Cluster-based network model

**Hao Li[1,2], Daniel Fernex[3], Richard Semaan[3], Jianguo Tan[1],†,
Marek Morzyński[4] and Bernd R. Noack[5,2],†**

[1]Science and Technology on Scramjet Laboratory, National University of Defense Technology,
Changsha 410073, Hunan Province, PR China

[2]Hermann-Föttinger-Institut, Technische Universität Berlin, Müller-Breslau-Straße 8,
D-10623 Berlin, Germany

[3]Institut für Strömungsmechanik, Technische Universität Braunschweig, Hermann-Blenk-Straße 37,
D-38108 Braunschweig, Germany

[4]Chair of Virtual Engineering, Poznań University of Technology, Jana Pawła II 24 street,
60-965 Poznań, Poland

[5]Center for Turbulence Control, Harbin Institute of Technology, Shenzhen, Room 312, Building C,
University Town, Xili, Shenzhen 518058, PR China

We propose an automatable data-driven methodology for robust nonlinear reduced-order modelling from time-resolved snapshot data. In the kinematical coarse-graining, the snapshots are clustered into a few centroids representing the whole ensemble. The dynamics is conceptualized as a directed network, where the centroids represent nodes and the directed edges denote possible finite-time transitions. The transition probabilities and times are inferred from the snapshot data. The resulting cluster-based network model constitutes a deterministic–stochastic grey-box model resolving the coherent-structure evolution. This model is motivated by limit-cycle dynamics, illustrated for the chaotic Lorenz attractor and successfully demonstrated for the laminar two-dimensional mixing layer featuring Kelvin–Helmholtz vortices and vortex pairing, and for an actuated turbulent boundary layer with complex dynamics. Cluster-based network modelling opens a promising new avenue with unique advantages over other model-order reductions based on clustering or proper orthogonal decomposition.

Key words: low-dimensional models, shear layers, turbulent boundary layers

1. Introduction

We propose a cluster-based network model (CNM) from time-resolved snapshot data exemplified for a laminar mixing layer and an actuated turbulent boundary layer. The goal is purely data-driven reduced-order modelling trading the physical insights from first principles, e.g. the Galerkin method (see e.g. Holmes *et al.* 2012), with simplicity, robustness and closeness to the original data.

† Email addresses for correspondence: jianguotan@nudt.edu.cn, bernd.noack@hit.edu.cn

The mixing layer is an archetypical flow configuration associated with many academic and industrial applications. The flow is discussed virtually in any textbook of fluid mechanics. In the early stage, the laminar mixing layer gives rise to periodic, spatially growing Kelvin–Helmholtz vortices as described in stability theory (Michalke 1964), by vortex models (Hama 1962) or by a proper orthogonal decomposition (POD) Galerkin model (Noack, Papas & Monkewitz 2005). At later stages, multiple vortex pairings induce the inverse cascade at lower wavenumbers and frequencies (Coats 1997). In addition, three-dimensional instabilities enrich the coherent structures by rib vortices and spanwise waviness (see e.g. Liu 1989). These mixing-layer structures may be seen in the near-field region of wakes and jets. Moreover, control of most shear flows, including bluff-body wakes and jets, is based on an effective manipulation of the mixing layer (Fiedler & Fernholz 1990).

Another fundamental flow configuration is the turbulent boundary layer. Since Prandtl's (1904) discovery of the boundary layer theory, this flow is the cornerstone of practically every fluid and aerodynamic problem. In particular, skin-friction reduction through passive or active means has been the subject of research for many decades (Gad-el-Hak 2000; Fan & Dong 2016). Promising strategies include riblets (Walsh & Lindemann 1984), compliant surfaces (Luhar, Sharma & McKeon 2016), spanwise wall oscillations (Jung, Mangiavacchi & Akhavan 1992; Quadrio, Ricco & Viotti 2009) and spanwise travelling waves with a Lorentz force (Du & Karniadakis 2000) or wall-normal deflection (Klumpp, Meinke & Schröder 2011; Albers *et al.* 2020). In this study, a spatio-temporal surface deformation with transverse travelling waves is chosen targeting aerodynamic applications. Thus, a drag reduction of 4.5 % was experimentally achieved for a turbulent boundary layer (Li *et al.* 2015). In a numerical partner study, the actuation parameters were improved, yielding 31 % drag reduction (Albers *et al.* 2020; Fernex *et al.* 2020). The actuation was also applied over a wing section (Albers, Meysonnat & Schröder 2019), where the pressure varies in the streamwise direction. Thus, the total drag was reduced by 7.5 % accompanied by a slight lift increase.

For many decades, the mixing layer and the turbulent boundary layer have been long-standing benchmarks for reduced-order modelling. For the mixing layer, employed methods include input–output transfer functions (Sasaki *et al.* 2017), parabolized stability equations (Sasaki *et al.* 2018), vortex filament models (Ashurst & Meiburg 1988), POD models (Delville *et al.* 1999; Ukeiley *et al.* 2001; Wei & Rowley 2009) and cluster-based reduced-order models (Kaiser *et al.* 2014). Already the laminar two-dimensional shear layer can give rise to multiple frequencies (Kasten *et al.* 2016). The early stages of the convectively unstable and nearly linear dynamics of mixing layers and jets are well resolved by parabolized stability equations requiring little empirical input (Jordan & Colonius 2013). After the three-dimensional transition, the accuracy of stability-based methods rapidly deteriorates or describes only a narrow frequency spectrum of the fluid dynamics. Stability methods combined with eddy-viscosity closure models may significantly extend the application range (Liu 1989). Alternatively, data-driven grey-box models from snapshot data distilling the coherent-structure dynamics become an attractive avenue (Taira *et al.* 2018).

