local metrics of classical variational DA, typically the square of the Euclidean distance (hence related to the L_2 norm), with the squared Wasserstein distance. This is intuitively what we are after in order to cope with mislocation errors mentioned in Sect. 1.1 in the context of DA. This should redefine the nature of the DA update step. Let us formalise this idea (Feyeux, 2016).

We will seize this opportunity to introduce some of our notation in the context of discrete DA, which is a widely adopted standpoint in the geosciences. Let us focus on a classical DA 3D-Var cost function (Daley, 1991):

$$G_{\mathrm{cl}}(x^{\mathrm{a}}) = \left\| y^{\mathrm{b}} - x^{\mathrm{a}} \right\|_2^2 + \left\| y^{\mathrm{o}} - \mathbf{H}x^{\mathrm{a}} \right\|_2^2, \qquad (6)$$

where $\|\cdot\|_2$ is the Euclidean norm, $y^{\mathrm{b}} \in \mathbb{R}^{N_{\mathrm{b}}}$ is the first guess, $y^{\mathrm{o}} \in \mathbb{R}^{N_{\mathrm{o}}}$ is the vector of observations, and $\mathbf{H}$ is the observation operator.[1] $x^{\mathrm{a}} \in \mathbb{R}^{N_{\mathrm{a}}}$ is the dummy variable of this optimisation problem whose optimal value corresponds to the DA state analysis. Now, the substitution of the Euclidean norm yields the new 3D-Var cost function:

$$G_{\mathrm{w}}(x^{\mathrm{a}}) = \mathcal{W}_2^2(y^{\mathrm{b}}, x^{\mathrm{a}}) + \mathcal{W}_2^2(y^{\mathrm{o}}, \mathbf{H}x^{\mathrm{a}}), \qquad (7)$$

where $\mathcal{W}_2$ is some discretisation of the Wasserstein distance based on the cost defined by the square of the Euclidean distance. Note that this 3D-Var case requires balancing two instances of a Wasserstein-based metric. The analysis state is known as a *Wasserstein barycentre* (abridged W-barycentre in the following).

Feyeux (2016) and Feyeux et al. (2018) explored the optimisation aspects of this DA problem. However, Feyeux (2016) ultimately pointed to a possible inconsistency in the definition of the DA problem formulated in Eq. (7), where the system is only partially observed (non-trivial $\mathbf{H}$). In the case where the system is fully observed, typically when $\mathbf{H}$ is the identity operator, the outcome of the optimisation problem, i.e. the analysis, matches our expectations. However, when the system is partially observed, inconsistencies are observed. Let us see why.

Figure 4a considers the DA problem based on Eq. (7), assuming that only half of the domain is observed. We have

---

[1]The notation y^{b} and y^{o} is at variance with the more familiar x^{b} and y notation of DA, respectively. However, this change will prove very useful in the following; it follows the idea that the full *information vector* is $y = [(y^{\mathrm{b}})^\top, (y^{\mathrm{o}})^\top]^\top$, whose components may benefit from homogeneous notation (Talagrand, 1997).

solved the corresponding mathematical and numerical problem as raised by Feyeux (2016) and displayed its solution. However, one observes that the mass of the solution concentrates on the observed subdomain and neglects the rest of the domain where the prior mainly concentrates, an outcome suspected by Feyeux (2016). Instead, we would have intuitively preferred a solution close to the one offered by Fig. 4b, whose formulation and numerical solution differ and follow the new approach developed in the present paper (how we obtained this solution will be described in Sect. 2).

The main caveat of Eq. (7) comes from the fact that the system is only partially observed, as well as the requirement that OT is *balanced*; i.e. the origin and target densities must have the same mass. This mass balance applies to both OT terms in Eq. (7), between y^{b} and x^{a} and between y^{o} and $\mathbf{H}x^{\mathrm{a}}$:

$$\mathrm{m}\left(x^{\mathrm{a}}\right) = \mathrm{m}\left(y^{\mathrm{b}}\right), \qquad \mathrm{m}\left(\mathbf{H}x^{\mathrm{a}}\right) = \mathrm{m}\left(y^{\mathrm{o}}\right). \qquad (8)$$

Here, the mass of a vector $x \in \mathbb{R}^N$ is defined by

$$\mathrm{m}(x) = \mathbf{1}^{\top} x = \sum_{i=1}^{N} x_i, \qquad (9)$$

with $\mathbf{1} \in \mathbb{R}^N$ hereafter defined as the vector of entries 1.

Now, if we further assume, for simplicity, that y^{b} and y^{o} have the same mass (which is the case in Fig. 4), then

$$\mathrm{m}\left(\mathbf{H}x^{\mathrm{a}}\right) = \mathrm{m}\left(y^{\mathrm{o}}\right) = \mathrm{m}\left(y^{\mathrm{b}}\right) = \mathrm{m}\left(x^{\mathrm{a}}\right). \qquad (10)$$

As a result, we obtain $\mathrm{m}(\mathbf{H}x^{\mathrm{a}}) = \mathrm{m}(x^{\mathrm{a}})$, which is an undesired prior piece of information as to where to find the mass of x^{a}. Simply put, unless the system is fully observed, this approach amounts to *the streetlight effect*. This is precisely what happens in Fig. 4a with the undesired concentration of the mass of x^{a} close to the edge of the observed subdomain.

To overcome this caveat and find a proper alternative to Eq. (7), we need (i) to renounce comparing the fields in observation space (in the observation discrepancy term of the cost function) and (ii) to introduce *unbalanced OT* (i.e. we need to be able to accommodate states of distinct masses). In the computer science context of pure OT, the latter has been discussed by Chizat et al. (2018). However, our solution differs formally and will be DA-centric.

1.6   Objective and outline

The objective of this paper is to lift the objection of Feyeux (2016) and propose a DA framework based on the Wasserstein distance, thereby offering a consistent way to bridge OT and classical DA. The new formalism will be referred to as *hybrid OTDA* (hybrid optimal transport data assimilation) in the rest of this paper (or *OTDA* for brevity). We will focus on the definition of a 3D-Var DA problem and how to obtain its analysis state and the associated analysis error covariance matrix.

At least within the perimeter of this paper, some restrictions apply. Firstly, the physical fields considered in the DA problem are *non-negative* (concentration of tracer, pollutants, water vapour, hydrometeors, chemical and biogeochemical species, etc.). However, as opposed to Feyeux (2016), the methods of this paper do not require the (possibly noisy) background state y^{b} and observation y^{o} to be non-negative. We stress once again that the states of our DA problem are physical fields onto which OT is applied and are not meant to be a pdf of a random variable. Secondly, the observation operator $\mathbf{H}$ is assumed to be *linear*. This is only meant for convenience and to obtain a rigorous correspondence between the primal and dual cost functions of the 3D-Var case. Making this assumption is very common in geophysical DA: $\mathbf{H}$ can indeed be seen as the tangent linear of a nonlinear observation operator within the inner loop of a 3D-Var or a 4D-Var case (see, for instance, Courtier, 1997).

The outline of the paper is as follows. After the present introduction (Sect. 1), Sect. 2 discloses our main idea and discusses two mathematical paths to solve the underlying optimisation problem; the first path is enlightening but not necessarily practical, whereas the alternative path is direct and robust but hides some of the concepts behind it. Section 3 provides one- and two-dimensional illustrations of a 3D-Var analysis based on the new hybrid OTDA formalism. These illustrations will show the possibilities and flexibility of the new framework. Importantly, Sect. 3 will also depict classical DA as a limit case of the formalism. In Sect. 4, the second-order analysis, i.e. the uncertainty quantification of the OTDA 3D-Var case, is derived, discussed, and illustrated. Conclusions and perspectives are given in Sect. 5.

2   The main proposal

2.1   Notation and conventions

Non-negative vectors x of size N are called *discrete measures*; they lie in the orthant $\mathcal{O}_N^+ \triangleq \mathbb{R}_+^N$. Although most mathematical OT theories work on *normalised* discrete measures, yielding *probability vectors*, this assumption will not be needed in this paper. The open subset of $\mathcal{O}_N^+$ of all the positive discrete measures will be denoted as $\mathcal{O}_N^{+,\star} \triangleq \mathbb{R}_{+,\star}^N$.

We will distinguish the *observations* $y^{\mathrm{b}} \in \mathbb{R}^{\mathcal{N}_{\mathrm{b}}}$ and $y^{\mathrm{o}} \in \mathbb{R}^{\mathcal{N}_{\mathrm{o}}}$ from the observable states $x^{\mathrm{b}} \in \mathcal{O}_{N_{\mathrm{b}}}^+$ and $x^{\mathrm{o}} \in \mathcal{O}_{N_{\mathrm{o}}}^+$. y^{b}, which corresponds to the first guess of conventional DA, and y^{o}, which corresponds to the traditional observation vector, are known before solving the 3D-Var problem. These vectors embody all of the information processed in the analysis. By contrast, the observables x^{b} and x^{o}, which are related to y^{b} and y^{o}, respectively, through an observation operator (the identity for y^{b} and x^{b}), are not known a priori. They will be estimated along with the analysis state $x^{\mathrm{a}} \in \mathcal{O}_{N_{\mathrm{a}}}^+$. Note that these vectors may well lie in distinct vector spaces of different dimensions; hence, the introduction of as many dimen-
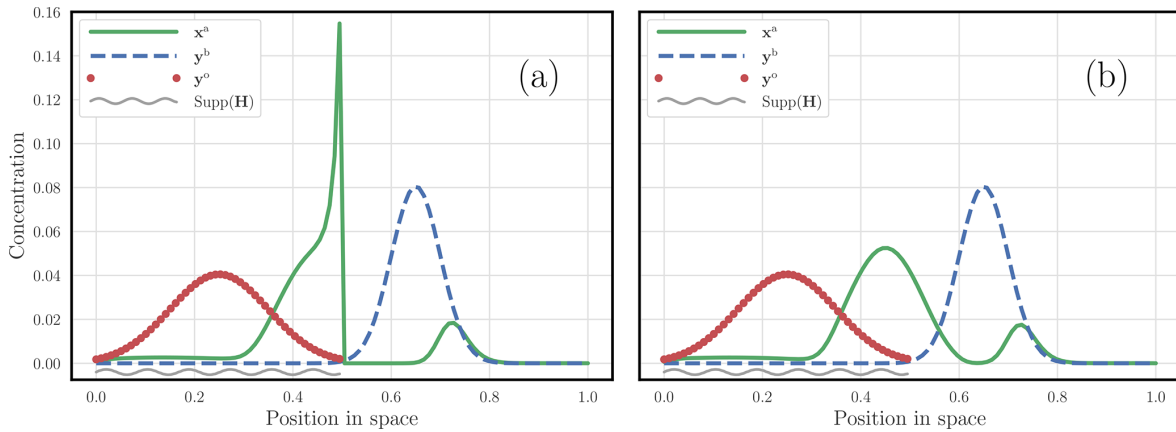
Figure 4. The panels illustrate the analysis of a 3D-Var case that relies on the Wasserstein distance rather than a local metric. The red dots represent the observations, while the dashed blue curve represents the background state. The observations are only focused on the left half of the domain. The solution of the optimisation problem in Eq. (7) is displayed as a solid green curve in panel **(a)**. The solution of the optimisation problem that we will propose in this paper is displayed as a solid green curve in panel **(b)**. The support of the observation is suggested using a wavy grey segment. These states are typically one-dimensional puff pollutant concentrations. They should not be confused with pdfs of a single random variable.

sions N_b, N_o, $\mathcal{N}_\mathrm{b}$, and $\mathcal{N}_\mathrm{o}$. $x_i^\star$ can be seen as the value taken by $\boldsymbol{x}^\star$ at site $\boldsymbol{r}_i^\star$, for $\star = \mathrm{b}, \mathrm{o}$, and a and $i \in [\![1, N_\star]\!]$. Mind that the distinction between $\boldsymbol{y}^\mathrm{b}$ and $\boldsymbol{x}^\mathrm{b}$ and the introduction of $\boldsymbol{x}^\mathrm{o}$ are novelties of OTDA compared with classical DA.

Like in classical DA, the vectors $\boldsymbol{y}^\mathrm{b}$ and $\boldsymbol{y}^\mathrm{o}$ are subject to (prior) errors whose statistics are specified by the likelihoods $p(\boldsymbol{y}^\mathrm{b}|\boldsymbol{x}^\mathrm{b})$ and $p(\boldsymbol{y}^\mathrm{o}|\boldsymbol{x}^\mathrm{o})$, respectively. Up to constants that do not depend on $\boldsymbol{x}^\mathrm{b}, \boldsymbol{x}^\mathrm{o}, \boldsymbol{y}^\mathrm{b}$, and $\boldsymbol{y}^\mathrm{o}$, we assume the existence of ζ_b and ζ_o, such that

$$\ln p(\boldsymbol{y}^\mathrm{b}|\boldsymbol{x}^\mathrm{b}) \overset{\triangle}{=} -\zeta_\mathrm{b}(\boldsymbol{y}^\mathrm{b} - \boldsymbol{x}^\mathrm{b}) + \mathrm{cst}, \tag{11a}$$

$$\ln p(\boldsymbol{y}^\mathrm{o}|\boldsymbol{x}^\mathrm{o}) \overset{\triangle}{=} -\zeta_\mathrm{o}(\boldsymbol{y}^\mathrm{o} - \mathbf{H}\boldsymbol{x}^\mathrm{o}) + \mathrm{cst}. \tag{11b}$$

Thus, various error statistics can be considered. These errors are hypothesised to be mutually independent. The observation operator $\mathbf{H} : \mathcal{O}_{N_\mathrm{o}}^+ \longmapsto \mathbb{R}^{\mathcal{N}_\mathrm{o}}$ used in the definition of ζ_o is assumed to be linear. This qualification is for convenience and could be lifted if necessary. It is further assumed that ζ_b and ζ_o are strictly convex functions. This is, for instance, the case if we choose Gaussian error statistics yielding

$$\zeta_\mathrm{b}(\boldsymbol{e}_\mathrm{b}) = \frac{1}{2} \|\boldsymbol{e}_\mathrm{b}\|_{\mathbf{B}^{-1}}^2, \qquad \zeta_\mathrm{o}(\boldsymbol{e}_\mathrm{o}) = \frac{1}{2} \|\boldsymbol{e}_\mathrm{o}\|_{\mathbf{R}^{-1}}^2. \tag{12}$$

Here, $\|\boldsymbol{e}\|_{\mathbf{A}} = \sqrt{\boldsymbol{e}^\top \mathbf{A} \boldsymbol{e}}$. $\mathbf{B}$ is the positive definite background error covariance matrix and $\mathbf{R}$ is the positive definite observation error covariance matrix. Finally, in the following, the $\mathrm{m}(\star)$ operator will act not only on vectors but also, more generally, on any tensor, and it will return the sum of all of its entries.

2.2 Formalism of discrete optimal transport

To discretise and solve the continuous Kantorovich optimisation problem introduced in Sect. 1.3, we will need two el-

ementary pieces of information about OT. These are not the only techniques that we will leverage, but both represent cornerstones towards a numerical solution to our proposal and, hence, require a proper introduction.

2.2.1 The primal cost function

Let us consider two discrete measures $\boldsymbol{x}^\mathrm{b} \in \mathcal{O}_{N_\mathrm{b}}^+$ and $\boldsymbol{x}^\mathrm{o} \in \mathcal{O}_{N_\mathrm{o}}^+$ with the same mass:

$$m \overset{\triangle}{=} \mathrm{m}\left(\boldsymbol{x}^\mathrm{b}\right) = \mathrm{m}\left(\boldsymbol{x}^\mathrm{o}\right). \tag{13}$$

For convenience, $\mathcal{O}_{\mathrm{b,o}}^+$ will be used as an alias for the set $\mathcal{O}_{N_\mathrm{b} \times N_\mathrm{o}}^+$. A cost matrix $\mathbf{C}_\mathrm{bo} \in \mathcal{O}_{\mathrm{b,o}}^+$ is given. The optimisation problem will be formulated using discrete Kantorovich transference plans $\mathbf{P}^\mathrm{bo} \in \mathcal{O}_{\mathrm{b,o}}^+$. The optimal discrete transference plan is given by the minimiser of the following optimisation problem:

$$W_{\mathbf{C}_\mathrm{bo}}(\boldsymbol{x}^\mathrm{b}, \boldsymbol{x}^\mathrm{o}) \overset{\triangle}{=} \min_{\mathbf{P}^\mathrm{bo} \in \mathcal{U}_\mathrm{bo}(\boldsymbol{x}^\mathrm{b}, \boldsymbol{x}^\mathrm{o})} \mathrm{Tr}\left(\mathbf{C}_\mathrm{bo}^\top \mathbf{P}^\mathrm{bo}\right). \tag{14a}$$

Here, the trace sums up the costs attached to each path, and the set of admissible transference plans is defined by

$$\mathcal{U}_\mathrm{bo} \overset{\triangle}{=} \left\{ \mathbf{P} \in \mathcal{O}_{\mathrm{b,o}}^+ : \quad \mathbf{P}\mathbf{1}_\mathrm{o} = \boldsymbol{x}^\mathrm{b}, \quad \mathbf{P}^\top \mathbf{1}_\mathrm{b} = \boldsymbol{x}^\mathrm{o} \right\}, \tag{14b}$$

which selects the discrete transference plans with the proper marginals. $W_{\mathbf{C}_\mathrm{bo}}$ could be viewed as a discrete equivalent to *the square of* the Wasserstein distance $\mathcal{W}_{\mathcal{C}_\mathrm{bo}}^2$ introduced in Eq. (5).

