

# A rough-and-ready cluster-based approach for extracting finite-time coherent sets from sparse and incomplete trajectory data

Gary Froyland<sup>1</sup> and Kathrin Padberg-Gehle<sup>2</sup>

<sup>1</sup>*School of Mathematics and Statistics, The University of New South Wales, Sydney NSW 2052, Australia*

<sup>2</sup>*Technische Universität Dresden, Fachrichtung Mathematik, Institut für Wissenschaftliches Rechnen, D-01062 Dresden, Germany*

(Dated: 24 June 2015)

We present a numerical method to identify regions of phase space that are approximately retained in a mobile compact neighbourhood over a finite time duration. Our approach is based on spatio-temporal clustering of trajectory data. The main advantages of the approach are the ability to produce useful results (i) when there are relatively few trajectories and (ii) when there are gaps in observation of the trajectories as can occur with real data. The method is easy to implement, works in any dimension, and is fast to run.

Keywords: coherent sets, Lagrangian coherent structure, spatio-temporal clustering.

Coherent features in time-dependent dynamical systems are difficult to identify, and considerable effort has been put into the development of identification algorithms. Most approaches require knowledge of the dynamical system or high-resolution trajectory information, which in applications may not be available. We present a trajectory-based method that is aimed squarely at the situation where the available information is poor: there are few trajectories, the available trajectories do not span the full time duration under consideration, and there are missing observations within trajectories. As our method is very simple to implement and fast to run, it also provides a rapid “first cut” coherent structure analysis even in situations where the full dynamical system or high-resolution trajectory data is available.

## I. INTRODUCTION

There are a number of different concepts that describe the notion of coherent behaviour in time-dependent dynamical systems. Probabilistic approaches define finite-time coherent sets<sup>12,17,18</sup> as regions of phase space that minimally mix with the surrounding phase space during a specified time duration of finite length. Lagrangian coherent structures can be defined as material lines that extremize a certain stretching or shearing quantity<sup>20,21</sup>, while another approach tries to identify curves on which local dynamics approximates local rigid-body motion<sup>25</sup>. There are also topological<sup>1</sup> and ergodicity-based<sup>6</sup> descriptions of coherence, although these are not designed for aperiodic dynamics. Finally, a recent geometric characterisation<sup>11</sup> defines finite-time coherent sets as those sets with boundary to volume ratios that remain minimal under the evolution of the dynamics, and proves that such a characterisation arises naturally as the advective limit of the probabilistic approaches<sup>12,18</sup>.

In the present paper we develop cluster-based tech-

niques to highlight distinct groups of trajectories that remain in compact, approximately spherical subregions of phase space over a finite time duration. Let  $\Phi : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$  denote the flow of a continuous time dynamical system on  $\mathbb{R}^d$ , i.e.  $\Phi(x, t)$  denotes the state of the system at time  $t$  with initial value  $x$  (at time 0). We define a dynamic metric

$$\mathbf{D}(x, y) := \int_0^T \rho(\Phi(x, t), \Phi(y, t))^2 dt, \quad (1)$$

based on some metric  $\rho$  on  $\mathbb{R}^d$ . For example, if  $\rho$  is the Euclidean metric, then  $x$  and  $y$  are close according to  $\mathbf{D}$  provided they remain close in a Euclidean sense averaged over the time interval  $[0, T]$ .

In this general setup, one is free to choose  $\rho$  and also how the terms  $\rho(\Phi(x, t), \Phi(y, t))$ ,  $t \in [0, T]$  are combined to form  $\mathbf{D}(x, y)$ . Setting  $\rho$  to be the Euclidean metric is a natural choice if the trajectory data lies in  $\mathbb{R}^d$  and shortly we will give geometric reasons for why this is a good choice. The sum-of-squares combination is a convenient form for the specific numerical clustering approach proposed below. One could alternatively define, for example,  $\mathbf{D}(x, y) := \int_0^T \rho(\Phi(x, t), \Phi(y, t))^p dt$ , for  $1 \leq p < \infty$ , or  $\mathbf{D}(x, y) := \max_{t \in [0, T]} \rho(\Phi(x, t), \Phi(y, t))$ .

In practice, suppose we have  $n$  trajectories given at discrete times  $\{0, 1, \dots, T\}$ , denoted  $x_{i,t} \in \mathbb{R}^d$ ,  $i = 1, \dots, n, t = 0, \dots, T$ . We wish to cluster the initial points  $x_{i,0} \in \mathbb{R}^d$  according to  $\mathbf{D}$ . The discrete-time version of (1) is

$$\mathbf{D}(x_{i,0}, x_{j,0}) = \sum_{t=0}^T \rho(x_{i,t}, x_{j,t})^2, \quad (2)$$

for  $1 \leq i, j \leq n$ . At this point, one could calculate  $n(n-1)/2$  interpoint distances  $\mathbf{D}(x_{i,0}, x_{j,0})$ ,  $1 \leq i < j \leq n$ . This general approach of clustering using the dynamic metric (1) or (2) is very flexible and in principle one could employ any suitable (according to the three properties outlined below) clustering method on  $\mathbb{R}^d$  from

the vast number available (see e.g. Ref. 34): centroid-based algorithms like k-means<sup>3,24</sup> and fuzzy *c*-means<sup>4,5</sup>, spectral methods<sup>9,37</sup>, density-based clustering<sup>2,8</sup>, and methods based on community detection<sup>10</sup> (for example, modularity<sup>30</sup>).

The sum-of-squares form of  $\mathbf{D}$  and choice of Euclidean metric for  $\rho$  allows us to further rewrite (2) as

$$\mathbf{D}(x_{i,0}, x_{j,0}) = \sum_{t=0}^T \rho(x_{i,t}, x_{j,t})^2 = \|X_i - X_j\|^2, \quad (3)$$

for  $1 \leq i, j \leq n$ , where  $X_i = (x_{i,0}, x_{i,1}, \dots, x_{i,T})$ . Thus, we have a convenient representation of the dynamic metric  $\mathbf{D}$  on  $\mathbb{R}^d$  as the squared Euclidean distance between trajectories in  $\prod_{t=0}^T \mathbb{R}^d = \mathbb{R}^{d(T+1)}$ .

As elaborated in the next section we use the fuzzy *c*-means clustering algorithm<sup>4,5</sup> on  $\mathbb{R}^{d(T+1)}$ . Our reasons for using fuzzy *c*-means are threefold (but not necessarily exclusive to fuzzy *c*-means):

1. When searching for  $K$  clusters<sup>31</sup>, fuzzy *c*-means produces  $K$  auxiliary “centres” and aims to allocate data to clusters by reducing the total squared distance from the data to their corresponding centre. If  $\rho$  is the Euclidean metric, then fuzzy *c*-means will favour clusters that are close to spherical at each time instant. Such clusters therefore will not “spread out” in phase space, will remain in an approximately tubular region in lower-dimensional space-time (phase space plus one time coordinate, see Figure 1), and will on average have low boundary size to volume at each time instant. The low boundary size relative to volume property is compatible with probabilistic approaches<sup>12,18</sup> and geometric approaches<sup>11</sup> to finite-time coherent sets.
2. Fuzzy *c*-means provides feedback in the form of the membership value describing the likelihood that a trajectory belongs to a cluster. Because finite-time coherent sets do not necessarily fully partition the phase space, we can identify non-coherent collections of trajectories as those with a low membership for all clusters.
3. Fuzzy *c*-means is computationally efficient, particularly for large numbers of trajectories.

Finally, we note that once the initial points  $\{x_{i,0}\}_{1 \leq i \leq n}$  have been clustered, the full trajectories are also clustered, as by definition trajectories remain within the same cluster for all  $t = 0, \dots, T$ .

Clustering trajectory data is a recent problem in the analysis of spatio-temporal datasets, with many contributions found in the data mining literature. We refer the reader to Ref. 23 for a recent review and to the literature review in Ref. 22 for a brief summary of the different approaches used for spatio-temporal clustering. Ref. 22 proposes an augmented fuzzy *c*-means algorithm, with different weights for the temporal and spatial components. Ref. 29 and the thesis 28 introduce a distance

measure essentially identical in form to (1). Ref. 29 uses this metric with density-based algorithms<sup>2</sup> to cluster trajectories in geo-referenced data sets and to find optimal time intervals for clustering. The papers 22, 28, and 29 do not consider how to treat incomplete trajectory data. Other distance measures have been proposed to account for application-specific purposes, e.g. for studying movement patterns in traffic<sup>32,33</sup>. However, to the best of our knowledge, spatio-temporal clustering approaches have not been employed for studying transport phenomena and coherent behaviour in time-dependent dynamical systems and an exploration of the tuning of clustering methods to this application has not been undertaken.

Our main contributions are (i) posing the problem of identifying finite-time coherent sets as an objective trajectory-based clustering problem, (ii) developing a methodology for handling incomplete data that consistently uses all available data, and (iii) indicating some rules of thumb for applying these techniques in practice.

An outline of the paper is as follows. We describe our approach first in the situation where there are a finite number of trajectories available, sampled at a finite number of times. We then consider combinations of a continuum of trajectories and a continuum of observation times. Section II concludes by showing that our clustering framework for coherent sets is objective and independent of isotropic scaling of space and time. Our method handles missing trajectory data naturally and we discuss this in Section III. A discussion of false positives, possibly inaccurate results, and how to identify these is in Section IV; we also outline some rules of thumb for parameter choices. Section V illustrates the approach for several examples: firstly in one-dimensional dynamics, where the geometry of the spatio-temporal clustering is more transparent, secondly for the well-known double-gyre flow and the transitory double gyre flow for comparison with existing coherent set identification approaches, and thirdly on ocean surface drifter data. We demonstrate how reasonable results can be achieved even when large percentages of trajectory observations are missing, and when the trajectory dataset is comprised of trajectories much shorter than the full time duration under analysis.

## II. FULL DATA CASE

We first describe the case where all trajectories span the finite time duration and there are no missing observations. We begin by describing our setup in the situation where there are a finite number of trajectories sampled at a finite collection of time instances, this means that all trajectories are sampled at all time instances; we then follow with versions that are continuous in space and/or time.

