The aim of the paper is to introduce an alternative to the Euclidian distance employed in variational formulation of data assimilation: the Wasserstein distance. The Wasserstein distance originates from the optimal transport and provides a better solution to the problem of phase error. The manuscript introduces this new metric and its use in DA at a theoretical level. Then numerical illustrations are provided in one dimension for a linear advection problem and for a nonlinear shallow water. This method relies on the restrictive assumption that part of the fields are “probability distribution” over compact support in geographical domain of physical space. Such “probability distribution” should not be confused by the classical probability distributions encountered in DA that represent the uncertainty in state space e.g. the forecast error distribution or the analysis error distribution.

The manuscript is well organized with an appropriate balance between the theoretical presentation from optimal-transport background and the numerical illustrations. However, it can be improved to facilitate its reading following the recommendations made in major comments.

Major comments:

1) The example introduced in Fig. 1 to illustrate the potential of the method is not clear and could be improve as follows:
   a) You should precise the distribution name within the paragraph: “This is illustrated in Fig. 1 which shows two densities $\rho_0$ and $\rho_1$. The second density $\rho_1$ can be seen as the first one $\rho_0$ with position error.”;
   b) I guess the terminology of density & distribution & probability distribution should be avoid to prevent from any confusion in DA application, and especially the probabilistic interpretation of DA (see next comments 2);
   c) You should introduce the formalism for $L^2$ cost functions saying that the minimum of the cost function $||\rho - \rho_0||^2 + ||\rho - \rho_1||^2$ is given by $\rho_* = \frac{1}{2}(\rho_0 + \rho_1)$; while the average in the sense of the Wasserstein distance is the one of the figure, that is in between the two densities – without detailing the Wasserstein distance, as it is in the present manuscript.

2) The work presented here is limited to the case where the state vector and observations are positive fields with finite and normalized integral – part of the state vector is assume to be a probability measure over the domain – this seems very restrictive compared with the diversity of fields usually considered in data assimilation but solution to manage this issue can be considered (especially for image data). However the restriction to being a probability measure is not my objection: the problem I see is the possible confusion between probability distribution of error (forecast and analysis error distributions) and the particular case where a field (or part of the state vector) is a probability distribution. I think it would help the reader to insist on the difference between the classical framework of DA (with generic vector state) and this particular case, so to avoid any confusion between the particular field property (probability in compact domain in the physical space) and classical
error distribution (probability in state space): while mathematically appropriate, I think the
terminology of probability densities \( P(\Omega) \) (section 2.2.1 and definition 2.1) should be
replaced by something far from “probability densities”. For instance in place of “probability
densities” (title section 2.2.1 & definition), you could introduce a particular class for the
fields, for instance it could be called “mass-class”, keeping this terminology all along the
manuscript, with a remark paragraph that would precise that in optimal transport what is
socalled mass-class is actually probability distribution, indicating that the terminology is
introduced to prevent from confusion with state/error probability distribution.

3) Kantorovitch potential (K-potential) plays a crucial role in the theoretical presentation as
well as in the numerical solution of the minimizing process, but very few is said about its
computation.
   - How the K-potential is it computed in this study: please give the detail of the algorithm
     used here, the indication provided in the manuscript about the construction of the K-
potential in 1D (line 1-6 p6) is not enough. Detail, at least within a paragraph, how the K-
potential can be computed in 2D/3D, even if only 1D example are considered here.
   - Illustrate what is the K-potential for the particular case of two gaussian distribution where
     \( \rho_0 \) (\( \rho_1 \)) is a Gaussian of mean \( m_0 \) (\( m_1 \)) and variance \( \sigma_0^2 \)
     (\( \sigma_1^2 \)). If it exists, give the analytical expression for the potential in this case?

4) p12,l1-2 and l14-15: Following the author and the numerical example developed in this
section, the minimizing problem Eq(14) leads to two different solutions depending the
choice of the dot product used along the minimizing process, but no detail is given
explaining why this situation occurs. This could be due to possible multiple minima of the
cost function or to a non-convergence of the minimizing process when using the \( L^2 \) dot
product. Authors mentioned the “success of the minimization of \( J_w \)” (l15) but without
clearly indicating if the convergence was successful, or not, for the \( L^2 \) dot product. In this
simple example, uniqueness of the minimum should be guaranted, indicating that the \( L^2 \)
dot-product is not able to provide a good path toward the minimum. If this is correct than the
author should mention it more clearly:
   “In this example, the minimizing process based on the \( L^2 \) scalar product fails to reach the
unique minimum of the cost function as shown on … (additional illustration)”

   An additional figure (or panel in Fig.3) is needed to observe the non-convergence toward
   the minimum for this situation: please shows the value of the cost function \( J_w \) along the
   iterations of the minimizing process when using the two dot-products.
   I think a discussion is missing concerning existence and unicity of the \( J_w \) cost function,
   this should be included at the end of section 3.1.
   Is it possible to replace the steppest descent by a conjugated gradient? Do you think that
   this replacement could improve the convergence for the \( L2 \) gradient?

Minor comments:

1) p1, l11: “To achieve that goal” → “... this goal”
2) p1, l17: “.. to be sought (the control vector) is ..” → “.. to be sought, the control vector, is ..”
3) p7, l9: \( \omega_b \) is not defined in Eq(13)
4) p3, l8: “Wasserstein distance is to compare” → “ Wasserstein distance to compare”
5) p3, l9: “data assimilation Actual” → “data assimilation, Actual”
6) p3, l32: Observational operator is denoted by “G” in place of the more classical “H”
   notation. Please replace G into H along the manuscript.
7) p5, l23: Precise the page/section number in Ambrosio et al. (2008).
8) p 10, l14-18: Remind the equation number associated with the cost function and gradient.
   \( L^2 \) cost function is related with Eq.(2), Wasserstein cost function with Eq.(14), and the
iteration steps are deduced from Eq.(18).

9) P9, l17: write the push-forward for a given $x \in \Omega$ as $\rho_1[T(x)] |\det \nabla T_x| = \rho_0(x)$.

10) P10, l19: “$\alpha^n$ is chosen as optimal”: explain how it is computed, and provide an appropriate reference.