Since the pioneering POD model of Aubry *et al.* (1988) for the unforced turbulent boundary layer, numerous advances of data-driven Galerkin models have been proposed. Podvin & Lumley (1998) proposed a low-dimensional model for the minimal channel flow unit for the purpose of physical understanding. Later, Podvin (2009) has developed an accurate high-dimensional POD model for the wall region of a turbulent channel flow. The drag-reducing effect of compliant walls has been included in POD models by Lumley, Rempfer & Blossey (1999, p. 147).

The POD Galerkin methods arguably constitute the most popular and best-investigated data-driven grey-box modelling methods; POD Galerkin methods are intimately tied with the Navier–Stokes equations. While the kinematics, the modal expansion, is distilled from data, the temporal dynamics may be derived from first principles. Yet, the modal expansion encapsulates a convection dominated dynamics in an elliptic approach. This mismatch between the modelling approach and dynamics is the root cause of the fragility of data-driven Galerkin models (Noack 2016). For the mixing layer, the lack of robustness is particularly pronounced as exhibited by modelled transients which are orders of magnitudes too large (Noack *et al.* 2005). Moreover, the time integration of the Galerkin model may easily lead to states far away from the training data, i.e. outside the region of model validity. This problem persists for other data-driven modal expansions, such as dynamic mode decomposition (Rowley *et al.* 2009; Schmid 2010).

This robustness challenge of elliptical approaches is avoided by cluster-based reduced-order models pioneered by Burkardt, Gunzburger & Lee (2006). Here, the state is coarse-grained to a small number of centroids representative of the whole ensemble of snapshots. Hence, modelled states will be close to the training data by construction. The potential of an extrapolation, e.g. predicting larger fluctuation amplitudes, is traded for robustness, i.e. staying close to the snapshot data.

In the cluster-based Markov model (CMM) for the mixing layer (Kaiser *et al.* 2014; Li & Tan 2020), the temporal evolution is modelled as a probabilistic Markov model of the transition dynamics. The state vector of cluster probabilities may initially start in a single centroid but eventually diffuses to a fixed point representing the post-transient attractor. This fixed point is well reproduced by CMM. In addition, CMM has provided valuable physical insights for the mixing layer and Ahmed body wake (Kaiser *et al.* 2014), for the turbulent boundary layer (Ishar *et al.* 2019), for combustion related mixing (Cao *et al.* 2014) and for control design (Kaiser *et al.* 2017; Nair *et al.* 2019).

A challenge for CMM is the temporal evolution: the state may quickly diffuse over the whole attractor, often within one typical time period. This study aims at a cluster-based network model (CNM) with improved dynamics resolution following Fernex *et al.* (2019). The dynamics is modelled by 'constant velocity flights' between the centroids as 'airports'. The transition probabilities and times are consistent with the snapshot data. The dynamics is thus restricted to a sparse network of routes between the centroids. Network models are enjoying increasing popularity in all mathematical modelling fields including biology, sociology and the computer sciences. Network models have also been employed to explain vortex dynamics (Nair & Taira 2015; Taira, Nair & Brunton 2016). Newman (2010) provides an excellent introduction to networks.

On the surface, the CNM, CMM and POD models look like similar data-driven reduced-order models from snapshot data. Yet, there are fundamental application differences which may be elucidated by an analogy to computational fluid mechanics. The traditional CMM might be compared with unsteady Reynolds-averaged Navier–Stokes simulations converging to the mean flow while resolving some dynamic features during the transient. In contrast, the proposed CNM mimics a large-eddy simulation designed to resolve unsteady coherent-structure dynamics. The applications of CNM are comparable with POD models. The POD model can be conceptualized as a data-driven version of the spectral method being routed in the traditional Galerkin methodology. In contrast, CNM is closer to a collocation method using the centroids as 'lighthouses' for the corresponding Voronoi cells.