### A. Discrete setting: a finite number of finitely-sampled trajectories.

Suppose we have  $n$  trajectories of maximum length  $T + 1$ , denoted  $x_{i,t} \in \mathbb{R}^d, i = 1, \dots, n, t = 0, 1, \dots, T$ . We consider each trajectory  $\{x_{i,t}\}_{0 \leq t \leq T}$  as a point  $X_i = (x_{i,0}, x_{i,1}, \dots, x_{i,T}) \in \mathbb{R}^{d(T+1)}$ , which we also refer to as a trajectory. We imagine  $\mathbb{R}^{d(T+1)}$  as  $\prod_{t=0}^T \mathbb{R}^d$ , where the product of copies of our phase space  $\mathbb{R}^d$  is ordered in increasing time  $t$ . The fuzzy  $c$ -means clustering algorithm<sup>4,5</sup> is a soft clustering based on the calculation of a centre for each cluster and a likelihood of membership of each data point to each centre. Suppose we have identified  $K$  cluster centres  $C_k \in \mathbb{R}^{d(T+1)}, k = 1, \dots, K$ . We may decompose the  $C_k$  as  $(c_{k,0}, c_{k,1}, \dots, c_{k,T}) \in \mathbb{R}^{d(T+1)}$ , so that each  $c_{k,t} \in \mathbb{R}^d$  can be regarded as a point in phase space  $\mathbb{R}^d$  at time  $t$ . Associated with each trajectory  $X_i$  is a likelihood  $0 \leq u_{k,i} \leq 1$  of  $X_i$  being associated with the cluster centre  $C_k$ .

Given trajectories  $X_i, i = 1, \dots, n$ , cluster centres  $C_k, k = 1, \dots, K$ , and membership likelihoods  $u_{k,i}, i = 1, \dots, n, k = 1, \dots, K$ , the total “goodness of fit” of the memberships of trajectories in clusters is measured by likelihood-weighted intracluster distances, which we wish to minimise:

$$\sum_{k=1}^K \sum_{i=1}^n u_{k,i}^m \|X_i - C_k\|^2 = \sum_{k=1}^K \sum_{i=1}^n u_{k,i}^m \sum_{t=0}^T \|x_{i,t} - c_{k,t}\|^2. \quad (4)$$

This minimisation is subject to the constraints that (i)  $\sum_{k=1}^K u_{k,i} = 1$  for  $i = 1, \dots, n$  and (ii)  $u_{k,i} \geq 0$  for all  $k = 1, \dots, K, i = 1, \dots, n$ . The parameter  $m > 1$  is the fuzziness exponent. Increasing  $m$  corresponds to softer clusters, while as  $m$  approaches 1, the membership likelihoods converge to either 0 or 1, resulting in a hard clustering<sup>5</sup> (this latter effect is most easily seen from the update rule (6) below). The basic fuzzy  $c$ -means algorithm<sup>4,5</sup> in our notation above proceeds as follows.

#### Algorithm 1:

1. Initialize membership values  $u_{k,i}$  either randomly or computed via step 3 based on an initial seeding of  $K$  centres (e.g. randomly or by the  $k$ -means++ algorithm<sup>3</sup>)
2. Calculate centres:

$$C_k = \frac{\sum_{i=1}^n u_{k,i}^m X_i}{\sum_{i=1}^n u_{k,i}^m}, \quad (5)$$

$$k = 1, \dots, K.$$

3. Update membership values:

$$u_{k,i} = \frac{1/\|X_i - C_k\|^{2/(m-1)}}{\sum_{j=1}^K (1/\|X_i - C_j\|^{2/(m-1)})}, \quad (6)$$

$$k = 1, \dots, K, i = 1, \dots, n.$$

4. Evaluate objective (4). If the improvement in the objective is below a threshold, go to step 5; otherwise go to step 2.
5. Output cluster centres  $C_k \in \mathbb{R}^{d(T+1)}, k = 1, \dots, K$  and membership likelihoods  $u_{k,i} \in [0, 1], k = 1, \dots, K, i = 1, \dots, n$ .

The update rule in step 2 is constructed by fixing the  $u_{k,i}$  and choosing  $C_k$  so that the gradient of the objective (4) is zero. Similarly, the update rule in step 3 is constructed by fixing the  $C_k$  and choosing the  $u_{k,i}$  so that the gradient of the Lagrangian incorporating (4) and the constraint  $\sum_{k=1}^K u_{k,i} = 1$  is zero.

**Implementation in MATLAB:** If  $X$  is an  $n \times d(T+1)$  array of  $n$  trajectories in  $\mathbb{R}^d$  (i.e. the rows of  $X$  are the vectors  $X_i, i = 1, \dots, n$  discussed above), Algorithm 1 is implemented in MATLAB by the function `fcm` in the Fuzzy Logic Toolbox:

```
opts(1)=m;
[c,u]=fcm(X,K,opts);
```

If `fcm` is called without `opts`, the default value of  $m$  is 2. For  $d = 2$ , to display the membership values for cluster  $k$  at time slice  $t \in \{0, 1, \dots, T\}$ , one can use

```
scatter(X(:,2*t+1),X(:,2*t+2),[],u(k,:),'.');
```

For  $d = 3$ , one can similarly use `scatter3`.

We remark that the centres  $C_k = (c_{k,0}, c_{k,1}, \dots, c_{k,T})$ ,  $k = 1, \dots, K$ , are generally not true trajectories of the dynamical system, although they may be remarkably close to true trajectories in some cases. For each  $k = 1, \dots, K$ , one can identify the maximum likelihood trajectory  $X_{i_k^*}$  for the  $k^{\text{th}}$  cluster, where  $i_k^* = \arg \max_i u_{k,i}$ . The trajectory  $X_{i_k^*}$  is the most likely to belong to the  $k^{\text{th}}$  cluster and may be thought of as a “probabilistic centre” of the cluster. The probabilistic centers can be interpreted as “low dimensional representations” of the macroscopic behavior of the system, as they describe the coherent motion of trajectories in the corresponding cluster. We illustrate both the centre and maximum likelihood trajectory in Figure 1.

Note that there is potential for one to include weights as coefficients for the terms  $\|x_{i,t} - c_{k,t}\|^2$  in (4), which could depend on  $i, t$ , or  $k$ . A particularly important example is the inclusion of weights  $q_i \geq 0$  corresponding to the “mass” assigned to a point  $x_{i,0}$ . For example, if one is searching for coherent regions in an oil or chemical spill in the ocean, one is likely interested in the behavior of the oil or chemical, rather than the water. In order to obtain clusters that focus on the nonuniform distribution of oil or chemical, one can replace  $u_{k,i}^m$  with  $q_i u_{k,i}^m$  in (4) and (11).

Also note that at present, clustering into spheres is preferred by the Euclidean norm. If one wishes to favour clustering into ellipsoids, with orthogonal semi-axis vectors  $v_1, \dots, v_d$  and corresponding semi-axis

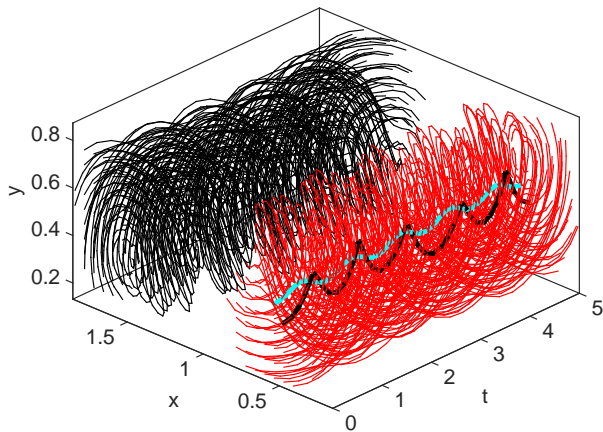


FIG. 1: Result of clustering of the double-gyre flow with  $n = 512$ ,  $K = 2$ ,  $m = 2$ ,  $d = 2$ ,  $T = 50$  (corresponding to the real-time interval  $[0, 5]$  in time steps of 0.1).

Trajectories  $X_i$  with membership values  $u_{k,i} > 0.9$ ,  $k = 1, 2$  are shown. Note that in this figure we display space-time using time as a third coordinate, whereas the clustering computations in Algorithm 1 occur in  $\mathbb{R}^{102}$ . One sees that the red and black trajectories remain in a compact region as time evolves. The probabilistic centre of the cluster corresponding to the red trajectories is shown as a meandering black curve, the light blue curve is the centre as computed by Algorithm 1.

lengths  $\ell_1, \dots, \ell_d$ , then one may simply scale the  $x_{i,t}$  data in  $\mathbb{R}^d$  along each  $v_j$  by  $1/\ell_j$ ,  $j = 1, \dots, d$ , use the Euclidean norm in the objective above, and then rescale the data along each  $v_j$  by  $\ell_j$ ,  $j = 1, \dots, d$ . Other distance functions could be used to replace Euclidean distance, but in the absence of specific replacement motivations based on known properties of the underlying dynamical system, Euclidean distance represents a natural isotropic default distance metric.

Sections II B–II D outline extensions of the above setup to situations where either one or both of the spatial data or temporal data are on a continuum. These constructions are mainly of a theoretical nature, but have been included to (i) demonstrate what the analogous objects are in a continuum setting if e.g. a full dynamical systems model were available, and (ii) indicate how the discrete “finite data” setting above is a special case (constructed by subsampling in space and/or time) of the continuum “full model” setting. Sections II B–II D could be omitted on a first reading.

### B. Semi-continuous setting #1: a continuum of initial points, with trajectories finitely-sampled in time.

Suppose we have a continuum of initial points in a set  $A \subset \mathbb{R}^d$ . We now write trajectories as  $x(x_0, t) \in \mathbb{R}^d$ ,  $x_0 \in A$ ,  $t = 0, 1, \dots, T$ . Individual trajectories  $\{x(x_0, t)\}_{0 \leq t \leq T}$  for fixed  $x_0 \in A$  are still regarded as

elements of  $\mathbb{R}^{d(T+1)}$  as before, and we write an individual trajectory as  $X(x_0) \in \mathbb{R}^{d(T+1)}$ . Note that the likelihoods  $u_{x_0,k}$  are also continuously parameterised by  $x_0 \in A$ , and we write these in functional form as  $u_k(x_0)$ , so that  $u_k : A \rightarrow [0, 1]$ ,  $k = 1, \dots, K$ . In the finite-trajectory setting, the initial points of trajectories need not be uniformly distributed over the phase space, nor be given a uniform weight. If one wishes to model the evolution of a passive tracer field with nonuniform density, one will either have a greater density of points in areas of high tracer density or apply weights to points with higher tracer concentration. To capture this effect in the continuum setting, we need a density function  $q : A \rightarrow \mathbb{R}$ , satisfying  $\int_A q(y) dy = 1$ . We interpret  $\int_{B_\epsilon(x_0)} q(y) dy$  as the fraction of initial points that belong to an  $\epsilon$ -neighbourhood of  $x_0$ . For example, if the initial  $x_0$  are uniformly sampled over  $A$ , then  $q \equiv 1/\text{vol}(A)$ .