2.2.2 Entropic regularisation

The optimisation problem in Eq. (14) is a *linear program* that is convex (Peyré and Cuturi, 2019, and references therein). However, it is not generally strictly convex and, hence, does not necessarily exhibit a single minimum. Adding to the difficulty, its cost function (Eq. 14a) is constrained. *Entropic regularisation* addresses these issues and is used here to lift the constraints and to render the problem strictly convex. In particular, it will force any state vector that is a solution of the problem to be positive. A comprehensive justification is given by Peyré and Cuturi (2019). More precisely, we will use a *Kullback–Leibler divergence* (KL) regularisation term that is inserted in Eq. (14a),

$$\mathrm{Tr}\left(\mathbf{C}_{\mathrm{bo}}^{\top}\mathbf{P}^{\mathrm{bo}}\right) \rightarrow \mathrm{Tr}\left(\mathbf{C}_{\mathrm{bo}}^{\top}\mathbf{P}^{\mathrm{bo}}\right) + \varepsilon\mathcal{K}(\mathbf{P}^{\mathrm{bo}}|\boldsymbol{\nu}^{\mathrm{bo}}), \tag{15}$$

which incorporates some prior transference plan $\boldsymbol{\nu}^{\mathrm{bo}}$ and does not require $\mathrm{m}\left(\mathbf{P}^{\mathrm{bo}}\right) = 1$, whereas Peyré and Cuturi (2019) opted for a basic entropy term. The KL term (Boyd and Vandenberghe, 2004) is defined by

$$\mathcal{K}(\boldsymbol{p}|\boldsymbol{q}) \overset{\Delta}{=} \sum_{i} p_i \ln \frac{p_i}{q_i} - p_i + q_i. \tag{16}$$

It can be checked that the Hessian of the regularised cost function (Eq. 15) is a diagonal matrix of coefficients $\varepsilon/P_{ij}^{\mathrm{bo}} \geq \varepsilon$ because $0 \leq P_{ij}^{\mathrm{bo}} \leq 1$, making the problem ε-strongly convex. We choose, for example, $\boldsymbol{\nu}^{\mathrm{bo}} = \boldsymbol{x}^{\mathrm{b}}(\boldsymbol{x}^{\mathrm{o}})^{\top}/m$ and $\varepsilon > 0$, the latter of which is the regularisation scalar parameter. Note that this particular $\boldsymbol{\nu}^{\mathrm{bo}}$ is an admissible transference plan, i.e. it belongs to $\mathcal{U}_{\mathrm{bo}}$, and can be interpreted as a complete statistical decoupling of the transference plan with respect to the origin and target discrete measures. In the limit $\varepsilon \rightarrow 0^{+}$ of vanishing regularisation, the solution should not depend on the choice of $\boldsymbol{\nu}^{\mathrm{bo}}$. However, the convergence to the solution at finite ε may depend on this choice. The primal cost function augmented with such an entropic regularisation is usually solved numerically with the iterative *Sinkhorn algorithm* (Sinkhorn, 1964). However, this is not the path followed in this paper, although we have used it as well.

Finally note that the technique to convexify such an optimisation problem with a KL term has been introduced in DA by Bocquet (2009) and Bocquet et al. (2011) following principles of statistical physics.

2.3 From classical data assimilation to hybrid optimal transport data assimilation

Figure 5 is a schematic representation of the flow of information in a classical DA update (and in particular 3D-Var), using the notation introduced above. In this case, the observables $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ and the analysis state $\boldsymbol{x}^{\mathrm{a}}$ are the same by construction; hence, $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ are not needed. This diagram, which could also be seen as a Bayesian network, corresponds
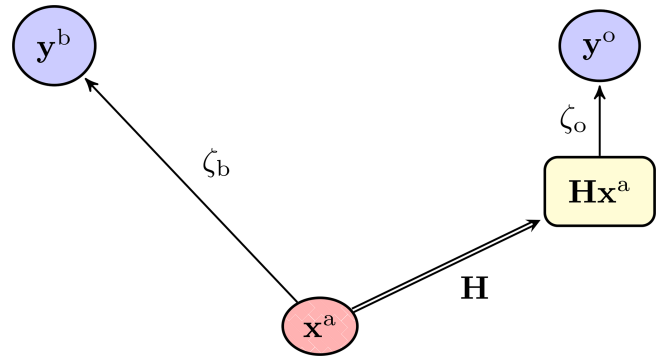
Figure 5. A diagrammatic representation of the classical 3D-Var update, with the observations $\boldsymbol{y}^{\mathrm{b}}$ (the first guess) and $\boldsymbol{y}^{\mathrm{o}}$ (the observation vector), the analysis state $\boldsymbol{x}^{\mathrm{a}}$, and the observed analysis $\mathbf{H}\boldsymbol{x}^{\mathrm{a}}$. A double-line arrow represents a deterministic map, whereas a single-line arrow represents a statistical binding between the origin and the target.

to the cost function

$$L_{\mathrm{cl}}(\boldsymbol{x}^{\mathrm{a}}) = \zeta_{\mathrm{b}}(\boldsymbol{y}^{\mathrm{b}} - \boldsymbol{x}^{\mathrm{a}}) + \zeta_{\mathrm{o}}(\boldsymbol{y}^{\mathrm{o}} - \mathbf{H}\boldsymbol{x}^{\mathrm{a}}) \tag{17}$$

to be minimised over $\boldsymbol{x}^{\mathrm{a}}$. Now let us make use of the observables $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ as new degrees of freedom but bind them by OTs to $\boldsymbol{x}^{\mathrm{a}}$, using the cost matrices $\mathbf{C}_{\mathrm{ba}}$ and $\mathbf{C}_{\mathrm{oa}}$, respectively.

This yields the diagram in Fig. 6, which corresponds to the cost function

$$L_{\mathrm{w}}(\boldsymbol{x}^{\mathrm{a}}) = \min_{\boldsymbol{x}^{\mathrm{b}}\in\mathcal{O}_{N_{\mathrm{b}}}^{+}, \boldsymbol{x}^{\mathrm{o}}\in\mathcal{O}_{N_{\mathrm{o}}}^{+}} \left\{ \zeta_{\mathrm{b}}(\boldsymbol{y}^{\mathrm{b}} - \boldsymbol{x}^{\mathrm{b}}) + \zeta_{\mathrm{o}}(\boldsymbol{y}^{\mathrm{o}} - \mathbf{H}\boldsymbol{x}^{\mathrm{o}}) \right.$$
$$\left. + W_{\mathbf{C}_{\mathrm{ba}}}(\boldsymbol{x}^{\mathrm{b}}, \boldsymbol{x}^{\mathrm{a}}) + W_{\mathbf{C}_{\mathrm{oa}}}(\boldsymbol{x}^{\mathrm{o}}, \boldsymbol{x}^{\mathrm{a}}) \right\}. \tag{18}$$

It must be minimised over $\boldsymbol{x}^{\mathrm{a}}$, yielding an analysis state $\boldsymbol{x}^{\mathrm{a}}$; this analysis state can also be seen as the W-barycentre between $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$. Note that $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ are discrete measures of unknown mass. For the optimisation problem, they lie in $\mathcal{O}_{N_{\mathrm{b}}}^{+}$ and $\mathcal{O}_{N_{\mathrm{o}}}^{+}$, respectively.

Moving from Eq. (17) to Eq. (18) following the principles and guidance of the introductory Sect. 1.4 is empirical, although no more than in Ning et al. (2014) and Feyeux (2016). Showing the merits of this move from Eq. (17) to Eq. (18) is the goal of the present paper. As opposed to Feyeux et al. (2018), it can deal with sparse and noisy observations, i.e. non-trivial $\mathbf{H}$. We will show that classical DA is embedded in this generalisation. Moreover, the merits of the new cost function will be a posteriori qualitatively supported by the outcome of the numerical experiments (to the expert's eyes), which improve over previous formalism's outcomes. We would like to point out that we have also developed a consistent probabilistic and Bayesian formalism fully supporting the introduction of Eq. (18). However, we felt that the derivation was too long and technical for this paper and would not be helpful in the exploration of the direct consequences of Eq. (18).
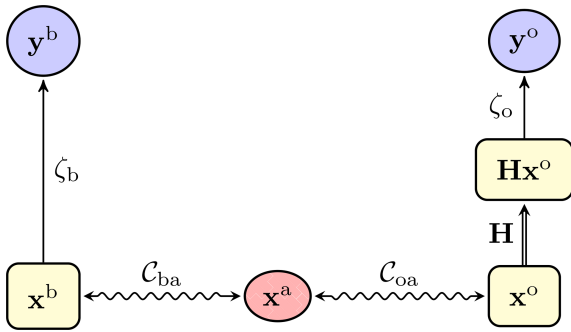
Figure 6. A diagrammatic representation of the hybrid OTDA 3D-Var update, with the observations y^b (the first guess) and y^o (the observation vector); the observables x^b and x^o; and x^a, which is the W-barycentre. A double-line arrow represents a deterministic map, a single-line arrow represents a statistical binding between the origin and the target, and a wavy line represents the *weaker* bindings of x^b with x^a and x^o with x^a through OTs. This diagram can be seen as an unfolding of that in Fig. 5.

We call Eq. (18) a *high-level* primal cost function because the metrics $W_{\mathbf{C}_\mathrm{ba}}$ and $W_{\mathbf{C}_\mathrm{oa}}$ have not yet been replaced by their transference plan expression, as opposed to, for example, Eq. (14a). Passing to a lower-level primal cost function would require expanding Eq. (18) using Eq. (14a) twice.

In the subsequent two subsections, we will investigate two pathways to solve the optimisation problem in Eq. (18). The first path (Sect. 2.4) unveils some of the key concepts behind its solution and partially disentangles the classical DA part from the W-barycentre part of the full analysis. This approach is enlightening but not necessarily practical. The second path is an alternative which is direct and robust but hides some of the fundamental principles underlying the solution. The busy reader could skip directly to the latter (i.e. Sect. 2.5).

2.4　Decomposition of the optimisation problem and effective cost metric

In this section, key ideas behind the minimisation of Eq. (18) are sketched and discussed. The level of mathematical rigour of this section is that of casual methodological DA in the geoscience literature. However, we stress that all of the algorithms discussed here have been successfully numerically tested on various configurations. The solution of Eq. (18) presented in this section is not necessarily robust, but it is enlightening and, hence, worth discussing.

Repeated contravariant indices – meaning the same tensor index is present as the upper and lower index – in tensor expressions will be understood as summed over, following Einstein's convention.

2.4.1　Dual formulation of the primal problem

One way, although not the only one, to write the explicit primal problem associated with Eq. (18) is through the use of a *gluing* transference plan $\mathbf{P}^\mathrm{boa} \in \mathcal{O}_{\mathrm{b,o,a}}^+$, where $\mathcal{O}_{\mathrm{b,o,a}}^+ = \mathbb{R}_+^{N_\mathrm{b} N_\mathrm{o} N_\mathrm{a}}$ (see pp. 11–12 of Vilani, 2009). $\mathbf{P}^\mathrm{boa} \in \mathcal{O}_{\mathrm{b,o,a}}^+$ is a 3-tensor whose marginals are x^b, x^o, and x^a and that glues the transference plans $\mathbf{P}^\mathrm{ba}$ between x^b and x^a and $\mathbf{P}^\mathrm{oa}$ between x^o and x^a:

$$\mathcal{L} = \min_{x^\mathrm{a} \in \mathcal{O}_\mathrm{a}^+} L_\mathrm{w}(x^\mathrm{a}) \tag{19a}$$

$$= \min_{x^\mathrm{b} \in \mathcal{O}_\mathrm{b}^+ \, x^\mathrm{o} \in \mathcal{O}_\mathrm{o}^+ \, x^\mathrm{a} \in \mathcal{O}_\mathrm{a}^+} \left[\zeta_\mathrm{b}(y^\mathrm{b} - x^\mathrm{b}) + \zeta_\mathrm{o}(y^\mathrm{o} - \mathbf{H}x^\mathrm{o}) \right.$$

$$\left. + \min_{\mathbf{P} \in \mathcal{U}_\mathrm{boa}} \left\{ P_{ijk} C_\mathrm{ba}^{ik} + P_{ijk} C_\mathrm{oa}^{jk} \right\} \right]. \tag{19b}$$

Here, the admissible set of (glued) transference plans, the set of all 3-tensors of non-negative entries whose marginals are x^b, x^o, and x^a, is defined by

$$\mathcal{U}_\mathrm{boa} \triangleq \left\{ \mathbf{P} \in \mathcal{O}_{\mathrm{b,o,a}}^+ : \quad \forall i, \; P_{ijk} 1_\mathrm{o}^j 1_\mathrm{a}^k = x_i^\mathrm{b}, \right.$$

$$\left. \forall j, \; P_{ijk} 1_\mathrm{b}^i 1_\mathrm{a}^k = x_j^\mathrm{o}, \quad \forall k, \; P_{ijk} 1_\mathrm{b}^i 1_\mathrm{o}^j = x_k^\mathrm{a} \right\}. \tag{19c}$$

Due to the hardly scalable dimensionality of the primal problem, based on either a 3-tensor or a couple of 2-tensors, we wish to derive a dual problem equivalent to the primal one, using Lagrange multipliers to lift the constraints with (as will be checked later) a significantly smaller dimensionality.

This leads to a series of transformations of the problem $\mathcal{L}$, from a Lagrangian to a dual cost function, which is reported in Appendix A for the mathematically inclined reader. The outcome is a dual problem which reads

$$\mathcal{L}^* = \max_{(f_\mathrm{b}, f_\mathrm{o}) \in \mathcal{U}_\mathrm{bo}^*(\mathbf{C}_\mathrm{ba}, \mathbf{C}_\mathrm{oa}, \mathbf{H})} \left\{ f_\mathrm{b}^\top y^\mathrm{b} + f_\mathrm{o}^\top y^\mathrm{o} \right.$$

$$\left. - \zeta_\mathrm{b}^*(f_\mathrm{b}) - \zeta_\mathrm{o}^*(f_\mathrm{o}) \right\}, \tag{20a}$$

where the $*$ symbol refers to *dual* and where the polyhedron $\mathcal{U}_\mathrm{bo}^*(\mathbf{C}_\mathrm{ba}, \mathbf{C}_\mathrm{oa}, \mathbf{H})$ is defined by

$$\mathcal{U}_\mathrm{bo}^*(\mathbf{C}_\mathrm{ba}, \mathbf{C}_\mathrm{oa}, \mathbf{H}) \triangleq \left\{ f_\mathrm{b} \in \mathbb{R}^{\mathcal{N}_\mathrm{b}}, \; f_\mathrm{o} \in \mathbb{R}^{\mathcal{N}_\mathrm{o}} : \right.$$

$$\left. \forall i, j, k, \quad f_\mathrm{b}^i + f_\mathrm{o}^l H_l^j \leq C_\mathrm{ba}^{ik} + C_\mathrm{oa}^{jk} \right\}. \tag{20b}$$

In Eq. (20), the maps ζ_b^* and ζ_o^* are the Legendre–Fenchel transforms of the maps ζ_b and ζ_o, respectively. Let us recall that the Legendre–Fenchel transform $f \longmapsto \zeta^*(f)$ of the map $e \longmapsto \zeta(e)$ is defined by $\zeta^*(f) = \sup_e \left\{ f^\top e - \zeta(e) \right\}$. For instance, in the case of Gaussian error statistics (as in Eq. 12), these transforms are given by

$$\zeta_\mathrm{b}^*(f_\mathrm{b}) = \frac{1}{2} \|f_\mathrm{b}\|_\mathbf{B}^2, \qquad \zeta_\mathrm{o}^*(f_\mathrm{o}) = \frac{1}{2} \|f_\mathrm{o}\|_\mathbf{R}^2. \tag{21}$$

Note that, in this section, we do not add the entropic regularisation to the cost functions for the sake of conciseness and

because it does not play a role in the key ideas developed in this section; however, it would likely be added and employed in numerical applications.

2.4.2 Decomposition of the dual problem

These transformations allow us to trade the primal for the dual problem. Due to the fact that there are N_a constraints indexed by $k \in [\![1, N_a]\!]$ for each f_b and f_o pair in $\mathcal{U}_{bo}^*$ and that the tightest of these constraints can account for the others, the problem in Eq. (20) should be equivalent to

$$\mathcal{L}^* = \max_{(f_b, f_o) \in \mathcal{U}_{bo}^*(C_{bo}, H)} \left\{ f_b^\top y^b + f_o^\top y^o - \zeta_b^*(f_b) - \zeta_o^*(f_o) \right\}, \tag{22a}$$

where the polyhedron $\mathcal{U}_{bo}^*(C_{bo}, H)$ is defined by

$$\mathcal{U}_{bo}^*(C_{bo}, H) \triangleq \left\{ f_b \in \mathbb{R}^{\mathcal{N}_b}, f_o \in \mathbb{R}^{\mathcal{N}_o} : \right.$$
$$\left. \forall i, j, \quad f_b^i + f_o^l H_l^j \leq C_{bo}^{ij} \right\} \tag{22b}$$

and the *effective cost metric* C_{bo} is given (in the absence of entropic regularisation) by

$$[C_{bo}]_{ij} \triangleq \min_k \left\{ [C_{ba}]_{ik} + [C_{oa}]_{jk} \right\}. \tag{22c}$$

According to Eq. (22c), this effective cost is given by the cost of the cheapest path(s), which is intuitive. The optimal transference glued plan, P, can be connected to the optimal transference plan P^{bo} between x^b and x^o with the cost C_{bo} in Eq. (22c), by marginalising on the intermediate density, i.e. the W-barycentre,

$$P_{ij}^{bo} = P_{ijk} 1_a^k. \tag{23}$$

The solution for the analysis state x^a is given by

$$x_k^a = P_{ijk} 1_b^i 1_o^j, \tag{24}$$

by the definition of the marginals of the gluing transference plan P (Eq. 19c). However, we do not have direct access to the optimal gluing P from the dual problem (Eq. 22). This will be made simpler later on when adding the entropic regularisation to the problem.

