

Banner appropriate to article type will appear here in typeset article

Dynamics-augmented cluster-based network model

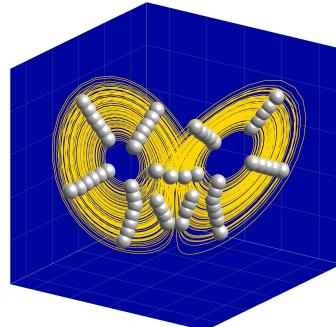
Chang Hou¹, Nan Deng^{1†} and Bernd R. Noack^{1,2‡}

¹School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen 518055, People's Republic of China

²Guangdong Provincial Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics, Harbin Institute of Technology, Shenzhen 518055, People's Republic of China

(Received xx; revised xx; accepted xx)

In this study, we propose a novel data-driven reduced-order model for complex dynamics, including nonlinear, multi-attractor, multi-frequency, and multiscale behaviours. The starting point is a fully automatable cluster-based network model (CNM) (Li *et al.* *J. Fluid Mech.* vol. 906, 2021, A21) which kinematically coarse-grains the state with clusters and dynamically predicts the transitions in a network model. In the proposed dynamics-augmented CNM (dCNM), the prediction error is reduced with trajectory-based clustering using the same number of centroids. The dCNM is first exemplified for the Lorenz system and then demonstrated for the three-dimensional sphere wake featuring periodic, quasi-periodic and chaotic flow regimes. For both plants, the dCNM significantly outperforms the CNM in resolving the multi-frequency and multiscale dynamics. This increased prediction accuracy is obtained by stratification of the state space aligned with the direction of the trajectories. Thus, the dCNM has numerous potential applications to a large spectrum of shear flows, even for complex dynamics.



Key words: Wakes/Jets: Wakes, Nonlinear dynamic systems: Low-Dimensional Models

1. Introduction

Advancements in computational capabilities and flow measurement technologies are producing a rapidly increasing amount of high-fidelity flow data. The coherent spatio-temporal structures of the flow data enable data-driven reduced-order models (ROMs). In terms of kinematics, ROMs furnish simplified descriptions that enrich our understanding of fundamental flow processes (Holmes *et al.* 1996), facilitated by increasingly powerful

† Email address for correspondence: dengnan@hit.edu.cn

‡ Email address for correspondence: bernd.noack@hit.edu.cn

machine learning methods (Brunton *et al.* 2020). ROMs may also allow the prediction of future states with acceptable accuracy. In the context of flow control, ROMs are serving as efficient tools for designing and testing control strategies, replacing costly high-fidelity simulations with an acceptable trade-off in accuracy (Bergmann & Cordier 2008).

First-principle-based ROMs have historically been the foundation of the ROM community, as only a limited number of large data sets were available at that time. The Galerkin framework is one of the most classic methods in this category. By projecting the Navier-Stokes equations onto a low-dimensional subspace, the Galerkin model elegantly describes the original dynamics, exhibiting self-amplified amplitude-limited dynamics (Stuart 1971; Busse 1991; Noack & Eckelmann 1994). Landau (1944) and Stuart (1958) pioneered the mean-field model, a major progress in first-principle-based ROMs that provides insight into flow instabilities and bifurcation theory. For instance, in the case of a supercritical Hopf bifurcation, mean-field models have been applied to the vortex shedding behind a cylinder (Strykowski & Sreenivasan 1990; Schumm *et al.* 1994; Noack *et al.* 2003) and high Reynolds number turbulent wake flow (Bourgeois *et al.* 2013). For more complex flows undergoing successive bifurcations, including both Pitchfork and Hopf bifurcations, weakly nonlinear mean-field analysis is applied to the wake of axisymmetric bodies (Fabre *et al.* 2008), the wake of a disk (Meliga *et al.* 2009) and the fluidic pinball (Deng *et al.* 2020). Furthermore, in the field of resolvent analysis, the mean-field theory also contributes by decomposing the system into time-resolved linear dynamics and a feedback term involving quadratic nonlinearity (McKeon *et al.* 2004; Gómez *et al.* 2016; Rigas *et al.* 2017).

In contrast to a first principle ROM, a data-driven version is based on a low-dimensional representation of flow snapshots. Proper orthogonal decomposition (POD) is a commonly used example. POD begins with the eigenvalue or singular value decomposition of the correlation matrix, yielding a low-dimensional subspace comprising leading orthogonal eigenvectors. This subspace provides an “optimal” Galerkin expansion with minimal average residual in the energy norm. Since Aubry *et al.* (1988) introduced the groundbreaking POD-Galerkin model for unforced turbulent boundary layers, numerous POD models have emerged for various configurations. Examples include POD models for channel flow (Podvin & Lumley 1998; Podvin 2009), the wake of a two-dimensional square cylinder Bergmann *et al.* (2009), laminar and turbulent vortex shedding (Iollo *et al.* 2000), and flow past a circular cylinder with dynamic subgrid-scale model and variational multiscale model Iollo *et al.* (2000). There are also various variations of the POD model, e.g. integrating the actuation terms into the projection system for control design (Bergmann & Cordier 2008; Luchtenburg *et al.* 2009) and balanced POD (Rowley 2005), which is derived from a POD approximation to the product of controllability and observability Gramians to obtain an approximately balanced truncation (Moore 1981). Increasingly powerful machine learning methods can make data-driven ROMs more automated. Examples include the sparse identification of nonlinear dynamics (SINDy) aim at human interpretable models (Brunton *et al.* 2016), ROMs with artificial neural networks (San & Maulik 2018; San *et al.* 2019; Zhu *et al.* 2019; Kou & Zhang 2021), turbulence modelling and flow estimation with multi-input multi-output by deep neural networks (Kutz 2017; Li *et al.* 2022), and manifold learning methods (Farzamnik *et al.* 2023).

In this work, we focus on automated data-driven modelling. The starting point is cluster-based ROMs (CROMs), pioneered by Burkardt *et al.* (2006). Clustering is an unsupervised classification of patterns into groups commonly used in data science (Jain & Dubes 1988; Jain *et al.* 1999; Jain 2010), it is popular in data mining, document retrieval, image segmentation, and feature detection (Kim *et al.* 2022). The foundation of the CROM lies in the cluster-based Markov model (CMM) proposed by Kaiser *et al.* (2014), which combines a cluster analysis of an ensemble of snapshots and a Markov model for transitions between different flow

states reduced by clustering. The CMM has provided a valuable physical understanding of the mixing layer, Ahmed body wakes (Kaiser *et al.* 2014), combustion-related mixing (Cao *et al.* 2014), and supersonic mixing layer (Li & Tan 2020). Nair *et al.* (2019) applied the cluster-based model to feedback control for drag reduction and first introduced a directed network for dynamical modelling. Building on this concept, Fernex *et al.* (2021) and Li *et al.* (2021) further proposed the cluster-based network model (CNM) with improved long-timescale resolution. Instead of the “stroboscopic” view of the CMM, the CNM focuses on non-trivial transitions. The dynamics are restricted to a simple network model between the cluster centroids, like a deterministic-stochastic flight schedule which allows only a few possible flights with corresponding probabilities and flight times consistent with the data set. Networks of complex dynamic systems have gained great interest, forming an increasingly important interdisciplinary field known as network science (Watts & Strogatz 1998; Albert & Barabási 2002; Börner *et al.* 2007; Barabási 2013). Network-based approaches are often used in fluid flows (Nair & Taira 2015; Hadjighasem *et al.* 2016; Taira *et al.* 2016; Yeh *et al.* 2021; Taira & Nair 2022), in conjunction with clustering analysis (Bollt 2001; Schlueter-Kuck & Dabiri 2017; Murayama *et al.* 2018; Krueger *et al.* 2019). The critical structures that modify the dynamical system can be identified by the intra- and inter-cluster interactions using community detection (Gopalakrishnan Meena *et al.* 2018; Gopalakrishnan Meena & Taira 2021).

CROMs are fully automated, robust, and physically interpretable, while the model accuracy is strongly related to the clustering process. The state space is equivalently discretised in the above-mentioned CROMs, leading to a lack of dynamic coverage. For example, the CNM struggles to capture multiscale behaviours such as the oscillations near attractors and the amplitude variations between trajectories. To address this issue, an effective solution is to employ dynamics-augmented clustering to determine the centroid distribution. Inspired by the hierarchical clustering (Deng *et al.* 2022) and the network sparsification (Nair & Taira 2015), we propose a dynamics-augmented cluster-based network model (dCNM) with an improved resolution of complex dynamics. In this case, the time-resolved dynamics are reflected by the evolution of trajectory segments after the state space is clustered. These segments are automatically identified from cluster transitions and are represented by centroids obtained through segment averaging. A second-stage clustering further refines the centroids, eliminating the network redundancy and also deepening the comprehension of underlying physical mechanisms. The proposed dCNM can systematically identify complex dynamics involved in the case of multi-attractor, multi-frequency, and multiscale dynamic systems. Figure 1 provides a comparative illustration of CNM and dCNM in terms of kinematics and dynamics, exemplified by an inward spiral trajectory in a two-dimensional state space.

The dCNM is first applied to the Lorenz (1963) system as an illustrative example. The Lorenz attractor is notable for the “butterfly effect”, showcasing the chaotic dynamics governed by only three ordinary differential equations. Subsequently, we demonstrate the dCNM on the sphere wake of the periodic, quasi-periodic, and chaotic flow regimes. The sphere wake is a well-investigated benchmark configuration, serving as a prototype flow of bluff body wakes commonly encountered in many modern applications, for instance, the design of drones and air taxis. Despite the simple geometry, the sphere wake can experience a series of bifurcations with increasing Reynolds number. Along the route to turbulence, the flow system exhibits steady, periodic, quasi-periodic, and chaotic flow regimes. The transient and post-transient flow dynamics, characterised by multi-frequency and multiscale behaviours, provide a challenging testing ground for reduced-order modelling.

This manuscript is organised as follows: In § 2, the clustering algorithm and the different perspectives on the dCNM strategy are described. In § 3, the dCNM is illustrated on the Lorenz system, and in § 4, it is demonstrated on the sphere wake of three flow regimes: the

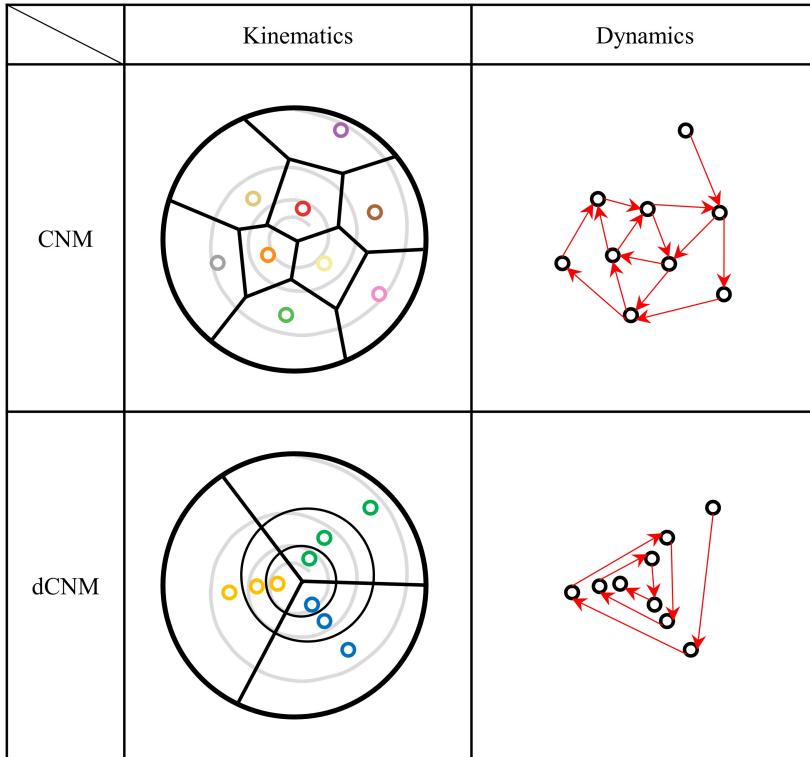


Figure 1: Principle sketches: The CNM and the dCNM are illustrated using an inward spiral trajectory in a two-dimensional state space with the same number of centroids. The thick solid lines denote cluster divisions, and the thin solid lines represent sub-cluster divisions. The centroids are represented by coloured dots, and their colours represent their cluster affiliation. The CNM centroids are derived from snapshot averages within each cluster and show uniform geometric coverage, whereas the dCNM centroids incorporate dynamic features and exhibit a weighted distribution. Consequently, dCNM accurately reconstructs the cycle-to-cycle variations and also ensures precise transition sequencing.

periodic flow, the quasi-periodic flow and the chaotic flow. In § 5, the main findings and improvements are summarised, and future directions are suggested.

2. Dynamics-augmented cluster-based network model

In this section, we detail the process of the dynamics-augmented cluster-based network model. In § 2.1, the k -means++ clustering algorithm and its demonstration on the state space are introduced. The second-stage clustering on the trajectory segments is further discussed in § 2.2. In § 2.3, the transition characteristics are described, and in § 2.4, different criteria are introduced to evaluate the performance of the proposed model. The variables used in this section are listed in table 1.

2.1. Clustering the state space

The dynamics-augmented clustering procedure is divided into two steps. Initially, the state space is clustered, yielding coarse-grained state transition dynamics with trajectory segments composed of time-continuous snapshots within each cluster. Subsequently, we cluster these

Variables	Description
\mathbf{u}^m	Time-resolved snapshots
M	Number of snapshots
Clustering the state space	
K	Number of clusters
C_k, C_i	Clusters obtained by the state space clustering
χ_k^m	Characteristic function of the state space clustering
M_k	Number of snapshots in cluster C_k
χ_{ik}^m	Characteristic function of transition from C_k to C_i
\mathbf{c}_k	Centroids of clusters
n_{ik}	Number of transitions from C_k to C_i
n_k	Total number of transitions from C_k
n_{traj}	Total number of transitions of the data set
Q_{ik}	Cluster transition probability from C_k to C_i
T_{ik}	Cluster transition time from C_k to C_i
\mathbf{Q}	Cluster transition probability matrix
\mathbf{T}	Cluster transition time matrix
\mathbf{R}^u	Cluster deviation on snapshots
Clustering the trajectory segments	
$\mathcal{T}_{(kl)}$	The l -th trajectory segment in C_k
$\chi_{(kl)}^m$	Characteristic function of the second-stage clustering
$M_{(kl)}$	Number of snapshots in trajectory segment $\mathcal{T}_{(kl)}$
$\mathbf{c}_{(kl)}, \mathbf{c}_{(ij)}$	Centroids of trajectory segments
$\mathbf{L} = [L_1, \dots, L_K]^\top$	Number of sub-clusters for the second-stage clustering
$n_{(ij)(kl)}$	Number of transitions from $\mathbf{c}_{(kl)}$ only to $\mathbf{c}_{(ij)}$
$Q_{(ij)(kl)}$	Centroid transition probability from $\mathbf{c}_{(kl)}$ to $\mathbf{c}_{(ij)}$
\mathbf{Q}_{ik}	Centroid transition probability matrix
Q_k	Centroid transition probability tensor
$\mathbf{R}^{\mathcal{T}}$	Cluster deviation on trajectory segments

Table 1: Table of variables. Subscripts k and i are related to the level of clusters from the state space clustering, and subscripts l and j are related to the level of trajectory segments.

trajectory segments, utilising centroids derived from the average of each segment. This step optimises the centroid distribution and eliminates the redundancy of the trajectory segments.

The first-stage clustering discretises the high-dimensional state space by grouping the snapshots. We first define a Hilbert space $\mathcal{L}^2(\Omega)$, in which the inner product of vector fields in the domain Ω is given by a square-integrable function:

$$(\mathbf{u}, \mathbf{v})_\Omega = \int_\Omega d\mathbf{x} \mathbf{u} \cdot \mathbf{v}, \quad (2.1)$$

where \mathbf{u} and \mathbf{v} represent snapshots of this vector field, also known as observations in the

machine learning context. The corresponding norm is defined as:

$$\|\mathbf{u}\|_{\Omega} := \sqrt{(\mathbf{u}, \mathbf{u})_{\Omega}}. \quad (2.2)$$

The distance D between two snapshots can be calculated as follows:

$$D(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_{\Omega}. \quad (2.3)$$

The unsupervised k -means++ algorithm (MacQueen 1967; Lloyd 1982; Arthur & Vassilvitskii 2007) is used for clustering. It operates automatically, devoid of assumptions or data categorisation. Serving as the foundation of cluster analysis, this algorithm partitions a set of M time-resolved snapshots \mathbf{u}^m , where $m = 1 \dots M$, into K clusters C_k , where $k = 1 \dots K$. Each cluster corresponds to a centroidal Voronoi cell, with the centroid defined as the average of the snapshots within the same cluster. The algorithm comprises the following steps:

(i) Initialisation: K centroids \mathbf{c}_k , where $k = 1 \dots K$, are randomly selected. In contrast to the k -means algorithm, k -means++ optimises the placement of these centroids to prevent sensitivity to initial conditions.

(ii) Assignment: Each snapshot \mathbf{u}^m is allocated to the nearest centroid by $\arg \min_k D(\mathbf{u}^m, \mathbf{c}_k)$. The characteristic function is used to mark their affiliation, and it is defined as follows:

$$\chi_k^m := \begin{cases} 1, & \text{if } \mathbf{u}^m \in C_k \\ 0, & \text{otherwise} \end{cases} \quad (2.4)$$

(iii) Update: Each centroid is recalculated by averaging all the snapshots belonging to the corresponding cluster as follows:

$$\mathbf{c}_k = \frac{1}{M_k} \sum_{\mathbf{u}^m \in C_k} \mathbf{u}^m = \frac{1}{M_k} \sum_{m=1}^M \chi_k^m \mathbf{u}^m, \quad (2.5)$$

where

$$M_k = \sum_{m=1}^M \chi_k^m. \quad (2.6)$$

(iv) Iteration: The Assignment and Update steps are repeated until convergence is reached. Convergence means that the centroids do not move or stabilise below a certain threshold. The algorithm minimises the intra-cluster variance and maximises the inter-cluster variance. The intra-cluster variance is computed as follows:

$$J(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{k=1}^K \sum_{m=1}^M \chi_k^m \|\mathbf{u}^m - \mathbf{c}_k\|_{\Omega}^2. \quad (2.7)$$

Each iteration reduces the value of the criterion J until convergence is reached.

The cluster probability distribution $\mathbf{P} = [P_1, \dots, P_K]$ is determined by $P_k = M_k/M$ for each cluster C_k , and satisfies the normalisation condition $\sum_{k=1}^K P_k = 1$.

The geometric properties of the clusters are quantified for further analysis. The cluster standard deviation on the snapshots R_k^u measures the cluster size, following Kaiser *et al.* (2014), as:

$$R_k^u = \sqrt{\frac{1}{M_k} \sum_{m=1}^M \chi_k^m \|\mathbf{u}^m - \mathbf{c}_k\|_{\Omega}^2}. \quad (2.8)$$

The time-resolved snapshots should be equidistantly sampled and cover a statistically representative time window of the coherent structure evolution. As a rule of thumb, at least

ten periods of the dominant frequency are needed to obtain reasonably accurate statistical moments and at least K snapshots per characteristic period to capture an accurate temporal evolution.