Equation (4) now reads

$$\begin{aligned} & \sum_{k=1}^K \int_A u_k(x_0)^m \|X(x_0) - C_k\|^2 q(x_0) dx_0 \\ &= \sum_{k=1}^K \sum_{t=0}^T \int_A u_k(x_0)^m \|x(x_0, t) - c_{k,t}\|^2 q(x_0) dx_0 \quad (7) \end{aligned}$$

Here is a simple example to help visualise what is going on. Let phase space be  $[0, 1]$ , and consider trajectories of length two, generated by a map  $S : [0, 1] \circlearrowleft$ . Geometrically, we look for clusters in data of the form  $(x(x_0, 0), x(x_0, 1))$ , for all  $x_0 \in [0, 1]$ , which is nothing but the (weighted, if  $q$  is not constant) graph of  $S$  considered as a one-dimensional subset of  $[0, 1]^2$ ; see Figure 2(a). Figure 2(b) shows clusters in data of the form  $(x(x_0, 0), x(x_0, 2))$ .

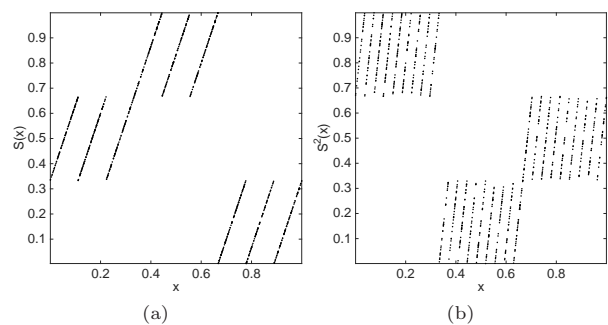


FIG. 2: Graphs of the interval map  $S$  (15), which permutes the intervals  $[0, 1/3]$ ,  $[1/3, 2/3]$ ,  $[2/3, 1]$ . (a)  $x$  vs  $S(x)$ ; (b)  $x$  vs  $S^2(x)$ . Considered as a time series of length 2, Algorithm 1 would seek clusters in these graphs, considered as a subset of  $[0, 1]^2$ .

### C. Semi-continuous setting #2: a finite number of continuously-sampled trajectories.

Suppose now that instead of a continuum of finitely-sampled trajectories, we now have a finite collection of trajectories observed continuously in time:  $x_i(t), i = 1, \dots, n, t \in [0, T]$ , with optional weights  $q_i, i = 1, \dots, n$ , representing the mass of the point  $x_i(0)$ . We consider  $x_i : [0, T] \rightarrow \mathbb{R}^d, i = 1, \dots, n$  as a finite number of continuous mappings from  $[0, T]$  to  $\mathbb{R}^d$ . Similarly, the cluster centres  $c_k : [0, T] \rightarrow \mathbb{R}^d, k = 1, \dots, K$  are a finite number of (not necessarily continuous) mappings from  $[0, T]$  to  $\mathbb{R}^d$ . The RHS of (4) becomes

$$\sum_{k=1}^K \sum_{i=1}^n u_{k,i}^m \int_0^T \|x_i(t) - c_k(t)\|^2 q_i dt. \quad (8)$$

To visualise the integral, imagine we have a one-dimensional (time-dependent) flow in a phase space  $\mathbb{R}$ , and  $x_i(t)$  is one trajectory from  $t = 0$  to  $t = T$ . The integral (8) computes the total ( $q$ -weighted) squared distance between the graphs of the functions  $x_i$  and  $c_k$  in  $[0, T] \times \mathbb{R}$ ; or in other words, the squared  $L^2$  distance  $\|x_i - c_k\|_{L^2([0,T])}^2$ .

### D. Fully continuous setting: a continuum of initial points, continuously-sampled trajectories.

Combining the constructions from the previous two paragraphs, we now have functions  $x : A \times [0, T] \rightarrow \mathbb{R}^d$ . Our likelihood functions remain as  $u_k : A \rightarrow \mathbb{R}^+, k = 1, \dots, K$ . The RHS of (4) becomes

$$\sum_{k=1}^K \int_0^T \left( \int_A u_k(x_0)^m \|x(x_0, t) - c_k(t)\|^2 q(x_0) dx_0 \right) dt. \quad (9)$$

### E. Isotropic scaling of space and time has no effect

Given that our clustering is occurring in the product space formed from as many copies of our phase space as there are time instants, it is pertinent to consider the effect, if any, of isotropically scaling space and time. We show that in fact, there is no real effect caused by such scaling.

In the fully discrete setting, by (4) if space were scaled isotropically by a factor  $\alpha$  and time by a factor  $\beta$ , then (4) would simply increase by  $\alpha^2$ . Thus the cluster centres and likelihood functions that minimise (4) are simply isotropically scaled and unchanged, respectively, under this isotropic scaling of space and/or time.

In the continuum setting, we again consider scaling space isotropically by a factor  $\alpha$  and time by a factor  $\beta$ . This amounts to defining new ‘‘primed’’ variables:  $x'_0 = \alpha x_0, t' = \beta t, A' = \alpha A, T' = \beta T, x'(x'_0, t') =$

$\alpha x(x_0, t), c'_k(t') = \alpha c_k(t), u'_k(x'_0) = u_k(x_0)$ , and  $q'(x'_0) = q(x_0)/\alpha^d$ . Then changing variables from  $x_0$  to  $x'_0$  and from  $t$  to  $t'$  we have

$$\begin{aligned} (9) &= \sum_{k=1}^K \int_0^{\beta T} \left( \int_{\alpha A} u_k(x'_0)^m \|x(x'_0, t') - c_k(t')\|^2 \right. \\ &\quad \left. q(x'_0) \alpha^{-d} \beta^{-1} dx'_0 \right) dt' \\ &= \alpha^{-2} \beta^{-1} \sum_{k=1}^K \int_0^{T'} \left( \int_{A'} u'_k(x'_0)^m \|x'(x'_0, t') - c'_k(t')\|^2 \right. \\ &\quad \left. q'(x'_0) dx'_0 \right) dt'. \end{aligned}$$

Thus, switching to the primed coordinates will simply increase the objective (9) by a constant factor  $\alpha^2 \beta$  over the original unprimed coordinates. Again, the cluster centres and likelihood functions that minimise (9) are isotropically scaled and unchanged, respectively, under this isotropic scaling of space and/or time. In particular, the clustering algorithm does not care how space is scaled against time.

### F. Frame-independence

To check frame-independence of an algorithm, one applies the algorithm to an original dataset, then subjects the dataset to a (possibly time-dependent) affine transformation, where the linear part is orthogonal. If the algorithm applied to the transformed dataset yields the transformed output of the original dataset, then the algorithm is frame-independent; see Ref. 38 for details.

We consider the situation where we have a finite collection of finitely-sampled trajectories; the arguments presented apply equally to the other situations discussed in Sections II B–II D.

**Proposition:** Algorithm 1 is frame-independent.

*Proof:* Let  $\{x_{i,t}\}_{1 \leq i \leq n, 0 \leq t \leq T}$  be an original collection of trajectories. Apply Algorithm 1 to  $\{x_{i,t}\}$  to obtain centres  $C_k = (c_{k,0}, \dots, c_{k,T}) \in \mathbb{R}^{d(T+1)}, k = 1, \dots, K$  and likelihoods  $u_{k,i}, k = 1, \dots, K, i = 1, \dots, n$  that minimise (4). Denote the transformed trajectories  $y_{i,t} := O_t x_{i,t} + o_t$ , where  $O_t$  is an orthogonal  $d \times d$  matrix and  $o_t \in \mathbb{R}^d$ . Form transformed centres  $c'_{k,t} := O_t c_{k,t} + o_t$ . Notice that (4) has the same value when evaluated with  $\{x_{i,t}\}, \{c_{k,t}\}$  and  $\{u_{k,i}\}$ , and with  $\{y_{i,t}\}, \{c'_{k,t}\}$  and  $\{u_{k,i}\}$ . This is because the transformation  $x \mapsto O_t x + o_t$  is an isometry with respect to the Euclidean norm for each  $t = 0, 1, \dots, T$ . Because  $\{c_{k,t}\}$  and  $\{u_{k,i}\}$  minimise (4) for the dataset  $\{x_{i,t}\}$ , one has  $\{c'_{k,t}\}$  and  $\{u_{k,i}\}$  minimise (4) for the dataset  $\{y_{i,t}\}$ .  $\square$

If we use a non-standard inner product  $\langle \cdot, \cdot \rangle := x^\top Q x$  for some symmetric positive-definite  $d \times d$  matrix  $Q$  to define a norm  $\|\cdot\|'$  on each phase space slice  $\mathbb{R}^d$ , then an analogous proposition would hold under transformations  $x \mapsto O_t x + o_t$  provided  $O_t^* = Q^{-1} O_t^\top Q = O_t^{-1}$ .

### III. TREATMENT OF MISSING DATA

Missing data can be treated naturally in our spatio-temporal clustering framework. Taking the finitely sampled, finite trajectory setting of Section II A, by *missing data*, we mean a trajectory  $\{x_{i,t}\}_{0 \leq t \leq T}$  where the values  $x_{i,t}$  are available only on a strict subset of time instances  $\mathcal{T}_i \subset \{0, \dots, T\}$ ; that is, only  $\{x_{i,t}\}_{t \in \mathcal{T}_i}$  is available. In terms of the abstract dynamic norm (3), we handle this by leaving out those terms in the sum over  $t$  in equation (3) that correspond to times at which data is unavailable. Thus, the treatment of missing data we propose is not specific to fuzzy clustering. In the fuzzy clustering framework, this corresponds to excluding those time instants  $t \in \mathcal{T}_i^c$  for which  $x_{i,t}$  is unavailable from both the centre update and membership likelihood update rules. Thus, only data that is available at a particular time instant  $t$  is used to calculate cluster centre coordinates at that time  $t$ .

To do this efficiently, we consider the known portion of trajectory  $i$ , namely  $\{x_{i,t}\}_{t \in \mathcal{T}_i}$ , as a point in the lower-dimensional space  $\mathbb{R}^{d|\mathcal{T}_i|}$  for the purposes of computing Euclidean distances in the clustering algorithm. This projection to a lower-dimensional space is easily incorporated into Algorithm 1. For  $i = 1, \dots, n$ , we define  $\pi_i : \mathbb{R}^{d(T+1)} \rightarrow \mathbb{R}^{d|\mathcal{T}_i|}$  by  $\pi_i X_i = \hat{X}_i = (\hat{x}_{i,0}, \hat{x}_{i,1}, \dots, \hat{x}_{i,T})$ , where

$$\hat{x}_{i,t} = \begin{cases} x_{i,t}, & \text{if } t \in \mathcal{T}_i; \\ 0, & \text{if } t \notin \mathcal{T}_i. \end{cases} \quad (10)$$

To exclude unavailable observations from centre updates, for each time instant  $t = 0, 1, \dots, T$ , we define  $\mathcal{I}_t = \{i : t \in \mathcal{T}_i\} \subset \{1, \dots, n\}$ , namely the indices of all trajectories with observations available at time  $t$ .