For now, let us find an alternative solution bypassing the need for the gluing P and define the map

$$\kappa^{bo} : [\![1, N_b]\!] \times [\![1, N_o]\!] \longmapsto \mathcal{P}([\![1, N_a]\!])$$
$$(i, j) \longmapsto \kappa_{ij}^{bo} = \arg\min_k \left(C_{ba}^{ik} + C_{oa}^{jk} \right), \tag{25}$$

where $\mathcal{P}(S)$ is defined as the set of all subsets of S. The set κ_{ij}^{bo} lists all of the indices k that are *relays* to the transport in between the sites corresponding to index i and index j. That is why the W-barycentre can be obtained from P^{bo}:

$$x_k^a = P_{ijk} 1_b^i 1_o^j = \sum_{ij} P_{ij}^{bo} \delta_{k \in \kappa_{ij}^{bo}}. \tag{26}$$

In the next section, we will show how to estimate P^{bo} using entropic regularisation and, hence, leverage Eq. (26) to compute x^a. κ^{bo} is reminiscent of the so-called *McCann interpolant* in OT theory, as it is only related to the OT between x^b and x^o, bypassing x^a and, hence, the transference plan P^{bo}. Please refer to Remark 7.1 by Peyré and Cuturi (2019) and to Gangbo and McCann (1996) for a description of the McCann interpolant, even when there is no Monge map. This suggests that the analysis x^a is not an interpolation of x^b and x^o in the space of values, as for classical DA, but along a geodesic in a Riemannian space built on a metric derived from the Wasserstein distance.

Nonetheless, the above derivation shows that we can trade a W-barycentre problem characterised by a couple of OT problems for a single OT problem defined by an effective metric C_{bo}. This principle is schematically illustrated in Fig. 7.

This suggests a simpler two-step algorithm, where the steps consist of the following: (i) solving a hybrid OTDA problem but with a single OT problem under an effective cost metric, which yields the analysed observables x^b and x^o, and (ii) computing the W-barycentre of x^b and x^o. To avoid making an overly large detour, the derivation of this algorithm is presented in Appendix B.

2.4.3 Classical data assimilation as a particular case

The primal problem (Eq. 17) of classical DA reads as follows:

$$\mathcal{L}_{cl} = \min_{x^a \in \mathcal{O}_{N_a}^+} \left\{ \zeta_b(y^b - x^a) + \zeta_o(y^o - H x^a) \right\}. \tag{27}$$

Let us see how the OTDA formalism in Eq. (22) can account for classical DA. In the context of classical DA, the observable spaces for x^b, x^o, and x^a are assumed to be the same by construction. Let us then define the cost matrices

$$\left[C_{ba}^\infty\right]_{ij} \triangleq \left[C_{oa}^\infty\right]_{ij} \triangleq \begin{cases} 0 & \text{if } i = j \\ +\infty & \text{if } i \neq j \end{cases} ; \tag{28}$$

i.e. it is assumed that the cost of moving masses is as large as it can be. Looking back at Eq. (19) but with these costs, it is clear that, in order to avoid the primal cost function being $+\infty$, the transference plan P_{ijk} must always be 0 unless $i = j = k$. However, this implies, from the definition of $\mathcal{U}_{boa}$, that the observables coincide, $x^b = x^o = x^a$, and that their mass is given by $m(P)$. In this limit where the specific cost matrices are equal to C_{ba}^∞ and C_{oa}^∞, the OTDA primal problem become mathematically equivalent to classical DA. Hence, classical DA is a limit case of OTDA. Note that, from its definition (Eq. 22c), the effective cost C_{bo} obtained from C_{ba}^∞ and C_{oa}^∞ coincides with $C_{bo}^\infty \triangleq C_{ba}^\infty = C_{oa}^\infty$.

2.5 A direct algorithmic solution

The two-step approach of Sect. 2.4 has merit in that it connects to the traditional W-barycentre problem, by first esti-
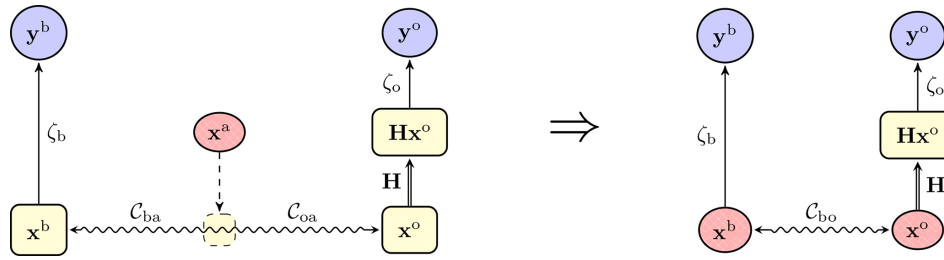
Figure 7. Trading a full hybrid OTDA problem, characterised by a W-barycentre defined by the cost metrics $\mathbf{C}_{\mathrm{ba}}$ and $\mathbf{C}_{\mathrm{oa}}$, with a simplified hybrid OTDA problem, characterised by a single OT problem defined by an effective cost metric $\mathbf{C}_{\mathrm{bo}}$.

mating $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$, and later computes the W-barycentre in between both states. It also suggests the existence of the effective cost metric of the problem. However, going through its consecutive steps may not be necessary for pure computational purposes. Here, we describe a direct approach that yields the analysis of the OTDA problem. It is less enlightening, but it is practical and will be used in the subsequent illustrations of the present paper.

An alternative formulation to the primal problem in Eq. (19) relies on two transference plans $\mathbf{P}^{\mathrm{ba}}$ and $\mathbf{P}^{\mathrm{oa}}$ corresponding to the two transports of the underlying W-barycentre problem, instead of the gluing one. Moreover, entropic regularisation is now enforced via $\mathcal{K}(\mathbf{P}^{\mathrm{ba}}|\boldsymbol{v}^{\mathrm{ba}})$ and $\mathcal{K}(\mathbf{P}^{\mathrm{oa}}|\boldsymbol{v}^{\mathrm{oa}})$. The corresponding optimisation problem reads as follows:

$$
\mathcal{L} = \min_{\boldsymbol{x}^{\mathrm{b}}\in\mathcal{O}_{\mathrm{b}}^{+}\,\boldsymbol{x}^{\mathrm{o}}\in\mathcal{O}_{\mathrm{o}}^{+}\,\boldsymbol{x}^{\mathrm{a}}\in\mathcal{O}_{\mathrm{a}}^{+}} \Bigg[\zeta_{\mathrm{b}}(\boldsymbol{y}^{\mathrm{b}}-\boldsymbol{x}^{\mathrm{b}}) + \zeta_{\mathrm{o}}(\boldsymbol{y}^{\mathrm{o}}-\mathbf{H}\boldsymbol{x}^{\mathrm{o}})
$$
$$
+ \min_{\mathbf{P}^{\mathrm{ba}}\in\mathcal{U}_{\mathrm{ba}}\,\mathbf{P}^{\mathrm{oa}}\in\mathcal{U}_{\mathrm{oa}}} \Big\{ \varepsilon\mathcal{K}(\mathbf{P}^{\mathrm{ba}}|\boldsymbol{v}^{\mathrm{ba}}) + \varepsilon\mathcal{K}(\mathbf{P}^{\mathrm{oa}}|\boldsymbol{v}^{\mathrm{oa}})
$$
$$
+ P_{ik}^{\mathrm{ba}}C_{\mathrm{ba}}^{ik} + P_{jk}^{\mathrm{oa}}C_{\mathrm{oa}}^{jk} \Big\} \Bigg], \tag{29a}
$$

where the admissible sets of the respective transference plans $\mathbf{P}^{\mathrm{ba}}$ and $\mathbf{P}^{\mathrm{oa}}$ are defined by

$$
\mathcal{U}_{\mathrm{ba}} \triangleq \Big\{ \mathbf{P}\in\mathcal{O}_{\mathrm{b,a}}^{+}: \quad \mathbf{P}\mathbf{1}_{\mathrm{a}}=\boldsymbol{x}^{\mathrm{b}}, \quad \mathbf{P}^{\top}\mathbf{1}_{\mathrm{b}}=\boldsymbol{x}^{\mathrm{a}} \Big\}, \tag{29b}
$$

$$
\mathcal{U}_{\mathrm{oa}} \triangleq \Big\{ \mathbf{P}\in\mathcal{O}_{\mathrm{o,a}}^{+}: \quad \mathbf{P}\mathbf{1}_{\mathrm{a}}=\boldsymbol{x}^{\mathrm{o}}, \quad \mathbf{P}^{\top}\mathbf{1}_{\mathrm{o}}=\boldsymbol{x}^{\mathrm{a}} \Big\}. \tag{29c}
$$

Following the same type of derivation as reported in the previous sections and Appendix B, the corresponding dual problem to be minimised is obtained as follows:

$$
\mathcal{J}_{\varepsilon}^{*} = \min_{\boldsymbol{f}_{\mathrm{b}}\in\mathbb{R}^{N_{\mathrm{b}}}\,\boldsymbol{f}_{\mathrm{o}}\in\mathbb{R}^{N_{\mathrm{o}}}\,\boldsymbol{f}_{\mathrm{a}}\in\mathbb{R}^{N_{\mathrm{a}}}} J_{\varepsilon}^{*}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{o}},\boldsymbol{f}_{\mathrm{a}}). \tag{30a}
$$

Here, discarding the constant $-\varepsilon\mathrm{m}(\boldsymbol{v}^{\mathrm{ba}}) - \varepsilon\mathrm{m}(\boldsymbol{v}^{\mathrm{oa}})$, the associated regularised Lagrangian is

$$
J_{\varepsilon}^{*}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{o}},\boldsymbol{f}_{\mathrm{a}}) = \varepsilon Z_{\varepsilon}^{\mathrm{ba}}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{a}}) + \varepsilon Z_{\varepsilon}^{\mathrm{oa}}(\boldsymbol{f}_{\mathrm{o}},\boldsymbol{f}_{\mathrm{a}})
$$
$$
+ \zeta_{\mathrm{b}}^{*}(\boldsymbol{f}_{\mathrm{b}}) + \zeta_{\mathrm{o}}^{*}(\boldsymbol{f}_{\mathrm{o}}) - \boldsymbol{f}_{\mathrm{b}}^{\top}\boldsymbol{y}^{\mathrm{b}} - \boldsymbol{f}_{\mathrm{o}}^{\top}\boldsymbol{y}^{\mathrm{o}}, \tag{30b}
$$

with a *partition function* associated with each transport:

$$
Z_{\varepsilon}^{\mathrm{ba}} \triangleq \sum_{ik} P_{ik}^{\mathrm{ba}}, \tag{30c}
$$

$$
Z_{\varepsilon}^{\mathrm{oa}} \triangleq \sum_{jk} P_{jk}^{\mathrm{oa}}, \tag{30d}
$$

where

$$
P_{ik}^{\mathrm{ba}} = v_{ik}^{\mathrm{ba}} e^{\left(f_{\mathrm{b}}^{i}+f_{\mathrm{a}}^{k}-C_{\mathrm{ba}}^{ik}\right)/\varepsilon}, \tag{30e}
$$

$$
P_{jk}^{\mathrm{oa}} = v_{jk}^{\mathrm{oa}} e^{\left(f_{\mathrm{o}}^{l}H_{l}^{j}-f_{\mathrm{a}}^{k}-C_{\mathrm{oa}}^{jk}\right)/\varepsilon}. \tag{30f}
$$

It turns out that the optimal $\boldsymbol{f}_{\mathrm{a}}$ can be obtained analytically as a function of $\boldsymbol{f}_{\mathrm{b}}$ and $\boldsymbol{f}_{\mathrm{o}}$, which we checked makes the optimisation numerically more efficient and robust. Indeed, let us introduce $\psi_{k} \triangleq e^{f_{\mathrm{a}}^{k}/\varepsilon}$. We could optimise $J_{\varepsilon}^{*}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{o}},\boldsymbol{f}_{\mathrm{a}}=\varepsilon\ln\boldsymbol{\psi})$ on $\boldsymbol{\psi}$:

$$
0 = \partial_{\psi_{k}}J_{\varepsilon}^{*}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{o}},\boldsymbol{f}_{\mathrm{a}}) \tag{31a}
$$

$$
= \sum_{i} v_{ik}^{\mathrm{ba}} e^{\left(f_{\mathrm{b}}^{i}-C_{\mathrm{ba}}^{ik}\right)/\varepsilon} - \frac{1}{\psi_{k}^{2}} \sum_{j} v_{jk}^{\mathrm{oa}} e^{\left(f_{\mathrm{o}}^{l}H_{l}^{j}-C_{\mathrm{oa}}^{jk}\right)/\varepsilon}, \tag{31b}
$$

yielding the solution

$$
\psi_{k}^{2} = \frac{Z_{\varepsilon,k}^{\mathrm{oa}}}{Z_{\varepsilon,k}^{\mathrm{ba}}}, \tag{32a}
$$

$$
Z_{\varepsilon,k}^{\mathrm{oa}} \triangleq \sum_{j} v_{jk}^{\mathrm{oa}} e^{\left(f_{\mathrm{o}}^{l}H_{l}^{j}-C_{\mathrm{oa}}^{jk}\right)/\varepsilon}, \tag{32b}
$$

$$
Z_{\varepsilon,k}^{\mathrm{ba}} \triangleq \sum_{i} v_{ik}^{\mathrm{ba}} e^{\left(f_{\mathrm{b}}^{i}-C_{\mathrm{ba}}^{ik}\right)/\varepsilon}. \tag{32c}
$$

Up to irrelevant constants, the resulting effective cost function using the optimal ψ_{k} is

$$
J_{\varepsilon}^{*}(\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{o}}) = 2\varepsilon\sum_{k}\sqrt{Z_{\varepsilon,k}^{\mathrm{ba}}Z_{\varepsilon,k}^{\mathrm{oa}}} + \zeta_{\mathrm{b}}^{*}(\boldsymbol{f}_{\mathrm{b}}) + \zeta_{\mathrm{o}}^{*}(\boldsymbol{f}_{\mathrm{o}})
$$
$$
- \boldsymbol{f}_{\mathrm{b}}^{\top}\boldsymbol{y}^{\mathrm{b}} - \boldsymbol{f}_{\mathrm{o}}^{\top}\boldsymbol{y}^{\mathrm{o}}. \tag{33}
$$

Now, the optimal W-barycentre $\boldsymbol{x}^{\mathrm{a}}$ is given by either $x_{k}^{\mathrm{a}}=P_{ik}^{\mathrm{ba}}1_{\mathrm{b}}^{i}$ or $x_{k}^{\mathrm{a}}=P_{jk}^{\mathrm{oa}}1_{\mathrm{o}}^{j}$, i.e.

$$
x_{k}^{\mathrm{a}} = \psi_{k}Z_{\varepsilon,k}^{\mathrm{ba}} = \frac{1}{\psi_{k}}Z_{\varepsilon,k}^{\mathrm{oa}}, \tag{34}
$$

from which we can infer the ψ_k-free expression

$$x_k^a = \sqrt{Z_{\varepsilon,k}^{ba} Z_{\varepsilon,k}^{oa}}. \tag{35}$$

It is also useful to retrieve the optimal value of $\boldsymbol{f}_a$ and obtain

$$f_a^k = \varepsilon \ln \psi_k = \frac{\varepsilon}{2} \ln \left(\frac{Z_{\varepsilon,k}^{oa}}{Z_{\varepsilon,k}^{ba}} \right), \tag{36}$$

so that we can compute the other two analysed observables, $\boldsymbol{x}^b$ and $\boldsymbol{x}^o$, using

$$x_i^b = P_{ik}^{ba} 1_a^k = \sum_k \psi_k v_{ik}^{ba} e^{\left(f_b^i - C_{ba}^{ik}\right)/\varepsilon}$$
$$= e^{f_b^i/\varepsilon} \sum_k v_{ik}^{ba} e^{\left(f_a^k - C_{ba}^{ik}\right)/\varepsilon}, \tag{37a}$$

$$x_j^o = P_{jk}^{oa} 1_a^k = \sum_k \frac{1}{\psi_k} v_{jk}^{oa} e^{\left(f_o^l H_l^j - C_{oa}^{jk}\right)/\varepsilon}$$
$$= e^{\left(f_o^l H_l^j\right)/\varepsilon} \sum_k v_{jk}^{oa} e^{\left(-f_a^k - C_{oa}^{jk}\right)/\varepsilon}. \tag{37b}$$

Note that most of these expressions can be assessed in a robust way in the log domain. For instance, in practice, we use, equivalently to Eqs. (35) and (37),

$$\varepsilon \ln x_k^a = \frac{\varepsilon}{2} \ln \sum_i v_{ik}^{ba} e^{\left(f_b^i - C_{ba}^{ik}\right)/\varepsilon}$$
$$+ \frac{\varepsilon}{2} \ln \sum_j v_{jk}^{oa} e^{\left(f_o^l H_l^j - C_{oa}^{jk}\right)/\varepsilon}, \tag{38a}$$

$$\varepsilon \ln x_i^b = f_b^i + \varepsilon \ln \sum_k v_{ik}^{ba} e^{\left(f_a^k - C_{ba}^{ik}\right)/\varepsilon}, \tag{38b}$$

$$\varepsilon \ln x_j^o = f_o^l H_l^j + \varepsilon \ln \sum_k v_{jk}^{oa} e^{\left(-f_a^k - C_{oa}^{jk}\right)/\varepsilon}. \tag{38c}$$

3 Numerical illustrations

In this section, we showcase a selection of OTDA 3D-Var analyses. These are meant to stress the versatility of the formalism and the diverse solutions it offers, with significantly more degrees of freedom than in classical DA. The OTDA state analysis is carried out using the process in Sect. 2.5 and its formulas. Unless specifically discussed, entropic regularisation is used with $\varepsilon = 10^{-3}$. The dual cost function in Eq. (33) is minimised using the quasi-Newton method L-BFGS-B (Liu and Nocedal, 1989), which yields the optimal $\boldsymbol{f}_b$ and $\boldsymbol{f}_o$. Then, Eq. (38) is employed to compute $\boldsymbol{x}^b$, $\boldsymbol{x}^o$, and $\boldsymbol{x}^a$.