2.2. Clustering the trajectory segments

After the state space is discretized, the trajectory is also divided into segments. We use the cluster transition information to identify the trajectory segments that pass through a cluster.

Based on the temporal information from the given data set, the nonlinear dynamics between snapshots are modelled as linear transitions between clusters, known as the classic CNM (Fernex *et al.* 2021; Li *et al.* 2021). We infer the probability of cluster transition from the data as follows:

$$Q_{ik} = \frac{n_{ik}}{n_k}, \quad i, k = 1, \dots, K, \quad (2.9)$$

where Q_{ik} is the direct cluster transition probability from cluster C_k to C_i and n_{ik} is the number of transitions from C_k only to C_i :

$$n_{ik} = \sum_{m=1}^M \chi_{ik}^m, \quad (2.10)$$

where

$$\chi_{ik}^m = \begin{cases} 1, & \text{if } \mathbf{u}^m \in C_k \text{ \& } \mathbf{u}^{m+1} \in C_i \\ 0, & \text{otherwise} \end{cases} \quad (2.11)$$

n_k is the total number of transitions from C_k regardless of the destination cluster:

$$n_k = \sum_{i=1}^K n_{ik}, \quad i, k = 1, \dots, K. \quad (2.12)$$

If $Q_{ik} \neq 0$, it can be inferred that in cluster C_k there exists at least one trajectory segment that is bound for cluster C_i . We assign distinct labels to each trajectory segment corresponding to all destination clusters, denoted as $\mathcal{T}_{(kl)}$, where k and l represent the l -th segment in C_k . Therefore the snapshots are marked according to their trajectory affiliations by a characteristic function:

$$\chi_{(kl)}^m = \begin{cases} 1, & \text{if } \mathbf{u}^m \in \mathcal{T}_{(kl)} \\ 0, & \text{otherwise} \end{cases} \quad (2.13)$$

where k represents the cluster affiliation, and l represents the trajectory segment affiliation. The total number of trajectory segments in C_k equals n_k . Note that the final trajectory segment of the data set will not be considered as it will not lead to any destination cluster and is usually incomplete. The total number of trajectory segments in the data set can be obtained by the sum of n_k as follows:

$$n_{\text{traj}} = \sum_{k=1}^K n_k. \quad (2.14)$$

Analogous trajectory segments within the same cluster will be merged in the subsequent clustering stage. Operations on the trajectories can often be costly. Efficiency in clustering can be achieved by mapping the operations performed on trajectory segments to their corresponding averages, i.e., the trajectory segment centroids, given their topological relationship. Additionally, the propagation of our model relies on centroids, rendering the trajectory information essentially unnecessary. We define the centroids $\mathbf{c}_{(kl)}$ as the average

of snapshots belonging to the same trajectory segment:

$$\mathbf{c}_{(kl)} = \frac{1}{M_{(kl)}} \sum_{\mathbf{u}^m \in \mathcal{T}_{(kl)}} \mathbf{u}^m = \frac{1}{M_{(kl)}} \sum_{m=1}^M \chi_{(kl)}^m \mathbf{u}^m, \quad (2.15)$$

where

$$M_{(kl)} = \sum_{m=1}^M \chi_{(kl)}^m. \quad (2.16)$$

The subsequent question pertains to how to determine the number of sub-clusters. The allocation of sub-clusters within each cluster can be automatically learnt from the data. To maintain the spatial resolution, more sub-clusters should be assigned to clusters with a larger transverse size. We first introduce a transverse cluster size vector $\mathbf{R}^{\mathcal{T}}$, which is defined by the standard deviation of the n_k centroids $\mathbf{c}_{(kl)}$ with respect to the cluster centroid \mathbf{c}_k as follows:

$$R_k^{\mathcal{T}} = \sqrt{\frac{1}{n_k} \sum_{l=1}^{n_k} \|\mathbf{c}_{(kl)} - \mathbf{c}_k\|_{\Omega}^2}. \quad (2.17)$$

Next, we denote the number of sub-clusters as L_k for clustering the centroids in cluster C_k . A K -dimensional vector $\mathbf{L} = [L_1, \dots, L_K]^{\top}$ records the numbers of sub-clusters in each cluster, with L_k determined by:

$$L_k = \min(\lfloor \hat{R}_k^{\mathcal{T}} n_{\text{traj}} (1 - \beta) \rfloor + 1, n_k). \quad (2.18)$$

Here, the vector $\mathbf{R}^{\mathcal{T}}$ is normalised with the sum $\sum_{k=1}^K R_k^{\mathcal{T}}$, denoted as $\hat{\mathbf{R}}^{\mathcal{T}}$, which ensures a suitable distribution of sub-clusters for the ensemble of n_{traj} trajectories. To increase the flexibility of the model, we introduce a sparsification controller $\beta \in [0, 1]$ in this clustering process. For the extreme value of $\beta = 1$, all the centroids are merged into one centroid, and the dCNM is identical to a classic CNM, with the maximum sparsification. For the other extreme $\beta = 0$, the dCNM is minimally sparsified according to the transverse cluster size. For periodic or quasi-periodic systems, the dCNM with a large β can capture most of the dynamics, while for complex systems such as chaotic systems, a small β may be needed. In addition, the minimum function prevents the possibility that the left-hand side of the equation exceeds the number of centroids n_k when β is too small, causing the second-stage clustering to not be performed. The choice of β is discussed in Appendix C.

The refined centroids are obtained by averaging a series of centroids related to analogous trajectory segments. The redundancy of the n_{traj} centroids is mitigated, and the corresponding transition network becomes sparse. The k -means++ algorithm is also used in the second-stage clustering. It will iteratively update the centroids $\mathbf{c}_{(kl)}$ and the characteristic function $\chi_{(kl)}^m$ until convergence or the maximum number of iterations is reached. The overall clustering process of the dCNM is summarised in Algorithm 1.

2.3. Characterising the transition dynamics

We use the centroids obtained from § 2.2 as the nodes of the network and the linear transitions between these centroids as the edges of the network. First, we introduce two transition properties: the centroid transition probability $Q_{(ij)(kl)}$ and the transition time T_{ik} .

Figure 2 illustrates the definition of the subscripts in the centroid transition probability $Q_{(ij)(kl)}$, which can contain all possible transitions between the refined centroids of clusters C_k and C_l . Considering the transitions between these centroids, we define $Q_{(ij)(kl)}$ as:

Algorithm 1: Pseudocode for the dynamics-augmented clustering procedure

Input: \mathbf{u}^m : Snapshots;
 K : Number of clusters;
 β : Sparsification index ($0 \leq \beta \leq 1$);
Output: $\mathbf{c}_{(kl)}$: Refined centroids;
 R_k^u, R_k^T : Geometric properties;
 $\chi_{(kl)}^m$: Characteristic function

```

1 Apply  $k$ -means++ algorithm with  $K$  clusters to  $\mathbf{u}^m$ 
2 Save the characteristic function as  $\chi_k^m$ 
3 for  $k \leftarrow 1$  to  $K$  do
4   for  $i \leftarrow 1$  to  $K$  do
5     Compute the transition probability  $Q_{ik}$ 
6     if  $Q_{ik} \neq 0$  then
7       Locate the time-continuous snapshots in cluster  $C_k$  on each trajectory
        segment to  $C_i$ 
8       Save the characteristic function  $\chi_{(kl)}^m$  accordingly
9     end
10   end
11   Compute and save the centroids  $\mathbf{c}_{(kl)}$  by  $\chi_{(kl)}^m$ , compute the geometric properties
       $R_k^u$  and  $R_k^T$ 
12 end
13 Compute  $\mathbf{L}$  by  $R_k^T$  and  $\beta$ 
14 for  $k \leftarrow 1$  to  $K$  do
15   Locate the centroids  $\mathbf{c}_{(kl)}$  in cluster  $C_k$ 
16   Apply  $k$ -means++ algorithm with  $L_k$  clusters directly to  $\mathbf{c}_{(kl)}$ 
17   Update the characteristic function  $\chi_{(kl)}^m$  and the centroids  $\mathbf{c}_{(kl)}$ .
18 end
```

$$Q_{(ij)(kl)} = \frac{n_{(ij)(kl)}}{n_k}, \quad i, k = 1, \dots, K, \quad j = 1, \dots, L_i, \quad l = 1, \dots, L_k, \quad (2.19)$$

where $n_{(ij)(kl)}$ is the number of transitions from $\mathbf{c}_{(kl)}$ only to $\mathbf{c}_{(ij)}$. This definition differs from that of the CNM, which uses the cluster transition Q_{ik} in 2.9 to define the probability. In fact, we can compute Q_{ik} by summing up $Q_{(ij)(kl)}$ as follows:

$$Q_{ik} = \sum_{j=1}^{L_i} \sum_{l=1}^{L_k} Q_{(ij)(kl)}. \quad (2.20)$$

The definition of the transition time T_{ik} is identical to the CNM, as shown in figure 3. This property is not further investigated in the present work, as the transition time crossing the same clusters varies little in most dynamic systems.

Let t^n be the instant when the first snapshot enters, and t^{n+1} be the instant when the last snapshot leaves on one trajectory segment passing through cluster C_k . The residence time τ_k^n is the duration of staying in cluster C_k on this segment, which is given by:

$$\tau_k^n = t^{n+1} - t^n. \quad (2.21)$$

For an individual transition from C_k to C_i , the transition time is defined as τ_{ik}^n , which can be

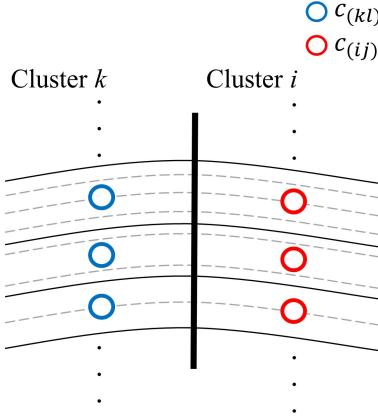


Figure 2: Illustration of the subscripts in the refined centroid transitions. After the state space is clustered, only one subscript is needed to distinguish the different clusters, such as C_k and C_i . After the trajectory segments are clustered, two subscripts are needed to represent the refined centroids, such as $c_{(kl)}$ in C_k and $c_{(ij)}$ in C_i .

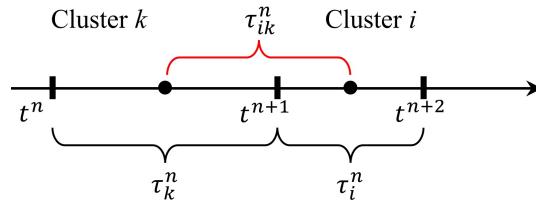


Figure 3: Individual transition time τ_{ik}^n for the transition from cluster C_k to C_i

obtained by the average of the residence times from both clusters:

$$\tau_{ik}^n = (\tau_k^n + \tau_i^n)/2. \quad (2.22)$$

By averaging τ_{ik}^n for all the individual transitions from C_k to C_i , the transition time can be expressed as follows:

$$T_{ik} = \frac{\sum_{n=1}^{n_{ik}} \tau_{ik}^n}{n_{ik}}. \quad (2.23)$$

The essential dynamics can also be summarised into single entities as in the CNM, since the cluster-level information is still retained in the current model. For completeness, we introduce the cluster transition probability matrix \mathbf{Q} and the cluster transition time matrix \mathbf{T} as:

$$\begin{aligned} \mathbf{Q} &= Q_{ik} \in \mathbb{R}^{K \times K}, \quad i, k = 1, \dots, K \\ \mathbf{T} &= T_{ik} \in \mathbb{R}^{K \times K}, \quad i, k = 1, \dots, K. \end{aligned} \quad (2.24)$$

The cluster indices are reordered in both matrices to enhance readability. C_1 is the cluster with the highest distribution probability, C_2 is the cluster with the highest transition probability leaving from C_1 , and C_3 is the cluster with the highest transition probability leaving from C_2 , so on and so forth. If the cluster with the highest probability is already assigned, we choose the cluster with the second highest probability. If all the clusters with nonzero transition probabilities are already assigned, we choose the next cluster with the highest distribution probability among the rest.

By analogy with \mathbf{Q} , the centroid transition probability $Q_{(ij)(kl)}$ for given affiliations of the

departure cluster k and destination cluster i can form a centroid transition matrix \mathbf{Q}_{ik} that captures all possible centroid dynamics between the two clusters:

$$\mathbf{Q}_{ik} = \mathcal{Q}_{(ij)(kl)} \in \mathbb{R}^{L_i \times L_k}, \quad j = 1, \dots, L_i, \quad l = 1, \dots, L_k. \quad (2.25)$$

Moreover, to summarise the centroid transition dynamics, the centroid transition probability $\mathcal{Q}_{(ij)(kl)}$ for a given affiliation k of only the departure cluster can form a centroid transition tensor \mathbf{Q}_k that captures all the possible centroid dynamics from this cluster, as:

$$\mathbf{Q}_k = \mathcal{Q}_{(ij)(kl)} \in \mathbb{R}^{K \times L_i \times L_k}, \quad i = 1, \dots, K, \quad j = 1, \dots, L_i, \quad l = 1, \dots, L_k. \quad (2.26)$$

The dCNM propagates the state motion based on the centroids $\mathbf{c}_{(kl)}$ for the reconstruction. To determine the transition dynamics, we first use \mathbf{Q}_k to find the centroid transitions from the initial centroid $\mathbf{c}_{(kl)}$ to the destination $\mathbf{c}_{(ij)}$. As the destination centroids are determined, the cluster-level dynamics are determined correspondingly. Then, \mathbf{T} is used to identify the related transition time.

We assume a linear state propagation between the two centroids $\mathbf{c}_{(kl)}$ and $\mathbf{c}_{(ij)}$ obtained from the tensors, as follows:

$$\mathbf{u}^m(t) = \alpha_{ik}(t)\mathbf{c}_{(ij)} + [1 - \alpha_{ik}(t)]\mathbf{c}_{(kl)}, \quad \alpha_{ik} = \frac{t - t_k}{T_{ik}}. \quad (2.27)$$

Here t_k is the time when the centroid $\mathbf{c}_{(kl)}$ is left. Note that we can use splines (Fernex *et al.* 2021) or add the trajectory supporting points (Hou *et al.* 2022) to interpolate the motion between the centroids for smoother trajectories.

Intriguingly, we observe that the trajectory-based clustering of the dCNM enhances the resolution of the cluster transitions. Now each centroid only has a limited number of destination centroids, often within the same cluster. This minimises the likelihood of selecting the wrong destination cluster based solely on the cluster transition probability matrix, as is the case in classic CNM. Consequently, it becomes feasible to accurately resolve long-term cluster transitions without the need for historical information. It can be argued that dCNM effectively constrains cluster transitions, leading to outcomes similar to those obtained with the higher-order CNM (Fernex *et al.* 2021). This improvement is attained by replacing higher-order indexing with higher-dimensionality dual indexing. Specifically, the dual indexing also results in a substantial reduction in the model complexity. While the complexity of the high-order CNM is defined as $K^{\tilde{L}}$, where K is the number of clusters and \tilde{L} is the order, the model complexity of the dCNM is expressed as $\sum_{k=1}^K L_k$, which is a significantly lower value, particularly when \tilde{L} is relatively large. In terms of computational efficiency, dCNM with $\beta = 0.80$ reduces the computational time by 40% as compared to CNM with the same number of centroids. This improvement is primarily attributed to the hierarchical clustering. The computational load of the first-stage clustering on the state space is reduced by a small number of clusters K . The second-stage clustering on the trajectory segments accounts only for 20% of the total computation time.

2.4. Validation

The auto-correlation function and the representation error are used for validation. We examine the prediction errors for cluster-based models considering both spatial and temporal perspectives. The spatial error arises from the inadequate representation by cluster centroids, as evidenced by the representation error and the auto-correlation function. The temporal error arises due to the imprecise reconstruction of intricate snapshot transition dynamics. This can be observed directly through the temporal evolution of snapshot affiliations and, to some extent, through the auto-correlation function.

The auto-correlation function is a practical tool for evaluating ROMs, as it can statistically

reflect the prediction errors. Additionally, the auto-correlation function circumvents the problem of directly comparing two trajectories with finite prediction horizons, which may suffer from phase mismatch (Fernex *et al.* 2021). This is particularly relevant for chaotic dynamics, whereby minor differences in initial conditions can lead to divergent trajectories, making the direct comparison of time series meaningless. The unbiased auto-correlation function of the state vector (Protas *et al.* 2015) is given by:

$$R(\tau) = \frac{1}{T - \tau} \int_0^{T-\tau} (\mathbf{u}(x, t), \mathbf{u}(x, t + \tau))_{\Omega} dt, \quad \tau \in [0, T]. \quad (2.28)$$

In this study, $R(\tau)$ will be normalised by $R(0)$ (Deng *et al.* 2022). This function can also infer the spectral behaviour by computing the fluctuation energy at the vanishing delay.

The representation error can be numerically computed as:

$$E_r = \frac{1}{M} \sum_{m=1}^M D_{\mathcal{T}}^m, \quad (2.29)$$

where $D_{\mathcal{T}}^m$ is the minimum distance from the snapshot \mathbf{u}^m to the states on the reconstructed trajectory \mathcal{T} :

$$D_{\mathcal{T}}^m = \min_{\mathbf{u}^n \in \mathcal{T}} \|\mathbf{u}^m - \mathbf{u}^n\|_{\Omega}. \quad (2.30)$$

3. Lorenz system as an illustrative example

In this section, we apply the dCNM to the Lorenz (1963) system to illustrate its superior spatial resolution in handling multiscale dynamics. We also compare it with the CNM (Li *et al.* 2021; Fernex *et al.* 2021) of the same rank as a reference.

The Lorenz system is a three-dimensional autonomous system with non-periodic, deterministic, and dissipative dynamics that exhibit exponential divergence and convergence to strange fractal attractors. The system is governed by three coupled nonlinear differential equations:

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x), \\ \frac{dy}{dt} &= x(\rho - z) - y, \\ \frac{dz}{dt} &= xy - \beta z. \end{aligned} \quad (3.1)$$

The system parameters are set as $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. These equations emulate the Rayleigh-Bénard convection. The trajectory of the system revolves around two weakly unstable oscillatory fixed points, forming two sets of attractors, that are loosely called ‘‘ears’’. These two ears have similar but not identical shapes, with the left ear being rounder and thicker in the toroidal region. The region where the ears overlap is called the branching region. The Lorenz system has two main types of dynamics. One is that the inner loop in each ear varies and oscillates for several cycles. The other is that the inner loop may randomly switch from one ear to another in the branching region and resume oscillatory motion.

We numerically integrate the system using the fourth-order explicit Runge-Kutta method. The time-resolved 10000 snapshots data with $\mathbf{u}^m = [x, y, z]^{\top}$ are collected at a sampling time step of $\Delta t = 0.015$ with an initial condition of $[-3, 0, 31]^{\top}$ (Fernex *et al.* 2021). This time step corresponds to approximately one-fiftieth of a typical cycle period. The first 5% of the snapshots are neglected to reserve only the post-transient dynamics.