The paper is organized as follows. Section 2 elaborates the methodology of the cluster-based network model. The limit-cycle dynamics and Lorenz attractor are employed as illustrating examples. Two flow configurations are chosen for the numerical analysis, an
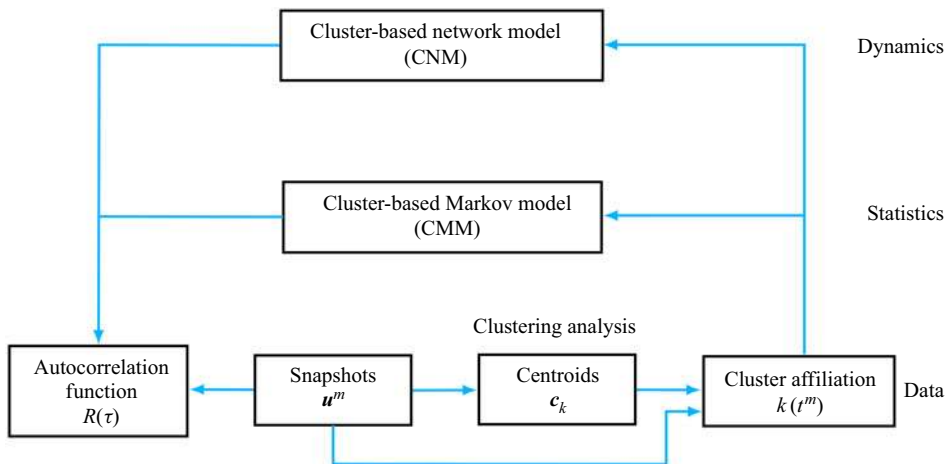
FIGURE 1. Principle sketch of cluster-based modelling. The time-resolved snapshots are partitioned into a predetermined number of clusters represented by centroids in an unsupervised manner. Thereafter, each snapshot has a cluster affiliation $k(t^m)$, being the index of the closest centroid. The cluster-based Markov model describes the evolution of the population of these clusters. The solution of CMM quickly converges against the asymptotic probability distribution. The proposed cluster-based network model resolves the dynamic transitions between the clusters by a deterministic–stochastic network.

incompressible laminar mixing layer and a turbulent boundary layer. For the mixing layer (§ 3), the proposed CNM is benchmarked against the cluster-based Markov model. In § 4, CNM is performed for the three-dimensional actuated turbulent boundary layer featuring a more complex dynamics and § 5 summarizes this study and outlines future directions of research.

2. Cluster-based modelling

In this section, we propose a novel cluster-based reduced-order model (ROM) for the coherent structure dynamics starting at the time-resolved snapshots. In §§ 2.1 and 2.2 clustering and the cluster-based Markov model is recapitulated. Section 2.3 proposes a novel data-driven dynamic network resolving the transition dynamics between the clusters. In § 2.4, the time-discrete cluster-based ROM is enhanced for a continuous-time velocity prediction. The model validation includes the autocorrelation function of the flow as discussed in § 2.5. Figure 1 previews the methodology and will be explained later in the section. The relative advantages of CMM and CNM are illustrated for the Lorenz attractor in § 2.6.

2.1. *Clustering as coarse-graining*

We consider velocity fields in a steady domain Ω which may be obtained from experiments or from numerical simulations. Starting point is an ensemble of M statistically representative, time-resolved snapshots as employed for cluster-based models (Kaiser *et al.* 2014), Dynamic mode decomposition (DMD) (Rowley *et al.* 2009; Schmid 2010) or POD (see e.g. Holmes *et al.* 2012). The velocity field is equidistantly sampled with time step Δt, i.e. the mth instant reads $t^m = m\Delta t$. The corresponding snapshot velocity field is denoted by $\boldsymbol{u}^m(\boldsymbol{x}) := \boldsymbol{u}(\boldsymbol{x}, t^m)$, $m = 1, \ldots, M$.
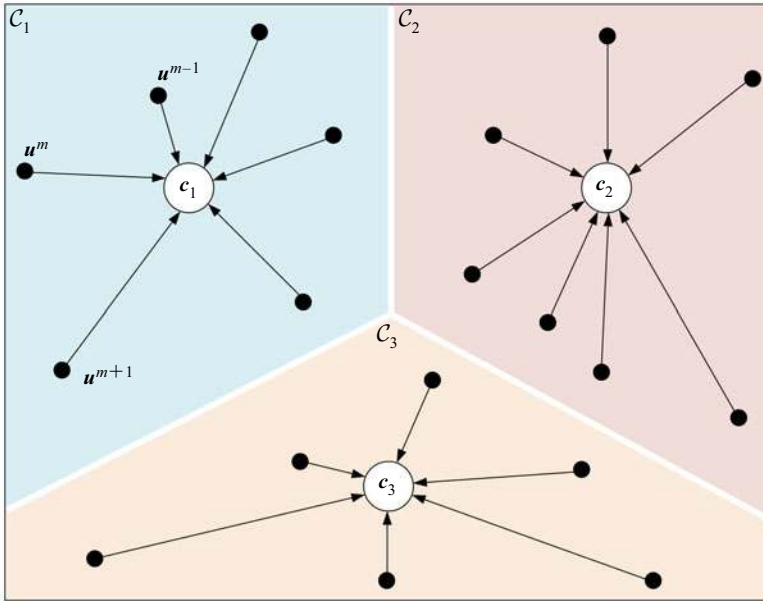
FIGURE 2. Clustering exemplified in a two-dimensional state space. The snapshots $\{u^m\}_{m=1}^M$ are coarse-grained into three clusters. Each centroid c_k, $k = 1, 2, 3$, is the averaged field of all snapshots belonging to this cluster. Every centroid c_k is associated with a Voronoi cell $\mathcal{C}_k$, i.e. a region in which the points are closer to c_k than any other centroid. The cluster affiliation for a snapshot is the cluster index of the closest centroid which has been indicated by an arrow. By definition, the cluster affiliation is the index of the corresponding Voronoi cell.