3.1 One-dimensional examples

Considering the case in which the physical space of the fields is one-dimensional, we build bell-shaped observations $\boldsymbol{y}^b$

and $\boldsymbol{y}^o$, related to an observable space of size $N_b = N_o = N_a = 10^2$ shared by $\boldsymbol{x}^b$, $\boldsymbol{x}^o$, and $\boldsymbol{x}^a$. As $\boldsymbol{y}^b$ is a fully observed instance of $\boldsymbol{x}^b$, we have $\mathcal{N}_b = N_b = 10^2$, while $\mathcal{N}_o$ may differ from N_o depending on the definition of the observation operator $\mathbf{H}$. We choose (Gaussian statistics)

$$\zeta_b(\boldsymbol{e}_b) = \frac{1}{2\sigma_b^2} \|\boldsymbol{e}_b\|^2, \qquad \zeta_b(\boldsymbol{e}_o) = \frac{1}{2\sigma_o^2} \|\boldsymbol{e}_o\|^2. \tag{39}$$

Here, $\sigma_b = \sigma_o = 10^{-2}$. The states are discretised over the interval $[0, 1]$ at sites/grid cells $r_i^\star = (i - \frac{1}{2})/N_\star$ for $i \in [\![1, N_\star]\!]$, with $\star = $ b, o, and a. Unless otherwise specified, the cost metric has a quadratic dependence on the distance between sites, i.e. $[\mathbf{C}_{ba}]_{ik} = |r_i^b - r_k^a|^2$ and $[\mathbf{C}_{oa}]_{jk} = |r_j^o - r_k^a|^2$. This is our reference set-up. The observation operator and the mass of the observations $\boldsymbol{y}^b$ and $\boldsymbol{y}^o$ will be specified for each experiment.

We consider four experiments in which we choose to vary key parameters in the OTDA set-up.

3.1.1 Varying the imbalance of the observation states

In the first experiment, the system is fully observed with $\mathbf{H} = \mathbf{I}$. We choose $m(\boldsymbol{y}^b) = 1$ and the mass of $\boldsymbol{y}^o$ to be in the set $m(\boldsymbol{y}^o) \in \{0.5, 1, 1.5\}$, with all the other parameters being fixed to the reference. The results are displayed in Fig. 8. Figure 8a corresponds to the case $m(\boldsymbol{y}^o) = 0.5$. The resulting mass of the analysed observables is then $m(\boldsymbol{x}^a) = m(\boldsymbol{x}^b) = m(\boldsymbol{x}^o) = 0.79$. The adjustment of $\boldsymbol{x}^b$ compared with $\boldsymbol{y}^b$ and the adjustment of $\boldsymbol{x}^o$ compared with $\boldsymbol{y}^o$, which are required to balance $\boldsymbol{x}^b$ and $\boldsymbol{x}^o$, are patent. Figure 8b corresponds to the case $m(\boldsymbol{y}^o) = 1$. The resulting mass of the analysed observables is then $m(\boldsymbol{x}^a) = m(\boldsymbol{x}^b) = m(\boldsymbol{x}^o) = 1$. No adjustment is required here because $m(\boldsymbol{y}^o) = m(\boldsymbol{y}^b)$, and $\boldsymbol{x}^o$ and $\boldsymbol{y}^o$ as well as $\boldsymbol{x}^b$ and $\boldsymbol{y}^b$ coincide. Finally, the mass of $\boldsymbol{y}^o$ is set to $m(\boldsymbol{y}^o) = 1.5$ in Fig. 8c. The resulting mass of the analysed observables is then $m(\boldsymbol{x}^a) = m(\boldsymbol{x}^b) = m(\boldsymbol{x}^o) = 1.34$. The adjustment of $\boldsymbol{x}^b$ compared with $\boldsymbol{y}^b$ and the adjustment of $\boldsymbol{x}^o$ compared with $\boldsymbol{y}^o$, which are required to balance $\boldsymbol{x}^b$ and $\boldsymbol{x}^o$, are visually obvious, but the balancing goes in the opposite direction compared with Fig. 8a, as expected.

3.1.2 Varying the sparseness of the observation operator

In this second experiment, with all of the other parameters being fixed to their reference value, only a fraction of the domain is observed, over $\left[0, \frac{1}{4}\right]$, $\left[0, \frac{1}{2}\right]$, and $\left[0, \frac{3}{4}\right]$, where $\mathbf{H} \in \mathcal{O}_{\mathcal{N}_o \times N_o}^+$ with $\mathcal{N}_o = N_o/4$, $N_o/2$, $3N_o/4$, and $H_l^j = \delta_{l,j}$ for $l \in [\![1, \mathcal{N}_o]\!]$ and $j \in [\![1, N_o]\!]$.

The masses of the states that are built to generate $\boldsymbol{y}^b$ and $\boldsymbol{y}^o$, before applying any observation operator, are set to 1 and 1.5, respectively. As a result, we have $m(\boldsymbol{y}^b) = 1$, but $m(\boldsymbol{y}^o)$ may depart from 1.5 depending on $\mathbf{H}$. The fully observed case corresponds to Fig. 8c. The results are displayed
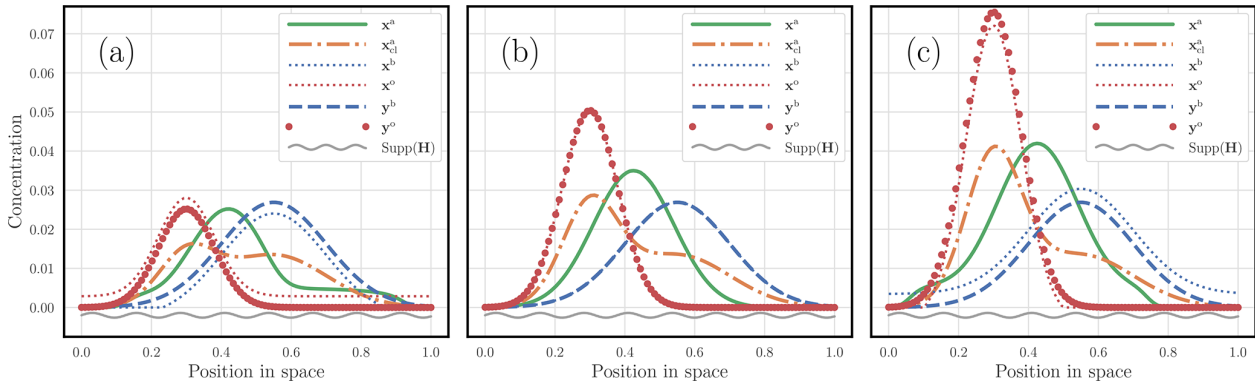
Figure 8. A hybrid OTDA 3D-Var analysis with one-dimensional physical states, where only the mass of y^o is varied. Its mass is $\mathrm{m}(y^o) = 0.5$ in panel (**a**), $\mathrm{m}(y^o) = 1$ in panel (**b**), and $\mathrm{m}(y^o) = 1.5$ in panel (**c**). The dashed blue curve corresponds to the first guess y^b; the red dots correspond to the observations y^o; the analysis state x^a is the solid green curve; and the analysed observables x^b and x^o are blue and red dotted curves, respectively. The support of the observation is underlined by a wavy grey segment. The corresponding classical analysis is also plotted with a dot-dash orange curve. The x axis corresponds to the position in space; the y axis corresponds to the concentration value of the fields.

in Fig. 9. It shows how smooth the OTDA solution can be compared with that of classical DA. However, as in Fig. 9a, OTDA can also handle obviously diverging sources of information, as is the case when the support of $\mathbf{H}$ is $\left[0, \frac{1}{4}\right]$ and when y^o and y^b can be seen to be barely consistent. In that case, the OTDA solution is smooth but bimodal.

3.1.3 Changing the nature of the cost metric

In this third experiment, we choose the cost metric to be of the form $[\mathbf{C}_{ba}]_{ik} = |r_i^b - r_k^a|^\alpha$ and $[\mathbf{C}_{oa}]_{jk} = |r_j^o - r_k^a|^\alpha$. Only half of the domain is observed over $\left[0, \frac{1}{2}\right]$, as in the case of Fig. 9b. As the mass of the state used to produce y^o is 1.5, we have a slightly different $\mathrm{m}(y^o) = 1.49$, with the rest of the mass being located in the unobserved part of the domain. All of the other parameters follow the reference set-up. The results are displayed in Fig. 10. For Fig. 10a, α is set to 0.5. For Fig. 10b, α is set to 1. For Fig. 10c, the cost metric is piecewise; it is quadratic, i.e. $\alpha = 2$, for pairs of sites separated by less than 10^{-1}, i.e. $|r_i^b - r_k^a| = |r_j^o - r_k^a| \leq 10^{-1}$, whereas the costs are chosen to be infinite for pairs of sites beyond this range. Hence, transport is prohibited beyond a distance of 10^{-1}. The case of a pure quadratic cost corresponds to Fig. 9b. The impact on the shape of the OTDA analysis is very significant and suggests that one could easily tailor their own cost to suit their specific DA problem.

3.1.4 Classical data assimilation as a sub-case of the hybrid optimal transport data assimilation

In the fourth experiment, we would like to numerically check the theoretical prediction of Sect. 2.4.3. Consider again the reference configuration; however, only half of the domain,

over $\left[0, \frac{1}{2}\right]$, is observed, $\mathbf{H} \in \mathcal{O}_{\mathcal{N}_o \times N_o}^+$ with $\mathcal{N}_o = N_o/2$ and $H_l^j = \delta_{l,j}$ for $l \in [\![1, \mathcal{N}_o]\!]$ and $j \in [\![1, N_o]\!]$. Most importantly, the cost metric has a quadratic dependence on the distance between sites, i.e. $[\mathbf{C}_{ba}]_{ik} = \lambda |r_i^b - r_k^a|^2$ and $[\mathbf{C}_{oa}]_{jk} = \lambda |r_j^o - r_k^a|^2$. The case $\lambda = 1$ corresponds to Fig. 9b. Figure 11 shows the results corresponding to $\lambda = 10^3$ for panel (a), $\lambda = 10^4$ for panel (b), and $\lambda = 10^6$ for panel (c). When λ is increased, the OTDA analysis should tend to the classical DA solution. This is indeed corroborated by Fig. 11 and supports the claim of Sect. 2.4.3. Note that, as opposed to the three earlier experiments, we had to tune ε here, as the wide range of λ has a significant impact on the balance of the key terms of the cost function (transport cost, discrepancy errors, and regularisation).

3.2 Two-dimensional examples

Considering the case in which the physical space of the fields is two-dimensional, we perform a couple of 3D-Var analyses on concentration fields (puffs of a pollutant). The states are discretised in the domain $[0, 1]^2$ at sites/grid cells $r_{i,j}^\star = \left((i - \frac{1}{2})/N_\star^x, (j - \frac{1}{2})/N_\star^y\right)$ for $(i, j) \in [\![1, N_\star^x]\!] \times [\![1, N_\star^y]\!]$, with $\star = b, o$, and a. We choose $N_b^x = N_b^y = N_o^x = N_o^y = N_a^x = N_a^y = 10^2$, such that $N_b = N_o = N_a = 10^4$. Hence, the number of control variables is 3×10^4. The observation vectors are y^b and y^o. As y^b is a fully observed instance of x^b, we have $\mathcal{N}_b = N_b$, while $\mathcal{N}_o$ may differ from N_o depending on the definition of the observation operator $\mathbf{H}$. Moreover, we choose (Gaussian statistics)

$$\zeta_b(e_b) = \frac{1}{2\sigma_b^2}\|e_b\|^2, \qquad \zeta_b(e_o) = \frac{1}{2\sigma_o^2}\|e_o\|^2. \qquad (40)$$
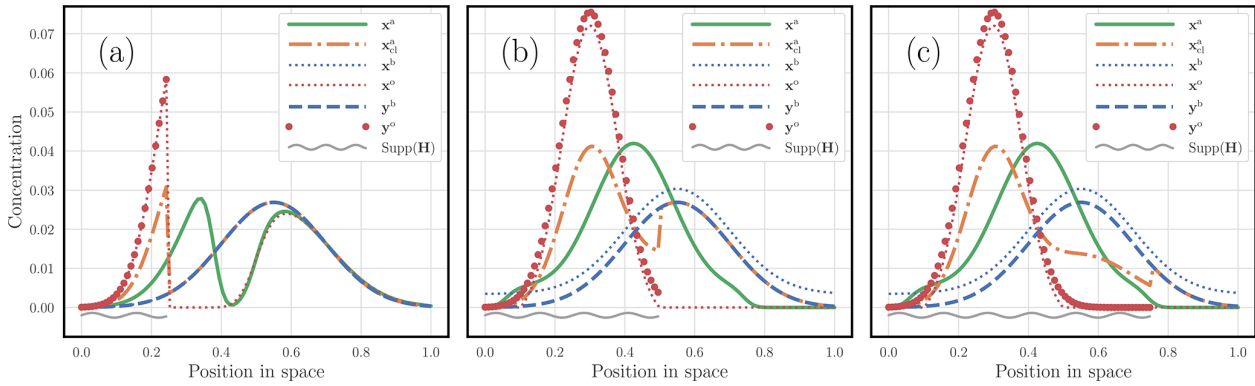
Figure 9. A hybrid OTDA 3D-Var analysis with one-dimensional physical states, where the observation operator is increasingly sparse. The support of $\mathbf{H}$ is $\left[0, \frac{1}{4}\right]$ for panel **(a)**, $\left[0, \frac{1}{2}\right]$ for panel **(b)**, and $\left[0, \frac{3}{4}\right]$ for panel **(c)**. See Fig. 8 for a description of the legend.
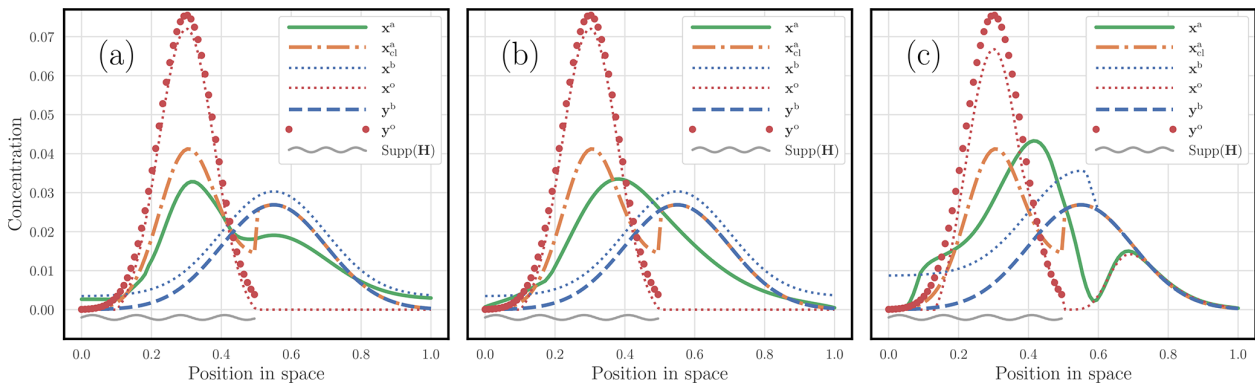
Figure 10. A hybrid OTDA 3D-Var analysis with one-dimensional physical states, where the cost metrics are changed. See the bulk of the text for a definition of the three cost metrics. See Fig. 8 for a description of the legend.

Here, $\sigma_b = \sigma_o = 10^{-2}$. The entropic regularisation parameter is set to $\varepsilon = 10^{-3}$.

The first analysis is displayed in Fig. 12. The observation operator $\mathbf{H}$ is the identity, but its support is restricted to the subdomain $[0, 0.6]^2$. The plumes of pollutants y^b and y^o are generated from states formed as combinations of bell-like puffs. The system is unbalanced with $m(y^b) = 1.35$ and $m(y^o) = 0.73$. The cost metric has a quadratic dependence on the distance between sites, i.e. $[\mathbf{C}_{ba}]_{ik} = \left\| r_i^b - r_k^a \right\|_2^2$ and $[\mathbf{C}_{oa}]_{jk} = \left\| r_j^o - r_k^a \right\|_2^2$. The OTDA analysis is clearly smoother than the classical solution. The classical solution does not cope very well with the seemingly disagreeing sources of information y^b and y^o, generating sharp transitions in the classical analysis. If y^b and y^o were consistently obtained from a truth perturbed with errors with short-range correlation, i.e. if they were drawn from the true prior distribution and in the absence of mislocation errors, then the classical analysis would be as good as it can be, whereas the OTDA solution may be too safe, i.e. too smooth. However, if one believes that structural errors and, in particular, loca-

tion errors can impact y^b and y^o, then the classical solution is improper and the OTDA analysis preferable.