#### Algorithm 2: Clustering with missing data

1. Initialize membership values  $u_{k,i}$ .
2. Calculate centres:

$$c_{k,t} = \frac{\sum_{i \in \mathcal{I}_t} u_{k,i}^m x_{i,t}}{\sum_{i \in \mathcal{I}_t} u_{k,i}^m}, \quad (11)$$

$k = 1, \dots, K$ ,  $t = 1, \dots, T$ . Note that we take a convex combination over only those observations available at time  $t$ .

3. Update membership values:

$$u_{k,i} = \frac{1/\|\pi_i X_i - \pi_i C_k\|^{2/(m-1)}}{\sum_{j=1}^K (1/\|\pi_i X_i - \pi_i C_j\|^{2/(m-1)})}, \quad (12)$$

$k = 1, \dots, K$ ,  $i = 1, \dots, n$ . Note that when computing Euclidean distances, we project onto only those temporal copies of phase space in which trajectory data for  $X_i$  is available.

4. Evaluate the objective

$$\sum_{k=1}^K \sum_{i=1}^n u_{k,i}^m \|\pi_i X_i - \pi_i C_k\|^2. \quad (13)$$

If the improvement in the objective is below a threshold, go to step 5; otherwise go to step 2.

5. Output cluster centres  $C_k \in \mathbb{R}^{dT}$ ,  $k = 1, \dots, K$  and membership likelihoods  $u_{k,i} \in [0, 1]$ ,  $k = 1, \dots, K$ ,  $i = 1, \dots, n$ .

Algorithm 2 is also frame-independent; the proof is identical to the proof of frame-independence of Algorithm 1.

We remark that Algorithm 2 will have a preference for clusters that each contain a similar total amount of data; for example, one cluster comprising 20 trajectories of length ten and another comprising 40 trajectories of length five both contain the same amount of data. In some problems, one may wish Algorithm 2 to have a preference for clusters with similar numbers of trajectories, irrespective of the amount of available data in each trajectory. To achieve this, one can replace  $u_{k,i}^m$  with  $u_{k,i}^m/|\mathcal{T}_i|$  in (11) in Step 2 and (13) in Step 4. The reasoning behind this replacement is that with the factor  $1/|\mathcal{T}_i|$ , (13) computes the *average* weighted squared distances from centres (per trajectory), whereas without this factor, the *total* squared distances from centres is computed. With this altered objective function, one constructs the correspondingly altered update rules (11)–(12) as outlined below Algorithm 1. We tested Algorithm 2 with and without this factor in the examples in Section V and found little difference; we report the results without this factor.

### IV. WHAT CAN GO WRONG?

Before we begin to outline some guidelines to avoid potential pitfalls in sections IV B–IV E, we introduce a quantity that (along with the likelihoods  $u_{k,i}$ ) can be useful for assessing confidence in the clustering reported by Algorithms 1 or 2.

#### A. Entropy and classification uncertainty

Each trajectory  $x_{i,t}$ ,  $t \in \mathcal{T}_i$  has relative probabilities  $u_{k,i} \in [0, 1]$ , of belonging to cluster  $C_k$ ,  $k = 1, \dots, K$ , respectively. We can now define an overall measure of certainty of cluster assignment of trajectory  $i$  via the normalised entropy of the probability vector  $[u_{1,i}, \dots, u_{K,i}]$ , namely

$$h_i := \frac{-\sum_{k=1}^K u_{k,i} \log u_{k,i}}{\log K}. \quad (14)$$

The quantity  $h_i$  takes values between 0 and 1, with  $h_i = 0$  representing certain classification of trajectory  $i$  to one

of the  $K$  clusters and  $h_i = 1$  representing complete uncertainty of classification of trajectory  $i$  to one of the  $K$  clusters, see also Ref. 4.

A collection of trajectories that are retained in a compact region of phase space over the time duration should correspond to a single cluster in  $\mathbb{R}^{d(T+1)}$ . Each of these trajectories should therefore have a low value of  $h_i$ . A spatial plot of the field  $h_i$  over the phase space  $\mathbb{R}^d$  is therefore useful for identifying the strength with which trajectories belong to clusters. Finer, cluster-by-cluster spatial information can be obtained by producing  $K$  spatial plots of the likelihoods  $u_{k,i}$  separately for each  $k = 1, \dots, K$ .

## B. False positives

Algorithms 1 and 2 will always produce centres and clusters, even if the system under consideration has no features that could be considered to be coherent. Thus, there is the possibility of Algorithms 1 and 2 reporting false positives. There are some easy ways to inspect the reported clusters and check for false positives. If the phase space is in one, two, or three dimensions, then one can visually inspect the clusters at each time instant to check if the clusters do indeed mostly remain in separate compact regions. This can be done by plotting  $u_{k,i}$  against  $x_{i,t}$  for  $k = 1, \dots, K$  and  $t = 0, 1, \dots, T$  (using e.g. the `scatter` command in MATLAB) to check the certainty of classification for individual clusters. If the phase space is not low-dimensional, one can plot  $u_{k,i}$  against  $i$  (or  $h_i$  against  $i$ ) and inspect how many trajectories have high confidence of classification. A low classification confidence is indicative of the cluster not corresponding to a coherent set.

## C. Choice of trajectory output times and choice of $m$

Clustering with respect to the Euclidean metric becomes less meaningful in high dimensions, with the distribution of interpoint distances becoming increasingly tight. This can be partly mitigated by using an  $\ell_p$  norm rather than the Euclidean  $\ell_2$  norm, but we have found the following rules of thumb very helpful, and have achieved good results with the standard Euclidean norm.

Firstly, one should choose the time between  $x_{i,t}$  and  $x_{i,t+1}$  to represent some nontrivial dynamics. If the increment  $t \rightarrow t+1$  is too short, the dynamics is close to the identity transformation, and one adds  $d$  dimensions to the clustering problem (making it more difficult) for no information gain. On the other hand, the increment from  $t \rightarrow t+1$  should not be so long that the underlying dynamics appears random over one time step; a group of nearby points at time  $t$  should remain in a “connected” region at time  $t+1$ , even though this region may be stretched and folded. Secondly, the total time duration  $T$  should not be so long that the entire phase space is

thoroughly mixed; for such  $T$  there is no chance of finding coherent sets. Once the step  $t \rightarrow t+1$  and total duration  $T$  have been selected as above, one should obtain reasonable results. Finally, to fine tune the value of  $m$  to ensure robust results, we suggest the following rule. Begin with  $m = 2$  and decrease  $m$ . For each value of  $m$ , record the locations of  $x_{i_k^*,0}$ , the maximum likelihood trajectories at time  $t = 0$  (the choice of  $t = 0$  is arbitrary). Find a range of  $m$  for which the locations of the maximum likelihood trajectories are stable (i.e. approximately fixed). Note that the centres  $c_{k,0}$  at time  $t = 0$  will tend to continue to vary with  $m$  so they are not good indicators of cluster stability with  $m$ .

## D. Centre collapse

If two or more of the reported cluster centres are all very close to one another in space, there are at least three possibilities. Firstly, it could be that there are no coherent structures in the trajectory data. Secondly, it could be that the choice of the step  $t \rightarrow t+1$  and/or  $T$  are unsuitable. Thirdly, even if the choice of the step  $t \rightarrow t+1$  and/or  $T$  are reasonable, it could be that the value of  $m$  is too high. In our experiments we have found that the larger  $d(T+1)$  is (the larger the total dimension), the smaller  $m$  needs to be to avoid centre collapse. This is not surprising because with higher dimension, the interpoint distances distribution is more tight, and a lower value of  $m$  emphasises differences in distance more. This is the reason behind our suggestion in the previous paragraph to start with  $m = 2$  and decrease  $m$  until the maximum likelihood trajectories are stable.

## E. Other inaccurate results

For systems that do contain finite-time coherent sets, there are some points to bear in mind to increase the accuracy of the reported clusters. If a finite-time coherent set is small relative to the domain size and few clusters are sought, because Algorithms 1 and 2 favour clusters containing approximately the same number of trajectories, the clusters may be much larger than the true coherent region. In such a situation, an inspection of the likelihood functions may reveal the small coherent regions as “high likelihood”. On the other hand, if there are few, large coherent sets, but one chooses a large value of  $K$ , then the coherent regions will likely be subdivided into several clusters.

These effects can be studied by varying the number of clusters  $K$  (which is cheap to experiment with). For each  $K$  one can visually inspect the clustering confidence according to  $u_{k,i}$  and  $h_i$ , as discussed in Section IV B. If a regime of cluster stability can be found for a number of consecutive  $K$ , this gives some confidence to the results. Finally, if sufficient data is available, the results

can be checked against the classical finite-time coherent set identification methods<sup>12,17,18</sup>.

## V. NUMERICAL EXPERIMENTS

### A. One-dimensional examples

We start with three one-dimensional maps on  $S^1$ , which we think of as the unit interval  $[0, 1]$  with the end-points identified. Because we are on  $S^1$  and not  $[0, 1]$ , the distance computation and the center updating are modified in the obvious way. The first example is given by

$$S(x) = \begin{cases} 3x \pmod{\frac{1}{3}} + 1/3, & x < \frac{1}{3}, \\ 3x - \frac{1}{3} \pmod{\frac{1}{3}} + 2/3, & \frac{1}{3} \leq x < \frac{2}{3}, \\ 3x - \frac{2}{3} \pmod{\frac{1}{3}}, & x \geq \frac{2}{3}. \end{cases} \quad (15)$$

The map  $S$  cyclically permutes the three intervals  $[0, \frac{1}{3})$ ,  $[\frac{1}{3}, \frac{2}{3})$  and  $[\frac{2}{3}, 1)$  and mixes each interval internally. Thus the graph of  $S$  features three equally sized blocks that are cyclically permuted, see Figure 2.

To test Algorithm 1 we select 1000 random initial conditions from  $[0, 1]$  and iterate them nine times by the mapping  $S$ . We want to find clusters in 1000 data points in  $\mathbb{R}^{10}$ . We choose  $K = 3$  and a very small fuzziness parameter of  $m = 1.1$ . The membership functions of the three clusters are shown in Figure 3 (a). As expected, the three coherent sets obtained are comprised of the three intervals; the evolution of these intervals is visualized in Figure 4. The cluster centers are the centers of the intervals and the  $u_{k,i}$  describe a very sharp trajectory-cluster membership. Increasing the fuzziness to  $m = 2$  gives a fuzzier result, but still has clear clusters; see Figure 3 (b).

