Supporting Information for "Towards inverse modelling of landscapes using the Wasserstein distance"

M. J. Morris[1], A. G. Lipp[2], G. G. Roberts[1]

[1]Department of Earth Science and Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK

[2]Merton College, University of Oxford, Oxford, OX1 4JD, UK

Contents of this file

1. Calculation of two dimensional Wasserstein distances for synthetic landscapes
2. Figure S1: Two dimensional Wasserstein distance for a synthetic landscape

Two Dimensional Wasserstein

As stated in the main text, the Wasserstein distance can be applied to distributions of N dimensions. We explain here the approach to calculate a Wasserstein distance for $N = 2$, using a pair of two dimensional density functions, $f(x, y)$ and $g(x, y)$. First, a cost matrix is calculated, representing the distance from each coordinate point in the source function (f) to every coordinate point in the target function (g). A choice must be made about the method used to define the distance between these points, known as

the cost function. Several approaches exist including the squared Euclidean distance and Manhattan distance. The cost c_{ij} to transport point i to point j using a squared Euclidean distance is given by

$$c_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2, \tag{1}$$

where x_i and y_j are the x and y coordinates of the ith and jth discrete points in the distributions respectively. A map of transport cost, known as the cost matrix may be constructed by computing the cost function for all coordinates.

There are a number of ways to transport f to g, however some are more efficient than others. Any chosen method of transport is given by a transport plan, π_{ij}, defined as the mapping of the source distribution onto the target distribution. i.e., entry i, j corresponds to how much of the source distribution f is translated from point i onto point j. Thus the total cost, C, of any given transport plan is

$$C = \sum_{i=1,j=1}^{n_f,n_g} \pi_{ij} c_{ij}. \tag{2}$$

We seek the most efficient, or optimal, transport plan, which results in the least cost to transport f to g (i.e. minimises C). This transport plan can be given generally, as per Kantorovich (1942), by the Linear Programming problem

$$W_p^p = \min_{\pi_{ij}} \sum_{i,j} \pi_{ij} c_{ij}, \tag{3}$$

where p is dependent on the exponent of the cost function. In our case for a squared euclidean distance, $p = 2$. Equation 3 is subject to three constraints. First, and secondly, the row and column sums of the transport plan are equal to the number of elements in f and

g respectively, and thirdly, that the optimal transport plan is greater than or equal to zero.

Examples of Wasserstein distances calculated using two dimensional density functions (Equation 3) are shown in Figure S1. This figure shows the results of an experiment similar to that shown in Figure 2 of the main manuscript. Due to the computational expense involved with this approach, only 7 translations of the source landscape are presented, rather than the 200 translations used for the marginals approach. Nonetheless, Figure S1c shows that the calculated misfit function correctly identifies the optimal location of the source landscape. We provide a python script which calculates the Wasserstein distance between two synthetic landscapes using the methodology described above at `github.com/MatthewJMorris/landscape-wasserstein`.

References

Kantorovich, L. V. (1942). On the Translocation of Masses. *Dokl. Acad. Nauk. USSR*, *37*, 7–8, 227–229.

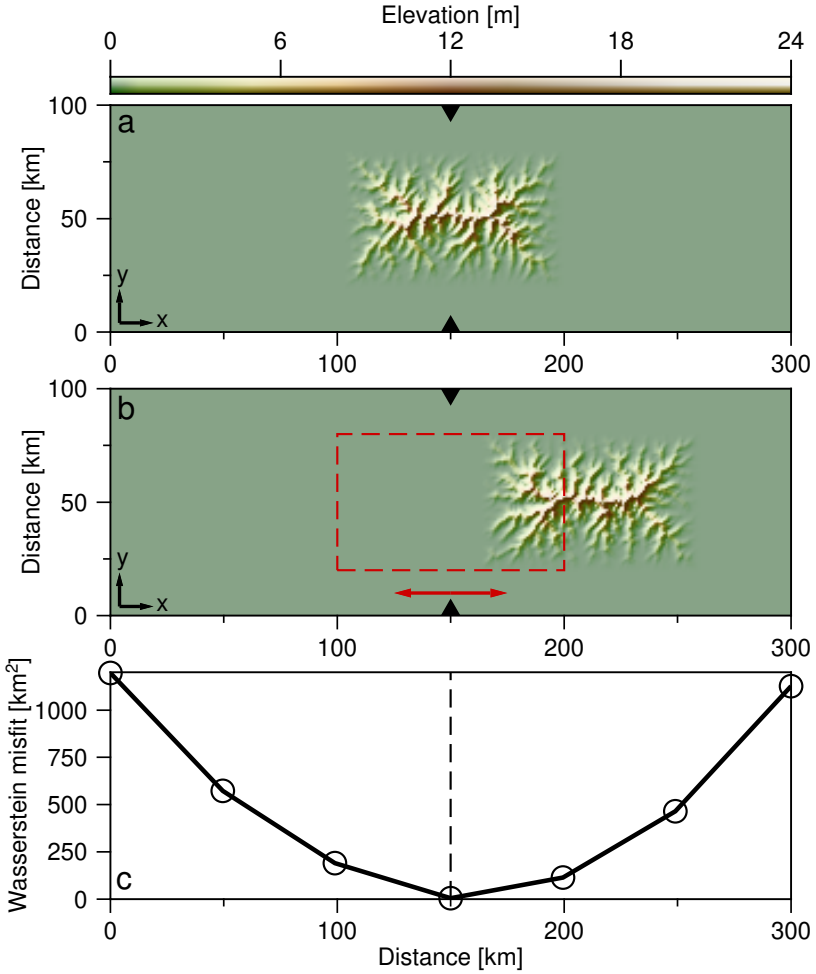

Figure S1. Two-dimensional Wasserstein distances for synthetic landscapes. (a) Target landscape, as per Figure 2b of the main manuscript. (b) Source landscape, as per Figure 2e of the main manuscript. (c) Two dimensional Wasserstein distance calculated between the target and the source landscapes for 7 translations (circles) of the source landscape across the domain using the two dimensional approach (Equation 3).