The second analysis is displayed in Fig. 13. The support of the observation operator $\mathbf{H}$ is again contained within the subdomain $[0, 0.6]^2$, but only one of four grid cells is actually observed in this area. The observation states y^b and y^o are generated from the same states as for Fig. 12. The system is unbalanced with $m(y^b) = 1.35$ and $m(y^o) = 0.18$. The cost metric is defined to be the same as in Fig. 12. The OTDA analysis is even smoother in this case compared with the classical DA analysis. It is much less impacted by the sparseness of the observation operator. The classical solution has to account for the staggered observations in the top left corner of the domain because the first guess in that region is very uncertain. By contrast, the OTDA solution assumes that location errors are possible; hence, it moves around the mass corresponding to these observations so that the structure of the observation operator is not as impactful on the analysis.
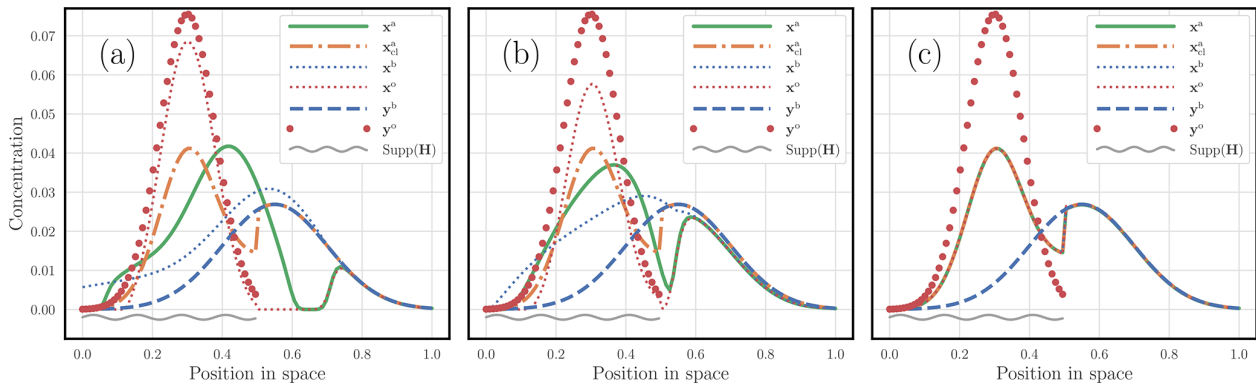
Figure 11. Scaling up the cost metrics $\lambda \mathbf{C}_{ba}$ and $\lambda \mathbf{C}_{oa}$ with increasing λ, the OTDA analysis converges to the classical DA analysis. Panels **(a)**, **(b)**, and **(c)** correspond to the scaling values $\lambda = 10^3, 10^4$, and 10^6, respectively. See Fig. 8 for a description of the legend.
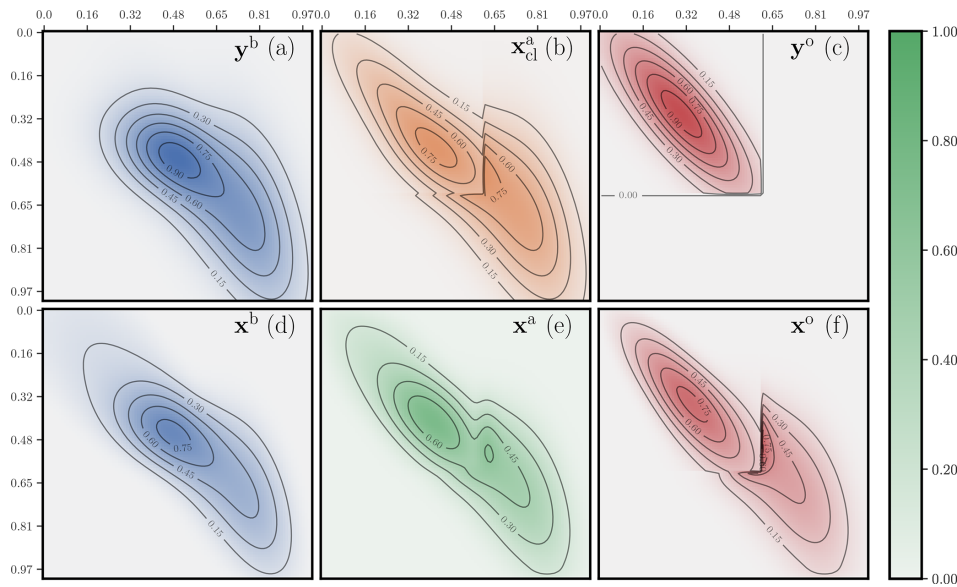
Figure 12. Two-dimensional concentration maps (plumes) of a hybrid OTDA analysis for the first configuration. The observations y^b and y^o; the analysed observables x^b and x^a, i.e. the state analysis; x^o; and the corresponding classical DA analysis x_{cl}^a are displayed. All fields are rescaled so that their joint maximum is 1. All concentration maps use the same scale. The colour bar represents a unified contrast scale for the diverse field concentrations.

4 Uncertainty quantification

In this section, we compute the posterior error covariance matrix $\mathbf{P}^a$ associated with the state analysis x^a, in order to complete the OTDA 3D-Var analysis description. There are many ways to proceed depending on the chosen regularisation and on the targeted degree of generality. Here, for the sake of consistency, we report on the way to derive $\mathbf{P}^a$ following the computation of the analysis state x^a proposed in Sect. 2.5.

4.1 Mathematical results

Let us denote the compounded vectors of the observations, of the Lagrange multipliers, and of the observables as well as the compounded observation operator by

$$y \triangleq \begin{bmatrix} y^b \\ y^o \end{bmatrix}, \qquad f \triangleq \begin{bmatrix} f_b \\ f_o \end{bmatrix},$$

$$x \triangleq \begin{bmatrix} x^b \\ x^o \end{bmatrix}, \qquad \mathcal{H} \triangleq \begin{bmatrix} \mathbf{I}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{bmatrix}, \tag{41}$$

of size $\mathcal{N}_b + \mathcal{N}_o, \mathcal{N}_b + \mathcal{N}_o, N_b + N_o$, and $(\mathcal{N}_b + \mathcal{N}_o) \times (N_b + N_o)$, respectively. Similarly, we define the sum of the error statistics by $\zeta(f) \triangleq \zeta_b(f_b) + \zeta_o(f_o)$, whose Legendre–Fenchel transform is $\zeta^*(f) = \zeta_b^*(f_b) + \zeta_o^*(f_o)$. Using this notation, we can recapitulate the key results of Sect. 2.5: the effective dual cost function is
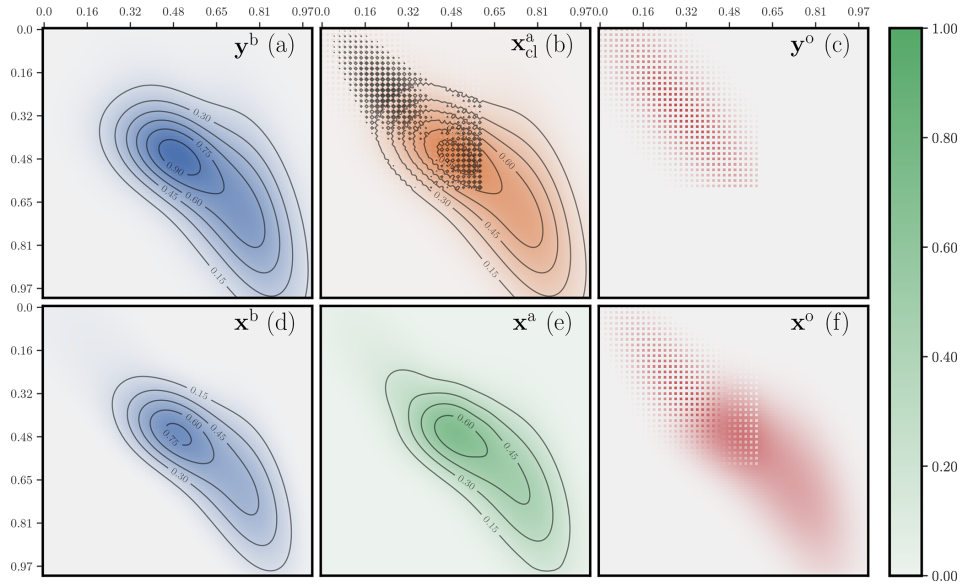
Figure 13. Two-dimensional concentrations maps of a hybrid OTDA analysis for the second configuration. The observations y^{b} and y^{o}; the analysed observables x^{b} and x^{a}, i.e. the state analysis; x^{o}; and the corresponding classical DA analysis are displayed. Compared to Fig. 12, only **H** has changed. The level sets in panels **(c)** and **(f)** are omitted because they are driven by the staggered observation operator.

$$J_\varepsilon^*(\boldsymbol{f}) \triangleq \varepsilon Z_\varepsilon(\boldsymbol{f}) + \zeta^*(\boldsymbol{f}) - \boldsymbol{f}^\top \boldsymbol{y}, \tag{42a}$$

$$Z_\varepsilon(\boldsymbol{f}) \triangleq 2 \sum_k \sqrt{Z_{\varepsilon,k}^{\mathrm{ba}}(\boldsymbol{f}_{\mathrm{b}}) Z_{\varepsilon,k}^{\mathrm{oa}}(\boldsymbol{f}_{\mathrm{o}})}. \tag{42b}$$

Here, the analysis state reads

$$x_k^{\mathrm{a}}(\boldsymbol{f}) = \sqrt{Z_{\varepsilon,k}^{\mathrm{ba}}(\boldsymbol{f}_{\mathrm{b}}) Z_{\varepsilon,k}^{\mathrm{oa}}(\boldsymbol{f}_{\mathrm{o}})}, \tag{43}$$

where the dependence of the analysis state and the partition functions on $\boldsymbol{f}$, $\boldsymbol{f}_{\mathrm{b}}$, and $\boldsymbol{f}_{\mathrm{o}}$ is now emphasised and made explicit.

Any prior source of error in the system stems from the information vector $\boldsymbol{y}$ and hence drives the posterior error in the analysis $\boldsymbol{x}^{\mathrm{a}}$. That is why we are interested in the sensitivity of $\boldsymbol{x}^{\mathrm{a}}$ with respect to $\boldsymbol{y}$, i.e. $\delta \boldsymbol{x}^{\mathrm{a}} = \partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}} \delta \boldsymbol{y}$. Denoting the expectation operator by $\mathbb{E}$, the error covariance matrix is then defined by

$$\mathbf{P}^{\mathrm{a}} = \mathbb{E}\left[\delta \boldsymbol{x}^{\mathrm{a}} (\delta \boldsymbol{x}^{\mathrm{a}})^\top\right], \tag{44a}$$

$$= (\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}}) \mathbb{E}\left[\delta \boldsymbol{y} \delta \boldsymbol{y}^\top\right] (\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}})^\top, \tag{44b}$$

$$= (\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}}) (\partial_{\boldsymbol{f}}^2 \zeta^*) (\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}})^\top, \tag{44c}$$

from which a matrix factor $\mathbf{X}^{\mathrm{a}}$ of $\mathbf{P}^{\mathrm{a}}$, i.e. which satisfies $\mathbf{P}^{\mathrm{a}} = \mathbf{X}^{\mathrm{a}} (\mathbf{X}^{\mathrm{a}})^\top$ and whose expressions are usually much shorter than those of $\mathbf{P}^{\mathrm{a}}$, can be extracted, up to the multiplication by an orthogonal matrix on the right:

$$\mathbf{X}^{\mathrm{a}} = \partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}} (\partial_{\boldsymbol{f}}^2 \zeta^*)^{\frac{1}{2}}. \tag{45}$$

To compute the sensitivity matrix $\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}}$, we leverage the stationarity of the dual cost function at the minimum:

$$\partial_{\boldsymbol{f}} J_\varepsilon^*(\boldsymbol{f}(\boldsymbol{y}), \boldsymbol{y}) = \mathbf{0}. \tag{46}$$

We resort to the implicit function theorem:

$$\mathbf{0} = \mathrm{d}_{\boldsymbol{y}} \partial_{\boldsymbol{f}} J_\varepsilon^*(\boldsymbol{f}(\boldsymbol{y}), \boldsymbol{y}) = \partial_{\boldsymbol{f}}^2 J_\varepsilon^* \partial_{\boldsymbol{y}} \boldsymbol{f} + \partial_{\boldsymbol{f}} \partial_{\boldsymbol{y}} J_\varepsilon^*, \tag{47}$$

which yields

$$\partial_{\boldsymbol{y}} \boldsymbol{f} = -\left[\partial_{\boldsymbol{f}}^2 J_\varepsilon^*\right]^{-1} \partial_{\boldsymbol{f}} \partial_{\boldsymbol{y}} J_\varepsilon^* = \left[\partial_{\boldsymbol{f}}^2 J_\varepsilon^*\right]^{-1}, \tag{48}$$

as $\partial_{\boldsymbol{f}} \partial_{\boldsymbol{y}} J_\varepsilon^* = -\mathbf{I}_{bo}$, where $\mathbf{I}_{bo}$ is the identity matrix in the compounded observation space $\mathbb{R}^{\mathcal{N}_{\mathrm{b}}+\mathcal{N}_{\mathrm{o}}}$. The sensitivity $\partial_{\boldsymbol{y}} \boldsymbol{x}^{\mathrm{a}}$ can now be computed using the Leibniz chain rule and Eq. (48):

$$\frac{\partial \boldsymbol{x}^{\mathrm{a}}}{\partial \boldsymbol{y}} = \frac{\partial \boldsymbol{x}^{\mathrm{a}}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} = \partial_{\boldsymbol{f}} \boldsymbol{x}^{\mathrm{a}} \left[\partial_{\boldsymbol{f}}^2 J_\varepsilon^*\right]^{-1}. \tag{49}$$

Let us now compute the Jacobian and Hessian in the right-hand side of Eq. (49). To that end and in order to externalise the observation operator, we introduce $\hat{Z}_\varepsilon$ and $\hat{\boldsymbol{x}}_{\mathrm{a}}$, such that

$$\hat{Z}_\varepsilon(\boldsymbol{\eta} = \mathcal{H}^\top \boldsymbol{f}) \triangleq Z_\varepsilon(\boldsymbol{f}), \qquad \hat{\boldsymbol{x}}_{\mathrm{a}}(\boldsymbol{\eta} = \mathcal{H}^\top \boldsymbol{f}) \triangleq \boldsymbol{x}^{\mathrm{a}}(\boldsymbol{f}), \tag{50}$$

and the related Jacobian and Hessians,

$$\boldsymbol{\Omega}_{\mathrm{bo,a}} \triangleq \partial_{\boldsymbol{\eta}} \hat{\boldsymbol{x}}_{\mathrm{a}}, \tag{51a}$$

$$\boldsymbol{\Omega}_{\mathrm{bo,bo}} \triangleq \varepsilon \partial_{\boldsymbol{\eta}}^2 \hat{Z}_\varepsilon, \tag{51b}$$

$$\boldsymbol{\Lambda}_{\mathrm{bo,bo}} \triangleq \partial_{\boldsymbol{f}}^2 \zeta^*. \tag{51c}$$

$\hat{Z}_\varepsilon$ and $\hat{\boldsymbol{x}}_{\mathrm{a}}$ can be shown to exist; they can be read off from the explicit expressions of Z_ε and $\boldsymbol{x}^{\mathrm{a}}$ as functions of $\boldsymbol{f}$.

These Jacobian and Hessians depend on the choice of the regularisation operator and need to be computed analytically, which is simple but tedious; this is not reported here, as it is a regularisation-dependent calculation. The Hessian of the dual cost function Eq. (42a) can then be written as the sum

$$\partial_f^2 J_\varepsilon^* = \mathbf{\Lambda}_{\mathrm{bo,bo}} + \mathcal{H}\mathbf{\Omega}_{\mathrm{bo,bo}}\mathcal{H}^\top, \tag{52}$$

while the sensitivity matrix now reads

$$\partial_f \boldsymbol{x}^{\mathrm{a}} = \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathcal{H}^\top. \tag{53}$$

Note that $\mathbf{\Omega}_{\mathrm{bo,bo}}$ can be interpreted as the covariance matrix of $\boldsymbol{x}$, the compounded observable vector as defined in Eq. (41) (although seen as a random vector), under the assumption that $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ are connected via the W-barycentre $\boldsymbol{x}^{\mathrm{a}}$ and the optimal transference plans $\mathbf{P}^{\mathrm{ba}}$ and $\mathbf{P}^{\mathrm{oa}}$ (all seen as random vectors). Combining Eqs. (52) and (53) with Eq. (45), we finally obtain the expression for a factor $\mathbf{X}^{\mathrm{a}}$ of $\mathbf{P}^{\mathrm{a}}$:

$$\mathbf{X}^{\mathrm{a}} = \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathcal{H}^\top \Big[\mathbf{\Lambda}_{\mathrm{bo,bo}} + \mathcal{H}\mathbf{\Omega}_{\mathrm{bo,bo}}\mathcal{H}^\top \Big]^{-1} \mathbf{\Lambda}_{\mathrm{bo,bo}}^{\frac{1}{2}}. \tag{54}$$

Alternatively, we can use the Sherman–Morrison–Woodbury transformation, under the assumption that $\mathbf{\Omega}_{\mathrm{bo,bo}}$ is invertible:

$$\mathbf{X}^{\mathrm{a}} = \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathbf{\Omega}_{\mathrm{bo,bo}}^{-1} \Big[\mathbf{\Omega}_{\mathrm{bo,bo}}^{-1} + \mathcal{H}^\top \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-1} \mathcal{H} \Big]^{-1} \mathcal{H}^\top \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}. \tag{55}$$

These formulas are similar to the normal equations of classical DA. However, it should be noted that, in Eqs. (54) and (55), all of the prior error statistics are encapsulated in $\mathbf{\Lambda}_{\mathrm{bo,bo}}$, whereas the impact of OT is encoded in $\mathbf{\Omega}_{\mathrm{bo,bo}}$. To be concrete, note that, when using Gaussian statistics (Eq. 12), $\mathbf{\Lambda}_{\mathrm{bo,bo}}$ would simply read

$$\mathbf{\Lambda}_{\mathrm{bo,bo}} = \begin{bmatrix} \mathbf{\Lambda}_{\mathrm{bb}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{\mathrm{oo}} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}. \tag{56}$$

4.2 Interpretation

Further, we can perform a block decomposition of $\mathbf{\Omega}$ onto the spaces of $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$:

$$\mathbf{\Omega}_{\mathrm{bo,bo}} \triangleq \begin{bmatrix} \mathbf{\Omega}_{\mathrm{bb}} & \mathbf{\Omega}_{\mathrm{bo}} \\ \mathbf{\Omega}_{\mathrm{bo}}^\top & \mathbf{\Omega}_{\mathrm{oo}} \end{bmatrix}. \tag{57}$$

It can be shown that $\mathbf{\Omega}_{\mathrm{bo}}$ is proportional to the optimal transference plan of the effective transport between $\boldsymbol{x}^{\mathrm{o}}$ and $\boldsymbol{x}^{\mathrm{b}}$ and that the blocks of the diagonal are themselves diagonal and depend on the observable states:

$$\mathbf{\Omega}_{\mathrm{bo}} = \frac{1}{\varepsilon}\mathbf{P}^{\mathrm{bo}}, \tag{58a}$$

$$\mathbf{\Omega}_{\mathrm{bb}} = \frac{1}{\varepsilon}\mathrm{diag}(\boldsymbol{x}^{\mathrm{b}}), \tag{58b}$$

$$\mathbf{\Omega}_{\mathrm{oo}} = \frac{1}{\varepsilon}\mathrm{diag}(\boldsymbol{x}^{\mathrm{o}}). \tag{58c}$$

For instance, this could be shown by the explicit computation of $\mathbf{\Omega}_{\mathrm{bo,bo}} = \varepsilon\partial_{\boldsymbol{\eta}}^2 \hat{Z}_\varepsilon$.