Figure 4 shows the phase portrait of the clustered Lorenz system from the CNM and dCNM. We set $K = 10$ for the state space clustering of the dCNM, which is consistent with previous studies (Kaiser *et al.* 2014; Li *et al.* 2021). This number is

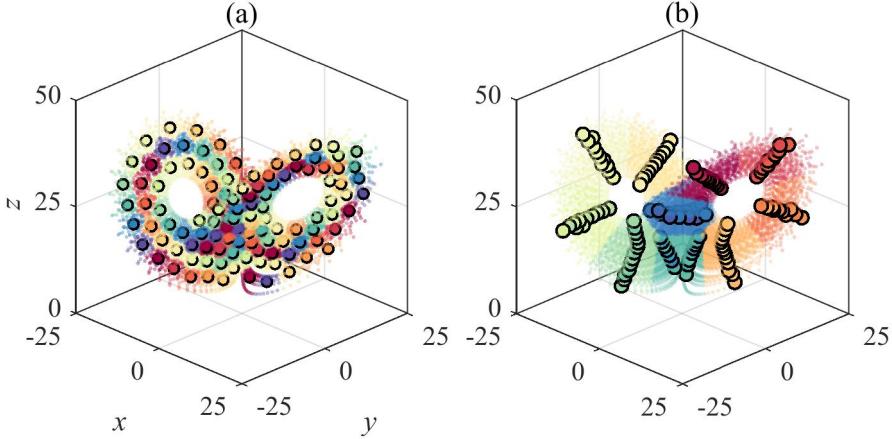


Figure 4: Phase portrait of the clustered Lorenz system from the CNM and dCNM. The small dots represent the snapshots, and the large dots represent the centroids. Snapshots and centroids with the same colour belong to the same cluster. As a comparison, the CNM result in (a) is shown with the same number of centroids as the corresponding dCNM result. The dCNM result in (b) is shown with $K = 10$ and $\beta = 0.90$.

large enough for the further subdivision of transition dynamics and is also small enough to obtain a simple structure for understanding. The sparsification index β is chosen with large numbers as $\beta = 0.90$ to allow for a distinct visualisation of the centroids. In addition, since the trajectory in each “ear” is confined to a two-dimensional surface, a high value of β is deemed suitable. The normalized transverse cluster size vector $\hat{\mathbf{R}}^\top = [0.1163, 0.1262, 0.1164, 0.0921, 0.0943, 0.0908, 0.1116, 0.0840, 0.0866, 0.0817]^\top$ corresponds to the number of sub-clusters $\mathbf{L} = [13, 15, 13, 11, 11, 11, 13, 10, 10, 10]^\top$.

The two models exhibit notable differences in centroid distribution. CNM clustering relies solely on the spatial topology in the phase space, evenly dividing the entire attractor and dispersing centroids uniformly throughout the phase portrait. It can be inferred that increasing the number of centroids under this uniform distribution does not lead to substantial changes, merely resulting in a denser centroid distribution. This uniform distribution possesses certain disadvantages regarding the dynamics. First, it unnecessarily complicates the transition rhythm as the deterministic large-scale transition may be fragmented into several stochastic transitions. Second, even with many centroids, it fails to capture the increasing oscillation amplitude between the loops in one ear, as the uniform distribution provides only a limited number of centroid orbits. The same result occurs for the branching region where these limited numbers of centroids usually oversimplify the switch between ears. In contrast, the distribution of the dCNM centroids resembles a weighted reallocation. For the Lorenz system, the state space is stratified along the trajectory direction, leading to a concentrated distribution of the dCNM centroids in the radial direction of the attractor and the branching region, which correspond to the system’s primary dynamics. Additionally, varying quantities of the centroids can be observed in the radial direction in the toroidal region, depending on its thickness. In thinner toroidal regions with smaller variations between trajectory segments, the second-stage clustering assigns fewer sub-clusters and, consequently, builds fewer centroids.

The cluster transition matrices, which are a distinctive feature of cluster modelling, are preserved because the dCNM maintains the coarse-grained transitions at the cluster level. Figure 5 illustrates the cluster transition probability matrix \mathbf{Q} and the corresponding transition

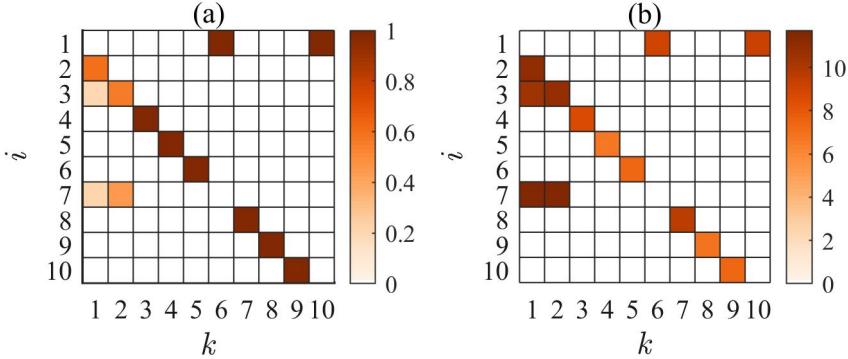


Figure 5: Transition matrices of the Lorenz system. The colour bar indicates the values of the terms. (a) Transition probability matrix \mathbf{Q} . (b) Transition time matrix \mathbf{T} .

time matrix \mathbf{T} to illustrate the significant dynamics of the Lorenz system. It is worth noting that in the case of the CNM with an equivalent number of centroids, the matrices become considerably larger, which diminishes their readability and interpretability. The matrices reveal three distinct cluster groups. The first group comprises clusters C_1 and C_2 , which resolve the branching region and exhibit similar transition probabilities to clusters C_3 and C_7 . The branching region is further linked to different ears and is crucial to the attractor oscillation. Clusters C_1 and C_2 can be referred to as flipper clusters (Kaiser *et al.* 2014), representing a switch between the different groups. The equivalent transition probability from C_2 is consistent with the random jumping behaviour of the two ears. The other two groups demonstrate an inner-group circulation corresponding to the main components of the two ears, exemplified by the cluster chains $C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_6$ and $C_7 \rightarrow C_8 \rightarrow C_9 \rightarrow C_{10}$. These chains exhibit deterministic transition probabilities that resolve the cyclic behaviour. In the second-stage clustering, these two groups are further categorised into numerous centroid orbits. Moreover, the transition time matrix resolves the variance in the transition times, with significantly shorter transition times observed in the cyclic groups compared to transitions involving the flipper clusters.

The original and reconstructed trajectories in the phase space are directly compared. We focus solely on the spatial resolution, disregarding phase mismatches during temporal evolution. Figure 6 shows the original Lorenz system and the reconstruction by the CNM and dCNM with the same parameters as in figure 4. To ensure clarity, we select a time window from $t = 0$ to $t = 30$ for the trajectories and employ spline interpolation for a smooth reconstruction. Inaccurate or non-physical centroid transitions, along with incomplete dynamic coverage, can lead to substantial deformations in the reconstructed trajectory. As expected, the dCNM provides a more accurate reconstruction than the CNM. The CNM uses a finite number of centroid orbits to represent oscillating attractors, converting slow and continuous amplitude growth into limited and abrupt amplitude jumps. Furthermore, the CNM may group one continuous snapshot loop into clusters belonging to different centroid orbits, often when these clusters are adjacent to each other. This can lead to unnecessary orbit-crossing centroid transitions and result in nonphysical radial jumps in the reconstructed trajectory. In contrast, the dCNM provides more comprehensive dynamic coverage, resolving more cyclic behaviour with additional centroid orbits. Dual indexing also guarantees accurate centroid transitions. The radial jumps are eliminated, as departing centroids can only transition to destination centroids within the same centroid orbits. Consequently, oscillations

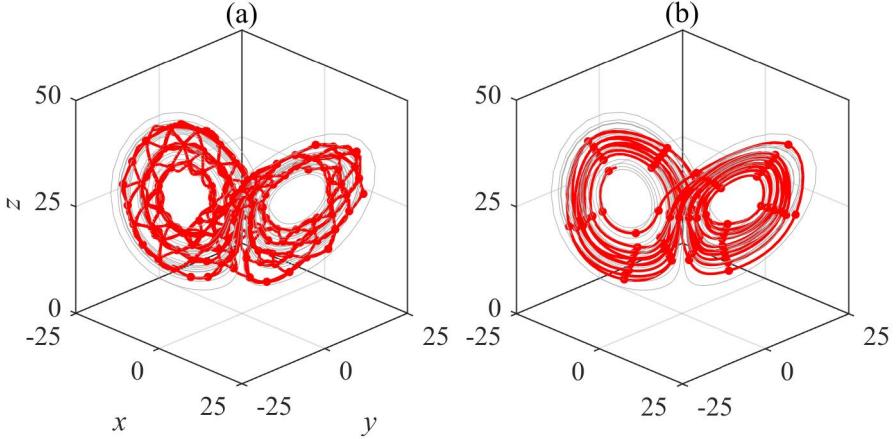


Figure 6: Trajectory of the Lorenz system. The thin grey curve represents the original trajectory, the thick red curve represents the reconstructed trajectory, and the red dots represent the centroids. (a) The CNM reconstruction and (b) the dCNM reconstruction are performed with the same parameters as in figure 4.

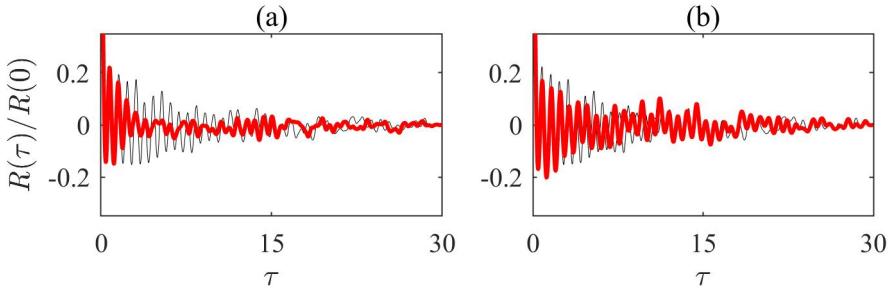


Figure 7: Auto-correlation function for $\tau \in [0, 30]$ of the Lorenz system. The thin black curves represent the original data set, and the thick red curves represent the models: (a) CNM and (b)dCNM.

are effectively resolved by the centroid orbits, and transitions between them are constrained by densely distributed centroids in the branching region, ensuring a smoothly varied oscillation.

The auto-correlation function is computed to reflect the model accuracy, as shown in figure 7. In the original data set, the normalised auto-correlation function $R(\tau)/R(0)$ vanishes smoothly as τ increases, and the variance between the periodic behaviour can be clearly observed. However, the CNM reconstruction captures only the first four periods of oscillation dynamics. As τ increases, there is a sudden amplitude decay accompanied by a phase mismatch. This can be attributed to amplitude jumps between the centroid loops and commonly occurring orbit-crossing transitions. In contrast, the dCNM reconstruction accurately captures both the amplitude and frequency of the oscillation dynamics, demonstrating robust and precise long-timescale behaviours.

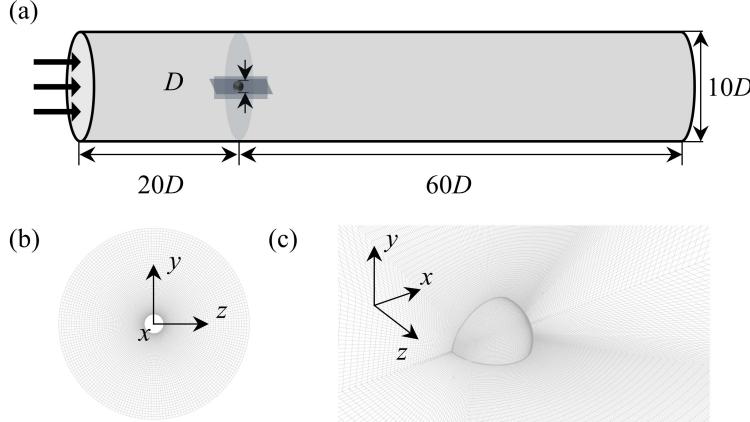


Figure 8: Numerical sketch of the sphere wake.

4. Dynamics-augmented modelling of the sphere wake

In this section, we demonstrate the dCNM for the transient and post-transient flow dynamics of the sphere wake. The numerical method for obtaining the flow field data set and the flow characteristics is presented in § 4.1. The performance of the dCNM for the periodic, quasi-periodic and chaotic flow regimes is evaluated in § 4.2, § 4.3 and § 4.4, respectively. The physical interpretation of the modelling strategy is discussed in § 4.5.

4.1. Numerical methods and flow features

Numerical simulation is performed to obtain the data set, as shown in figure 8. A sphere with a diameter D is placed in a uniform flow with a streamwise velocity U_∞ . The computational domain takes the form of a cylindrical tube, with its origin at the centre of the sphere and its axial direction along the streamwise direction (x -axis). The dimensions of the domain in the x , y , and z directions are $80D$, $10D$, and $10D$, respectively. The inlet is located $20D$ upstream from the sphere. These specific domain parameters are chosen to minimise any potential distortion arising from the outer boundary conditions while also mitigating computational costs (Pan *et al.* 2018; Lorite-Díez & Jiménez-González 2020). The fluid flow is governed by the incompressible Navier–Stokes equations:

$$\begin{aligned} \partial \mathbf{u} / \partial t + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p - \nabla^2 \mathbf{u} / Re &= 0, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \quad (4.1)$$

where \mathbf{u} denotes the velocity vector (u_x, u_y, u_z) , p is the static pressure, and Re is the Reynolds number, which is defined by:

$$Re = U_\infty D / \nu, \quad (4.2)$$

ν is the kinematic viscosity.

The net forces on the sphere have three components F_α , $\alpha = x, y, z$, and the corresponding force coefficients C_α are defined as:

$$C_\alpha = \frac{2F_\alpha}{\rho U_\infty^2 S}, \quad (4.3)$$

where $S = \pi D^2 / 4$ is the projected surface area of the sphere in the streamwise direction. The total drag force coefficient is $C_D = C_x$. Since the lift coefficient can have any direction in the

yz plane on the axisymmetric sphere, the total lift force coefficient C_L is given by:

$$C_L = \sqrt{C_y^2 + C_z^2}. \quad (4.4)$$

The flow parameters are non-dimensionalised based on the characteristic length D and the free-stream velocity U_∞ . This implies that the time unit scales are D/U_∞ , and the pressure scales are ρU_∞^2 , where ρ is the density. The Strouhal number St is correspondingly expressed as:

$$St = f, \quad (4.5)$$

where f is the characteristic frequency.

ANSYS Fluent 15.0 is used as the CFD solver for the governing equations with the cell-centred finite volume method (FVM). We impose a uniform streamwise velocity $\mathbf{u} = [U_\infty, 0, 0]$ at the inlet boundary and an outflow condition at the outlet boundary. The outflow condition is set as a Neumann condition for the velocity, $\partial_x \mathbf{u} = [0, 0, 0]$, and a Dirichlet condition for the pressure, $p_{\text{out}} = 0$. We apply a no-slip boundary condition on the sphere surface and a slip boundary condition on the cylindrical tube walls to prevent wake-wall interpolations. The pressure-implicit split-operator (PISO) algorithm is chosen for pressure-velocity coupling. For the governing equations, the second-order scheme is used for the spatial discretization, and the first-order implicit scheme is used for the temporal term. To satisfy the Courant–Friedrichs–Levy (CFL) condition, a small integration time step is set as $\Delta t = 0.01$ non-dimensional time unit, such that the Courant number is below 1 for all simulations. For the periodic flow at $Re = 300$, the simulation starts in the vicinity of the steady solution and runs for $t = 200$ time units, incorporating the transient and post-transient dynamics. For the quasi-periodic flow, the simulations are performed for $t = 500$ time units and for the chaotic flow for $t = 700$ time units. The snapshots are collected at a sampling time step of $\Delta t_s = 0.2$ time units for all the test cases. Moreover, we discard the first 200 time units to eliminate any transient phases for the quasi-periodic and chaotic cases. The relevant numerical investigation approach can be found in Johnson & Patel (1999); Rajamuni *et al.* (2018). For the convergence and validation studies, see Appendix A.

The wake of a sphere exhibits different flow regimes as Re increases, ultimately transitioning to a chaotic state. At $Re = 20 \sim 24$, flow separation occurs, forming a steady recirculating bubble, as observed in previous studies (Sheard *et al.* 2003; Eshbal *et al.* 2019). The length of this wake grows linearly with $\ln(Re)$. When Re surpasses 130 (Taneda 1956), the wake bubble starts oscillating in a wave-like manner, while the flow maintains axisymmetry. The first Hopf bifurcation takes place at approximately $Re \approx 212$ (Fabre *et al.* 2008), leading to a loss of axisymmetry and the emergence of a planar-symmetric double-thread wake with two stable and symmetric vortices. The orientation of the symmetry plane can vary (Johnson & Patel 1999). At a subsequent Hopf bifurcation around $Re = 270 \sim 272$ (Johnson & Patel 1999; Fabre *et al.* 2008), the flow becomes time-dependent, initiating periodic vortex shedding with the same symmetry plane as before. In the range $272 < Re < 420$ (Eshbal *et al.* 2019), periodicity and the symmetry plane diminish, with the vortex shedding becoming quasi-periodic and then fully three-dimensional. Beyond $Re = 420$, shedding becomes irregular and chaotic (Ormières & Provansal 1999; Eshbal *et al.* 2019; Pan *et al.* 2018), due to the azimuthal rotation of the separation point and lateral oscillations of the shedding.