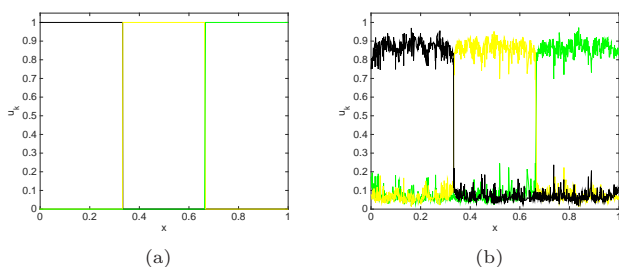


FIG. 3: Membership functions for the three clusters of the one-dimensional map (15) plotted against the initial conditions of the trajectories under consideration.

(a)  $m = 1.1$ , (b)  $m = 2$ .

If we reduce the number of desired clusters to  $K = 2$  the algorithm will either merge the first two or the second two clusters, depending on how the initial conditions are distributed. Trying to approximate  $K = 4$  coherent sets, one of the three clusters is divided into two clusters. Their centers are almost coinciding, an indication of false

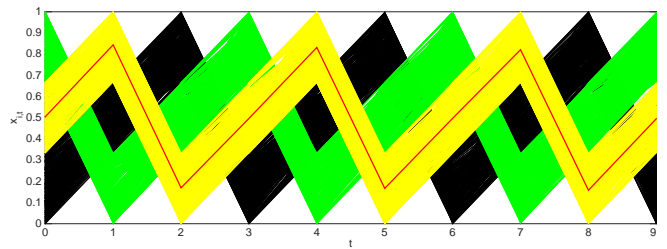


FIG. 4: Clustering of the 1d map (15): time evolution of the three clusters, one of the cluster centers as computed by Algorithm 1 in red.

positives, and the membership functions on this interval are very much fluctuating, see Figure 5.

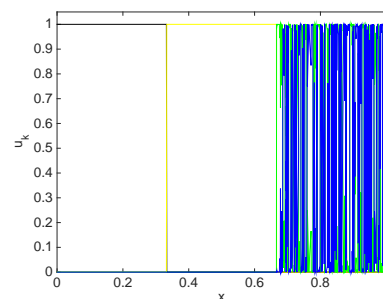


FIG. 5: Seeking four clusters of trajectories from (15) results in the partition of one of the clusters in two clusters with centers almost coinciding and highly fluctuating membership functions ( $m = 1.1$ ).

The map  $S$  in (15) has perfectly coherent sets: there is no transport between the three coherent sets. We now briefly consider two further one-dimensional systems on  $S^1$  to demonstrate the more common setting of leaking coherent sets. The first system, which will be referred to as (FLQ10), is a repeated cycle of three maps  $T_1, T_2, T_3$ , introduced in Ref. 13. It was shown in Ref. 13 that there are two coherent sets of different sizes that are cyclically permuted. Details of the model can be found in Ref. 13 (proof of Thm. 5.1 and Figure 1). Choosing again 1000 random initial conditions from  $[0, 1]$  and nine iterates of the maps (three cycles of  $T_3 \circ T_2 \circ T_1$ ), we seek to find two clusters in the ten-dimensional data. For this we choose a fuzziness constant of  $m = 1.5$ . In Figure 6 we show the two clusters in space-time, plotting only those points with a membership value of at least 95% (according to the  $u_{k,i}$ ) of belonging to one of the clusters. As expected, the cluster centers approximately cycle with period 3. We note that the two clusters at  $t = 0$  (and thus at  $t = 3, 6, 9$ ) are consistent with the coherent sets obtained in Ref. 13 (see in particular Figure 2 in Ref. 13, where the supports of the positive/negative parts of the eigenvector shown there are in good agreement with the two clusters at  $t = 0$ ).

A more general situation has been discussed in Ref. 14.



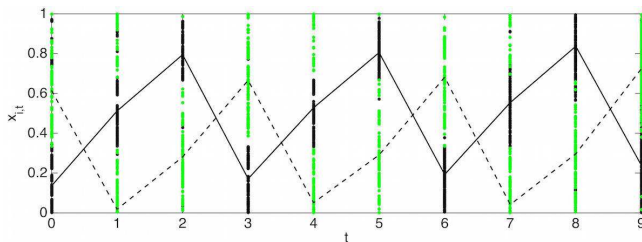


FIG. 6: Clustering of 1000 trajectories of length 10 of the system (FLQ10) in space-time ( $m = 1.5$ ). Plotted are only those points with  $u_{k,i} > 0.95$ . The solid and dashed lines indicate the centers of the two clusters as computed by Algorithm 1.

Two coherent sets were extracted that move in an aperiodic manner; see Example 1 in Ref. 14 for more details of the underlying model, which we will refer to as (FLS10) in the following. We can reproduce the coherent behaviour of (FLS10) using the same setting as described above. Figure 7 shows the two clusters in space-time, again plotting only those points with a membership value of at least 95% for one of the two clusters. The results are consistent to those in Ref. 14 (see in particular Figure 8 in Ref. 14, where the supports of the positive/negative parts of the Oseledets functions shown for iterates  $k = 0, \dots, 5$  are in good agreement with the two clusters at times  $t = 0, \dots, 5$ ). The membership functions of the two clusters (plotted for time  $t = 2$ ) for the choice  $m = 1.5$  and  $m = 2$  are shown in Figure 8 (one can also compare the form of the black membership function with the Oseledets function in Figure 8, Ref. 14 for  $k=2$ ). As anticipated, the clusters are not as clear-cut as in Figure 3. Eventually all trajectories will spread out over  $[0, 1]$ , so that spherical compact structures as detected by our approach cease to exist. We remark that in each of the one-dimensional examples, the maps have a uniform slope of 3, so that after the ninth iterate, nearby initial points have been separated by a factor of  $3^9 = 19683$ .

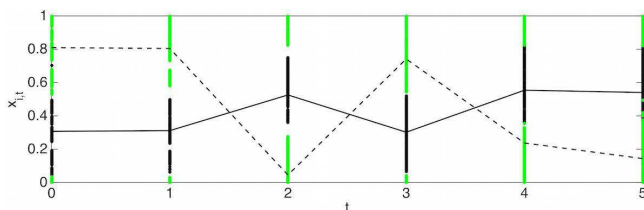


FIG. 7: Clustering of 1000 trajectories of the system (FLS10) in extended space ( $m = 1.5$ ). Plotted are only those points with  $u_{k,i} > 0.95$ . The solid and dashed lines indicate the probabilistic centers of the two clusters.

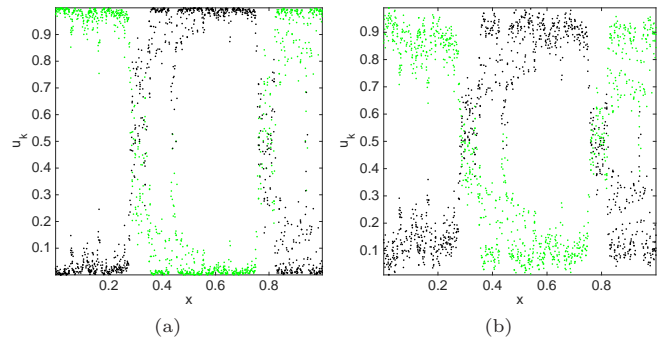


FIG. 8: Membership functions for the two clusters of the 1d map (FLS10) at time  $t = 2$ . (a)  $m = 1.5$ , (b)  $m = 2$ .

## B. Double gyre flow

We consider the time-dependent system of differential equations<sup>36</sup>

$$\begin{aligned} \dot{x} &= -\pi A \sin(\pi f(x, t)) \cos(\pi y) \\ \dot{y} &= \pi A \cos(\pi f(x, t)) \sin(\pi y) \frac{df}{dx}(x, t), \end{aligned} \quad (16)$$

where  $f(x, t) = \delta \sin(\omega t)x^2 + (1 - 2\delta \sin(\omega t))x$ .

For detailed discussions of the system we refer to Refs. 15, 17, and 36. As in Refs. 15 and 17 we fix parameter values  $A = 0.25$ ,  $\delta = 0.25$  and  $\omega = 2\pi$  and obtain a 1-periodic flow. In order to be able to compare our results with those in Ref. 17, where we have extracted two optimally coherent sets via transfer operator-based methods, we choose  $2^{15}$  initial points on a uniform grid on the invariant set  $[0, 2] \times [0, 1]$ . For each of these initial conditions we compute a trajectory on  $[0, \tau]$ , where  $\tau = 1, 5, 10$ . We output the trajectory data in increments of 0.1 time steps. Thus, for  $\tau = 1$  each trajectory is represented by a 22-dimensional vector ( $= (10 + 1) \times 2$ ), for  $\tau = 5$  and  $\tau = 10$  the corresponding vectors have length 102 and 202, respectively.

We start by extracting two clusters from the short trajectories ( $\tau = 1$ ). The upper panel of Figure 9 displays the membership values  $u_{1,i}$  (note that  $u_{2,i} = 1 - u_{1,i}$ ) with respect to the initial conditions in the two dimensional phase space. To study the influence of the fuzziness exponents on the results we choose  $m = 1.5$  (Figure 9(a)) and  $m = 2$  (Figure 9(b)). Both plots give a clear indication of the two coherent sets. To get a more detailed picture about the certainty of cluster membership we compute the entropy  $h$  from (14). The respective results are shown in the lower panel of Figure 9. For the smaller fuzziness exponent  $m = 1.5$  (Figure 9(c)) there are large regions of high certainty to belong to one of the two clusters, with some high uncertainty in the vicinity of the stable manifold of the hyperbolic periodic orbit on the  $x$ -axis. This uncertainty region increases significantly, when  $m = 2$  (Figure 9(d)) is used. Here only the two regular regions (corresponding to invariant tori in

the time-1 flow map) are highlighted as the most certain regions.

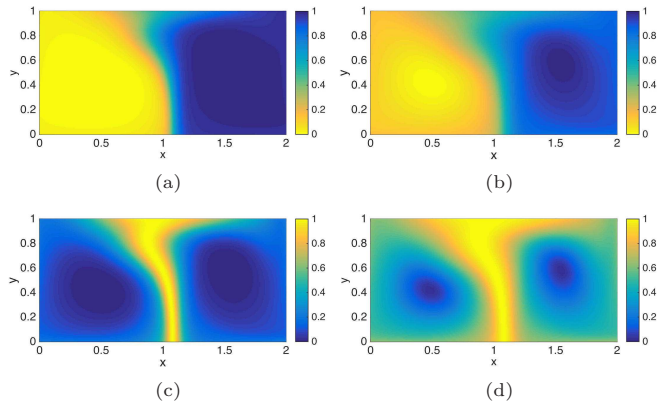


FIG. 9: 2-clustering of  $2^{15}$  trajectories in the double gyre flow (16) with flow time  $\tau = 1$  for  $m = 1.5$  (left) and  $m = 2$  (right). (a),(b) membership values  $u_{1,i}$  for  $m = 1.5$  and  $m = 2$ , respectively. (c),(d) corresponding entropy plots using (14).