Let us now examine the impact of OT on the analysis error covariance matrix. We first define

$$\mathbf{\Delta} = \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \mathcal{H}\mathbf{\Omega}_{\mathrm{bo,bo}}^{\frac{1}{2}}, \tag{59}$$

whose thin singular value decomposition is $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ is an orthogonal matrix of size $(\mathcal{N}_{\mathrm{b}} + \mathcal{N}_{\mathrm{o}}) \times (\mathcal{N}_{\mathrm{b}} + \mathcal{N}_{\mathrm{o}})$, $\mathbf{\Sigma}$ is a rectangular and diagonal matrix of size $(\mathcal{N}_{\mathrm{b}} + \mathcal{N}_{\mathrm{o}}) \times (N_{\mathrm{b}} + N_{\mathrm{o}})$, and $\mathbf{V}$ is an orthogonal matrix of size $(N_{\mathrm{b}} + N_{\mathrm{o}}) \times (N_{\mathrm{b}} + N_{\mathrm{o}})$. Then, we *standardise* Eq. (54) following, for example, Sect. 2.4.1 in Rodgers (2000):

$$\mathbf{X}^{\mathrm{a}} = \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathcal{H}^\top \Big[\mathbf{\Lambda}_{\mathrm{bo,bo}} + \mathcal{H}\mathbf{\Omega}_{\mathrm{bo,bo}}\mathcal{H}^\top \Big]^{-1} \mathbf{\Lambda}_{\mathrm{bo,bo}}^{\frac{1}{2}}, \tag{60a}$$

$$= \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathcal{H}^\top \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}$$
$$\times \Big[\mathbf{I}_{bo} + \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \mathcal{H}\mathbf{\Omega}_{\mathrm{bo,bo}}\mathcal{H}^\top \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \Big]^{-1}, \tag{60b}$$

$$= \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathbf{\Omega}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \mathbf{\Delta}^\top \Big[\mathbf{I}_{bo} + \mathbf{\Delta}\mathbf{\Delta}^\top \Big]^{-1}, \tag{60c}$$

$$= \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathbf{\Omega}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \mathbf{V}\mathbf{\Sigma}^\top \Big[\mathbf{I}_{bo} + \mathbf{\Sigma}\mathbf{\Sigma}^\top \Big]^{-1} \mathbf{U}^\top. \tag{60d}$$

Defining $\boldsymbol{\sigma} = (\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{\frac{1}{2}}$, which is a square diagonal of size $(\mathcal{N}_{\mathrm{b}} + \mathcal{N}_{\mathrm{o}}) \times (\mathcal{N}_{\mathrm{b}} + \mathcal{N}_{\mathrm{o}})$, we obtain, up to a multiplication by an irrelevant orthogonal matrix on the right, an equivalent factor for $\mathbf{P}^{\mathrm{a}}$:

$$\mathbf{X}^{\mathrm{a}} = \mathbf{\Omega}_{\mathrm{bo,a}}^\top \mathbf{\Omega}_{\mathrm{bo,bo}}^{-\frac{1}{2}} \mathbf{V} \frac{\boldsymbol{\sigma}}{\mathbf{I}_{bo} + \boldsymbol{\sigma}^2}. \tag{60e}$$

The diagonal values of $\boldsymbol{\sigma}$, denoted $\sigma_i \geq 0$, represent the independent degrees of freedom (dof) values of information that can be extracted from the observations, which, in our case, is the first guess $\boldsymbol{y}^{\mathrm{b}}$ and the traditional observations $\boldsymbol{y}^{\mathrm{o}}$, in contrast to Rodgers (2000), who only considers the dof values from $\boldsymbol{y}^{\mathrm{o}}$. The higher the σ_i, the more information attached to the dof of index i and the more squeezed the corresponding direction in $\mathbf{X}^{\mathrm{a}}$ and $\mathbf{P}^{\mathrm{a}}$. From Eq. (59) and, in particular, its transpose, $\mathbf{\Delta}^\top = \mathbf{\Omega}_{\mathrm{bo,bo}}^{\frac{1}{2}} \mathcal{H}^\top \mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}$, we can trace the flow of any piece of information. Such information stems from the observation vectors; hence, its flow starts in $\mathbf{\Delta}^\top$ from $\mathbf{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}$, the square root of the precision matrix $\mathbf{\Lambda}_{\mathrm{bo,bo}}^{-1}$. It is then transferred from the observation spaces to the observable spaces through $\mathcal{H}^\top$. It is finally optimally transported across the space of $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{o}}$ by $\mathbf{\Omega}_{\mathrm{bo,bo}}$, whose off-diagonal block is proportional to the transference plan $\mathbf{P}^{\mathrm{bo}}$. Hence, OT is not a primary source of uncertainty, as $\boldsymbol{y}^{\mathrm{b}}$ and $\boldsymbol{y}^{\mathrm{o}}$ can be, but moves information in between the observable spaces.

Let us now check the OTDA analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}$ in the classical DA limit. To that end, we study Eq. (54) in the classical limit. Similarly to $\mathbf{\Omega}_{\mathrm{bb}}$ and $\mathbf{\Omega}_{\mathrm{oo}}$ in Eq. (58a), $\mathbf{\Omega}_{\mathrm{aa}}$ is defined as the covariance matrix of $\boldsymbol{x}^{\mathrm{a}}$ when only accounting for both OTs, and it can be shown that it reads

$$\mathbf{\Omega}_{\mathrm{aa}} = \frac{1}{\varepsilon}\mathrm{diag}(\boldsymbol{x}^{\mathrm{a}}). \tag{61}$$

When the cost tends to $\mathbf{C}_{\mathrm{bo}}^{\infty}$, following the same arguments as in Sect. 2.4.3, x^{b}, x^{o}, and x^{a} must merge and, consequently, $\boldsymbol{\Omega}_{\mathrm{bo}} = \boldsymbol{\Omega}_{\mathrm{aa}} = \boldsymbol{\Omega}_{\mathrm{bb}} = \boldsymbol{\Omega}_{\mathrm{oo}}$. Hence, in this limit, $\boldsymbol{\Omega}_{\mathrm{bo,bo}} = \mathbf{1}_2 \boldsymbol{\Omega}_{\mathrm{aa}} \mathbf{1}_2^{\top}$ and $\boldsymbol{\Omega}_{\mathrm{bo,a}} = \mathbf{1}_2 \boldsymbol{\Omega}_{\mathrm{aa}}$, with $\mathbf{1}_2 = [1 \quad 1]^{\top}$. Then, substituting these expressions of $\boldsymbol{\Omega}_{\mathrm{bo,bo}}$ and $\boldsymbol{\Omega}_{\mathrm{bo,a}}$ into Eq. (54), we get

$$\mathbf{X}^{\mathrm{a}} = \boldsymbol{\Omega}_{\mathrm{aa}} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \left[\boldsymbol{\Lambda}_{\mathrm{bo,bo}} + \mathcal{H} \mathbf{1}_2 \boldsymbol{\Omega}_{\mathrm{aa}} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \right]^{-1} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{\frac{1}{2}}, \tag{62a}$$

$$= \boldsymbol{\Omega}_{\mathrm{aa}} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \left[\mathbf{I}_{bo} + \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-1} \mathcal{H} \mathbf{1}_2 \boldsymbol{\Omega}_{\mathrm{aa}} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \right]^{-1} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}, \tag{62b}$$

$$= \boldsymbol{\Omega}_{\mathrm{aa}} \left[\mathbf{I}_{\mathrm{a}} + \mathbf{1}_2^{\top} \mathcal{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-1} \mathcal{H} \mathbf{1}_2 \boldsymbol{\Omega}_{\mathrm{aa}} \right]^{-1} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}, \tag{62c}$$

$$= \left[\boldsymbol{\Omega}_{\mathrm{aa}}^{-1} + \mathbf{1}_2^{\top} \mathcal{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-1} \mathcal{H} \mathbf{1}_2 \right]^{-1} \mathbf{1}_2^{\top} \mathcal{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-\frac{1}{2}}. \tag{62d}$$

Here, $\mathbf{I}_{\mathrm{a}}$ is the identity matrix of size N_{a}. From Eq. (62b) to Eq. (62c), we relied on the shift matrix lemma (e.g. Asch et al., 2016). For $\boldsymbol{\Omega}_{\mathrm{aa}}^{-1}$ in Eq. (62d) to exist, it must be assumed that $x^{\mathrm{a}} \in \mathcal{O}_{N_{\mathrm{a}}}^{+,\star}$; i.e. all the entries of x^{a} are positive. This is verified when using entropic regularisation with $\varepsilon > 0$, no matter how small the entries of x^{a} are. Moreover, if x^{a} has zero entries, x^{a} can be represented as the limit of a sequence of positive discrete measures.

Now, as we have

$$\mathbf{A}^{-1} \triangleq \mathbf{1}_2^{\top} \mathcal{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{bo,bo}}^{-1} \mathcal{H} \mathbf{1}_2 = \boldsymbol{\Lambda}_{\mathrm{bb}}^{-1} + \mathbf{H}^{\top} \boldsymbol{\Lambda}_{\mathrm{oo}}^{-1} \mathbf{H}, \tag{63}$$

we conclude, from Eq. (62d), that the classical limit of the analysis error covariance matrix is

$$\mathbf{P}^{\mathrm{a}} = \mathbf{X}^{\mathrm{a}} \left(\mathbf{X}^{\mathrm{a}} \right)^{\top} \tag{64a}$$

$$= \left[\boldsymbol{\Omega}_{\mathrm{aa}}^{-1} + \mathbf{A}^{-1} \right]^{-1} \mathbf{A}^{-1} \left[\boldsymbol{\Omega}_{\mathrm{aa}}^{-1} + \mathbf{A}^{-1} \right]^{-1}. \tag{64b}$$

If the limit of x^{a} when $\varepsilon \to 0^{+}$ is in $\mathcal{O}_{N_{\mathrm{a}}}^{+,\star}$, then $\boldsymbol{\Omega}_{\mathrm{aa}}^{-1} = \varepsilon \mathrm{diag}(x^{\mathrm{a}})^{-1}$ must vanish. In this case,

$$\mathbf{P}^{\mathrm{a}} \xrightarrow[\varepsilon \to 0^{+}]{} \mathbf{A} = \left(\partial_{f_{\mathrm{b}}}^2 \zeta_{\mathrm{b}} + \mathbf{H}^{\top} \partial_{f_{\mathrm{o}}}^2 \zeta_{\mathrm{o}} \mathbf{H} \right)^{-1}, \tag{65}$$

which, assuming Gaussian errors, would read $\mathbf{P}^{\mathrm{a}} = \left(\mathbf{B}^{-1} + \mathbf{H}^{\top} \mathbf{R}^{-1} \mathbf{H} \right)^{-1}$, as expected from classical DA. However, if some of the entries of x^{a} vanish in the limit $\varepsilon \to 0^{+}$, we suspect that the limit of $\mathbf{P}^{\mathrm{a}}$ will be the classical analysis error covariance matrix $\mathbf{A}$ but with the columns and rows associated with the vanishing entries of x^{a} tapered to 0.

4.3 Numerical illustration

We consider the one-dimensional example where half of the domain is observed, over $\left[0, \frac{1}{2} \right]$. Here, $\mathbf{H} \in \mathcal{O}_{\mathcal{N}_{\mathrm{o}} \times N_{\mathrm{o}}}^{+}$ with $\mathcal{N}_{\mathrm{o}} = N_{\mathrm{o}}/2$ and $H_l^j = \delta_{l,j}$ for $l \in [\![1, \mathcal{N}_{\mathrm{o}}]\!]$ and $j \in [\![1, N_{\mathrm{o}}]\!]$. Further, the observations are unbalanced, $\mathrm{m}(y^{\mathrm{b}}) = 1$ and $\mathrm{m}(y^{\mathrm{o}}) = 1.49$; they have been generated through $\mathcal{H}$ by discrete measures of mass 1 and 1.5, respectively. Moreover, the

cost metric has a quadratic dependence on the distance between sites, i.e. $[\mathbf{C}_{\mathrm{ba}}]_{ik} = \lambda |r_i^{\mathrm{b}} - r_k^{\mathrm{a}}|^2$ and $[\mathbf{C}_{\mathrm{oa}}]_{jk} = \lambda |r_j^{\mathrm{o}} - r_k^{\mathrm{a}}|^2$, where $\lambda = 10^3$. We use the results of Sect. 4.1 to compute the analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}$, the transference plan $\mathbf{P}^{\mathrm{bo}}$, and the Jacobian $\boldsymbol{\Omega}_{\mathrm{bo,a}}$. The numerical results are displayed in Fig. 14. The OTDA analysis state is bimodal, with some mass being left over to the right of the domain to account for the long tail of the first guess, which is far from the observation support. Hence, there is a vanishing field region, roughly $[0.6, 0.7]$, which separates the two components of the analysis state. As expected from OT theory, $\mathbf{P}^{\mathrm{bo}}$ seems to converge towards a (non-trivial and barely differentiable) Monge map which, in this discrete context, has two branches, separated by the gap created by the vanishing field region. The analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}$ seems to converge to a diagonal matrix, with the exception of the vanishing field region. Indeed, there seems to be an uncertainty with respect to how much mass should be transferred from the first guess tail $[0.7, 1]$ to the main region $[0, 0.6]$. This is given away by variance peaks near the edges of the gap and by negative covariances between the two edge points of the gap.

5 Conclusions

In this paper, we have introduced a theoretical framework for integrating nonlocal optimal transport (OT) metrics into data assimilation (DA), which we refer to as *hybrid OTDA*. This framework addresses the inconsistencies initially identified by Feyeux et al. (2016) when local metrics in classical DA are replaced with nonlocal ones based on OT.

Our focus has been on defining a 3D-Var approach for hybrid OTDA and deriving the first- and second-order moments of its analysis. The hybrid OTDA 3D-Var method blends classical DA and its background and observation error statistics with a Wasserstein barycentre problem involving the observables associated with the first guess and the observation vector. Importantly, our work demonstrates that classical DA is encompassed within this theoretical framework.

We have shown that this optimisation problem can be decomposed and simplified into a hybrid OTDA problem with a single OT problem based on an effective cost. This first problem yields the estimated x^{b} and x^{o}, followed by a pure W-barycentre problem involving these states, whose solution is known as the McCann interpolant. This W-barycentre computation serves as the final analysis step.