Cluster analysis lumps similar objects into clusters as illustrated in figure 2. This lumping of data is performed in an unsupervised manner, i.e. no advance labelling or grouping of the data has been performed. In cluster-based models, the M snapshots $u^m(x)$ are coarse-grained into K clusters represented by the centroids $c_k(x)$, $k = 1, \ldots, K$, using the unsupervised k-means++ algorithm (Steinhaus 1956; MacQueen 1967; Lloyd 1982). The centroids characterize the typical flow patterns of each cluster, also called modes in the ROM community. The corresponding cluster-affiliation function maps a velocity field u to the index of the closest centroid,

$$k(u) = \arg\min_i \|u - c_i\|_\Omega, \tag{2.1}$$

where $\| \cdot \|_\Omega$ denotes the standard Hilbert space norm in the domain Ω (see appendix A). This function defines cluster regions as Voronoi cells around the centroids

$$\mathcal{C}_i = \left\{ u \in \mathcal{L}^2(\Omega) \colon k(u) = i \right\}. \tag{2.2}$$

This function can also be employed to map the snapshot index m to the representative cluster index $k(m) := k(u^m)$. Alternatively, the characteristic function

$$\chi_i^m := \begin{cases} 1, & \text{if } i = k(m), \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

describes whether the mth snapshot is affiliated with the lth centroid. The latter two quantities are equivalent.

The performance of a set of centroids $\{c_k\}_{k=1}^K$ with respect to a given set of snapshots $\{u^m\}_{m=1}^M$ is measured by the average variance of the snapshots with respect to their closest centroid. The corresponding inner-cluster variance reads

$$J(c_1, \ldots, c_K) = \frac{1}{M} \sum_{m=1}^M \left\| u^m - c_{k(m)} \right\|_\Omega^2.$$ (2.4)

The optimal centroids $\{c_k^\star\}_{k=1}^K$ minimize this inner-cluster variance:

$$\left(c_1^\star, \ldots, c_K^\star\right) = \underset{c_1, \ldots, c_K}{\arg\min} J(c_1, \ldots, c_K).$$ (2.5)

The argument is indeterminate with respect to a re-ordering. For CMM, we chose the first cluster as the one with the highest population, i.e. the largest number of associated snapshots. The $(k + 1)$th cluster, $k > 1$, has the largest transition probability from the kth one.

The optimization problem (2.5) is solved by the k-means++ algorithm. The K centroids are initialized randomly and then iterated until convergence is reached or when the variance J is small enough; k-means++ repeats the clustering process 30 times and take the best set of centroids.

The number of snapshots n_k in cluster k is given by

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

The centroids are the mean velocity field of all snapshots in the corresponding cluster. In other words,

$$c_k = \frac{1}{n_k} \sum_{u^m \in C_k} u^m = \frac{1}{n_k} \sum_{m=1}^M \chi_k^m u^m.$$ (2.7)

In the following centroid visualizations, we accentuate the vortical structures by displaying the fluctuations $c_k - \bar{u}$ around the snapshot mean $\bar{u}$ and not the full velocity field c_k.

2.2. Cluster-based Markov model

We briefly recapitulate CMM by Kaiser *et al.* (2014) as our benchmark cluster-based reduced-order model. In CMM, the state variable is the cluster population $p = [p_1, \ldots, p_K]^T$, where p_i represents the probability of being in cluster i and the superscript T denotes the transpose. The transition between clusters in a given time step Δt^c is described by the transition matrix $P = (P_{ij}) \in \mathcal{R}^{K \times K}$. The superscript '$c$' refers to cluster-based model. Here, P_{ij} is the transition probability of moving from cluster j to cluster i. Let p^l be the probability vector at time $t^l = l\Delta t^c$, then the change in one time step is described by

$$p^{l+1} = Pp^l.$$ (2.8)

With increasing iterations, the iteration (2.8) converges to the asymptotic probability $p^\infty := \lim_{l \to \infty} p^l$. In a typical case, (2.8) has a single fixed point p^∞. For completeness,

a continuous form of Markov models with a new transition matrix $\boldsymbol{P}^c$ is mentioned

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \boldsymbol{P}^c \boldsymbol{p}. \tag{2.9}$$

From the time-continuous form (2.9), the time-discrete one (2.8) can be derived. The opposite is not generally true. In the following, no continuous Markov models are employed.

A CMM of the time-resolved snapshots starts with cluster affiliation (2.1) which can also be considered function of time $k(t)$. We refer to the original paper for the determination of P_{ij} from $k(t)$. The time step Δt^c is a critical design parameter for CMM. A good choice is a value where the transition from one cluster to the next is likely. If the time step is too small, the Markov model idles many times in each cluster for a stochastic number of times before transitioning to the next cluster. The model-based transition time may thus significantly deviate from the deterministic data-driven trajectories through the clusters. If the time step is too large, one may miss intermediate clusters. We choose $\Delta t^c = T/10$, where T is the characteristic period of the flow. On a circular limit cycle with uniform rotation, this value is optimal for $K = 10$ clusters, enforcing the transition from one cluster to the next in each time step.