In this study, we examine three baseline flow regimes of the sphere wake: periodic flow at $Re = 300$, quasi-periodic flow at $Re = 330$ and chaotic flow at $Re = 450$. Figure 9 illustrates the flow characteristics of these regimes. Figure 9 (a) and (b) depict the lift coefficient C_L of the periodic flow. The oscillation of C_L starts after a few time units, with the amplitude gradually increasing in the cycle-to-cycle evolution, eventually forming the

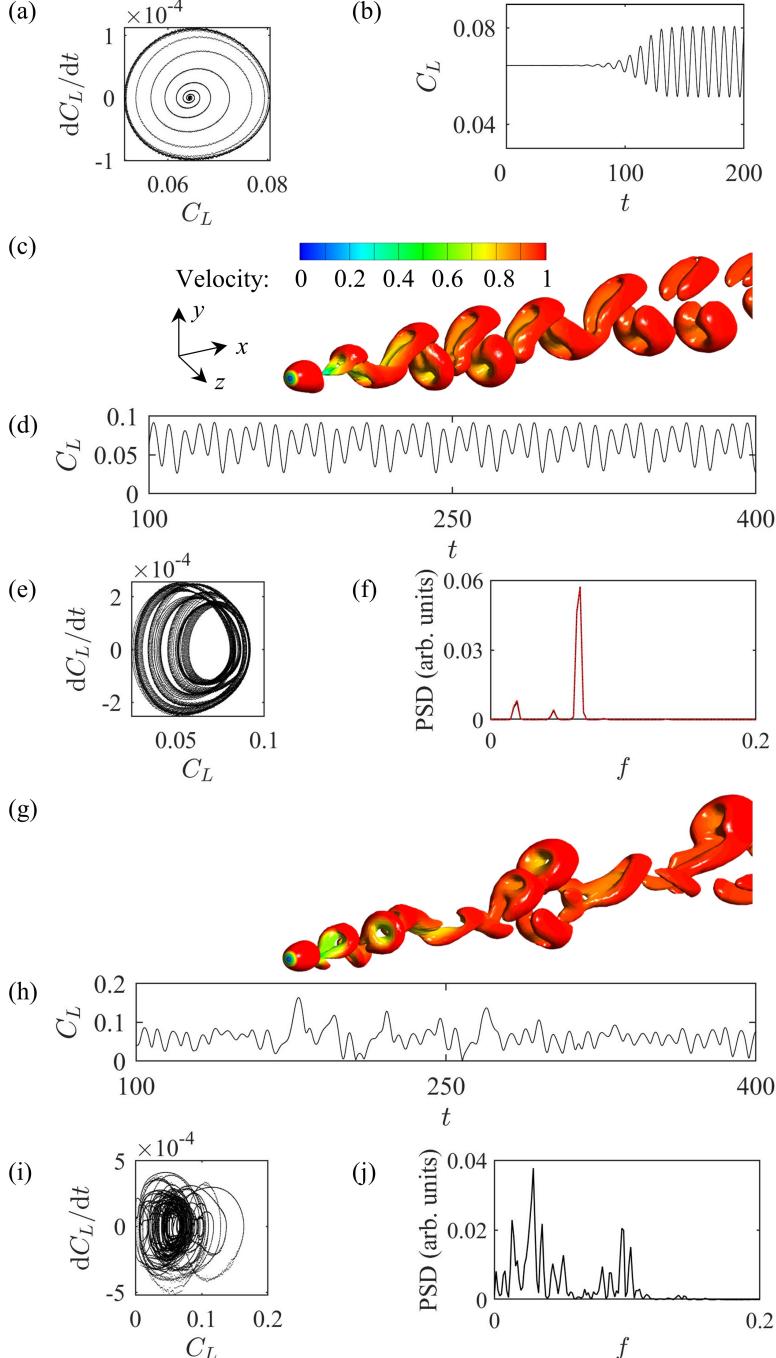


Figure 9: Flow characteristics of the sphere wake. The periodic flow at $Re = 300$ including the transient and post-transient dynamics is displayed by the (a) phase portrait of the lift coefficient C_L and (b) temporal evolution of C_L . The quasi-periodic flow at $Re = 330$ is displayed by the (c) vortex structures, where the vortexes are identified by the Q -criteria, and are colour-coded by the non-dimensional velocity U_∞ , (d) temporal evolution of C_L , (e) phase portrait of C_L and (f) power spectral density of C_L on time series of length $T_{\text{traj}} = 100$ (red curve) and $T_{\text{traj}} = 300$ (black curve). The chaotic flow at $Re = 450$ is displayed by the (g) vortex structures, (h) temporal evolution of C_L , (i) phase portrait of C_L and (j) power spectral density of C_L on a time series of length $T_{\text{traj}} = 500$.

limit cycle. Figure 9 (c) shows the instantaneous vortex structures of the quasi-periodic flow, identified by the Q -criterion and colour-coded by the non-dimensional velocity U_∞ . The vortex shedding forms hairpin vortices with slight variations between successive shedding events, signifying the absence of short-term periodicity while retaining long-term periodic behaviour. Figure 9 (d) and (e) show the temporal evolution of C_L and its phase portrait, respectively. The amplitude of C_L is strongly associated with the quasi-periodic dynamic, and the modulation also thickens the limit cycle of the oscillator on the phase portrait. The power spectral density in figure 9 (f) indicates two dominant frequencies: a higher frequency linked to natural shedding and a lower frequency associated with amplitude modulation resulting from variations between shedding events. For chaotic flow, periodicity entirely vanishes, and the flow regime displays the typical features of a chaotic system. The hairpin vortexes in figure 9 (g) shed irregularly, with varying separation angles and even double spirals. The temporal evolution of C_L in figure 9 (h) exhibits more complex dynamics, with the phase diagram in figure 9 (i) depicting many random loops that no longer exhibit circular patterns. Furthermore, the power spectral density of C_L in figure 9 (j) shows a broad peak, also indicating chaotic features.

We performed a *lossless* POD preprocessing on the snapshots to reduce the computational cost of clustering the three-dimensional flow field data set, as described in Appendix B. This preprocessing is optional and does not affect the distance measure in the clustering algorithm. For consistency, the notation snapshot is maintained in the following for the preprocessed data.

4.2. The periodic flow regime at $Re = 300$

We compare the dCNM to the CNM for the periodic flow regime of the sphere wake at $Re = 300$. The transient and post-transient dynamics are considered, providing insights into the mechanisms for the instability and nonlinear saturation.

The comparison between the CNM and the dCNM clustering for the periodic flow is presented in figure 10. In the dCNM, we set $K = 10$ for state space clustering and $\beta = 0.50$ for the sub-clustering. The value of β characterises the trade-off between a small number of sub-clusters and the model accuracy. The normalised transverse cluster size vector $\hat{\mathbf{R}}^\top = [0.1510, 0.1101, 0.0855, 0.1120, 0.1069, 0.1031, 0.0650, 0.0934, 0.0883, 0.0847]^\top$ corresponds to the number of sub-clusters $\mathbf{L} = [3, 3, 3, 4, 4, 4, 3, 3, 3, 3]^\top$. Classical multidimensional scaling (MDS) is applied to project the high-dimensional snapshots and centroids into a three-dimensional subspace $[a_1, a_2, a_3]^\top$ for visualisation. The snapshots form a conical surface in the three-dimensional subspace, where the trajectory spirals up from a fixed point to a periodic motion. This behaviour is indicative of a Hopf bifurcation, which involves an unstable steady solution and nonlinear saturation to a periodic limit cycle. Most of the CNM centroids are located on the limit cycle, and only a few resolve the transient phase. In contrast, the dCNM centroids offer a finer resolution of the amplitude growth.

The original and reconstructed trajectories of the CNM and dCNM for the periodic flow regime are shown in figure 11. The CNM fails to resolve the transient dynamics and only captures the stable limit cycle. In contrast, the dCNM can effectively resolve both the cyclic behaviour and the growing oscillation amplitude. Given the deterministic nature of these transient and post-transient dynamics, the centroid transition should yield a permutation matrix — each column and each row of the matrix should only have one element being unity. However, in the CNM, under-resolved transient dynamics result in a stochastic transition matrix. The transition uncertainty expressed within a column of the matrix often leads to prediction errors. Examples are the unphysical jumps between erratic cycles in the transient stage, as displayed in Figure 10 (a). The sub-clusters in dCNM are designed to capture a

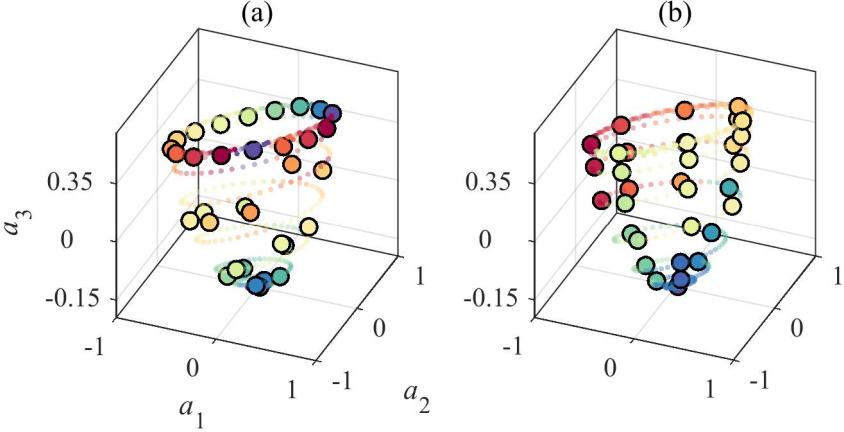


Figure 10: Three-dimensional visualisation of the clustered periodic flow regime of the sphere wake at $Re = 300$. Classical multidimensional scaling (MDS) is applied to the data set to visualise the high-dimensional snapshots and centroids in the subspace. The small dots represent the snapshots, and the large dots represent the centroids. Snapshots and centroids with the same colour belong to the same cluster. For comparison, the CNM result in (a) is shown with the same number of centroids as the corresponding dCNM result. The dCNM result in (b) is shown with $K = 10$ and $\beta = 0.50$.

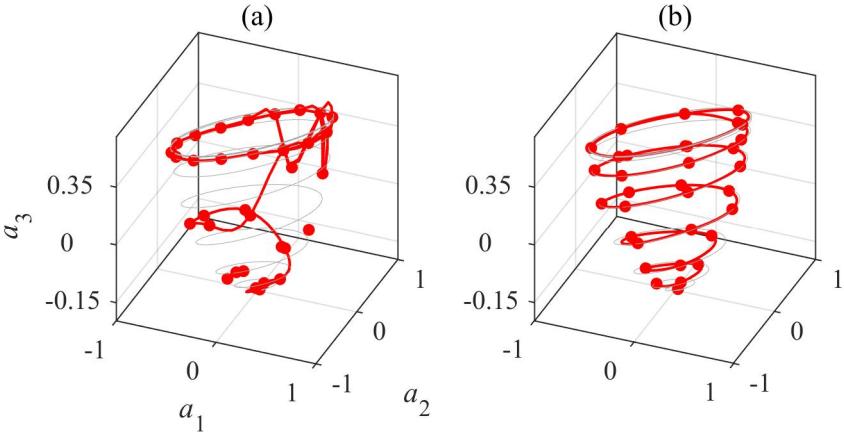


Figure 11: Trajectory of the periodic flow at $Re = 300$. The thin grey curve represents the original trajectory, the thick red curve represents the reconstructed trajectory, and the red dots represent the centroids. (a) The CNM reconstruction and (b) the dCNM reconstruction are obtained with the same parameters as in figure 10.

one-way forward transition from the starting point to the limit cycle, ensuring an accurate reconstruction of the deterministic dynamics.

The other extreme is the maximum transition uncertainty, which can be represented by the least informative transition matrix — a perfectly mixing matrix with elements $Q_{ik} \equiv 1/K$. The information entropy (Shannon 1948)

$$S(Q) = - \sum_{i=1}^K \sum_{k=1}^K Q_{ik} \ln Q_{ik} \quad (4.6)$$

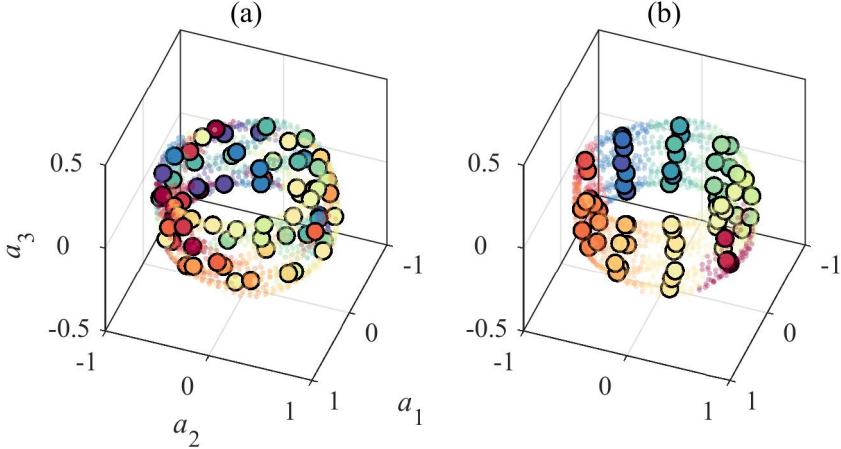


Figure 12: Same as figure 10, but for the quasi-periodic flow at $Re = 330$. (a) The CNM result with the same number of centroids as the dCNM. (b) The dCNM result with $K = 10$ and $\beta = 0.80$.

of the permutation matrix vanishes. In contrast, the maximum information entropy $K \ln K$ is obtained from the perfectly mixing transition matrix with equal elements $Q_{ik} \equiv 1/K$. Here, the knowledge of the current state has no predictive value for the future population. For the current case, the reference model has an entropy $S_{\text{CNM}} = 7.0586$, which is much smaller than the upper bound $33 \ln 33 \approx 115.38$. The proposed model minimizes entropy, $S_{\text{dCNM}} = 0$. Thus our novel clustering measurably increases the prediction accuracy.

4.3. The quasi-periodic flow regime at $Re = 330$

The clustered quasi-periodic flow of the CNM and dCNM are shown in figure 12. We set $K = 10$ for the state space clustering and $\beta = 0.80$ for the subsequent clustering, which proves adequate for capturing the quasi-periodic dynamics and ensuring clarity in visualisation. The choice of β and other results with different values of β are discussed in Appendix C. In this case, the normalized transverse cluster size vector $\hat{\mathbf{R}}^T = [0.1043, 0.1046, 0.1056, 0.1068, 0.0889, 0.1073, 0.1061, 0.1050, 0.1047, 0.0668]^T$, corresponds to the number of sub-clusters $\mathbf{L} = [7, 7, 7, 7, 6, 7, 7, 7, 7, 5]^T$. In the three-dimensional subspace, the snapshots collectively form a hollow cylinder. The system's dynamics are chiefly governed by two underlying physical phenomena: a cyclic behaviour that synchronises with natural vortex shedding and a quasi-stochastic component responsible for introducing variations between cycles, which is, in turn, synchronised with the oscillator amplitude.

The centroid distribution of the CNM reveals that the clustering algorithm fails to distinguish between the shedding dynamics and inter-cycle variations. It uniformly groups them based solely on spatial topology. Nonetheless, the CNM centroids effectively capture the cyclic behaviour, as there exist deterministic transitions between adjacent centroids within an orbit, forming a limit cycle structure akin to the “ear” of the Lorenz system. However, this centroid distribution inadequately models the quasi-stochastic component, as it overlooks the inter-cycle transitions. To comprehensively represent this dynamic, clear transitions between the limit cycles are essential. The clustering process obscures these transitions, causing the quasi-stochastic behaviour to resemble a random walk governed by a fully stochastic process. In essence, the clustering process cannot differentiate between the

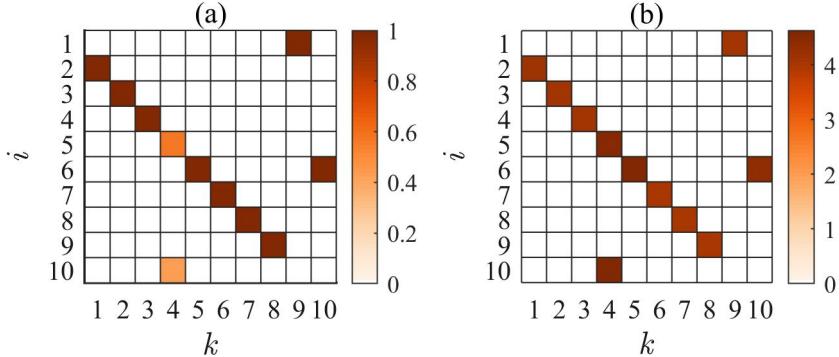


Figure 13: Same as figure 5, but for the quasi-periodic flow at $Re = 330$. (a) Transition probability matrix \mathbf{Q} . (b) Transition time matrix \mathbf{T} .

random jumps in the Lorenz system and the quasi-stochastic behaviour in this flow regime. This explains why the CNM often struggles with multifrequency problems. In contrast, the dCNM centroids automatically align along the axial direction of the cylinder with equidistant circumferential spacing, resulting in a greater number of centroid orbits compared to the CNM. This enhancement enables the accurate resolution of inter-cycle variations. For the quasi-stochastic behaviour, the denser and occasionally overlapping centroids in the axial direction ensure precise spatial representation of the transitions between the limit cycles. Additionally, this behaviour can be further constrained by the dual indexing approach for long-timescale periodicity, eliminating random jumps and ensuring accurate transitions between limit cycles.

The cluster transition matrices of the quasi-periodic flow regime are illustrated in figure 13. The quasi-periodic dynamics are evident from \mathbf{Q} , which displays dominant transition probabilities corresponding to deterministic cyclic behaviour and minor wandering transitions signifying inter-cycle variations. Cluster C_4 serves as the transition cluster with two destination clusters, C_5 and C_{10} . The two destination clusters have similar transition probabilities since they are visited for comparable times during the quasi-periodic transitions. The transition cluster bridges the deterministic cluster chains $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4$ and $C_6 \rightarrow C_7 \rightarrow C_9 \rightarrow C_1$ as two different limit cycles through two short cluster chains: $C_4 \rightarrow C_5 \rightarrow C_6$ and $C_4 \rightarrow C_{10} \rightarrow C_6$. These two limit cycles alternate with a fixed order, ultimately forming an extended cluster chain that constitutes the fundamental elements of the long-term periodicity. However, this characteristic is not effectively portrayed in the transition matrix. The purely probabilistic transitions from this matrix can result in arbitrary cluster transitions within the network model, introducing additional transition errors. Since the CNM relies on this cluster-level matrix, these transition errors present a notable challenge. While the transition tensors Q , which resolve the refined centroid transitions, mitigate this issue, we further discuss the transition tensors and the corresponding centroid transition matrices in Appendix D. The time matrix \mathbf{T} reveals that the transitions within a cyclic behaviour possess a generally similar time scale, with residence times in adjacent clusters changing smoothly, indicating the presence of a gradually evolving limit cycle.

The original and reconstructed trajectories using the CNM and dCNM for the quasi-periodic flow regime are displayed in figure 14. The reconstruction is achieved with the same parameters as in figure 12. As anticipated, the trajectory reconstructed by the CNM undergoes substantial deformation, featuring discontinuous cyclic behaviours and a serrated trajectory. Conversely, the dCNM produces cleaner cyclic behaviours with more noticeable

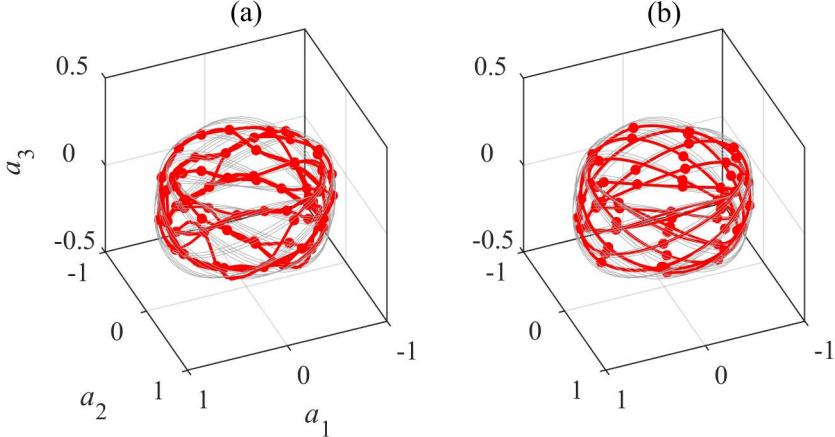


Figure 14: Same as figure 11, but for the quasi-periodic flow at $Re = 330$. (a) The CNM reconstruction and (b) the dCNM reconstruction are obtained with the same parameters as in figure 12.

variations. The reconstructed trajectory accurately replicates the intersecting limit cycles and guides the inter-cycle transition with reduced spatial errors. These observations highlight the ability of the dCNM centroids to capture significant dynamics without assuming any prior knowledge of the data set. A kinematic comparison between the POD reconstruction and the dCNM reconstruction is presented in Appendix E.