Our proposed method can be applied to sparsely and noisily observed systems, as expected from a robust DA method. It can also accommodate non-trivial error statistics typical of a 3D-Var approach. Furthermore, we have illustrated the method's flexibility in defining cost metrics through various one- and two-dimensional numerical examples. We have empirically checked how the OTDA analysis shifts towards the classical DA analysis, within the OTDA framework.
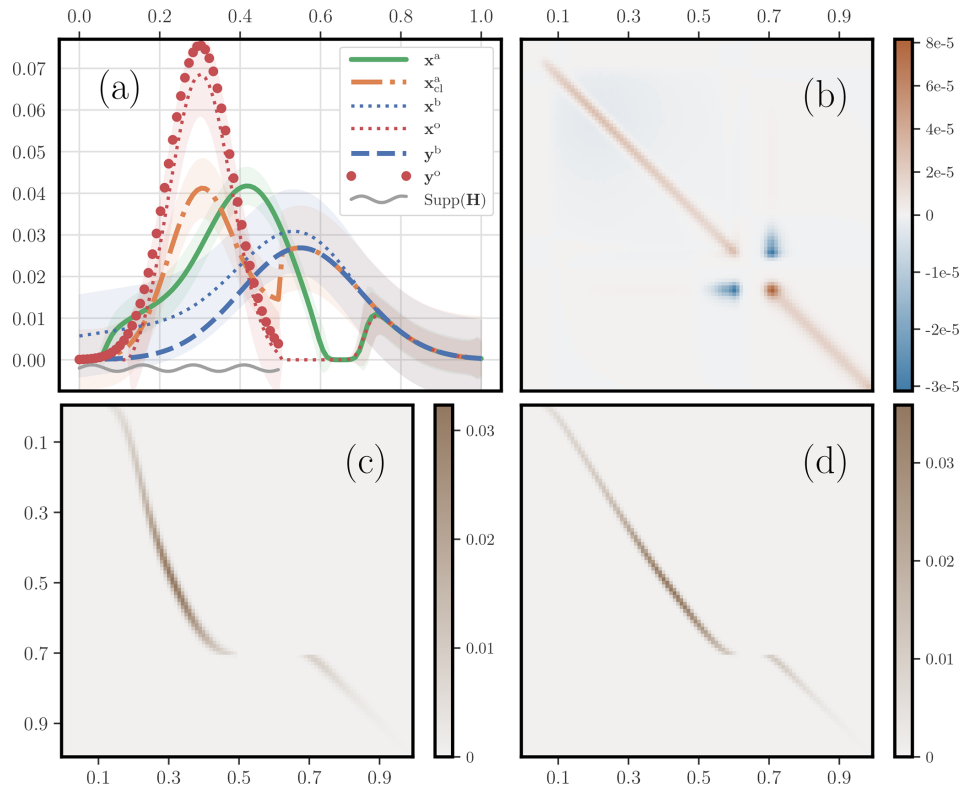
Figure 14. Illustration of the second-order analysis of an OTDA 3D-Var. Panel **(a)** shows the same plot as Fig. 11a but with the addition of shaded regions delineated by plus and minus the standard deviations about the estimates for x^a, x^a_{cl}, x^b, and x^o. These standard deviations are computed from the diagonal of the diagnosed posterior error covariance matrices associated with x^a, x^a_{cl}, x^b, and x^o. Panel **(b)** displays the analysis error covariance matrix $\mathbf{P}^a$. Panel **(c)** shows the optimal transference plan $\mathbf{P}^{bo}$. Panel **(d)** shows the a, b block part of the Jacobian matrix $\mathbf{\Omega}_{bo,a}$, which is denoted $\mathbf{\Omega}_{ba}$.

Note that, for now, some limitations apply; mainly, the framework is presently meant for non-negative fields.

While we have looked into several other promising developments regarding our methodology, we have chosen not to report them in this paper. These developments will be the subject of a future publication and include the following:

– the derivation of a Bayesian and probabilistic standpoint on OTDA;

– a generalised formalism enforcing physical regularisation, such as smoothness of the field, on the analysis state;

– a *stochastic matrix* formalism, which is a substitute to using transference plans but could offer more robustness in the presence of entropic regularisation;

– employing cost matrices defined across several spaces, which is useful for realistic application where x^b and x^o lie in very distinct spaces, such as the space of emission of a pollutant and the space of the pollutant concentrations, respectively.

While our primary focus in this paper was on the derivation and understanding of key cost functions within the hybrid OTDA framework, we did not delve much into the numerical challenges, algorithmic complexity, or computing acceleration. For this aspect of the developments, we would rather rely on developments from OT experts, who are continuously improving the efficiency of numerical OT (e.g. Flamary et al., 2021).

In addition to strengthening the developments mentioned above, our future research will explore the application of the hybrid OTDA formalism in a sequential DA framework, as this paper concentrated solely on static analysis. We are also interested in investigating the role played by error statistics and cost metrics $\{\zeta_b, \zeta_o, \mathbf{C}_{ba}, \mathbf{C}_{oa}\}$ and their balancing in the hybrid OTDA analysis as well as in developing their objective tuning.

Appendix A: From the primal to the dual cost function for the full problem

This appendix is dedicated to the derivation of the transformation from a Lagrangian variant of the primal problem to

the dual cost function (Eq. 20). It takes the form of the following series of transformations of the problem (from a Lagrangian to a dual cost function):

$$
\mathcal{L} = \min_{\boldsymbol{x}^b \, \boldsymbol{x}^o \, \boldsymbol{x}^a} \Bigg[\max_{\boldsymbol{f}_b} \Big\{ (\boldsymbol{y}^b - \boldsymbol{x}^b)^\top \boldsymbol{f}_b - \zeta_b^*(\boldsymbol{f}_b) \Big\}
$$
$$
+ \max_{\boldsymbol{f}_o} \Big\{ (\boldsymbol{y}^o - \mathbf{H}\boldsymbol{x}^o)^\top \boldsymbol{f}_o - \zeta_o^*(\boldsymbol{f}_o) \Big\}
$$
$$
+ \min_{\mathbf{P} \in \mathcal{O}_{b,o,a}^+} \Big\{ P_{ijk} C_{ba}^{ik} + P_{ijk} C_{oa}^{jk}
$$
$$
+ \max_{\boldsymbol{h}_b \, \boldsymbol{h}_o \, \boldsymbol{f}_a} \Big\{ h_i^b \Big(x_i^b - P_{ijk} 1_o^j 1_a^k \Big)
$$
$$
+ h_j^o \Big(x_j^o - P_{ijk} 1_b^i 1_a^k \Big)
$$
$$
+ f_a^k \Big(x_k^a - P_{ijk} 1_b^i 1_o^j \Big) \Big\} \Big\} \Bigg], \tag{A1a}
$$

$$
= \min_{\boldsymbol{x}^a \, \boldsymbol{x}^b \, \boldsymbol{x}^o} \Bigg[\max_{\boldsymbol{f}_b, \boldsymbol{f}_o} \Big\{ (\boldsymbol{y}^b - \boldsymbol{x}^b)^\top \boldsymbol{f}_b
$$
$$
- \zeta_b^*(\boldsymbol{f}_b) + (\boldsymbol{y}^o - \mathbf{H}\boldsymbol{x}^o)^\top \boldsymbol{f}_o
$$
$$
- \zeta_o^*(\boldsymbol{f}_o) \Big\} + \max_{\boldsymbol{h}_b \, \boldsymbol{h}_o \, \boldsymbol{f}_a} \min_{\mathbf{P} \in \mathcal{O}_{b,o,a}^+} \Big\{ P_{ijk} C_{ba}^{ik} + P_{ijk} C_{oa}^{jk}
$$
$$
+ h_i^b \Big(x_i^b - P_{ijk} 1_o^j 1_a^k \Big) + h_j^o \Big(x_j^o - P_{ijk} 1_b^i 1_a^k \Big)
$$
$$
+ f_a^k \Big(x_k^a - P_{ijk} 1_b^i 1_o^j \Big) \Big\} \Bigg], \tag{A1b}
$$

$$
= \max_{\substack{\boldsymbol{h}_b \, \boldsymbol{h}_o \, \boldsymbol{f}_a \, \boldsymbol{x}^a \, \boldsymbol{x}^b \, \boldsymbol{x}^o \\ \boldsymbol{f}_b \, \boldsymbol{f}_o \, \mathbf{P} \in \mathcal{O}_{b,o,a}^+}} \min
$$
$$
\Big[(\boldsymbol{y}^b - \boldsymbol{x}^b)^\top \boldsymbol{f}_b - \zeta_b^*(\boldsymbol{f}_b)
$$
$$
+ (\boldsymbol{y}^o - \mathbf{H}\boldsymbol{x}^o)^\top \boldsymbol{f}_o - \zeta_o^*(\boldsymbol{f}_o) + P_{ijk} C_{ba}^{ik} + P_{ijk} C_{oa}^{jk}
$$
$$
+ h_i^b \Big(x_i^b - P_{ijk} 1_o^j 1_a^k \Big) + h_j^o \Big(x_j^o - P_{ijk} 1_b^i 1_a^k \Big)
$$
$$
+ f_a^k \Big(x_k^a - P_{ijk} 1_b^i 1_o^j \Big) \Big], \tag{A1c}
$$

$$
= \max_{\boldsymbol{f}_b \, \boldsymbol{f}_o} \Big\{ \boldsymbol{f}_b^\top \boldsymbol{y}^b + \boldsymbol{f}_o^\top \boldsymbol{y}^o - \zeta_b^*(\boldsymbol{f}_b) - \zeta_o^*(\boldsymbol{f}_o)
$$
$$
+ \min_{\substack{\boldsymbol{h}_b \, \boldsymbol{h}_o \, \boldsymbol{f}_a \\ \boldsymbol{x}^a \, \boldsymbol{x}^b \, \boldsymbol{x}^o \, \mathbf{P} \in \mathcal{O}_{b,o,a}^+}}
$$
$$
\Big[(\boldsymbol{h}_b - \boldsymbol{f}_b)^\top \boldsymbol{x}^b + (\boldsymbol{h}_o - \mathbf{H}^\top \boldsymbol{f}_o)^\top \boldsymbol{x}^o + \boldsymbol{f}_a^\top \boldsymbol{x}^a
$$
$$
+ P_{ijk} \Big(C_{ba}^{ik} + C_{oa}^{jk} - h_i^b 1_o^j 1_a^k - h_j^o 1_b^i 1_a^k - f_a^k 1_b^i 1_o^j \Big) \Big] \Big\}, \tag{A1d}
$$

$$
= \max_{\boldsymbol{f}_b \, \boldsymbol{f}_o} \Bigg[\boldsymbol{f}_b^\top \boldsymbol{y}^b + \boldsymbol{f}_o^\top \boldsymbol{y}^o - \zeta_b^*(\boldsymbol{f}_b) - \zeta_o^*(\boldsymbol{f}_o)
$$
$$
+ \min_{\mathbf{P} \in \mathcal{O}_{b,o,a}^+} P_{ijk} \Big(C_{ba}^{ik} + C_{oa}^{jk} - f_b^i 1_o^j 1_a^k - H_l^j f_o^l 1_b^i 1_a^k \Big) \Bigg]. \tag{A1e}
$$

In Eq. (A1a), the maps ζ_b^* and ζ_o^* are the Legendre–Fenchel transforms of the maps ζ_b and ζ_o, respectively. From Eq. (A1d) to Eq. (A1e), taking the minimum over the observables $\boldsymbol{x}^b, \boldsymbol{x}^o$, and $\boldsymbol{x}^a$ implies enforcing $\boldsymbol{h}_b = \boldsymbol{f}_b, \boldsymbol{h}_o = \mathbf{H}^\top \boldsymbol{f}_o$, and $\boldsymbol{f}_a = \boldsymbol{0}$. Hence, we obtain the dual problem which only depends on the Lagrange multipliers:

$$
\mathcal{L}^* = \max_{(\boldsymbol{f}_b, \boldsymbol{f}_o) \in \mathcal{U}_{bo}^*(\mathbf{C}_{ba}, \mathbf{C}_{oa}, \mathbf{H})} \Big\{ \boldsymbol{f}_b^\top \boldsymbol{y}^b + \boldsymbol{f}_o^\top \boldsymbol{y}^o
$$
$$
- \zeta_b^*(\boldsymbol{f}_b) - \zeta_o^*(\boldsymbol{f}_o) \Big\}, \tag{A2a}
$$

where the $*$ symbol refers to *dual* and where the polyhedron $\mathcal{U}_{bo}^*(\mathbf{C}_{ba}, \mathbf{C}_{oa}, \mathbf{H})$ is defined by

$$
\mathcal{U}_{bo}^*(\mathbf{C}_{ba}, \mathbf{C}_{oa}, \mathbf{H}) \triangleq \Big\{ \boldsymbol{f}_b \in \mathbb{R}^{\mathcal{N}_b}, \, \boldsymbol{f}_o \in \mathbb{R}^{\mathcal{N}_o} :
$$
$$
\forall i, j, k, \quad f_b^i + f_o^l H_l^j \le C_{ba}^{ik} + C_{oa}^{jk} \Big\}. \tag{A2b}
$$

The inequality constraints of the polyhedron $\mathcal{U}_{bo}^*$ stem from the positivity constraint $P_{ijk} \ge 0$ in Eq. (A1e). Very importantly, we have the coincidence of the minimum of the primal problem with the maximum of the dual problem $\mathcal{L} = \mathcal{L}^*$, a property called *strong duality* (see Sect. 5.2 in Boyd and Vandenberghe, 2004). Strong duality can, for instance, be achieved if both the primal and dual cost functions are convex, which is the case here.

Appendix B: Derivation of the two-step hybrid optimal transport data assimilation algorithm

Here, we derive the two-step algorithm elaborated in Sect. 2.4.2. Moreover, entropic regularisation is added to the problem.

B1 First step: simplified hybrid optimal transport data assimilation problem

The first step of the full OTDA algorithm is a simplified OTDA problem based on a single OT problem driven by the cost $\mathbf{C}_{bo}$. The corresponding high-level primal cost function is

$$
\mathcal{L} = \min_{\boldsymbol{x}^b \in \mathcal{O}_b^+ \, \boldsymbol{x}^o \in \mathcal{O}_o^+} \Big\{ \zeta_b(\boldsymbol{y}^b - \boldsymbol{x}^b) + \zeta_o(\boldsymbol{y}^o - \mathbf{H}\boldsymbol{x}^o)
$$
$$
+ W_{\mathbf{C}_{bo}}(\boldsymbol{x}^b, \boldsymbol{x}^o) \Big\}. \tag{B1}
$$

The associated (lower-level) primal cost function, adding entropic regularisation ($\varepsilon > 0$), is then

$$\mathcal{L}_\varepsilon = \min_{x^b \in \mathcal{O}_b^+ \, x^o \in \mathcal{O}_o^+} \left[\zeta_b(y^b - x^b) + \zeta_o(y^o - Hx^o) \right.$$
$$\left. + \min_{P \in \mathcal{U}_{bo}} \left(\varepsilon \mathcal{K}(P|v) + P_{ij} C_{bo}^{ij} \right) \right]. \quad (B2a)$$

In this optimisation problem, the admissible set of transference plans, i.e. the set of all 2-tensors of non-negative entries whose marginals are x^b and x^o, is defined by

$$\mathcal{U}_{bo} \triangleq \left\{ P \in \mathcal{O}_{b,o}^+ : \quad P1_o = x^b, \quad P^\top 1_b = x^o \right\}. \quad (B2b)$$

As x^b and x^o are not predetermined, the prior transference plan v cannot be selected from $\mathcal{U}_{bo}$ a priori. Hence, the simplest choice, which we decided to implement, is to set v_{ij} to a constant, which assumes some statistical prior independence of x^b and x^o. A derivation of the dual problem equivalent to $\mathcal{L}_\varepsilon$ can be obtained in the exact same way as in the previous subsection, although it is now less cluttered because there is only one OT to account for, instead of two. The associated Lagrangian is

$$\mathcal{L}_\varepsilon = \max_{f_b \in \mathbb{R}^{N_b} \, f_o \in \mathbb{R}^{N_o}} \left[f_b^\top y^b + f_o^\top y^o - \zeta_b^*(f_b) - \zeta_o^*(f_o) \right.$$
$$+ \min_{P \in \mathcal{O}_{b,o}^+} \left(\varepsilon \sum_{ij} \left\{ P_{ij} \ln \frac{P_{ij}}{v_{ij}} - P_{ij} + v_{ij} \right\} \right.$$
$$\left. \left. + P_{ij} \left\{ C_{bo}^{ij} - f_b^i 1_o^j - H_l^j f_o^l 1_b^i \right\} \right) \right]. \quad (B3)$$

Again, the maps ζ_b^* and ζ_o^* are the Legendre–Fenchel transforms of the maps ζ_b and ζ_o, respectively. The variables f_b and f_o are Lagrange vectors; they are used to enforce the marginals of the transference plan associated with $W_{C_{bo}}$. The unconstrained minimisation over P, i.e. the inner minimisation problem in Eq. (B3), is obtained by cancelling the gradient with respect to P, which yields

$$P_{ij} = v_{ij} e^{\left(f_b^i + f_o^l H_l^j - C_{bo}^{ij} \right)/\varepsilon}. \quad (B4)$$

Substituting this solution into minus the Lagrangian $-\mathcal{L}_\varepsilon$ gives the regularised dual problem

$$\mathcal{J}_\varepsilon^* = \min_{f_b \in \mathbb{R}^{N_b} \, f_o \in \mathbb{R}^{N_o}} J_\varepsilon^*(f_b, f_o), \quad (B5a)$$

with the associated Lagrangian

$$J_\varepsilon^*(f_b, f_o) = \varepsilon (Z_\varepsilon - m(v)) + \zeta_b^*(f_b) + \zeta_o^*(f_o)$$
$$- f_b^\top y^b - f_o^\top y^o, \quad (B5b)$$

which relies on the *partition function*

$$Z_\varepsilon = \sum_{ij} P_{ij}. \quad (B5c)$$

The notation $\mathcal{J}_\varepsilon^*$ and J_ε^*, rather than $\mathcal{L}_\varepsilon^*$ and L_ε^*, signifies that we work on the opposite of $\mathcal{L}_\varepsilon^*$ and L_ε^* so as to obtain a dual problem to be minimised rather than maximised. Most importantly, we have, under conditions that will be satisfied in the following, the coincidence of the two minima $\mathcal{J}_\varepsilon^* = -\mathcal{L}_\varepsilon$, i.e. strong duality. Assuming one can obtain a proper correspondence between the optimal f_b and f_o of the dual problem and x^b and x^o of the primal problem, this implies, once again, that the primal problem can be traded for the dual problem.

Even though the regularised optimisation problem is slightly different from the unregularised one, a difference which is controlled by the value of ε, the new dual optimisation problem is free, i.e. without constraints. It can be solved as it is, using, for instance, the L-BFGS-B minimiser (Liu and Nocedal, 1989). The advantage of the regularised dual formulation is twofold: (1) the dual cost function is unconstrained (free optimisation), and (2) we will trade a minimisation over $N_b \times N_o$ variables for a minimisation over $N_b + N_o$ variables. This dual formulation can be viewed as a generalised *physical-space statistical analysis system* (PSAS) formalism (Courtier, 1997), an approach in which classical DA algebra is mostly carried out in observation space.