In figure 3(a,c,e), the effect of the suboptimal time step is illustrated for the CMM of a uniform rotation $u_1 = \cos(2\pi t)$, $u_2 = \sin(2\pi t)$. Here, four clusters and a time step $\Delta t^c = 1/16$ are chosen. The probability of staying in the cluster during one time step is $P_{11} = P_{22} = P_{33} = P_{44} = 3/4$ and the transition probability to the next counter-clockwise neighbour is $P_{14} = P_{21} = P_{32} = P_{43} = 1/4$. Thus, the probability of staying in one cluster for l steps exponentially decays, P_{11}^l. In contrast, the uniform rotation commands that the state is exactly three time steps in one cluster before it leaves in the fourth step. This example motivates the proposed cluster-based reduced-order model, foreshadowed in figure 3(b,d,f) and explained in the following section.

2.3. *Cluster-based network model*

For CMM, the time step Δt^c is, as mentioned, an important design parameter. This design parameter can be avoided by the new proposed CNM. The key idea is to abandon the 'stroboscopic' view of CMM and focus on non-trivial transitions from cluster j to cluster i. These transitions are characterized by two parameters: the probability Q_{ij} and a time scale T_{ij}. Evidently, no time step is needed for the description and the assumption of a constant transition time is found to be much more aligned with shear flow modes than assuming an exponential decay of residence time. Moreover, it could be relaxed by assuming a probability distribution of transition times.

In the following, the transition probability and transition time are inferred from the cluster affiliation function $k(t)$. The continuous form is convenient for discussion. The time-discrete affiliation function $k(m)$ can be made continuous by taking the cluster of the snapshot which is closest in time:

$$k(t) = k\left(\arg\min_m |t - t_m|\right). \tag{2.10}$$

The nth transition time t_n of the cluster affiliation is recursively defined as the first discontinuity of $k(t)$ for $t > t_{n-1}$. Here, $t_0 = 0$. The transition time t_n satisfies

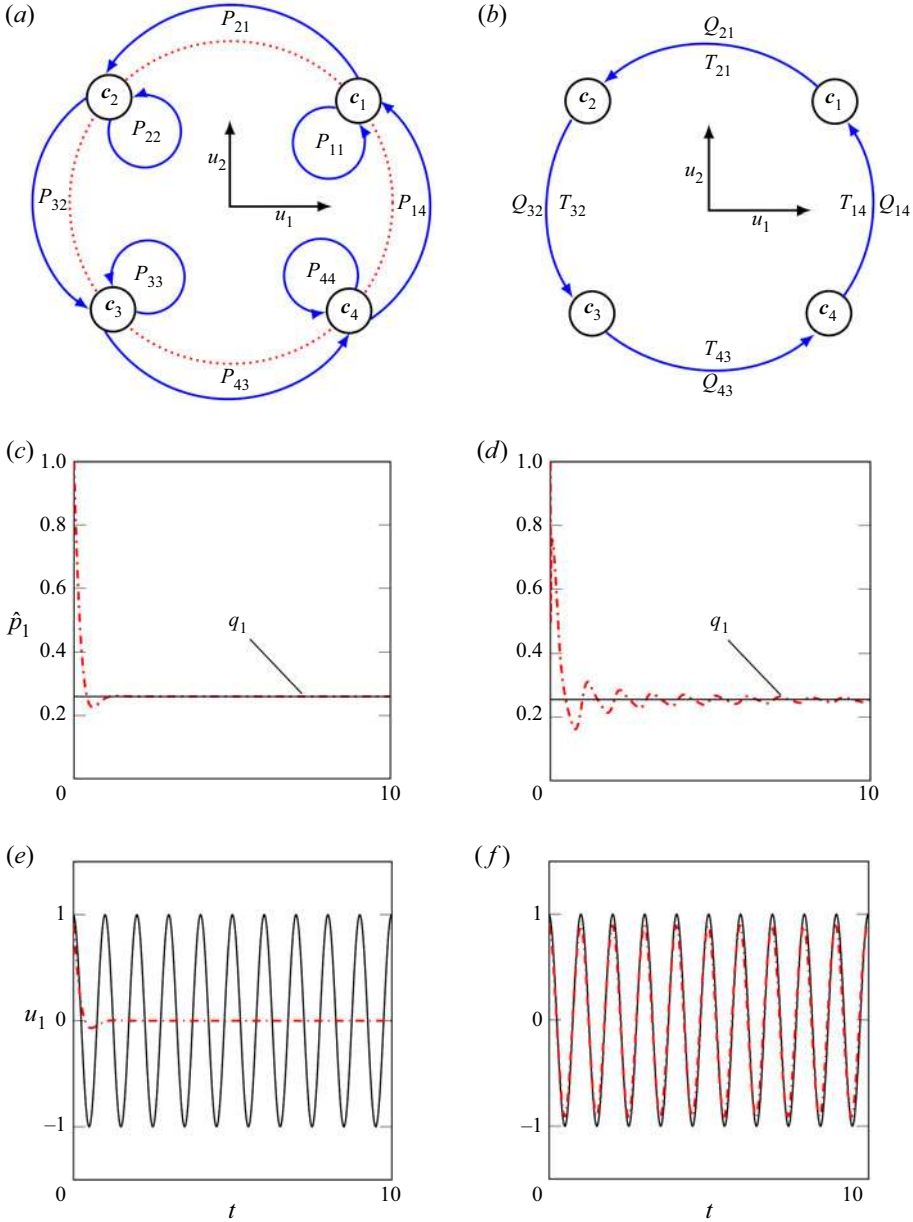$$k(t_n - \varepsilon) \neq k(t_n + \varepsilon) \tag{2.11}$$