Note that a maximum likelihood hard partition into two sets gives the same result for both  $m = 1.5$  and  $m = 2$ . The result is shown in Figure 11 (a), with parts of the stable manifold of the hyperbolic periodic orbit on the  $x$ -axis superimposed. This known dominant (infinite-time) transport barrier determines a large part of the boundary between the two extracted coherent sets. This compares very well to the observations made in Ref. 17 (see e.g. Figure 9.3 therein).

We now consider longer trajectories with flow times  $\tau = 5$  and  $\tau = 10$ . The respective results for  $m = 2$  are shown in Figure 10 (a) and (b). As expected from what we have seen in Ref. 17 the clustering of the initial conditions is again very much influenced by the stable manifold, see also Figure 11 (b) and (c) for the respective maximum likelihood partitions into two sets. This transport barrier also determines the regions of highest membership uncertainties, which is clearly visible in Figure 10 (c) and (d). We note that these entropy plots have striking similarity to the finite-time entropy fields obtained directly from the transfer operator<sup>16</sup>.

A visualization of the clusters in space-time for flow time  $\tau = 5$  is presented in Figure 1, where from 512 initial conditions we have plotted those trajectories for which the membership values  $u_{k,i} > 0.9$  ( $m = 2$ ).

So far we have used high-resolution and complete trajectory data. We now test our approach in the situation where the available information is poor. We use 512 initial conditions on a regular grid on  $[0, 2] \times [0, 1]$  and compute trajectories for flow time  $\tau = 5$ . We then destroy about 80% of the trajectory information by randomly setting the particle positions to NaN. This mimicks the situation that trajectories may not exist for the whole time span under consideration and additionally may have gaps in observation. Algorithm 2 produces two clusters

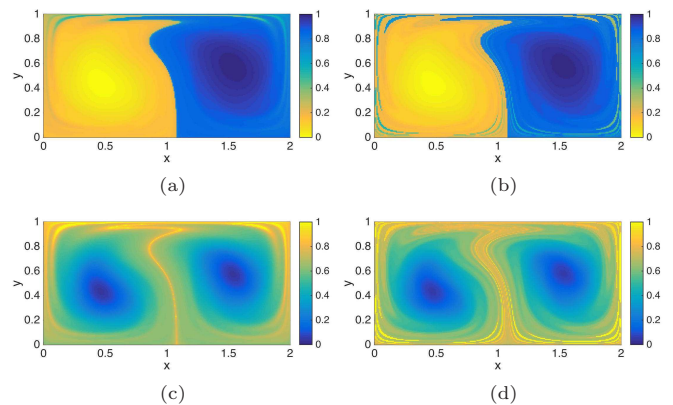


FIG. 10: 2-clustering of  $2^{15}$  trajectories in the double gyre flow (16) for  $m = 2$  and flow times  $\tau = 5$  (left) and  $\tau = 10$  (right). (a),(b) membership values  $u_{1,i}$  for  $\tau = 5$  and  $\tau = 10$ , respectively. (c),(d) corresponding entropy plots using (14).

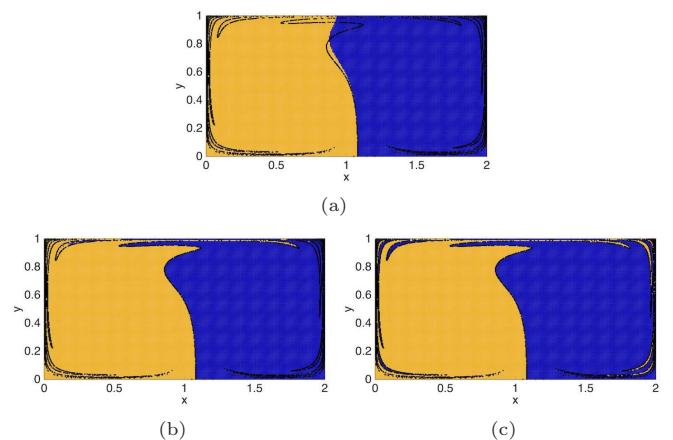


FIG. 11: Extraction of two clusters from  $2^{15}$  trajectories ( $m = 2$ ) for different flow times in (16) based on maximum likelihood of the membership values as in Figures 9 and 10. The dominant transport barrier is superimposed and bounds the two sets increasingly closely as  $\tau$  increases, as demonstrated in Ref. 17. (a) flow time  $\tau = 1$ ; (b)  $\tau = 5$ ; (c)  $\tau = 10$ .

from this highly incomplete trajectory data, as shown in Figure 12. Note that even with this severe data thinning, Algorithm 2 still classifies the remaining data points to the correct sides of the transport barriers.

Finally, we test what happens if we set  $K > 2$ . We restrict again to flow time  $\tau = 5$  and  $2^{15}$  trajectories and choose  $m = 2$ . If  $K = 3$  then compared to  $K = 2$  either the left or right cluster is subdivided as seen in the membership values in Figure 13 (a-c). The shapes of the resulting clusters in Figure 13 (a,b) do not have any similarity with known coherent structures for this system, but apparently the respective trajectory bundles stay coherent in our sense - with the cluster centers well separated.

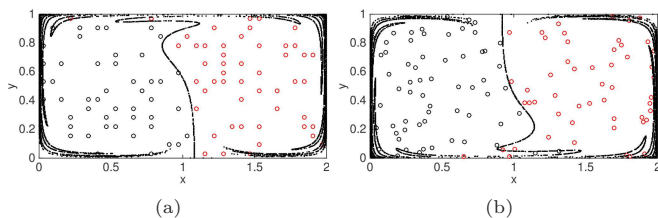


FIG. 12: 2-clustering of double gyre flow (16) with 80% of the trajectory data missing, based on  $2^9$  trajectories with flow time  $\tau = 5$  ( $m = 2$ ). (a) clusters at initial time, (b) clusters at final time. Corresponding transport barriers are superimposed.

However, Figure 13 (c) reveals that the left cluster, which is also present in the 2-clustering considered in Figure 10, is characterized by much higher membership function values compared to the other two clusters. We note that for  $K = 4$  we get a similar picture with the former two clusters both divided into two parts, and for  $K = 5$  one of the former two clusters is divided into two and the other into three parts.

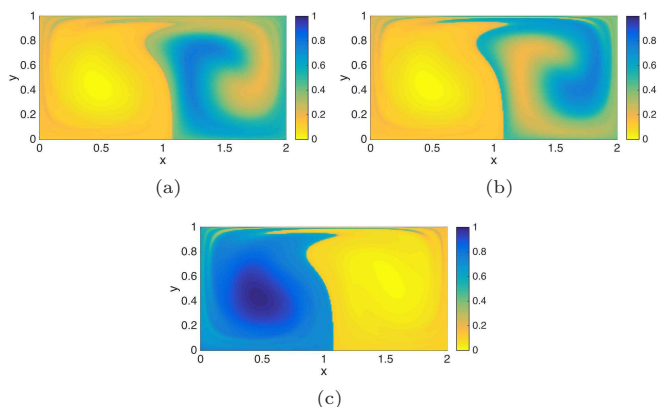


FIG. 13: 3-clustering of  $2^{15}$  trajectories in the double gyre flow for flow time  $\tau = 5$  and  $m = 2$ . (a-c) membership functions for the three clusters.

### C. Transitory double gyre flow

We consider the transitory dynamical system<sup>27</sup>

$$\dot{x} = -\frac{\partial}{\partial y}\Psi, \quad \dot{y} = \frac{\partial}{\partial x}\Psi \quad (17)$$

with stream function

$$\begin{aligned} \Psi(x, y, t) &= (1 - s(t))\Psi_P + s(t)\Psi_F \\ \Psi_P(x, y) &= \sin(2\pi x) \sin(\pi y) \\ \Psi_F(x, y) &= \sin(\pi x) \sin(2\pi y) \end{aligned}$$

and transition function

$$s(t) = \begin{cases} 0, & t < 0, \\ t^2(3 - 2t), & 0 \leq t \leq 1, \\ 1, & t > 1. \end{cases}$$

The nonautonomous dynamics, which rotates a double gyre pattern counter-clockwise by 90 degrees, is restricted to the time-interval  $[0, 1]$ . In Ref. 17 we have analysed this system using the transfer operator based coherent set framework. The unit square  $[0, 1]^2$  is invariant under the flow and we choose  $2^{14}$  initial conditions on a regular grid. We consider the flow on the transition interval  $[0, 1]$  and output the trajectory data in increments of 0.1 time steps. So  $T = 10$  and thus we represent every trajectory as a 22-dimensional vector.

Algorithm 1 with  $K = 2$  and  $m = 1.5$  returns  $u_{k,i}$  that take high values on the coherent sets observed in Ref. 17, as shown in Figure 14.

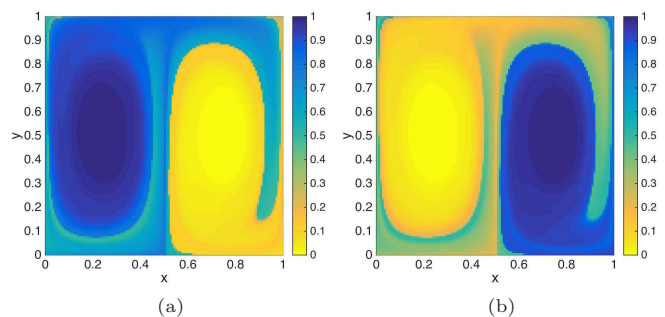


FIG. 14: 2-clustering of  $2^{14}$  trajectories in the transitory double gyre flow (17) for the time interval  $[0, 1]$ . (a),(b) membership values  $u_{k,i}$ ,  $k = 1, 2$  for  $m = 1.5$ .

A visualization of the two clusters in space-time is presented in Figure 15, where from 1024 initial conditions we have plotted those trajectories for which  $u_{k,i} > 0.95$  ( $m = 1.5$ ) together with the probabilistic cluster centers.

We study the influence of using information along a trajectory instead of only considering the initial and final points of a trajectory as many other identification algorithms do (i.e. taking 11 vs 2 time instances on  $[0, 1]$ ). The results of clustering  $2^{14}$  trajectories based only on the initial and end points of the trajectories are shown in Figure 16. The clusters are less smooth; an intuitive explanation for this is that Algorithm 1 only uses point information, as opposed to probability flow information as in Refs. 12, 17, and 18. Algorithm 1 needs to compensate for this by augmenting the point information with additional points over time.