Once the optimal values for f_b and f_o are obtained, the optimal discrete Kantorovich transference plan P can be computed using Eq. (B4). As a result, as marginals of this transference plan, the solutions for the observables are

$$x_i^b = P_{ij} 1_o^j = \sum_j P_{ij}, \qquad x_j^o = P_{ij} 1_b^i = \sum_i P_{ij}. \quad (B6)$$

B2   Second step: Wasserstein barycentre

Now that we have obtained the observables x^b and x^o via Eq. (B6), we would like to compute their W-barycentre. The joint mass m of these observables can be computed as follows:

$$m = m\left(x^b\right) = m\left(x^o\right). \quad (B7)$$

The high-level primal cost function of this W-barycentre problem is

$$J_w = \min_{x^a \in \mathcal{O}_{N_a}^+} \left\{ W_{C_{ba}}(x^b, x^a) + W_{C_{oa}}(x^o, x^a) \right\}. \quad (B8)$$

We have found and practised several ways to solve this problem. One way is to compute the McCann interpolant. This is theoretically elegant, but Eq. (26) did not leverage regularisation of the W-barycentre problem. Instead, the approach reported here is to use the dual optimisation problem, in conjunction with entropic regularisation at finite $\varepsilon > 0$. We leverage our knowledge of the mass m resulting from the first step of the algorithm by enforcing the mass in the cost function, $m(P) = m$. This seems redundant, but it actually yields, *by*

construction and very naturally, a numerically efficient algorithm comparable to the ad hoc log-domain scheme proposed in Sect. 4.4 of Peyré and Cuturi (2019).

Again, one way (although not the only way) to write the primal problem is to use a gluing transference plan, a 3-tensor whose marginals are x^b, x^o, and x^a:

$$L_\varepsilon = \min_{\substack{x^a \in \mathcal{O}_{N_a}^+ \\ \mathbf{P} \in \mathcal{U}_{\text{boa}}(x^a)}} \left\{ \mathbf{P} \cdot \mathbf{C}_{\text{boa}} + \varepsilon \mathcal{K}(\mathbf{P}|\nu) \right.$$

$$\left. + f_b^\top x^b + f_o^\top x^o \right\}, \tag{B9a}$$

where $[\mathbf{C}_{\text{boa}}]_{ijk} = C_{\text{ba}}^{ik} + C_{\text{oa}}^{jk}$, the binary operator $\cdot$ denotes the contraction of tensors, and

$$\mathcal{U}_{\text{boa}}(x^a) \triangleq \left\{ \mathbf{P} \in \mathcal{O}_{b,o,a}^+ : \quad \forall i, \ P_{ijk} 1_o^j 1_a^k = x_i^b, \right.$$

$$\left. \forall j, \ P_{ijk} 1_b^i 1_a^k = x_j^o, \quad \forall k, \ P_{ijk} 1_b^i 1_o^j = x_k^a \right\}. \tag{B9b}$$

The 3-tensor ν is chosen to be $\nu_{ijk} = x_i^b x_j^o / (m N_a)$, which is uniform in k and for which $\mathrm{m}(\nu) = m$. The resulting dual problem is

$$\mathcal{J}^* = \min_{f_b \in \mathbb{R}^{N_b} \ f_o \in \mathbb{R}^{N_o}} J^*(f_b, f_o), \tag{B10a}$$

where the associated Lagrangian is

$$J^*(f_b, f_o) = \varepsilon \left(m \ln \frac{Z_\varepsilon}{m} + m - \mathrm{m}(\nu) \right)$$

$$- f_b^\top x^b - f_o^\top x^o, \tag{B10b}$$

with the partition function

$$Z_\varepsilon = \sum_{ijk} \nu_{ijk} e^{\left(f_b^i + f_o^l H_l^j - C_{\text{ba}}^{ik} - C_{\text{oa}}^{jk} \right)/\varepsilon}. \tag{B10c}$$

This partition function is elegant but impractical because, with high dimensionality, a 3-tensor might be too large to store and compute with. However the partition function in Eq. (B10c) can be simplified by noticing that

$$Z_\varepsilon = \sum_{ij} \nu_{ij} e^{\left(f_b^i + f_o^l H_l^j - C_{\text{bo}}^{ij} \right)/\varepsilon}, \tag{B11}$$

where we introduced the effective cost metric

$$[\mathbf{C}_{\text{bo}}]_{ij} = -\varepsilon \ln \left(\sum_k \frac{\nu_{ijk}}{\nu_{ij}} e^{-\left(C_{\text{ba}}^{ik} + C_{\text{oa}}^{jk} \right)/\varepsilon} \right), \tag{B12}$$

which is the regularised cost – known in statistics and machine learning as a *soft-plus* transform – of Eq. (22c). The 2-tensor ν_{ij} plays the same role as that of the first step of the algorithm; we choose it as $\nu_{ij} = x_i^b x_j^o / m$, for which $\mathrm{m}(\nu) = m$. The dual problem now only involves 2-tensors and becomes numerically more efficient. Given the optimal f_b and f_o, the (glued) optimal transference plan $\mathbf{P}^{\text{boa}}$ is formally given by

$$P_{ijk}^{\text{boa}} = \frac{\nu_{ijk}}{Z_\varepsilon} e^{\left(f_b^i + f_o^l H_l^j - C_{\text{ba}}^{ik} - C_{\text{oa}}^{jk} \right)/\varepsilon}. \tag{B13}$$

The W-barycentre x^a is then given as a marginal of $\mathbf{P}^{\text{boa}}$:

$$x_k^a = P_{ijk} 1_b^i 1_o^j \tag{B14a}$$

$$= \frac{1}{Z_\varepsilon} \sum_{ij} \nu_{ijk} e^{\left(f_b^i + f_o^l H_l^j - C_{\text{ba}}^{ik} - C_{\text{oa}}^{jk} \right)/\varepsilon}. \tag{B14b}$$

Because of the normalisation of the transference plan to m, the entropic regularisation exhibits a $\varepsilon m \ln Z_\varepsilon$ instead of $\varepsilon Z_\varepsilon$. This systematically enforces normalisation in the computations of the gradients, as well as in the course of the numerical optimisation of the dual cost function, de facto working in the log domain. We experienced more stable computations and the ability to reach smaller ε, compared with the case without normalisation. This completes the solution through the two-step OTDA algorithm.

Review statement. This paper was edited by Olivier Talagrand and reviewed by two anonymous referees.

References

Amodei, M. and Stein, J.: Deterministic and fuzzy verification methods for a hierarchy of numerical models, Meteorol. Appl., 16, 191–203, https://doi.org/10.1002/met.101, 2009.

Asch, M., Bocquet, M., and Nodet, M.: Data Assimilation: Methods, Algorithms, and Applications, Fundamentals of Algorithms, SIAM, Philadelphia, ISBN 978-1-611974-53-9, https://doi.org/10.1137/1.9781611974546, 2016.

Bocquet, M.: Towards optimal choices of control space representation for geophysical data assimilation, Mon. Weather Rev., 137, 2331–2348, https://doi.org/10.1175/2009MWR2789.1, 2009.

Bocquet, M., Wu, L., and Chevallier, F.: Bayesian design of control space for optimal assimilation of observations. I: Consistent multiscale formalism, Q. J. Roy. Meteor. Soc., 137, 1340–1356, https://doi.org/10.1002/qj.837, 2011.

Boyd, S. P. and Vandenberghe, L.: Convex optimization, Cambridge university press, ISBN 978-0521833783, 2004.

Briggs, W. M. and Levine, R. A.: Wavelets and field forecast verification, Mon. Weather Rev., 125, 1329–1341, https://doi.org/1520-0493(1997)125<1329:WAFFV>2.0.CO;2, 1997.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data Assimilation in the Geosciences: An overview on methods, issues, and perspectives, WIREs Clim. Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X.: Scaling algorithms for unbalanced optimal transport problems, Math. Comput., 87, 2563–2609, https://doi.org/10.1090/mcom/3303, 2018.

Courtier, P.: Dual formulation of four-dimensional variational assimilation, Q. J. Roy. Meteor. Soc., 123, 2449–2461, https://doi.org/10.1002/qj.49712354414, 1997.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, New-York, ISBN 9780521458252, 1991.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Mon. Weather Rev., 134, 772–1784, https://doi.org/10.1175/MWR3146.1, 2006a.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems, Mon. Weather Rev., 134, 1785–1795, https://doi.org/10.1175/MWR3145.1, 2006b.

Duc, L. and Sawada, Y.: Geometry of rainfall ensemble means: from arithmetic averages to Gaussian-Hellinger barycenters in unbalanced optimal transport, J. Meteor. Soc. Jpn., 102, 35–47, https://doi.org/10.2151/jmsj.2024-003, 2024.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, https://doi.org/10.1002/met.25, 2008.

El Moselhy, T. A. and Marzouk, Y. M.: Bayesian inference with optimal maps, J. Comp. Phys., 231, 7815–7850, https://doi.org/10.1016/j.jcp.2012.07.022, 2012.

Evensen, G., Vossepoel, F. C., and van Leeuwen. P. J.: Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem, Springer Textbooks in Earth Sciences, Geography and Environment, Springer Cham, ISBN 978-3-030-96708-6, https://doi.org/10.1007/978-3-030-96709-3, 2022.

Farchi, A. and Bocquet, M.: Review article: Comparison of local particle filters and new implementations, Nonlin. Processes Geophys., 25, 765–807, https://doi.org/10.5194/npg-25-765-2018, 2018.

Farchi, A., Bocquet, M., Roustan, Y., Mathieu, A., and Quérel, A.: Using the Wasserstein distance to compare fields of pollutants: application to the radionuclide atmospheric dispersion of the Fukushima-Daiichi accident, Tellus B, 68, 31682, https://doi.org/10.3402/tellusb.v68.31682, 2016.

Feyeux, N.: Transport optimal pour l'assimilation de données images, Ph.D. thesis, Université Grenoble Alpes, https://inria.hal.science/tel-01480695 (last access: 7 July 2024), 2016.

Feyeux, N., Vidard, A., and Nodet, M.: Optimal transport for variational data assimilation, Nonlin. Processes Geophys., 25, 55–66, https://doi.org/10.5194/npg-25-55-2018, 2018.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., K., F., Fournier, N., Gautheron, L., Gayraud, N. T. H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T.: POT: Python Optimal Transport, J. Mach. Learn. Res., 22, 1–8, http://jmlr.org/papers/v22/20-451.html (last access: 7 July 2024), 2021.

Gangbo, W. and McCann, R. J.: The geometry of optimal transportation, Acta Math., 177, 113–1618, https://doi.org/10.1007/BF02392620, 1996.

Gilleland, E., Ahijevych, D. A., Brown, B. G., and Ebert, E. E.: Verifying forecasts spatially, B. Am. Meteorol. Soc., 91, 1365–1373, https://doi.org/10.1175/2010BAMS2819.1, 2010a.

Gilleland, E., Lindström, J., and Lindgren, F.: Analyzing the image warp forecast verification method on precipitation fields from the ICP, Weather Forecast., 25, 1249–1262, https://doi.org/10.1175/2010WAF2222365.1, 2010b.

Hoffman, R. N. and Grassotti, C.: A Technique for Assimilating SSM/I Observations of Marine Atmospheric Storms: Tests with ECMWF Analyses, J. Appl. Meteorol. Clim., 35, 1177–1188, https://doi.org/10.1175/1520-0450(1996)035<1177:ATFASO>2.0.CO;2, 1996.

Hoffman, R. N., Liu, Z., Louis, J.-F., and Grassotti, C.: Distortion representation of forecast errors, Mon. Weather Rev., 123, 2758–2770, https://doi.org/10.1175/1520-0493(1995)123<2758:DROFE>2.0.CO;2, 1995.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Q. J. Roy. Meteor. Soc., 144, 1257–1278, https://doi.org/10.1002/qj.3130, 2018.

Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, Cambridge, ISBN 9780521796293, 2003.

Keil, C. and Craig, G. C.: A displacement and amplitude score employing an optical flow technique, Weather Forecast., 24, 1297–1308, https://doi.org/10.1175/2009WAF2222247.1, 2009.

Lack, S. A., Limpert, G. L., and Fox, N. I.: An object-oriented multiscale verification scheme, Weather Forecast., 25, 79–92, https://doi.org/10.1175/2009WAF2222245.1, 2010.

Le Coz, C., Tantet, A., Flamary, R., and Plougonven, R.: Optimal transport for the multi-model combination of sub-seasonal ensemble forecasts, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-13445, https://doi.org/10.5194/egusphere-egu23-13445, 2023.

Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, Math. Programm., 45, 503–528, https://doi.org/10.1007/BF01589116, 1989.

Lledó, L., Skok, G., and Haiden, T.: Estimating location errors in precipitation forecasts with the Wasserstein and Attribution distances, EMS Annual Meeting 2023, Bratislava, Slovakia, 4–8 Sep 2023, EMS2023-602, https://doi.org/10.5194/ems2023-602, 2023.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A.: An introduction to sampling via measure transport, in: Handbook of Uncertainty Quantification, edited by: Ghanem, R., Higdon, D., and Owhadi, H., chap. 23, Springer International Publishing, Cham, 785–825, https://doi.org/10.1007/978-3-319-12385-1_23, 2017.

Monge, G.: Mémoire sur la théorie des déblais et des remblais, in: Histoire de l'Académie Royale des Sciences de Paris, 666–704, 1781.

Necker, T., Wolfgruber, L., Kugler, L., Weissmann, M., Dorninger, M., and Serafin, S.: The fractions skill score for ensemble forecast verification, Authorea [preprint], https://doi.org/10.22541/au.169169008.89657659/v1, 2023.

Ning, L. and 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 Resour. Res., 50, 5817–5830, https://doi.org/10.1002/2013WR014966, 2014.

Oliver, D. S.: Minimization for conditional simulation: Relationship to optimal transport, J. Comp. Phys., 265, 1–15, https://doi.org/10.1016/j.jcp.2014.01.048, 2014.

Optimal Transport: Github [code], https://github.com/PythonOT, last access: 7 July 2024.

Peyré, G. and Cuturi, M.: Computational Optimal Transport: With Applications to Data Science, Foundations and Trends in Machine Learning, 11, 355–607, https://doi.org/10.1561/2200000073, 2019.

Plu, M.: A variational formulation for translation and assimilation of coherent structures, Nonlin. Processes Geophys., 20, 793–801, https://doi.org/10.5194/npg-20-793-2013, 2013.

Ravela, S., Emanuel, K., and McLaughlin, D.: Data assimilation by field alignement, Physica D, 230, 127–145, https://doi.org/10.1016/j.physd.2006.09.035, 2007.

Rodgers, C. D.: Inverse methods for atmospheric sounding, vol. 2, World Scientific, Series on Atmospheric, Oceanic and Planetary Physics, ISBN 978-981-02-2740-1, https://doi.org/10.1142/3171, 2000.

SciPy: SciPy library main repository, Github [code], https://github.com/scipy/scipy, last access: 7 July 2024.

Sinkhorn, R.: A relationship between arbitrary positive matrices and doubly stochastic matrices, Ann. Math. Stat., 35, 876–879, 1964.

Skok, G.: Precipitation attribution distance, Atmos. Res., 295, 106998, https://doi.org/10.1016/j.atmosres.2023.106998, 2023.

Talagrand, O.: Assimilation of Observations, an Introduction, J. Meteor. Soc. Jpn., 75, 191–209, https://doi.org/10.2151/jmsj1965.75.1B_191, 1997.

Tamang, S. K., Ebtehaj, A., Zou, D., and Lerman, G.: Regularized variational data assimilation for bias treatment using the Wasserstein metric, Q. J. Roy. Meteor. Soc., 146, 2332–2346, https://doi.org/10.1002/qj.3794, 2020.

Tamang, S. K., Ebtehaj, A., van Leeuwen, P. J., Zou, D., and Lerman, G.: Ensemble Riemannian data assimilation over the Wasserstein space, Nonlin. Processes Geophys., 28, 295–309, https://doi.org/10.5194/npg-28-295-2021, 2021.

Tamang, S. K., Ebtehaj, A., van Leeuwen, P. J., Lerman, G., and Foufoula-Georgiou, E.: Ensemble Riemannian data assimilation: towards large-scale dynamical systems, Nonlin. Processes Geophys., 29, 77–92, https://doi.org/10.5194/npg-29-77-2022, 2022.

Vanderbecken, P. J., Dumont Le Brazidec, J., Farchi, A., Bocquet, M., Roustan, Y., Potier, É., and Broquet, G.: Accounting for meteorological biases in simulated plumes using smarter metrics, Atmos. Meas. Tech., 16, 1745–1766, https://doi.org/10.5194/amt-16-1745-2023, 2023.

Vilani, C.: Topics in Optimal Transportation, vol. 58 of Graduate Studies in Mathematics, American Mathematical Society, Providence, Rhode Island, ISBN 9780821833124, 2003.

Vilani, C.: Optimal Transport: Old and New, vol. 338 of Die Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin Heidelberg, ISBN 978-3-540-71049-3, 2009.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Weather Rev., 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Ying, Y.: A Multiscale Alignment Method for Ensemble Filtering with Displacement Errors, Mon. Weather Rev., 147, 4553–4565, https://doi.org/10.1175/MWR-D-19-0170.1, 2019.

Ying, Y., Anderson, J. L., and Bertino, L.: Improving Vortex Position Accuracy with a New Multiscale Alignment Ensemble Filter, Mon. Weather Rev., 151, 1387–405, https://doi.org/10.1175/MWR-D-22-0140.1, 2023.

Zhou, W., Bovik, A. C., Sheikh, H. R., and Simoncelli, E.: Image quality assessment: from error visibility to structural similarity, IEEE T. Image Process., 13, 600–612, https://doi.org/10.1109/TIP.2003.819861, 2004.