In the following sections, we shift our focus to the temporal aspects. The CNM uses the transition matrices to predict the next destination state for each step. The quasi-periodic feature will be obscured by the stochastic transition probability matrix and the missing historical information. In contrast, the dCNM preserves the transition sequence by embedding the sub-clusters (see Appendix D). Initially, we explore the cluster and trajectory segment affiliation for each snapshot in both the original data set and the dCNM reconstruction to illustrate the accuracy of transition dynamics, as depicted in figure 15. We maintain the same parameters as those used in figure 12 for the reconstruction. The affiliation of the original data reveals that the dual clustering effectively represents the quasi-periodic dynamics. The transition dynamics exhibit significant regularity, with centroids being sequentially and periodically visited, confirming deterministic transitions. Each period of centroid visits corresponds to an extended cluster chain, encompassing multiple centroid orbits and capturing cycle-to-cycle variations. The periodic visits of these extended cluster chains are instrumental in determining the long-timescale periodicity. These transition characteristics are fully preserved by the dCNM due to the dual indexing constraint. In this case, each departure centroid corresponds to only one destination centroid, eliminating the stochastic transition in the model and mitigating the transition errors.

Envelope demodulation can clearly reveal the long-timescale behaviour and is more efficient in reflecting the quasi-periodic dynamics. We analyse the envelope spectrum of the streamwise fluctuation velocity u'_x from the data set, CNM, high-order CNM, and dCNM, as depicted in figure 16. The spectrum of the data set exhibits a dominant frequency $f = 0.05$, representing long-timescale periodicity. However, the CNM spectrum shows significant noise and lacks a clear dominant frequency due to frequent transition errors. This observation supports the CNM's limitation in capturing multifrequency dynamics effectively. By incorporating the historical information, the high-order CNM demonstrates superior performance by producing a cleaner spectrum closely aligned with the CFD data's dominant

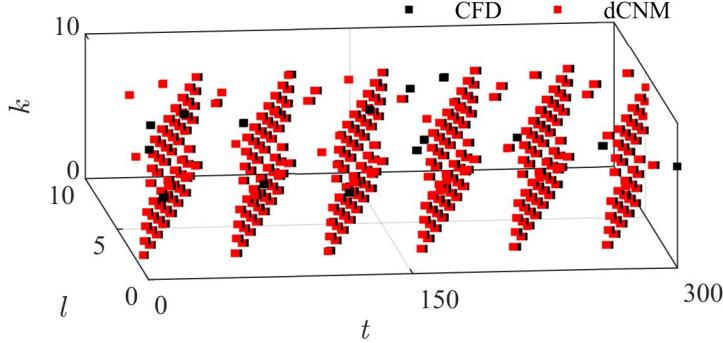


Figure 15: Transition illustrated with the temporal evolution of the cluster and trajectory segment affiliation of the quasi-periodic flow at $Re = 330$. The vertical direction represents the cluster-level transition and the horizontal direction represents the trajectory segments inside this cluster. The transition with black markers represents the CFD data, and the transition with red markers represents the reconstructed dynamics by the dCNM. The x axis is the non-dimensionalized time t , the y axis is the trajectory segment affiliation l , and the z axis is the cluster affiliation k . The reconstruction is achieved from the same parameters as in figure 12.

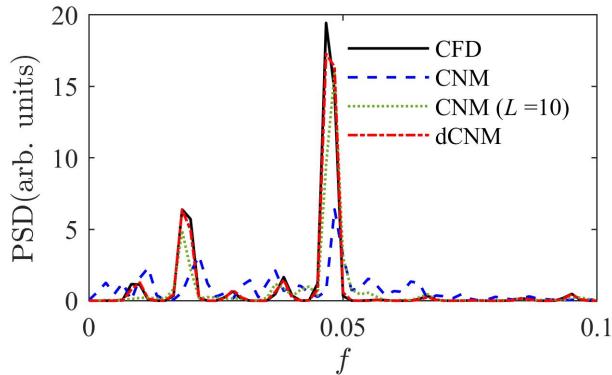


Figure 16: The envelope spectrum of the streamwise fluctuation velocity u'_x , obtained by the surface average in $x = 5D$. The dCNM reconstruction is achieved with the same parameters as in figure 12, and the CNM reconstruction is achieved with the same number of centroids as the dCNM

frequency. Remarkably, the dCNM outperforms all other models by precisely reconstructing both the frequency and amplitude while minimising noise.

4.4. The chaotic flow regime at $Re = 450$

The comparison between the CNM and dCNM clustering with $K = 10$ and $\beta = 0.40$ for the chaotic flow are illustrated in figure 17. This value of β is the sweet point between model complexity and model accuracy for this test case. Discussions with different values of β for this flow regime are presented in Appendix C. The normalized transverse cluster size vector $\hat{\mathbf{R}}^T = [0.1068, 0.1073, 0.1063, 0.1068, 0.1061, 0.0966, 0.1001, 0.0954, 0.0984, 0.0763]^\top$, corresponds to the number of sub-clusters $\mathbf{L} = [22, 22, 22, 22, 22, 20, 21, 20, 20, 8]^\top$. As the dynamics become more complex, the snapshots form a chaotic cloud, which is driven by numerous cyclic behaviours of different scales and indicates irregular three-dimensional

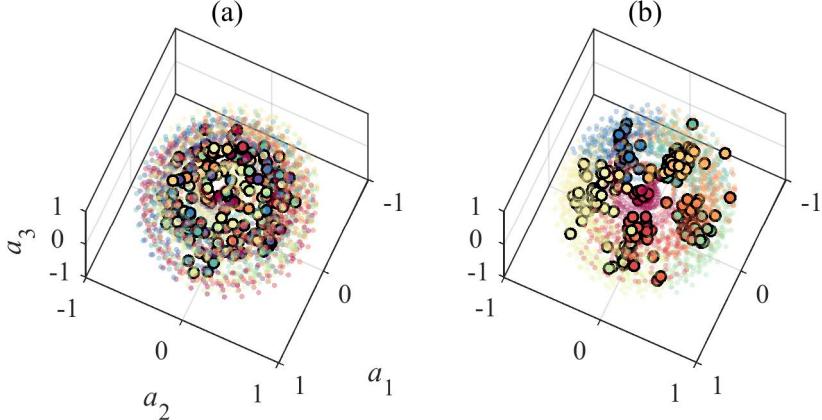


Figure 17: Same as figure 10, but for the chaotic flow of the sphere wake at $Re = 450$. (a) The CNM result with the same number of centroids as the dCNM. (b) The dCNM result with $K = 10$ and $\beta = 0.40$.

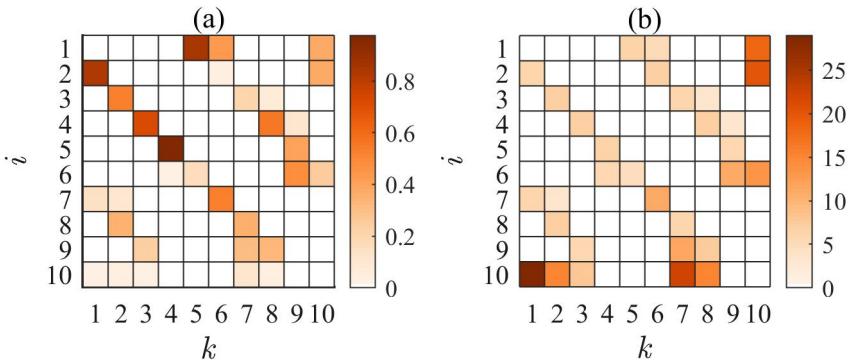


Figure 18: Same as figure 5, but for the chaotic flow at $Re = 450$. (a) Transition probability matrix \mathbf{Q} . (b) Transition time matrix \mathbf{T} .

vortex shedding. The CNM continues to cluster the data set primarily based on spatial properties, essentially dividing the chaotic cloud into different segments in an evenly distributed manner. Figure 17 (a) illustrates this process, with the uniformly spread centroids capturing only part of one whole cyclic behaviour, limiting their ability to resolve the multiscale dynamics. The dCNM centroids concentrate in regions of rich dynamics, enabling a more comprehensive resolution of the cyclic behaviours. These centroids, in various combinations, form the basis of multi-frequency and multiscale cyclic behaviour. Even after sparsification, the dCNM centroids can encompass a significant amount of scale diversity by merging only those that are spatially close to each other.

The cluster transition matrices of the chaotic flow are illustrated in figure 18. The probability matrix \mathbf{Q} in figure 18 (a) shows that most of the clusters have three or more destination clusters, indicating complex transition dynamics among them. Several dominant transition loops are identifiable, such as the large-size cluster chain: $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_6 \rightarrow C_1$, the mid-size cluster chains: $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_1$ and $C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_6 \rightarrow C_7 \rightarrow C_3$, and the small-size cluster chain: $C_6 \rightarrow C_7 \rightarrow C_8 \rightarrow C_6$.

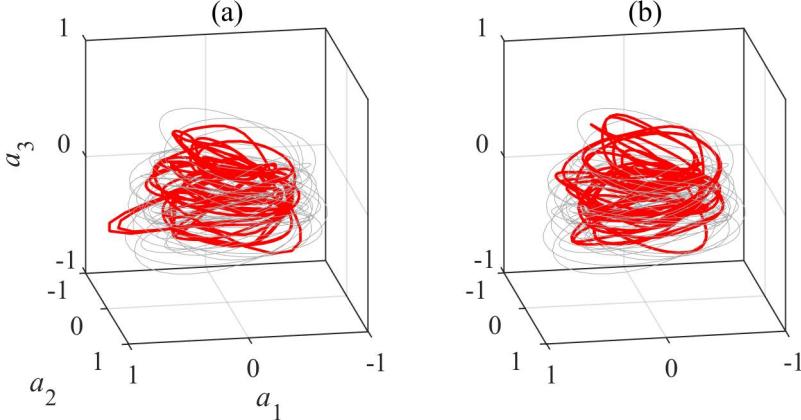


Figure 19: Same as figure 11, but for the chaotic flow at $Re = 450$. (a) The CNM reconstruction and (b) the dCNM reconstruction are obtained from the same parameters as in figure 17.

These cluster chains with different lengths represent cyclic loops at different scales. The small number of chains facilitates human understanding of the transition dynamics but is insufficient for accurately capturing the dynamics. The dominant loops have their key transition clusters inside, from which they can randomly jump into each other by choosing periodic or stochastic routes. This is where the transition error often occurs. The time matrix \mathbf{T} in figure 18 (b) shows the difference in the transition times between different types of transitions. Regarding the main loops, the time scale changes smoothly within its transitions. However, for the jumps between the loops, the time scale fluctuates considerably, and some transitions can be very large, showing the diversity of the dynamics. Moreover, this observation implies that the distribution density of snapshots differs among clusters. In other words, the distribution of the trajectory segments in different clusters also exhibits significant variations. This explains the necessity for determining the number of sub-clusters in the second-stage clustering based on the deviation \mathbf{R}^T . The refined transition matrices between the centroids are shown in Appendix D.

The original and reconstructed trajectories by the CNM and dCNM for this flow regime are shown in figure 19. The reconstruction is achieved with the same parameters as in figure 17. For a clear visualisation, only the trajectories from the first half of the entire time window are plotted. This selection suffices to analyse the precision of the current trajectory, as it contains ample dynamics. We exclude trajectory discrepancies triggered by phase mismatch and focus exclusively on the accuracy of the present trajectory. In the case of the CNM, noticeable disparities exist between the original trajectory and the reconstructed trajectory. These differences include variations in the shape, spatial location, and inclination angle of the cyclic loops. These disparities can be attributed to the elimination of small-scale structures and the blending of certain large-scale structures due to the uniform distribution of centroids. Regarding the dCNM, the reconstructed trajectory nearly occupies the entire chaotic cloud, closely resembling the original trajectory. The external and internal geometries are accurately reproduced, capturing both large-scale and small-scale structures. However, despite the improved accuracy, some deformations persist. These deformations arise from the interpolations between the limited centroids during one single cyclic loop. Notably, due to its complexity, achieving a superior reconstruction of a chaotic system often requires more

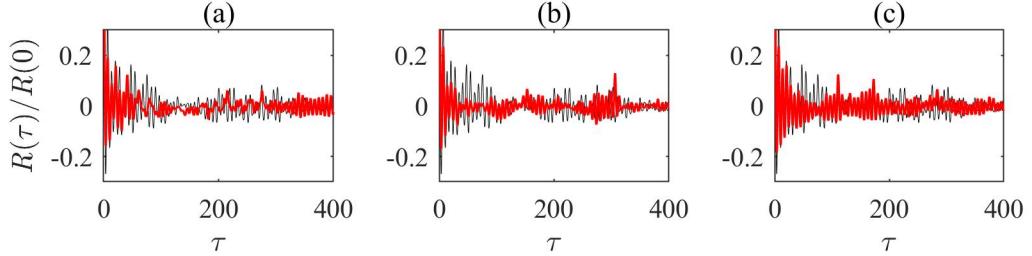


Figure 20: Auto-correlation function of the chaotic flow at $Re = 450$; here R is normalised by $R(0)$. The thin black curve represents the CFD data, and the thick red curve represents the reconstruction from different models. (a) The CNM reconstruction with the same number of centroids as the dCNM. (b) The high-order CNM reconstruction with $L = 10$ and the same number of centroids as the dCNM. (c) The dCNM reconstruction achieved with the same parameters as in figure 17.

refined centroids compared to a quasi-periodic system. The kinematic comparison with the POD reconstruction is also introduced in Appendix E.

Figure 20 shows the auto-correlation function of the CNM, high-order CNM, and dCNM. We still normalise this function by $R(0)$, and the time window is chosen from $t = 0$ to $t = 400$, which is sufficient for comparison. For the chaotic flow regime, $R(\tau)/R(0)$ denotes the kinetic energy level of the time window. Nonetheless, a notable discrepancy arises in the CNM, where the amplitude experiences a distinct decay after the initial few periods. It eventually stabilises with minimal variation, primarily due to the distorted reconstructed trajectory and transition errors. This limited variance is indicative of inaccuracies in capturing short-term dynamics, consistent with the absence of historical information. The high-order CNM, which incorporates this historical information, outperforms the CNM in this regard. Its amplitude decays gradually and exhibits variance akin to that of the data set. Additionally, it reveals some peaks with similar time delays, due to the potential introduction of unnecessary long-timescale periodicity into the reconstruction via the high-order cluster chain. The dCNM also surpasses the CNM with regard to accuracy. Both the amplitude and phase are faithfully retained, with a gradual amplitude decay and more pronounced variation. Eventually, the amplitude diminishes, similar to the original data set. As τ increases, all three models exhibit some degree of phase delay or lead. This is a consequence of averaged transition times introducing some errors into the model (Li *et al.* 2021).

4.5. Physical interpretation

One of the major advances of the cluster-based model is its strong physical interpretability. The dCNM also maintains and even enhances this nature while improving the model accuracy. In this section, we discuss the physical interpretation of the CROM exemplified for the sphere wake, with particular emphasis on the dCNM.

The cluster-based model spatially coarse-grains the snapshots into groups and represents them by centroids to reduce dimensionality. In contrast to the projection-based methodology, such as the POD-Galerkin model, the cluster-based model uses cluster centroids which are linear combinations of several snapshots, and thus reflects the representative patterns. This feature contributes to its high interpretability. The snapshot dynamics are mapped into the pattern dynamics, followed by the construction of a probabilistic mechanism to reduce temporal dimensionality. The network model, with centroids as nodes and centroid transitions as edges, converts the complex dynamics into pure data analysis. The centroids act as a bridge

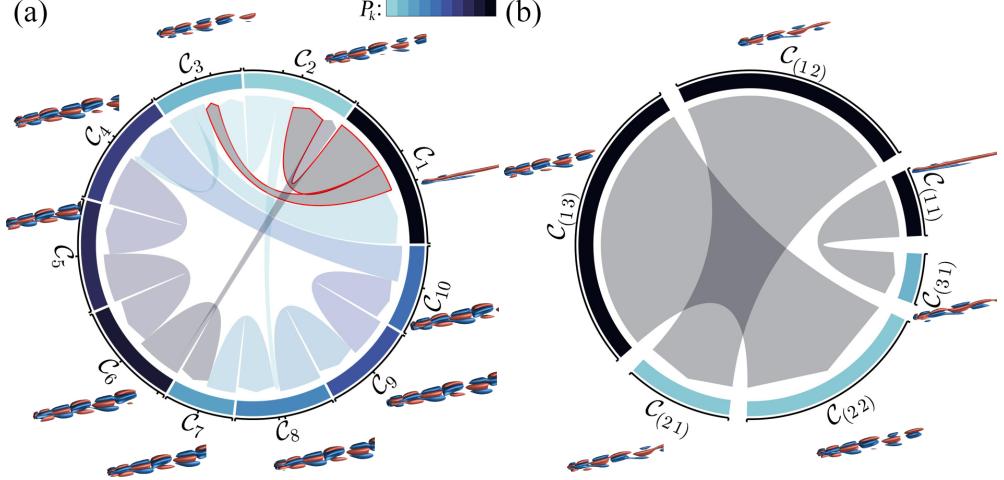


Figure 21: Transition diagram of the quasi-periodic flow at $Re = 300$. The centroids are depicted by the vortex distribution. The vortices are identified by the iso-surfaces of z -vorticity, with -1 for the negative vortices coloured in blue and 1 for the positive vortices coloured in red. The transition dynamics are depicted by the directed arrows, the size of the arrow tail represents the transition probability and the colour is consistent with the departure block. (a) Cluster transitions. Different blocks represent different clusters, the colour of the block represents the corresponding cluster probability distribution P_k , and the size of the block represents the cluster size R^u . (b) Sub-cluster transitions of $\beta = 0.50$, with transitions specifically departing from C_1 , corresponding to the red-bordered cluster transition in (a). Blocks with the same colour belong to the same cluster, the colour still represents the cluster probability distribution P_k , and the size of each block represents the sub-cluster size R_{sub}^u .

between the data-driven model and its underlying physical background. Furthermore, it is conceivable that the same model can be easily transferred to analogous pattern dynamics, even with distinct backgrounds, through adjustments to the centroids' backstory.

The sphere wake offers a concise physical interpretation based on the centroids. The coherent structure evolution governs the flow field and manifests as vortex shedding events with diverse dynamics. These shedding events can be captured well by a limit cycle, with a set of centroids representing flow patterns at different shedding phase as foundational elements. The cyclic transitions between these specific flow patterns collectively characterise the entire shedding process. The deterministic-stochastic transitions between different shedding events contribute to the overall periodic-chaotic dynamics.