### D. Drifter data

We demonstrate the efficacy of our approach on real-world data, namely drifter data from the Global Ocean Drifter Program available at

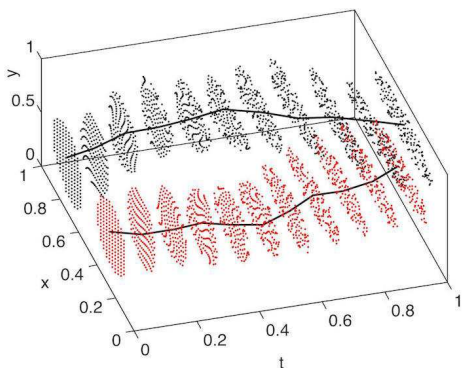


FIG. 15: Space-time plot of the two clusters obtained from  $2^{10}$  trajectories in the transitory double gyre flow.

Only trajectories with  $u_{k,i} > 0.95$  are shown, where  $m = 1.5$ . The solid curves indicate the position of the probabilistic cluster centers.

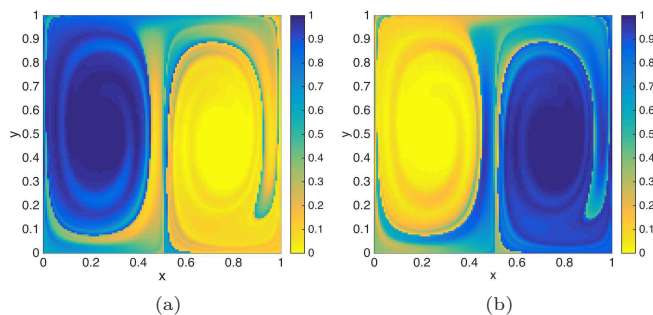


FIG. 16: Membership functions for the 2-clustering of transitory double gyre flow – using only initial and end points of the trajectories ( $m = 1.5$ ).

AOML/NOAA Drifter Data Assembly Center (<http://www.aoml.noaa.gov/envids/gld/>). The entire dataset spans the years 1979–2014, with drifter positions given every six hours. The area of observation is the global ocean (latitude  $[90, -78]$  and longitude  $[-180, 180]$ ). We focus on the years 2005–2009 and restrict to those drifters that have a minimum lifetime of one year within this five-year time span. We output the position of these 2267 trajectories (in longitude, latitude coordinates) every month, i.e. the length of our trajectories is 60 months.

We note that a typical drifter does not operate over the whole five years; that is, many terminate prior to December 2009 and many begin later than January 2005. There are also gaps in observations when there is a failure in recording the drifter location, so the data is highly incomplete. Figure 17 summarises two statistics: the distribution of drifter lifetimes and the number of drifters actively recording each month. The average lifetime of a drifter in this data set is about 23 months, with many drifters operating only for a year and only very few drifters for 4–5 years, see Figure 17 (a). On average, 869 trajectories (or 38% of all drifters in the period 2005–2009) are

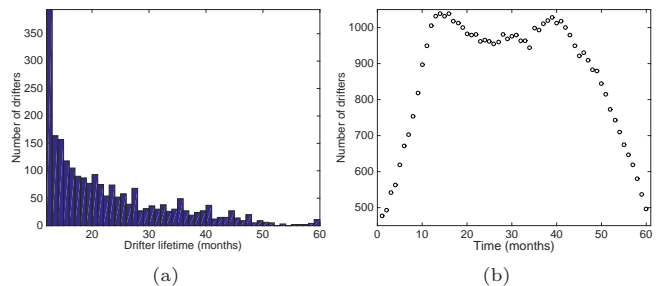


FIG. 17: Drifter statistics. (a) histogram of the drifter lifetimes, (b) number of drifters available at a certain time instance.

available at a given time instant, with less data at the beginning and the end of the considered five year time span; see Figure 17 (b).

As we consider the global ocean we have to respect distances on a sphere (we assume the surface of the ocean to be approximately spherical). We also have to ensure that we restrict cluster centers to the surface of this sphere. To achieve both of these requirements we use a cosine distance function, and update centers only on the sphere<sup>7</sup>. Every drifter trajectory is represented as a vector in  $3 \times 60 = 180$ -dimensional space. In contrast to our calculations in Algorithm 2, we simply display our results in cartesian longitude-latitude coordinates.

We first look for two clusters; Figure 18 shows results of the clustering algorithm for  $K = 2$ . Figure 18 (a) shows all drifter positions available on January 2005, coloured according to their maximum likelihood membership in one of the two clusters. Figure 18 (b) shows all drifter positions available on December 2009, again coloured according to their most likely cluster membership. Thus, we expect the red (resp. green) cluster in Figure 18 (a) to evolve coherently to the red cluster in Figure 18 (b). Of course, many of the drifters in Figure 18 (a) do not correspond to the same physical drifter in Figure 18 (b) because the lifetimes of many drifters are shorter than five years. Nevertheless, as physical drifters enter and leave the dataset over the five-year duration, the drifters tagged red (resp. green) move as a coherent cloud. This is illustrated in a video attached to our electronic submission, see Figure 18 (Multimedia view).

In Figure 18, one sees a separation of the Pacific Ocean (red) from the Atlantic and Indian Oceans (green), which are grouped together. Here, continental obstructions play an obvious role in the dynamical separation of the ocean surface flow. Figure 18 (a) ascribes the southern part of the Indian Ocean to the Pacific Ocean. This is in line with recent research<sup>19</sup> (see Figure 6 in Ref. 19) based on transfer operator analysis of the Ocean General Circulation Model for the Earth Simulator (OFES model)<sup>26,35</sup>, and consistent with a general eastward flow of water in high southern latitudes. One observes that the red drifters in Figure 18 (a) have flowed eastwards to rejoin the Pacific in Figure 18 (b).

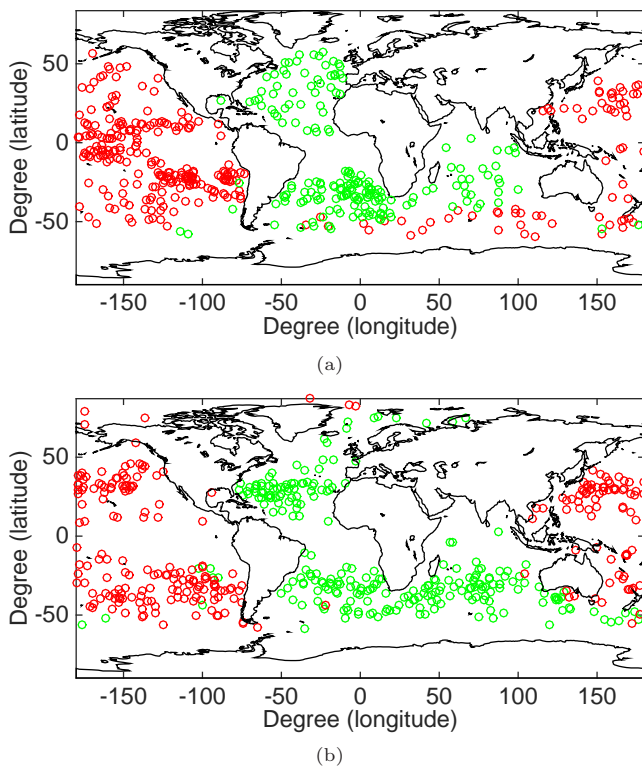


FIG. 18: 2-clustering of drifter data ( $m = 1.5$ ). (a) drifter positions January 2005, (b) drifter positions December 2009. Animation of the cluster motion 2005–2009 in online version (Multimedia view).

Figure 19 shows the results of Algorithm 2 with  $K = 5$  at January 2005 (a), July 2007 (b), and December 2009 (c). An animation is available online (Figure 19 (Multimedia view)). We choose  $K = 5$  in order to attempt to delineate the five major oceans: the North and South Atlantic Oceans, the North and South Pacific Oceans, and the Indian Ocean. Broadly, we see that the clustering does find the appropriate equatorial separations of the Atlantic and Pacific Oceans, and also separates the Indian Ocean.

Some of these separations are highlighted by investigating the certainty of membership of the individual drifters based on an entropy calculation (14). In Figure 20 those drifters (positions as of July 2007) are marked black when their relative entropy is  $> 0.1$ , corresponding to a maximum membership value of less than  $\approx 0.96$ . Figure 20 and in particular the time evolution of the drifters (Figure 20 (Multimedia view)) shows that the uncertain regions correspond to the major ocean barriers in the Atlantic and Pacific, and the Southern Ocean.

Our results in Figure 19 are strikingly similar to those shown in Figure 6 in Ref. 19, which have been derived using transfer operator methods and the (wind-forced) OFES model. For example in Figure 19 (a), when comparing with Figure 6 in Ref. 19 we see: the separation of the Pacific Ocean becoming more southerly as one

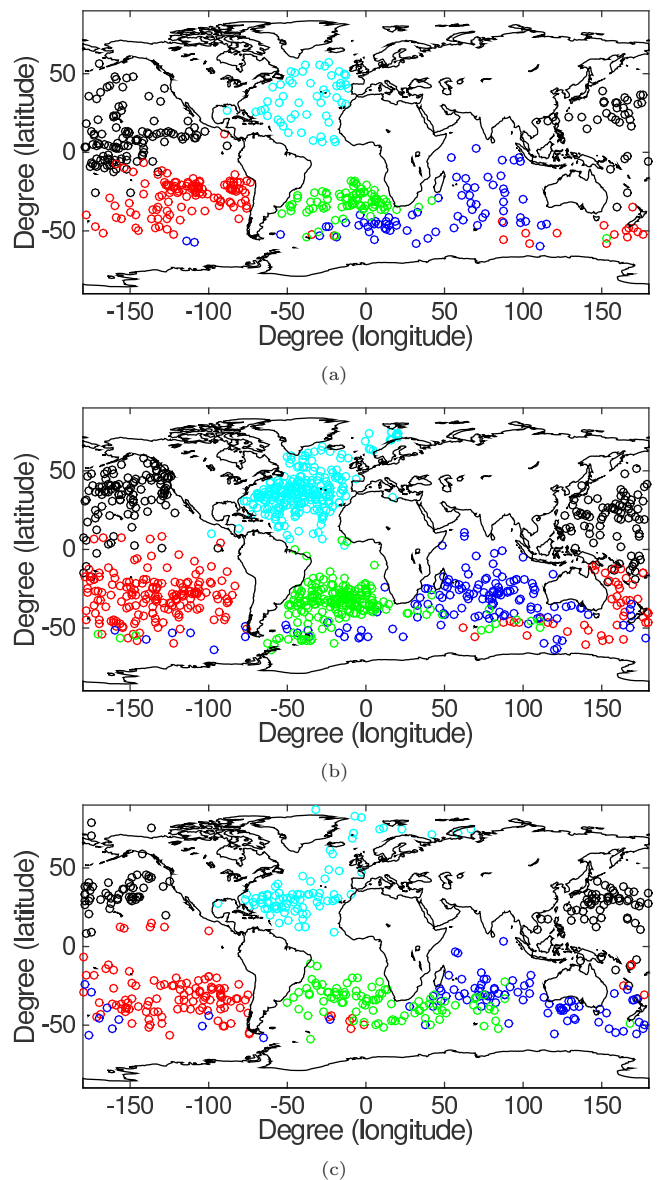


FIG. 19: 5-clustering of drifter data ( $m = 1.5$ ). (a) drifter positions January 2005, (b) drifter positions July 2007, (c) drifter positions December 2009. Animation of the cluster motion 2005–2009 in online version (Multimedia view).

proceeds westwards toward Australia; the Indian Ocean spilling westwards at its southerly boundary; and the South Atlantic forcing its way around the east coast of southern Africa. As described in Ref. 19, the Ekman dynamics of the ocean surface circulation guarantees that each of the five major oceans contains an attracting region associated to the great oceanic gyres and their corresponding garbage patches. The separations seen in Figure 19 (a) and Figure 6 in Ref. 19 and the features described above are associated with the basins of attraction of these five attracting regions.