FIGURE 3. Introduction of cluster-based models for a limit-cycle example. The CMM and CNM are displayed in (a,c,e) and (b,d,f), respectively. The uniform rotation $u_1 = \cos(2\pi t)$, $u_2 = \sin(2\pi t)$ in a two-dimensional plane is discretized by four centroids c_k, $k = 1, \ldots, 4$. (a) Phase portrait of CMM with time step $\Delta t = 1/16$. The centroids are near the limit cycle (red dashed line). The state vector residing in centroid c_i has the probability P_{ii} of staying in its state and P_{ji} of transitioning to centroid c_j in the considered time step. (b) Phase portrait of the CNM. The state in centroid c_i moves uniformly to its counter-clockwise neighbour taking a quarter period $T_{14} = T_{21} = T_{32} = T_{43} = 1/4$. Here, $Q_{14} = Q_{21} = Q_{32} = Q_{43} = 1$ and $Q_{ij} = 0$ otherwise. The estimated probability evolution starting in cluster $i = 1$ at $t = 0$ is illustrated for CMM (c) and CNM (d). Panels (e) and (f) present the model-based evolution of the first coordinate u_1 for CMM and CNM, respectively. In $(c–f)$, the solid black lines correspond to the uniform rotation and the dashed red line to the model.
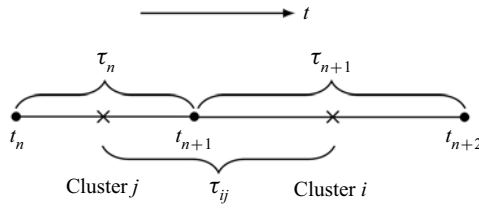
FIGURE 4. Sketch of times and periods employed in the cluster-based network model; × marks the centre of the cluster residence time, while • denotes the transition between clusters.

for any sufficiently small positive ε. For $t \in (t_n, t_{n+1})$, the data-based trajectory is assumed to stay in cluster k at the averaged time $(t_{n+1} + t_n)/2$ (see figure 4). The residence time in this cluster is defined by

$$\tau_n := t_{n+1} - t_n. \tag{2.12}$$

Let j and i be the indices of the clusters after t_n and t_{n+1} respectively. Then the transition time from j to i is defined as half of the residence time of both clusters:

$$\tau_{ij} := \frac{\tau_n + \tau_{n+1}}{2} = \frac{t_{n+2} - t_n}{2}. \tag{2.13}$$

This definition may appear arbitrary but is the least-biased guess consistent with the available data. The sum of all residence times from a given data set add up to the total investigated time period T_0.

The direct transition probability Q_{ij} and transition time T_{ij} can be inferred from the data. Then,

$$Q_{ij} = \frac{n_{ij}}{n_j}, \quad i, j = 1, \ldots, K, \tag{2.14}$$

where n_{ij} is the number of transitions from c_j to c_i and n_j the number of transitions departing from c_j regardless of the destination point

$$n_j = \sum_{i=1}^{K} n_{ij}, \quad i, j = 1, \ldots, K. \tag{2.15}$$

We emphasize that $n_{ii} = 0$ for $i = 1, \ldots, K$ by very definition of a direct transition. The direct transition matrix (DTM) $\boldsymbol{Q} = (Q_{ij}) \in \mathcal{R}^{K \times K}$ lumps these probabilities into a single entity.

Similarly, the direct transition time T_{ij} from cluster j to cluster i is taken to be the average of all values. This average is symbolically denoted by

$$T_{ij} = \langle \tau_{ij} \rangle. \tag{2.16}$$

These values are lumped into the matrix $\boldsymbol{T} = (T_{ij}) \in \mathcal{R}^{K \times K}$.

It should be noted that a given trajectory may repeatedly pass through the same clusters (Voronoi cells) with different residence and transition times. With enough data, this variability may be incorporated into the model. Our goal is to compare the Markov model with the most simple network model where constant (averaged) transition times are assumed.

The CNM predicts the asymptotic cluster probability p_i^∞ in cluster i. Let $[0, T_0]$ be a sufficiently long time horizon simulated by the model. Then, the probability of staying in cluster i is the cumulative residence time normalized by the simulation time

$$p_i^\infty = \frac{\sum \tau_i}{T_0}. \tag{2.17}$$

We return to the introductory example depicted in figure 3. The CNM is seen to accurately describe the uniform rotation (figure 3f) and correctly yields the asymptotic cluster probability $p_i^\infty = 1/4$, $i = 1, \ldots, 4$. In contrast, the prediction horizon of the CMM is limited to approximately one period. After this time, the initial condition is forgotten and the asymptotic distribution is reached – rendering CMM unsuitable for dynamic prediction. However, CMM predicts the asymptotic state faster than the CNM. For this particular example, the CMM could be made equivalent to the CNM by choosing $\Delta t^c = 1/4$. However, the Markov model will inevitably diffuse the state with a range of cluster transition times, e.g. for non-uniform rotation or more complex dynamics.