To explain the physical mechanisms of the flow regimes, we propose a chord transition diagram for the cluster transitions and sub-cluster transitions along with centroid visualisation, which provides a comprehensive view of the flow regime. We start with the periodic flow, as shown in figure 21. The cluster probability distribution P_k and the cluster size R^u used for visualising the blocks in figure 21 (a) are shown in figure 22 (a) and (b). The blocks in figure 21 (b) are split based on the sub-clusters, the transverse cluster size is shown in figure 22 (c) and the sub-cluster size is shown in figure 22 (d). The cluster transition diagram is capable of clearly distinguishing the dynamic behaviour categories. The circumferential arrows along the boundary represent the cyclic behaviours, with centroids transitioning to adjacent destination centroids. The radial arrows crossing the graph signify cycle-to-cycle transitions, with the centroids transitioning to non-adjacent destinations. These arrows usually

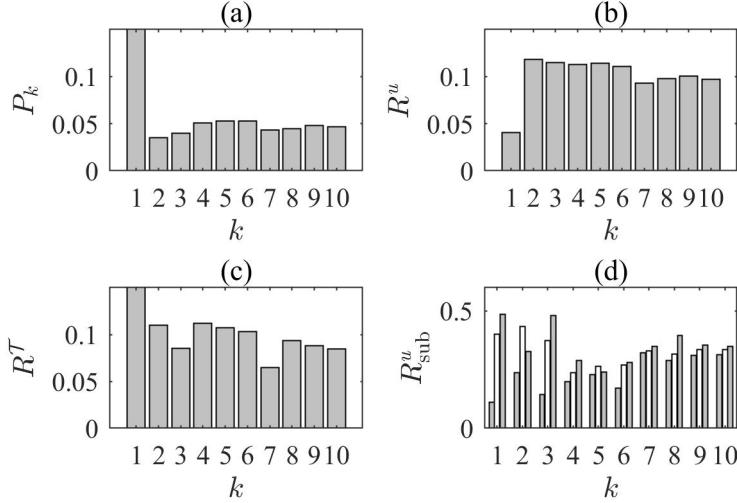


Figure 22: Cluster and centroid properties of the periodic flow at $Re = 300$. (a) Cluster probability distribution. (b) Normalised cluster size. (c) Normalised transverse cluster size. (d) Normalised sub-cluster size, where the elements from the same cluster sum to unity.

originate from the transition clusters. The number of radial arrows indicates the dynamic characteristics, with more arrows indicating more chaotic features.

The cluster dynamics is illustrated in figure 21 (a). The limit cycle is captured by the clusters C_4 to C_{10} , shown as the deterministic transitions between adjacent clusters. The transient phase is resolved by clusters from C_1 to C_3 . However, the stochastic transitions between the first three clusters are in contrast to the slowly varying amplitude in the transient state and are insufficient to represent this deterministic process. The distinct vortex structures of the three centroids also indicate a need for higher resolution. The sub-cluster transitions leaving from C_1 are shown in figure 21 (b). The sub-cluster centroids manifest as varying vortex structures, corresponding to the growing amplitudes. Each sub-cluster has only one destination, resulting in deterministic transitions. The stochastic cluster transitions from C_1 to C_2 and C_3 are now terminated into a chain of deterministic sub-cluster transitions, effectively reducing the prediction error.

The transition graph of the quasi-periodic flow regime is shown in figure 23, with the corresponding cluster and centroid properties given in figure 24. In figure 23 (a), the flow regime is characterised by the cyclic cluster transitions, with only three non-adjacent transitions, i.e., C_4 to C_{10} , C_{10} to C_6 , and C_9 to C_1 . The clusters involved in the cyclic transitions exhibit varying vortex structures within one shedding period. Their relatively higher probability distribution, as shown in figure 24 (a), suggests dominant flow patterns. For the bifurcating cluster C_4 , its destination clusters C_5 and C_{10} manifest visible differences in the far wake. Further distinction of transitions from C_4 is provided by the sub-clusters, as shown in figure 23 (b). The centroids belonging to the same cluster are roughly at the same shedding phase, but exhibit different vortex structures, exemplified by C_{41} and C_{42} . This difference leads to distinct shadings, such as C_{101} , and C_{51} . Consequently, finer dynamic resolution originating from C_4 is captured, enabling the deterministic transitions to C_5 and C_{10} , respectively.

The chaotic flow regime exhibits a more complex transition graph, as shown in figure 25. The relative information is illustrated in figure 26. In figure 25 (a), similar to other flow regimes, adjacent cluster transitions continue to dominate the flow field, reflecting

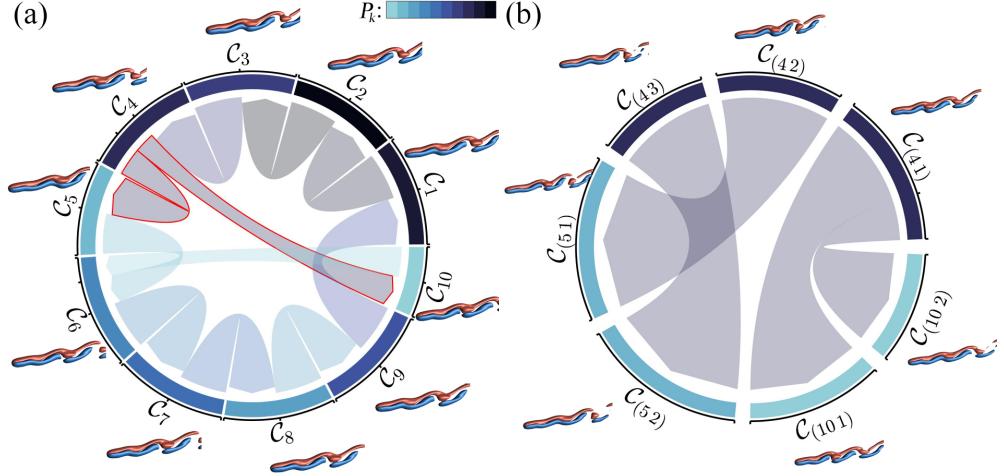


Figure 23: Same as figure 21, but for the quasi-periodic flow at $Re = 330$. (a) Cluster transitions. (b) Sub-cluster transitions of $\beta = 0.95$, with transitions specifically departing from C_4 , corresponding to the marked cluster transition in (a).

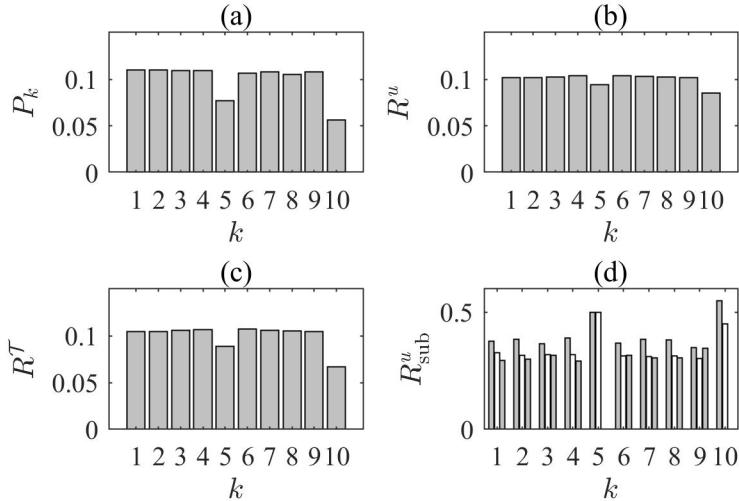


Figure 24: Same as figure 22, but for the quasi-periodic flow at $Re = 330$. (a) Cluster probability distribution. (b) Normalised cluster size. (c) Normalised transverse cluster size. (d) Normalised sub-cluster size, where the elements from the same cluster sum to unity.

cyclic behaviours. However, the increasing number of radial arrows with varying transition probabilities indicates chaotic features. Each centroid represents a distinct flow field with different scales of vortex structures, even within the same cyclic cluster transition. This discrepancy indicates that the current flow patterns are inadequate for capturing the entire shedding dynamic. Concerning the sub-cluster transitions in figure 25 (b), the increased arrows maintain the transition rhythm but offer more specificity. The flow states can be inferred from the vortex structures surrounding the cyclic diagram. The centroids within the same cluster also represent vortex structures sharing the same shedding phase but exhibiting different scales. Only centroids with similar scales of vortex structures are connected by

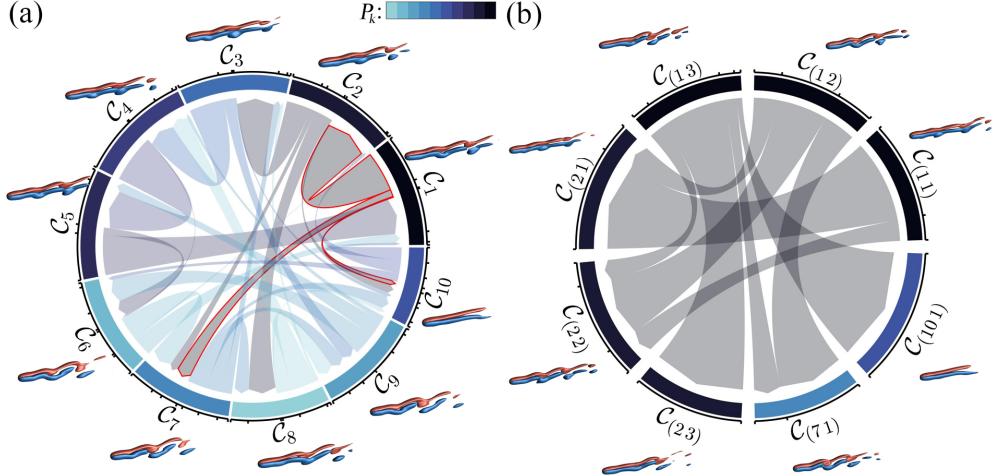


Figure 25: Same as figure 21, but for the chaotic flow at $Re = 450$. (a) Cluster transitions. (b) Sub-cluster transitions of $\beta = 0.95$, with transitions specifically departing from C_1 .

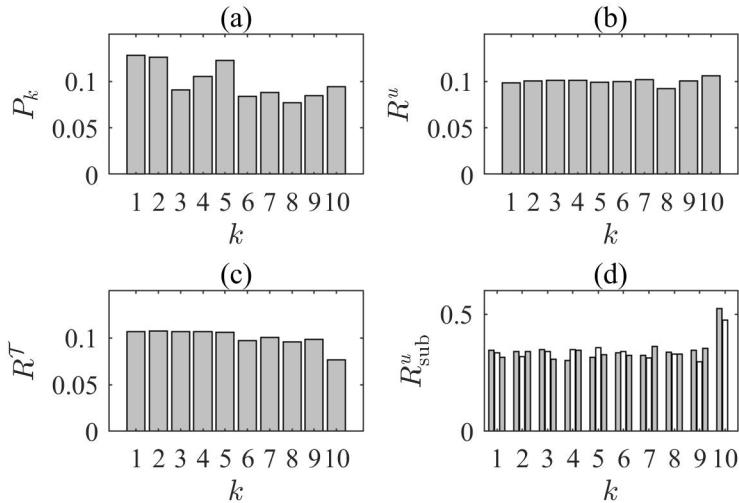


Figure 26: Same as figure 22, but for the chaotic flow at $Re = 450$. (a) Cluster probability distribution. (b) Normalised cluster size. (c) Normalised transverse cluster size. (d) Normalised sub-cluster size, where the elements from the same cluster sum to unity.

sub-cluster transitions. The departing sub-clusters are thus restricted, for instance, $C_{(23)}$ and $C_{(101)}$ each have only one departing sub-cluster. The diversity of the centroid transitions guarantees diverse flow scales, while simultaneously maintaining a consistent scale within the same cluster loop. Therefore, the dCNM facilitates multi-scale fidelity and smoother cyclic behaviours, significantly enhancing the representation capacity of the model.

When comparing the cluster transition and the sub-cluster transition in dCNM, it is evident that the state space clustering can be seen to automatically introduce prior knowledge into the model. This prior knowledge includes the inner-state kinematic information, as resolved by the trajectory segments, and the inter-state dynamic information, as resolved by the transitions between the trajectory segments. The incorporation aids in the automatic assignment of

refined centroids within each cluster and constrains the probabilistic transition dynamics. In essence, the dCNM can be regarded as having a built-in unsupervised physics-informing process, which results in superior model accuracy.

5. Conclusions and outlook

We propose an automatable data-driven reduced-order model for nonlinear dynamics. This model can resolve the periodic, quasi-periodic and chaotic dynamics of the sphere wake featuring multi-frequency and multiscale behaviours. The starting point is the cluster-based network model (CNM) (Fernex *et al.* 2021; Li *et al.* 2021), which is an automated framework employing clustering and network science. The dynamics within the CNM are described using a deterministic-stochastic approach on a network, where centroids act as nodes, and transitions serve as edges. However, the clustering process in the CNM relies on a uniform geometric coverage of the snapshot data, agnostic of the temporal dynamic relevance. For multi-frequency dynamics, this can result in large prediction errors. One example is the long transient to a limit cycle. Here, the slow increase in the radius requires a finer resolution than the robust angular dynamics. Hence, the CNM can be expected to be more accurate if the centroids are much denser in the radial direction than in the angular motion. This idea is incorporated in the proposed dynamics-augmented CNM (dCNM). The model can automatically stratify the state space along the trajectory direction.

The dCNM was applied to the Lorenz system (in § 3) and the three-dimensional sphere wake (in § 4), with $K = 10$ clusters for the coarse-graining of the state space. The Lorenz system features oscillatory dynamics, presented as two “ears” consisting of many unstable orbits, and stochastic dynamics, presented as random switching between the “ears”. For the future state, the phase can be accurately predicted, but the amplitude requires a higher resolution. The CNM is only capable of reconstructing limited loops of the cyclic behaviours and their related transitions in the branching area. Non-physical radial jumps also occur due to transition errors. On the other hand, the dCNM coarsely resolves the deterministic phases but accurately resolves the slowly varying amplitude. The attractor oscillations are distinctly defined, and the transitions in the branching region are subsequently constrained. For the transient and post-transient dynamics of the periodic sphere wake, the dCNM accurately resolves the slowly growing amplitude between the cyclic behaviours. Regarding the quasi-periodic sphere wake, the dCNM successfully captures both the periodic behaviour and cycle-to-cycle variations. Notably, it discerns intrinsic deterministic transition behaviours, which are often misinterpreted as stochastic transitions by the CNM. For the chaotic flow dominated by unstable periodic orbits with varying scales, the dCNM accurately distinguishes between these orbits and captures their transitions. Even after sparsification, chaotic features remain preserved, with transition dynamics demonstrating stochastic characteristics. Overall, these findings underscore the notable improvement of the dCNM in capturing and accurately representing multi-frequency and multiscale dynamics.

The dCNM offers several advantages over other reduced-order modelling strategies. It preserves the advantages of previous cluster-based approaches and adds new noteworthy features.

- (i) The prediction error is minimized. The slow evolution of amplitude oscillations, the deterministic quasi-periodic dynamics, and the stochastic chaotic dynamics can be automatically resolved without any prior knowledge.
- (ii) The model complexity is significantly reduced as the number of non-trivial transitions is mitigated by design. The CNM often requires more clusters and a higher order to achieve similar accuracy, with more complex cluster transition relationships.
- (iii) The physical interpretability of the model is enhanced.

Our results suggest entropy as a guiding principle of future cluster-based models. We characterize the prediction accuracy of the cluster-based network model with the Kullback-Leibler entropy of the transition matrix, called *transition entropy* for brevity. This transition entropy is significantly reduced for the dCNM as compared to the CNM for the same number of centroids. Further improvements may be expected by optimizing the β parameter. Thus, the dCNM development from snapshots can be fully automated. The results even inspire a new clustering based on the prediction uncertainty expressed with the transition entropy. An intrusive framework with Navier-Stokes propagator may be a further avenue for improvement.

The dCNM may be compared with the POD-based Galerkin method. By construction, centroids are physically interpretable as coherent structures. In contrast, POD models have no intrinsic meaning and typically mix different frequencies. However, in select cases, the Galerkin method may yield deep insights into linear and nonlinear dynamics. Examples are the Galerkin mean-field models for the effect of forcing on a vortex shedding (Semaan *et al.* 2016), for a single oscillator (Noack *et al.* 2003) and for frequency cross-talk (Luchtenburg *et al.* 2009). The authors work on combining the advantages of clustering and POD in human-interpretable dynamic models.

Acknowledgements. The authors appreciate the valuable discussions with Steven Brunton, Antonio Colanera, Guy Yoslan Cornejo Maceda, Stefano Discetti, Andrea Ianiro, François Lusseyran, Luc R. Pastur and Xin Wang.

Funding. This work is supported by the National Natural Science Foundation of China under grants 12172109, 12172111, and 12202121, by the China Postdoctoral Science Foundation under grants 2023M730866 and 2023T160166, by the Guangdong Basic and Applied Basic Research Foundation under grant 2022A1515011492, and by the Shenzhen Science and Technology Program under grant JCYJ20220531095605012.

Declaration of Interests. The authors report no conflict of interest.

Author ORCIDs. C. Hou, <https://orcid.org/0000-0001-7477-4242>; N. Deng, <https://orcid.org/0000-0001-6847-2352>; B. R. Noack, <https://orcid.org/0000-0001-5935-1962>

Author contributions. C. Hou: Methodology, Data Curation, Validation, Writing-Original draft preparation.

N. Deng: Supervision, Methodology, Validation, Writing-Reviewing and Editing, Funding acquisition.

B. R. Noack: Methodology, Conceptualisation, Supervision, Funding acquisition, Writing-Reviewing and Editing.

Appendix A. Convergence and validation studies on the simulation of the sphere wake

To determine an optimal grid size for the numerical analysis, grid convergence studies were conducted at a $Re = 300$. For a set of grids with different numbers of grid cells, the values of the typical flow characteristics are compared to obtain grid-independent results, including the time-averaged drag coefficient $\overline{C_D}$ and its standard deviation C'_D , the time-averaged lift coefficient $\overline{C_L}$ and its standard deviation C'_L , and the Strouhal number St .

The grid refinement is specifically applied to the surface of the sphere and the wake region. Across all grid configurations, the boundary layer thickness is adjusted to ensure that the y^+ value on the sphere's surface remains below 1. This adjustment implies that the first layer of the near-wall grid has a thickness of $0.01D$ (Pan *et al.* 2018) with a spacing ratio of 1.1.

The related flow characteristics of the simulations using different grids are listed in Table 2. Here, n_s refers to the number of nodes along the circumference of the sphere within one of

Cases	n_s	n_r	Grid cells	$\overline{C_D}$	C'_D	$\overline{C_L}$	C'_L	St
Grid (a)	32	61	1.93 million	0.6637	0.00183	0.0674	0.00960	0.1363
Grid (b)	49	61	4.61 million	0.6615	0.00194	0.0666	0.01062	0.1363
Grid (c)	49	70	5.07 million	0.6624	0.00185	0.0674	0.01031	0.1363
Grid (d)	49	100	6.58 million	0.6623	0.00175	0.0664	0.01051	0.1363
Grid (e)	64	61	8.12 million	0.6607	0.00193	0.0661	0.01084	0.1363

Table 2: Grid independence test at $Re = 300$.

	$\overline{C_D}$	$\overline{C_L}$	St
Present study	0.662	0.067	0.136
Johnson & Patel (1999)	0.656	0.069	0.137
Kim <i>et al.</i> (2001)	0.657	0.067	0.137
Giacobello <i>et al.</i> (2009)	0.658	0.067	0.134
Rajamuni <i>et al.</i> (2018)	0.665	0.070	0.137

Table 3: Validation of the numerical method at $Re = 300$, compared to the listed literature.

the ‘O’-blocks. This parameter is interconnected with the grid elements along the streamwise direction and the circumference of the cylinder. On the other hand, n_r signifies the number of elements along the radial direction originating from the surface of the sphere. Consequently, n_s governs the resolution of the wake region, whereas n_r dictates the resolution of the sphere surface region. Comparing grids (a), (b), and (e) reveals a relatively smaller difference between grids (b) and (e), especially in terms of standard deviations. As a result, we select $n_s = 49$ for further analysis concerning the sphere surface region. Examining grids (b), (c), and (d) leads to similar conclusions, given that there is a more significant increase in the number of grid cells from (c) to (d) than from (b) to (c), despite limited variations in the flow characteristics. Consequently, based on these comparisons, it can be concluded that grid (c) is suitable for conducting efficient simulations with sufficient accuracy in this study.