We remark that while the results in Ref. 19 are of a

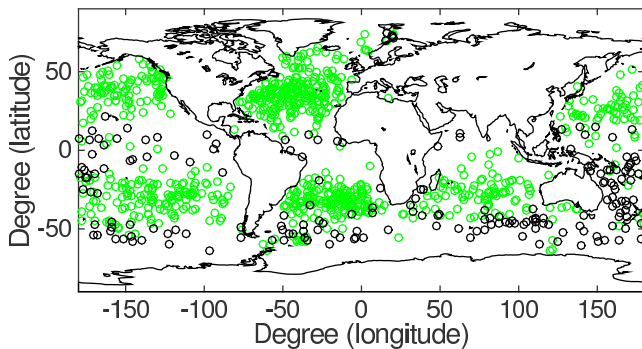


FIG. 20: Difficult to classify drifters (positions as of July 2007) are marked black when their relative entropy is  $h > 0.1$ . Animation of the cluster motion 2005–2009 in online version (Multimedia view).

higher spatial resolution than those obtained here, the experiments in Ref. 19 used just over  $10^6$  trajectories, recorded every eight weeks for a period of 48 weeks (a total of  $6.14 \times 10^6$  data points), while here we have  $869 \times 60 = 5.2 \times 10^4$  data points, comprised of 2267 incomplete trajectories. We also remark that while we have drawn comparisons between Ref. 19 and the present study, the former computed ocean boundaries as basins of attraction based on a repeating 48-week ocean circulation, while our present study seeks to compute estimates of coherent sets based on five years of non-repeating drifter data.

## VI. DISCUSSION

We have introduced a “rough-and-ready” general cluster-based approach for analysing coherent structures in time-dependent dynamical systems. Our method assigns individual trajectories membership in regions that retain a compact extent over a specified finite time duration. Our approach has several advantages.

First, the ability to work directly with a small number of trajectories, including the situations where the trajectories do not span the entire time duration of interest and where observations may be missing from within trajectories. Second, initial implementation is rapid (using e.g. the built-in MATLAB function `fcm` to perform the fuzzy clustering for the case of complete data), and the runtimes are fast (on the order of fractions of seconds for the one-dimensional maps in Section V A to less than 10 seconds to cluster a dataset of 32768 trajectories in 202 dimensions, as in the case of the double gyre flow with flow duration  $\tau = 10$  in Section V B). Third, our method considers entire trajectories (not just the endpoints) and automatically outputs clusters at every time instant in the trajectory data; thus a frame-by-frame description of the temporal evolution of the clusters is immediately obtained. Fourth, the use of fuzzy clustering provides feedback in the form of membership likelihoods and entropy,

which provide the user with an estimate of confidence with which a trajectory has been assigned to a particular compact region. Finally, the soft clustering approach is relatively insensitive to noise in the data. We note that the same methodology can be used to estimate coherent regions for SDEs, by simply generating stochastic trajectories and applying Algorithm 1.

## ACKNOWLEDGMENTS

The research of GF is supported by an ARC Future Fellowship (FT120100025). GF also thanks the University of Canterbury’s Erskine Fellowship scheme for partial financial support and the Department of Mathematics and Statistics at the University of Canterbury for providing excellent working conditions during part of the time this research was undertaken. KPG acknowledges support from an ARC Discovery Project (DP110100068). She also thanks the School of Mathematics and Statistics at the University of New South Wales for hospitality as well as the University of Canterbury for hospitality and financial support.

## REFERENCES

- <sup>1</sup>M. R. Allshouse and J.-L. Thiffeault. Detecting coherent structures using braids. *Physica D*, 241(2):95–105, 2012.
- <sup>2</sup>M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: ordering points to identify the clustering structure. In *ACM SIGMOD international conference on Management of data*, pages 49–60, 1999.
- <sup>3</sup>D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.
- <sup>4</sup>J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum, New York, 1981.
- <sup>5</sup>J. C. Bezdek, R. J. Hathaway, M. J. Sabin, and W. T. Tucker. Convergence theory for fuzzy c-means: counterexamples and repairs. *IEEE Transactions on Systems, Man and Cybernetics*, 17(5):873–877, 1987.
- <sup>6</sup>M. Budišić and I. Mezić. Geometry of the ergodic quotient reveals coherent structures in flows. *Physica D: Nonlinear Phenomena*, 241(15):1255 – 1269, 2012.
- <sup>7</sup>I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, Jan 2001.
- <sup>8</sup>M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.
- <sup>9</sup>M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematics*, 23(2):298–305, 1973.
- <sup>10</sup>S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
- <sup>11</sup>G. Froyland. Dynamical isoperimetry and the geometry of lagrangian coherent structures. Submitted, November 2014. Available at <http://arxiv.org/abs/1411.7186>.
- <sup>12</sup>G. Froyland. An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. *Physica D*, 250:1–19, 2013.
- <sup>13</sup>G. Froyland, S. Lloyd, and A. Quas. Coherent structures and

- isolated spectrum for Perron-Frobenius cocycles. *Ergodic Theory and Dynamical Systems*, 30:729–756, 2010.
- <sup>14</sup>G. Froyland, S. Lloyd, and N. Santitissadeekorn. Coherent sets for nonautonomous dynamical systems. *Physica D*, 239:1527–1541, 2010.
- <sup>15</sup>G. Froyland and K. Padberg. Almost-invariant sets and invariant manifolds – connecting probabilistic and geometric descriptions of coherent structures in flows. *Physica D*, 238:1507–1523, 2009.
- <sup>16</sup>G. Froyland and K. Padberg-Gehle. Finite-time entropy: A probabilistic approach for measuring nonlinear stretching. *Physica D: Nonlinear Phenomena*, 241:1612 – 1628, 2012.
- <sup>17</sup>G. Froyland and K. Padberg-Gehle. Almost-invariant and finite-time coherent sets: directionality, duration, and diffusion. In W. Bahsoun, C. Bose, and G. Froyland, editors, *Ergodic Theory, Open Dynamics, and Coherent Structures*, volume 70 of *Proceedings in Mathematics and Statistics*, chapter 9, pages 171–216. Springer, 2014.
- <sup>18</sup>G. Froyland, N. Santitissadeekorn, and A. Monahan. Transport in time-dependent dynamical systems: Finite-time coherent sets. *Chaos*, 20:043116, 2010.
- <sup>19</sup>G. Froyland, R. Stuart, and E. van Sebille. How well-connected is the surface of the global ocean? *Chaos*, 24:033126, 2014.
- <sup>20</sup>G. Haller. A variational theory of hyperbolic Lagrangian Coherent Structures. *Physica D*, 240:574–598, 2011.
- <sup>21</sup>G. Haller and F. Beron-Vera. Geodesic theory of transport barriers in two-dimensional flows. *Physica D*, 241(20):1680–1702, 2012.
- <sup>22</sup>H. Izakian, W. Pedrycz, and I. Jamal. Clustering spatio-temporal data: An augmented fuzzy C-Means. *IEEE Transactions on Fuzzy Systems*, 21(5):855–868, 2013.
- <sup>23</sup>S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo. Spatio-temporal clustering. In L. Rokach and O. Maimon, editors, *The Data Mining and Knowledge Discovery Handbook*, chapter 6, pages 855–874. Springer, 2 edition, 2010.
- <sup>24</sup>S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.
- <sup>25</sup>T. Ma and E. M. Bollt. Differential geometry perspective of shape coherence and curvature evolution by finite-time nonhyperbolic splitting. *SIAM Journal on Applied Dynamical Systems*, 13(3):1106–1136, 2014.
- <sup>26</sup>Y. Masumoto, H. Sasaki, T. Kagimoto, N. Komori, A. Ishida, Y. Sasai, T. Miyama, T. Motoi, H. Mitsudera, K. Takahashi, et al. A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of OFES (OGCM for the Earth Simulator). *J. Earth Simulator*, 1:35–56, 2004.
- <sup>27</sup>B. Mosovsky and J. Meiss. Transport in transitory dynamical systems. *SIAM J. Dyn. Syst.*, 10:35–65, 2011.
- <sup>28</sup>M. Nanni. *Clustering methods for spatio-temporal data*. Phd thesis, Dipartimento di Informatica, Università di Pisa, 2002.
- <sup>29</sup>M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3):267–289, 2006.
- <sup>30</sup>M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- <sup>31</sup>If  $K$  is not known *a priori*, to determine  $K$  one can employ any of several existing methods which automatically produce an optimal number of clusters from datasets.
- <sup>32</sup>N. Pelekis, I. Kopanakis, I. Ntoutsis, G. Marketos, G. Andrienko, and Y. Theodoridis. Similarity search in trajectory databases. In *Proceedings of the 14th IEEE International Symposium on Temporal Representation and Reasoning (TIME 2007)*, pages 129–140. IEEE CS Press, 2007.
- <sup>33</sup>S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3-4):225–239, 2008.
- <sup>34</sup>L. Rokach. A survey of clustering algorithms. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer US, 2010.
- <sup>35</sup>H. Sasaki, M. Nonaka, Y. Masumoto, Y. Sasai, H. Uehara, and H. Sakuma. An eddy-resolving hindcast simulation of the quasiglobal ocean from 1950 to 2003 on the earth simulator. In *High resolution numerical modelling of the atmosphere and ocean*, pages 157–185. Springer, 2008.
- <sup>36</sup>S. C. Shadden, F. Lekien, and J. E. Marsden. Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. *Physica D*, 212:271–304, 2005.
- <sup>37</sup>J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- <sup>38</sup>C. Truesdell and W. Noll. *The non-linear field theories of mechanics*. Springer, 3rd edition, 2004.