2.4. *Velocity fields associated with the cluster-based reduced-order models*

The CMM describes the cluster population

$$\boldsymbol{p} = \left[p_1, \ldots, p_K \right]^{\mathrm{T}} \tag{2.18}$$

at discrete times $t = l\Delta t^c$. In the following, this population is considered to be continuous in time, e.g. by using linear or higher-order interpolation. The corresponding velocity field $\boldsymbol{u}(\boldsymbol{x}, t)$ at time t is defined as the expectation value,

$$\boldsymbol{u}(\boldsymbol{x}, t) = \sum_{i=1}^{K} p_i(t) \boldsymbol{c}_i(\boldsymbol{x}), \tag{2.19}$$

where $\boldsymbol{c}_i$ is the ith centroid.

The CNM is based on centroid visits at discrete times. The clusters $k_0, k_1, k_2, \ldots$ are visited at times

$$t_0 = 0, \quad t_1 = T_{k_1 k_0}, \quad t_2 = t_1 + T_{k_2 k_1}, \ldots \tag{2.20}$$

consistent with the direct transition matrix (Q_{ij}) and the transition times T_{ij}. A uniform motion is assumed between these visits. In other words, for $t \in [t_n, t_{n+1}]$ the velocity field reads

$$\boldsymbol{u}(\boldsymbol{x}, t) = \alpha_n(t) \boldsymbol{c}_{k_n}(\boldsymbol{x}) + [1 - \alpha_n(t)] \boldsymbol{c}_{k_{n+1}}(\boldsymbol{x}), \quad \alpha_n = \frac{t_{n+1} - t}{t_{n+1} - t_n}. \tag{2.21}$$

We note that a smoother motion may be achieved with splines.

The actual flow computations are based on a lossless POD, as elaborated in appendix A. The interpolations are performed with the mode amplitudes $\boldsymbol{a} = [a_1, \ldots, a_N]^{\mathrm{T}}$ before being transcribed into velocity fields via the POD expansion.

Figure 5 compares the classical CMM with stroboscopic temporal prediction of discrete states and the proposed CNM with time-continuous uniform motion on a network of routes between two centroids. In the top row, the possible states are illustrated. In case of the CMM, the states (2.19), denoted by red dots, quickly converge to the mean flow, like Reynolds-averaged Navier–Stokes simulations. The CNM-predicted state (2.21) moves on the directed network, marked by red arrows, and is reminiscent of large-eddy simulations.
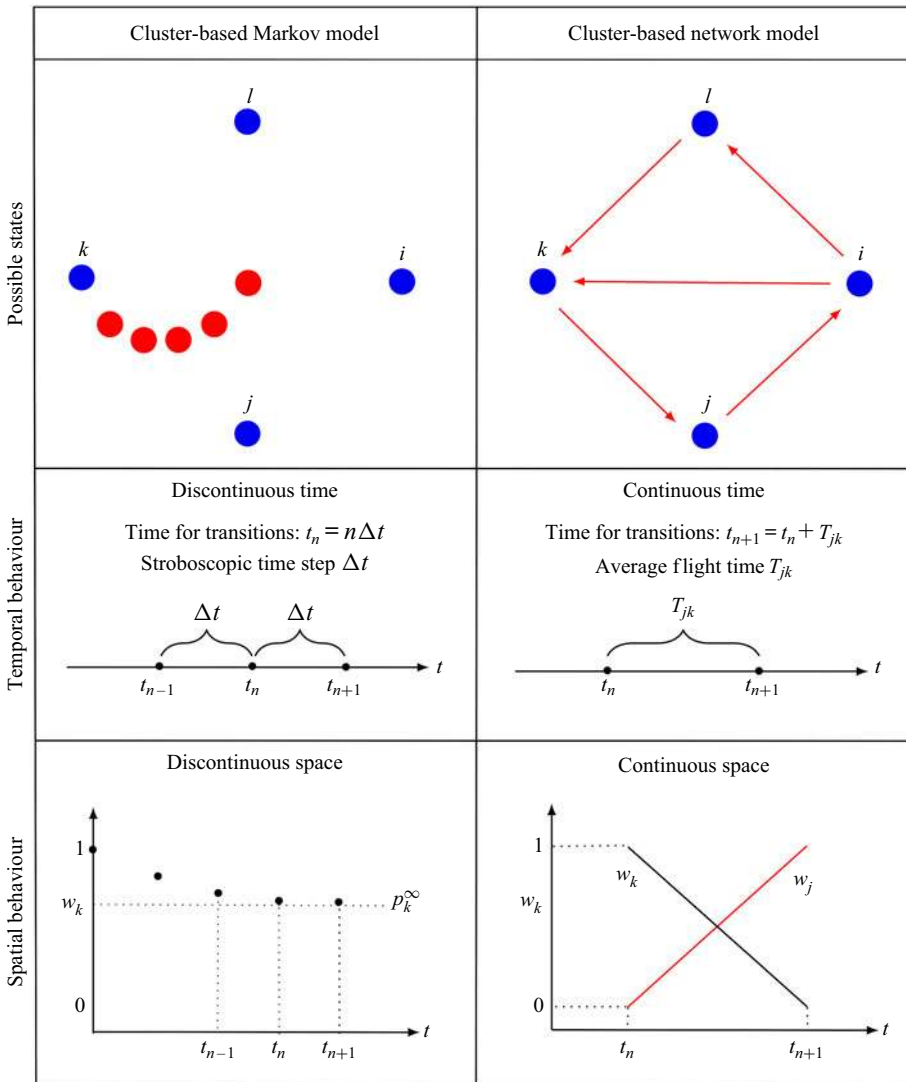
FIGURE 5. Comparison of the CMM and CNM. For reasons of simplicity, an example with four centroids i, j, k and l is shown. The CNM exemplifies the evolution of the weight of centroid k. For the CMM, the transition from cluster k to j is shown. The predicted state (2.21) is determined by the weights $w_k(t) = \alpha(t)$ and $w_j = 1 - \alpha(t)$, $\alpha = (t_{n+1} - t)/T_{jk}$ in the time interval $[t_n, t_{n+1}]$. For details, see text.