To validate the numerical method, we compare our results with available data from related studies. Table 3 presents a comparison between the time-averaged drag coefficient $\overline{C_D}$, the time-averaged lift coefficient $\overline{C_L}$, and the Strouhal number St obtained in this study and those reported in other work for $Re = 300$. The results obtained from various studies exhibit a high degree of similarity. This consistency indicates that our study aligns well with these flow characteristics, as the values are all small and sensitive.

The convergence and validation studies presented here instil confidence that our computational grid and selected numerical schemes are adequate for the wake simulations and for testing the reduced-order modelling method.

Appendix B. Optional POD before clustering

The computational burden of clustering algorithms becomes a concern when dealing with high-dimensional flow field data. Utilising a *lossless* proper orthogonal decomposition (POD) can effectively compress the dataset. Implementing the clustering algorithm on the compressed data rather than the high-dimensional velocity fields can significantly reduce the computational time.

Here we introduce the snapshot POD methodology for the completeness of our work. The

M snapshots of the flow field can be decomposed into spatial POD modes with temporal amplitudes, where the m -th snapshot can be expressed as:

$$\mathbf{u}^m(\mathbf{x}) \approx \mathbf{u}_0(\mathbf{x}) + \sum_{i=1}^{M-1} a_i^m \mathbf{u}_i(\mathbf{x}), \quad (\text{B } 1)$$

where \mathbf{u}_0 is the mean flow, a_i is the mode amplitude and \mathbf{u}_i is the related mode. For the three-dimensional sphere flow in this work, we maintain the leading 500 POD modes for a *loseless* POD, which can resolve more than 99.9% of the fluctuation energy from all the flow regimes.

The distance between the snapshots translates into the distance between the corresponding mode amplitudes as follows:

$$D(\mathbf{u}^m, \mathbf{u}^n) = D(\mathbf{a}^m, \mathbf{a}^n). \quad (\text{B } 2)$$

With this transformation, the reduction in the computational time can be one or two orders of magnitude, and the statistical description has been formulated in Fernex *et al.* (2021) and Li *et al.* (2021).

Appendix C. Modelling with different values of β

In the dCNM framework, a sparsification controller $\beta \in [0, 1]$ is set to determine the number of sub-clusters in each cluster. As in Eq. (2.18), the number of sub-clusters is decided by the transverse cluster size and the number of trajectory segments in each cluster. With a given β , the number of sub-clusters in each cluster can be decided, then the second-stage clustering algorithm will automatically search the centroids. Decreasing β will lead to more sub-clusters, which means higher model complexity and also higher accuracy in the dynamic reconstruction. The optimal choice of β can be determined by searching for a sweet point which balances the model complexity and the model accuracy.

We demonstrate the impact of β on the modelling of the post-transient sphere wake. Figure 27 displays the clustering results for the quasi-periodic flow, where β takes values of 1, 0.95, 0.80, and 0. Figure 28 illustrates the results for the chaotic flow, with β values of 1, 0.95, 0.40, and 0. $\beta = 1$ means fully sparse, and the centroids are equivalent to the cluster averages, yielding results identical to those of the CNM. Conversely, when $\beta = 0$, the model is minimally sparse, resulting in the highest model accuracy, but also the highest model complexity. As β decreases, the centroids try to cover more cyclic behaviours, gradually outlining the entire structure. This expansion involves more trajectory segments and, consequently, increases the model resolution. For the quasi-periodic flow regime, there are limitations to this enhancement. Due to the finite axial length of the cylinder, the trajectory segments often overlap. Consequently, increasing the number of sub-clusters to a certain extent results in extensive centroid overlap, offering minimal contributions to the resolution improvement. This situation is evident when comparing figure 27 (c) and (d), where the centroid distributions are very similar, and where centroid overlap is prevalent. In contrast, centroid overlap is rare in chaotic flows, allowing for noticeable accuracy improvements with smaller β values. However, using a small β will result in lengthy and complex centroid transition information. Therefore, for chaotic dynamics, it is advisable to strike a balance between the model accuracy and complexity by adjusting β based on specific purposes.

From a spatial perspective, we evaluated the representation error using different β for the two flow regimes, which is also relevant to determine the appropriate β , where a sweet point of β can be found considering the model complexity and the model accuracy, as shown in figure 29. The representation error exhibits different trends for the two flow regimes as β

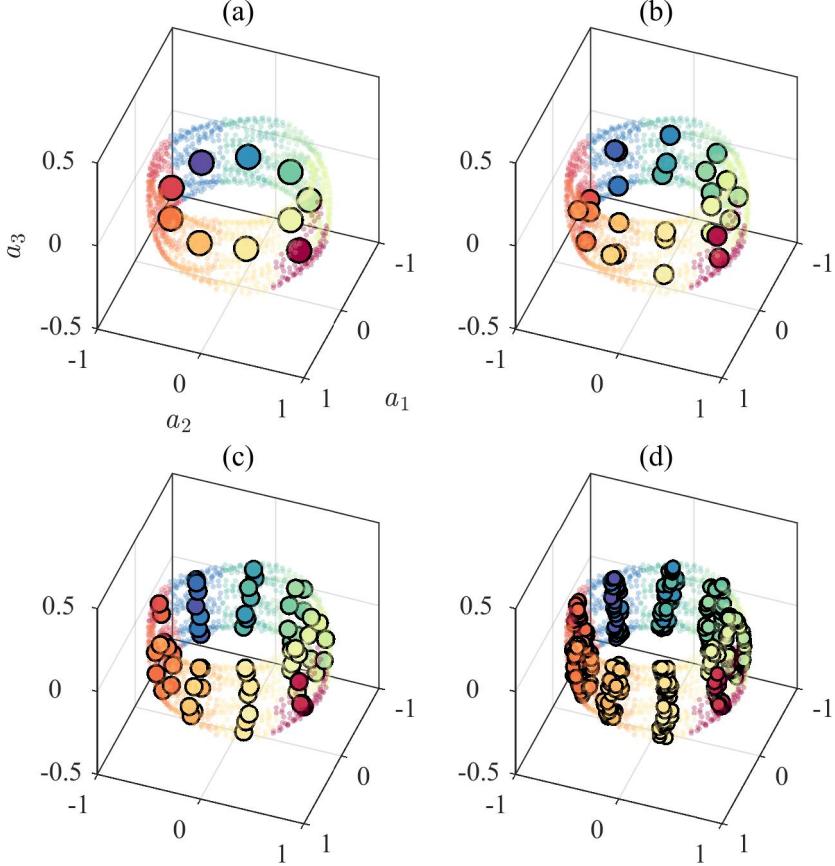


Figure 27: Clustering results with different β on the quasi-periodic flow. (a) $\beta = 1$. (b) $\beta = 0.95$. (c) $\beta = 0.80$. (d) $\beta = 0$.

increases. In the quasi-periodic flow, the representation error remains relatively constant over a wide range of β values and then sharply increases near $\beta = 1$. This abrupt rise suggests that sparsification eliminates the cycle-to-cycle variations. For the chaotic flow, the representation error changes smoothly from $\beta = 0$ to $\beta = 1$, indicating the loss of diversity of the main loop. We can expect a Pareto Optimality from the spatial representation error for these two cases, i.e. $\beta = 0.80$ for the quasi-periodic case and $\beta = 0.40$ for the chaotic case.

From a temporal perspective, the assignments of each snapshot to the clusters and centroids of the chaotic flow are illustrated in figure 30. When $\beta = 1$, the centroid affiliation is disregarded, and only the cluster-level transitions can be observed, this is the same with the CNM. The temporal evolution of centroids relies solely on the stochastic cluster transition probabilities, with each centroid visited multiple times, as shown in figure 30 (a). Conversely, for $\beta = 0$, most centroids are visited only once, leading to the minimum transition error, as seen in figure 30 (d). From figure 30 (b) and (c), we can conclude that even with sparsification, varied cyclic behaviours can still be effectively captured by the dCNM. This is because different centroid combinations in the dCNM reconstruction constitute extended cluster chains mentioned in § 4.3, and the occurrence of extended cluster chains affirms the capability of the dCNM to effectively resolve the multiscale dynamics. The generally similar visiting sequences in the extended cluster chains from the dCNM reconstruction and the

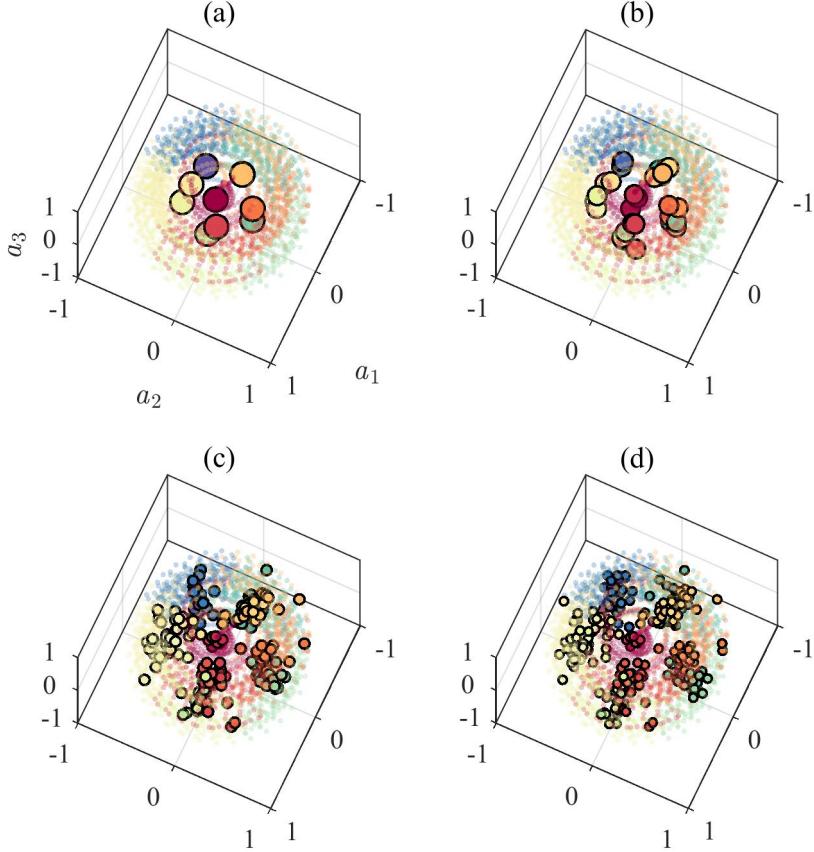


Figure 28: Clustering results with different β on the chaotic flow. (a) $\beta = 1$. (b) $\beta = 0.95$. (c) $\beta = 0.40$. (d) $\beta = 0$.

data set ensure the model accuracy and the difference highlights that the stochastic transition characteristics of chaotic dynamics are also reserved.

Appendix D. The centroid transition matrix

For the nonzero terms in the cluster transition probability matrix, we can embed a corresponding centroid transition matrix based on the sub-clusters and then record all the dual indexing centroid transitions by the transition tensors Q .

The centroid transition matrices of the quasi-periodic flow, as discussed in § 4.2, departing from C_4 are shown in figure 31. The matrices of the chaotic flow, as discussed in § 4.3, departing from C_1 are shown in figure 32. For the quasi-periodic flow regime, the matrices are sparse and clear, with centroids having only one destination, indicating deterministic transitions. Moreover, these transitions impose specific constraints on the quasi-stochastic dynamics. Once the departing centroid is determined, all destination centroids belong to the same destination cluster, and the nonzero terms in this column appear only in one matrix. The stochastic cluster transition can therefore become deterministic. Compared to the quasi-periodic flow, the transition probabilities in the matrices of the chaotic flow exhibit stochastic centroid transitions. Some departing centroids have destination centroids within

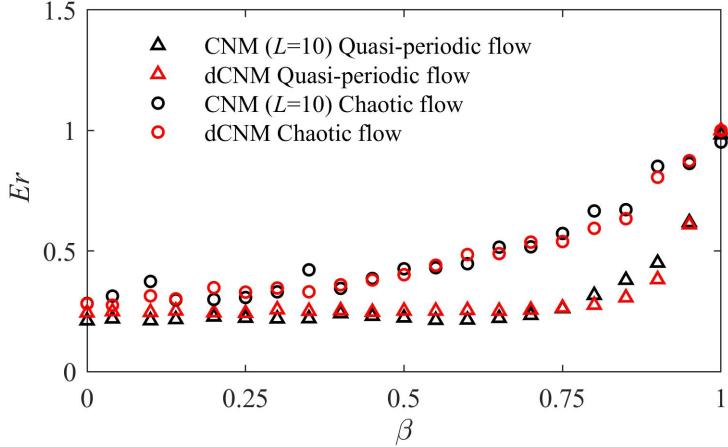


Figure 29: Representation error versus the sparsification index β for the quasi-periodic flow and the chaotic flow. The results of dCNM are marked with red, and the corresponding results of high-order CNM with the same number of centroids are marked with black. All values have been normalised using the representation error of classical CNM with 10 clusters. The marginally lower error of the high-order CNM for the quasi-periodic case is due to the more numerous distribution of the centroids in one limit cycle, which constitutes a smoother cyclic trajectory. The dCNM centroids also focus on the variation between loops, thus with fewer centroids in each loop.

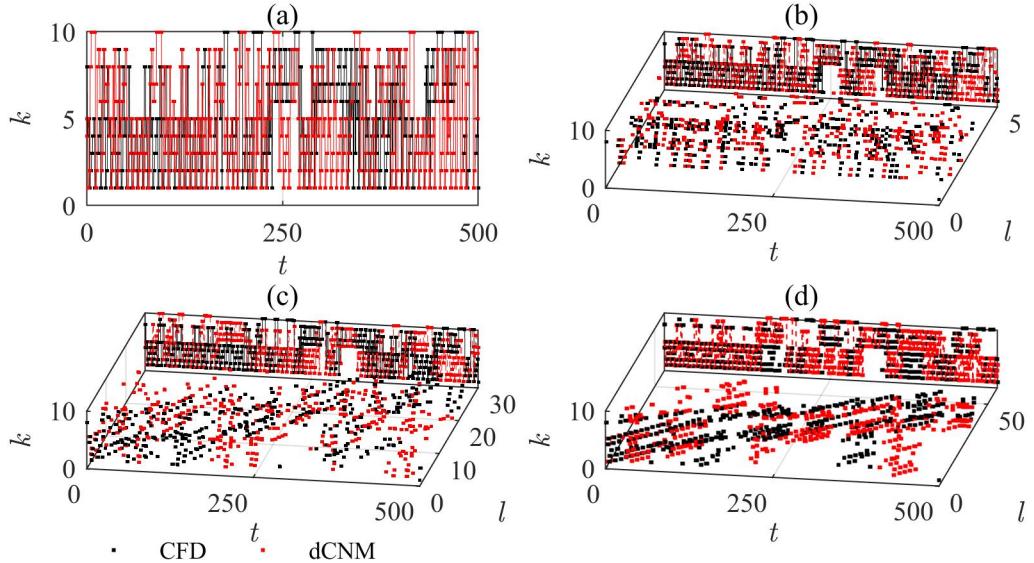


Figure 30: Temporal evolution of cluster and trajectory segment affiliation with different β for the chaotic flow. (a) $\beta = 1$, (b) $\beta = 0.95$, (c) $\beta = 0.40$, and (d) $\beta = 0$.

the same cluster, while others do not. Consequently, some centroids participate solely in deterministic cluster loops, while others also engage in random jumps between cluster loops. This distinction separates the cluster transitions from periodic and stochastic routes and serves as a constraint that distinguishes multiscale loops and their associated cycle-to-cycle transitions.

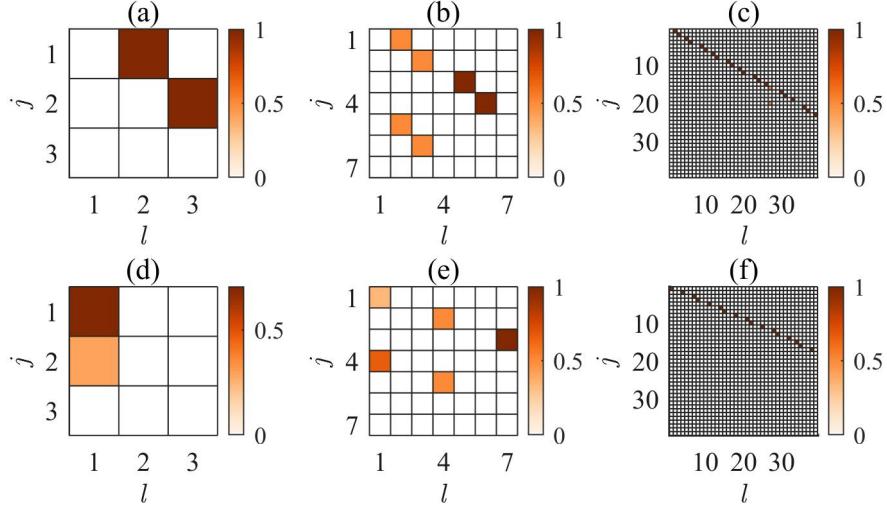


Figure 31: Centroid transition matrices departing from C_4 with different β for the quasi-periodic flow: (a) $\beta = 0.95$, (b) 0.80, and (c) 0 for the cluster transition $C_4 \rightarrow C_5$; and (d) $\beta = 0.95$, (e) 0.80, and (f) 0 for $C_4 \rightarrow C_{10}$.

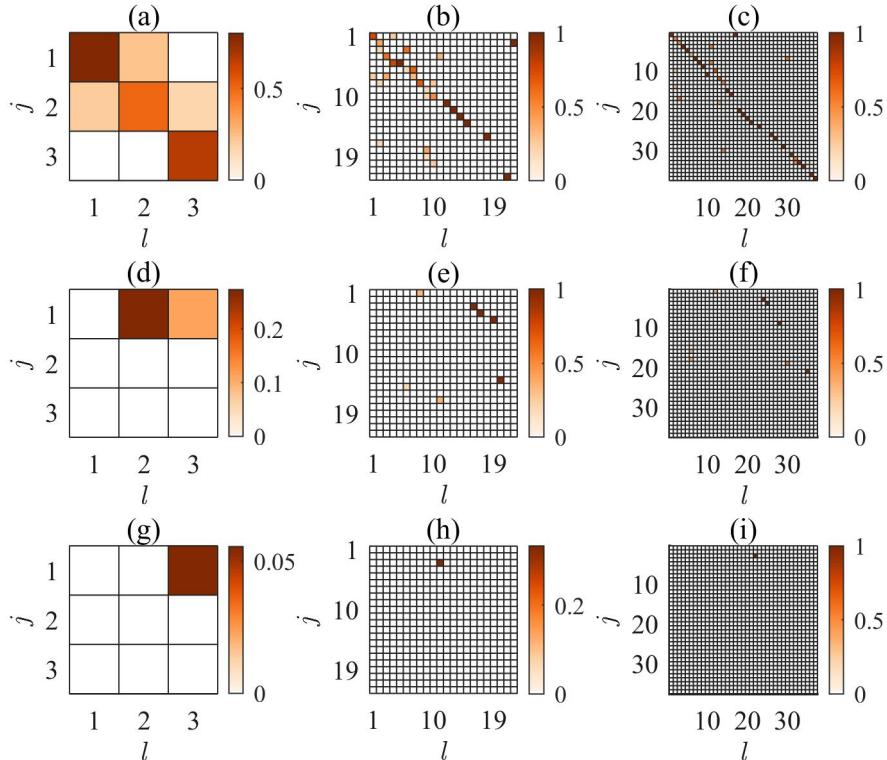


Figure 32: Centroid transition matrices departing from C_1 with different β for the chaotic flow: (a) $\beta = 0.95$, (b) 0.40, and (c) 0 for the cluster transition $C_1 \rightarrow C_2$; (d) $\beta = 0.95$, (e) 0.40, and (f) 0 for $C_1 \rightarrow C_7$; and (g) $\beta = 0.95$, (h) 0.40, and (i) 0 for $C_1 \rightarrow C_{10}$.

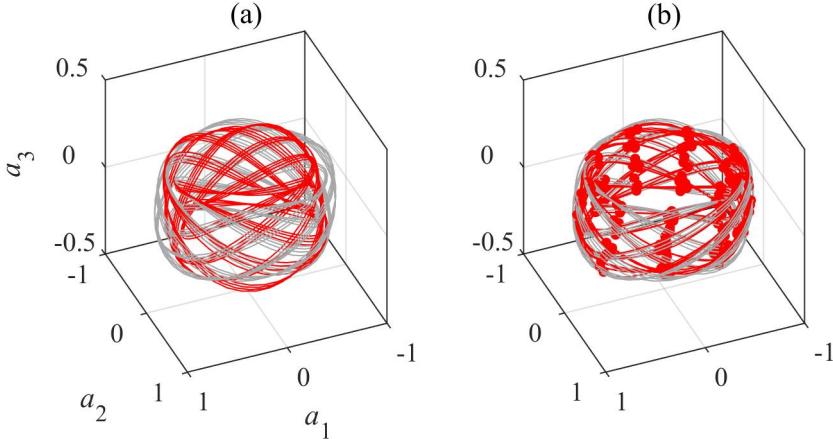


Figure 33: Comparison between the POD reconstruction and the dCNM reconstruction for the quasi-periodic flow at $Re = 330$: (a) The POD reconstruction resolving 50% of the fluctuation energy and (b) the dCNM reconstruction with $\beta = 0.5$.

Appendix E. The POD reconstruction and the dCNM reconstruction

In this section, we compared the flow kinematics reconstructed by POD and the dCNM. The POD reconstruction uses the leading POD modes and their mode amplitudes to reconstruct the flow. The number of POD modes is chosen so that the resolved fluctuation energy is equal to the value of $1 - \beta$ from the dCNM, i.e. 90% of the fluctuation energy resolved by the POD reconstruction is comparable to the dCNM reconstruction with $\beta = 0.1$. For the quasi-periodic flow at $Re = 330$, the POD reconstruction with 50% of the fluctuation energy needs the 5 leading POD modes. For the chaotic flow at $Re = 450$, the POD reconstruction with 60% of the fluctuation energy takes the 19 leading modes. The dCNM outperforms the POD in resolving the key features of the data under the same standard. For the quasi-periodic flow in figure 33, the POD results better resolve the tiny variation between trajectories while exhibiting a larger deformation of the overall geometry. The dCNM results cover the whole geometry better, while the tiny variation between trajectories is averaged by the centroids. This elaborates the dCNM with the advantage of being more robust to noise. This trend is similarly observed in the chaotic flow in figure 34, where the dCNM better outlines the geometry and resolves the prominent features.

REFERENCES

- ALBERT, R. & BARABÁSI, A.-L. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74** (1), 47–97.
- ARTHUR, D. & VASSILVITSKII, S. 2007 K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics.
- AUBRY, N., HOLMES, P., LUMLEY, J. L. & STONE, E. 1988 The dynamics of coherent structures in the wall region of a turbulent boundary layer. *J. Fluid Mech.* **192**, 115–173.
- BARABÁSI, A.-L. 2013 Network science. *Phil. Trans. R. Soc. A* **371** (1987), 20120375.
- BERGMANN, M., BRUNEAU, C.-H. & IOLLO, A. 2009 Enablers for robust POD models. *J. Comput. Phys.* **228** (2), 516–538.
- BERGMANN, M. & CORDIER, L. 2008 Optimal control of the cylinder wake in the laminar regime by trust-region methods and POD reduced-order models. *J. Comput. Phys.* **227** (16), 7813–7840.

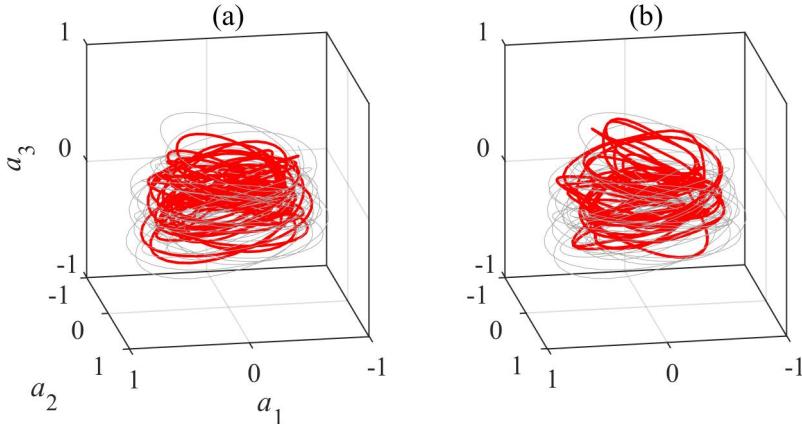


Figure 34: Same as figure 33, but for the chaotic flow at $Re = 450$. (a) The POD reconstruction resolving 60% of the fluctuation energy and (b) the dCNM reconstruction with $\beta = 0.4$.

- BOLLT, E. M. 2001 Combinatorial control of global dynamics in a chaotic differential equation. *Int. J. Bifurcat. Chaos* **11** (08), 2145–2162.
- BÖRNER, K., SANYAL, S. & VESPIGNANI, A. 2007 Network science. *Annu. Rev. Inf. Sci. Technol.* **41** (1), 537–607.
- BOURGEOIS, J. A., NOACK, B. R. & MARTINUZZI, R. J. 2013 Generalized phase average with applications to sensor-based flow estimation of the wall-mounted square cylinder wake. *J. Fluid Mech.* **736**, 316–350.
- BRUNTON, S. L., NOACK, B. R. & KOUMOUTSAKOS, P. 2020 Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52** (1), 477–508.
- BRUNTON, S. L., PROCTOR, J. L. & KUTZ, J. N. 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113** (15), 3932–3937.
- BURKARDT, J., GUNZBURGER, M. & LEE, H.-C. 2006 Centroidal voronoi tessellation-based reduced-order modeling of complex systems. *SIAM J. Sci. Comput.* **28** (2), 459–484.
- BUSSE, F. H. 1991 Numerical analysis of secondary and tertiary states of fluid flow and their stability properties. *Appl. Sci. Res.* **48** (3-4), 341–351.
- CAO, Y., KAISER, E., BORÉE, J., NOACK, B. R., THOMAS, L. & GUILAIN, S. 2014 Cluster-based analysis of cycle-to-cycle variations: application to internal combustion engines. *Exp. Fluids* **55** (11).
- DENG, N., NOACK, B. R., MORZYŃSKI, M. & PASTUR, L. R. 2020 Low-order model for successive bifurcations of the fluidic pinball. *J. Fluid Mech.* **884**, A37.
- DENG, N., NOACK, B. R., MORZYŃSKI, M. & PASTUR, L. R. 2022 Cluster-based hierarchical network model of the fluidic pinball – cartographing transient and post-transient, multi-frequency, multi-attractor behaviour. *J. Fluid Mech.* **934**, A24.
- ESHBAL, L., RINSKY, V., DAVID, T., GREENBLATT, D. & VAN HOUT, R. 2019 Measurement of vortex shedding in the wake of a sphere at $re = 465$. *J. Fluid Mech.* **870**, 290–315.
- FABRE, D., AUGUSTE, F. & MAGNAUDET, J. 2008 Bifurcations and symmetry breaking in the wake of axisymmetric bodies. *Phys. Fluids* **20** (5).
- FARZAMNIK, E., IANIRO, A., DISCETTI, S., DENG, N., OBERLEITHNER, K., NOACK, B. R. & GUERRERO, V. 2023 From snapshots to manifolds – a tale of shear flows. *J. Fluid Mech.* **955**, A34.
- FERNEX, D., NOACK, B. R. & SEMAAN, R. 2021 Cluster-based network modeling—from snapshots to complex dynamical systems. *Sci. Adv.* **7** (25), eabf5006.
- GIACOBELLO, M., OOI, A. & BALACHANDAR, S. 2009 Wake structure of a transversely rotating sphere at moderate reynolds numbers. *J. Fluid Mech.* **621**, 103–130.
- GÓMEZ, F., BLACKBURN, H. M., RUDMAN, M., SHARMA, A. S. & McKEON, B. J. 2016 A reduced-order model of three-dimensional unsteady flow in a cavity based on the resolvent operator. *J. Fluid Mech.* **798**, R2.

- GOPALAKRISHNAN MEENA, M., NAIR, A. G. & TAIRA, K. 2018 Network community-based model reduction for vortical flows. *Phys. Rev. E* **97** (6).
- GOPALAKRISHNAN MEENA, M. & TAIRA, K. 2021 Identifying vortical network connectors for turbulent flow modification. *J. Fluid Mech.* **915**, A10.
- HADJIGHASEM, A., KARRASCH, D., TERAMOTO, H. & HALLER, G. 2016 Spectral-clustering approach to lagrangian vortex detection. *Phys. Rev. E* **93** (6).
- HOLMES, P., LUMLEY, J. L. & BERKOOZ, G. 1996 *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press.
- HOU, C., DENG, N. & NOACK, B. R. 2022 Trajectory-optimized cluster-based network model for the sphere wake. *Phys. Fluids* **34** (8).
- IOLLO, A., LANTERI, S. & DÉSIDÉRI, J.-A. 2000 Stability properties of POD-galerkin approximations for the compressible navier-stokes equations. *Theor. Comput. Fluid Dyn.* **13** (6), 377–396.
- JAIN, A. K. 2010 Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.* **31** (8), 651–666.
- JAIN, A. K. & DUBES, R. C. 1988 *Algorithms for clustering data*. Prentice-Hall, Inc.
- JAIN, A. K., MURTY, M. N. & FLYNN, P. J. 1999 Data clustering. *ACM Comput. Surv.* **31** (3), 264–323.
- JOHNSON, T. A. & PATEL, V. C. 1999 Flow past a sphere up to a reynolds number of 300. *J. Fluid Mech.* **378**, 19–70.
- KAISER, E., NOACK, B. R., CORDIER, L., SPOHN, A., SEGOND, M., ABEL, M., DAVILLER, G., ÖSTH, J., KRAJNOVIĆ, S., NIVEN, R. K. & ET AL. 2014 Cluster-based reduced-order modelling of a mixing layer. *J. Fluid Mech.* **754**, 365–414.
- KIM, J., KIM, D. & CHOI, H. 2001 An immersed-boundary finite-volume method for simulations of flow in complex geometries. *J. Comput. Phys.* **171** (1), 132–150.
- KIM, M., LIU, F., JAIN, A. K. & LIU, X. 2022 Cluster and aggregate: Face recognition with large probe set. In *Advances in Neural Information Processing Systems* (ed. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh), , vol. 35, pp. 36054–36066. Curran Associates, Inc.
- KOU, J. & ZHANG, W. 2021 Data-driven modeling for unsteady aerodynamics and aeroelasticity. *Prog. Aerosp. Sci.* **125**, 100725.
- KRUEGER, P. S., HAHSLER, M., OLINICK, E. V., WILLIAMS, S. H. & ZHARFA, M. 2019 Quantitative classification of vortical flows based on topological features using graph matching. *Proc. R. Soc. A* **475** (2228), 20180897.
- KUTZ, J. N. 2017 Deep learning in fluid dynamics. *J. Fluid Mech.* **814**, 1–4.
- LANDAU, L. D. 1944 On the problem of turbulence. In *Dokl. Akad. Nauk USSR*, , vol. 44, p. 311. Dokl. Akad. Nauk USSR.
- LI, H., FERNEX, D., SEMAAN, R., TAN, J., MORZYŃSKI, M. & NOACK, B. R. 2021 Cluster-based network model. *J. Fluid Mech.* **906**, A21.
- LI, H. & TAN, J. 2020 Cluster-based Markov model to understand the transition dynamics of a supersonic mixing layer. *Phys. Fluids* **32** (5).
- LI, S., LI, W. & NOACK, B. R. 2022 Machine-learned control-oriented flow estimation for multi-actuator multi-sensor systems exemplified for the fluidic pinball. *J. Fluid Mech.* **952**, A36.
- LLOYD, S. 1982 Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28** (2), 129–137.
- LORENZ, E. N. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** (2), 130–141.
- LORITE-DÍEZ, M. & JIMÉNEZ-GONZÁLEZ, J. 2020 Description of the transitional wake behind a strongly streamwise rotating sphere. *J. Fluid Mech.* **896**, A18.
- LUCHTENBURG, D. M., GÜNTHER, B., NOACK, B. R., KING, R. & TADMOR, G. 2009 A generalized mean-field model of the natural and high-frequency actuated flow around a high-lift configuration. *J. Fluid Mech.* **623**, 283–316.
- MACQUEEN, J. 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, , vol. 1, pp. 281–297. Oakland, CA, USA.
- MCKEON, B. J., LI, J., JIANG, W., MORRISON, J. F. & SMITS, A. J. 2004 Further observations on the mean velocity distribution in fully developed pipe flow. *J. Fluid Mech.* **501**, 135–147.
- MELIGA, P., CHOMAZ, J.-M. & SIPP, D. 2009 Global mode interaction and pattern selection in the wake of a disk: a weakly nonlinear expansion. *J. Fluid Mech.* **633**, 159–189.
- MOORE, B. 1981 Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automat. Contr.* **26** (1), 17–32.
- MURAYAMA, S., KINUGAWA, H., TOKUDA, I. T. & GOTODA, H. 2018 Characterization and detection of

- thermoacoustic combustion oscillations based on statistical complexity and complex-network theory. *Phys. Rev. E* **97** (2).
- NAIR, A. G. & TAIRA, K. 2015 Network-theoretic approach to sparsified discrete vortex dynamics. *J. Fluid Mech.* **768**, 549–571.
- NAIR, A. G., YEH, C.-A., KAISER, E., NOACK, B. R., BRUNTON, S. L. & TAIRA, K. 2019 Cluster-based feedback control of turbulent post-stall separated flows. *J. Fluid Mech.* **875**, 345–375.
- NOACK, B. R., AFANASIEV, K., MORZYŃSKI, M., TADMOR, G. & THIELE, F. 2003 A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. *J. Fluid Mech.* **497**, 335–363.
- NOACK, B. R. & ECKELMANN, H. 1994 A global stability analysis of the steady and periodic cylinder wake. *J. Fluid Mech.* **270**, 297–330.
- ORMIÈRES, D. & PROVANSAL, M. 1999 Transition to turbulence in the wake of a sphere. *Phys. Rev. Lett.* **83** (1), 80–83.
- PAN, J., ZHANG, N. & NI, M. 2018 The wake structure and transition process of a flow past a sphere affected by a streamwise magnetic field. *J. Fluid Mech.* **842**, 248–272.
- PODVIN, B. 2009 A proper-orthogonal-decomposition-based model for the wall layer of a turbulent channel flow. *Phys. Fluids* **21** (1).
- PODVIN, B. & LUMLEY, J. 1998 A low-dimensional approach for the minimal flow unit. *J. Fluid Mech.* **362**, 121–155.
- PROTAS, B., NOACK, B. R. & ÖSTH, J. 2015 Optimal nonlinear eddy viscosity in galerkin models of turbulent flows. *J. Fluid Mech.* **766**, 337–367.
- RAJAMUNI, M. M., THOMPSON, M. C. & HOURIGAN, K. 2018 Transverse flow-induced vibrations of a sphere. *J. Fluid Mech.* **837**, 931–966.
- RIGAS, G., SCHMIDT, O. T., COLONIUS, T. & BRÈS, G. A. 2017 One-way navier-stokes and resolvent analysis for modeling coherent structures in a supersonic turbulent jet. *AIAA Paper* pp. 2017–4046.
- ROWLEY, C. W. 2005 Model reduction for fluids, using balanced proper orthogonal decomposition. *Int. J. Bifurcat. Chaos* **15** (03), 997–1013.
- SAN, O. & MAULIK, R. 2018 Extreme learning machine for reduced order modeling of turbulent geophysical flows. *Phys. Rev. E* **97** (4).
- SAN, O., MAULIK, R. & AHMED, M. 2019 An artificial neural network framework for reduced order modeling of transient flows. *Commun. Nonlinear Sci. Numer. Simul.* **77**, 271–287.
- SCHLUETER-KUCK, K. L. & DABIRI, J. O. 2017 Coherent structure colouring: identification of coherent structures from sparse data using graph theory. *J. Fluid Mech.* **811**, 468–486.
- SCHUMM, M., BERGER, E. & MONKEWITZ, P. A. 1994 Self-excited oscillations in the wake of two-dimensional bluff bodies and their control. *J. Fluid Mech.* **271**, 17–53.
- SEMAAN, R., KUMAR, P., BURNAZZI, M., TISSOT, G., CORDIER, L. & NOACK, B. R. 2016 Reduced-order modelling of the flow around a high-lift configuration with unsteady coanda blowing. *J. Fluid Mech.* **800**, 72–110.
- SHANNON, C. E. 1948 A mathematical theory of communication. *The Bell system technical journal* **27** (3), 379–423.
- SHEARD, G. J., THOMPSON, M. C. & HOURIGAN, K. 2003 From spheres to circular cylinders: the stability and flow structures of bluff ring wakes. *J. Fluid Mech.* **492**, 147–180.
- STRYKOWSKI, P. J. & SREENIVASAN, K. R. 1990 On the formation and suppression of vortex ‘shedding’ at low reynolds numbers. *J. Fluid Mech.* **218**, 71–107.
- STUART, J. T. 1958 On the non-linear mechanics of hydrodynamic stability. *J. Fluid Mech.* **4** (1), 1–21.
- STUART, J. T. 1971 Nonlinear stability theory. *Annu. Rev. Fluid Mech.* **3** (1), 347–370.
- TAIRA, K. & NAIR, A. G. 2022 Network-based analysis of fluid flows: Progress and outlook. *Prog. Aerosp. Sci.* **131**, 100823.
- TAIRA, K., NAIR, A. G. & BRUNTON, S. L. 2016 Network structure of two-dimensional decaying isotropic turbulence. *J. Fluid Mech.* **795**, R2.
- TANEDA, S. 1956 Experimental investigation of the wake behind a sphere at low reynolds numbers. *J. Phys. Soc. Japan* **11** (10), 1104–1108.
- WATTS, D. J. & STROGATZ, S. H. 1998 Collective dynamics of ‘small-world’ networks. *Nature* **393** (6684), 440–442.
- YEH, C.-A., GOPALAKRISHNAN MEENA, M. & TAIRA, K. 2021 Network broadcast analysis and control of turbulent flows. *J. Fluid Mech.* **910**, A15.
- ZHU, L., ZHANG, W., KOU, J. & LIU, Y. 2019 Machine learning methods for turbulence modeling in subsonic flows around airfoils. *Phys. Fluids* **31** (1).