As displayed in the middle row, the CMM is discrete in time while the CNM dynamics is time continuous with cluster visits after pre-specified transition times T_{jk}. The bottom row shows another difference: CMM describes averages over all centroids while the CNM only allows for linear interpolations between two neighbouring centroids. This interpolation is consistent with the purpose of accurately resolving evolving coherent structures. Averaging over many centroids acts like a low-pass filter mitigating the fluctuation level and thus underresolving the coherent structures.

2.5. *Validation of the cluster-based reduced-order models*

Following Protas, Noack & Östh (2015), the cluster-based model is validated based on the computed and predicted autocorrelation function of the velocity field. The unbiased autocorrelation function reads

$$R(\tau) := \frac{1}{T - \tau} \int_{\tau}^{T} \int_{\Omega} \boldsymbol{u}(\boldsymbol{x}, t - \tau) \cdot \boldsymbol{u}(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t, \quad \tau \in [0, T]. \tag{2.22}$$

This function reveals the turbulent fluctuation level $R(0)$ and the frequency spectrum. Moreover, the problem of comparing two trajectories with finite dynamic prediction horizons due to the increasing phase mismatch is avoided (Pastoor *et al.* 2005).

In case of the CNM, the modelled autocorrelation function $\hat{R}$ is based on the modelled velocity field (2.21). In case of the CMM, the time integration quickly leads to the average flow and is not indicative of the range of possible initial conditions. Hence, K trajectories are considered starting with $p_k = 1$ for each cluster k, or, equivalently, $\boldsymbol{p}^{\circ k}(t = 0) = [\delta_{1k}, \ldots, \delta_{Kk}]^{\mathrm{T}}$. These cluster-specific autocorrelation functions are weighted with the cluster probability p_i^{∞}

$$\hat{R}(\tau) := \sum_{k=1}^{K} p_i^{\infty} \int_{0}^{T} \int_{\Omega} \boldsymbol{u}^{\circ k}(\boldsymbol{x}, t) \cdot \boldsymbol{u}^{\circ k}(\boldsymbol{x}, t + \tau) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t, \quad \tau \in [0, T], \tag{2.23}$$

where $\boldsymbol{u}^{\circ k}$ denotes the CMM-predicted velocity field starting in cluster k.

2.6. *Lorenz system as an illustrating example*

Following the original CMM paper by Kaiser *et al.* (2014), the CNM is illustrated for the celebrated Lorenz (1963) system, arguably the first demonstration of a chaotic dynamics in low-dimensional dynamics. The Lorenz system is a three-dimensional autonomous system of nonlinear ordinary differential equations. The derivation was inspired by a Galerkin model of Rayleigh–Bénard convection, but typically selected parameters clearly exceed the range of model validity (Sparrow 1982). The system features non-periodic, deterministic, dissipative dynamics associated with exponential divergence and convergence to a fractal strange attractor. The three coupled nonlinear differential equations read

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y - x), \tag{2.24a}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = x(r - z) - y, \tag{2.24b}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - bz, \tag{2.24c}$$

with the system parameters $\sigma = 10$, $b = 8/3$ and $r = 28$. For these parameters, there are three unstable fixed points at $(0, 0, 0)$ and $(\pm\sqrt{72}, \pm\sqrt{72}, 27)$, denoted by F^{+} and F^{-}, respectively. The attractor of the Lorenz system resembles two butterfly wings around F^{+} and F^{-} in phase space. The trajectory typically oscillates for several periods with increasing amplitude around a fixed point (F^{+} or F^{-}) before it moves to the other wing. The number of revolutions made on either side varies unpredictably from one cycle to the next.

The Lorenz equations (2.24) are solved employing an explicit fourth-order Runge–Kutta scheme with an initial condition on the attractor. The time-resolved snapshot data $\boldsymbol{x}(t_m)$
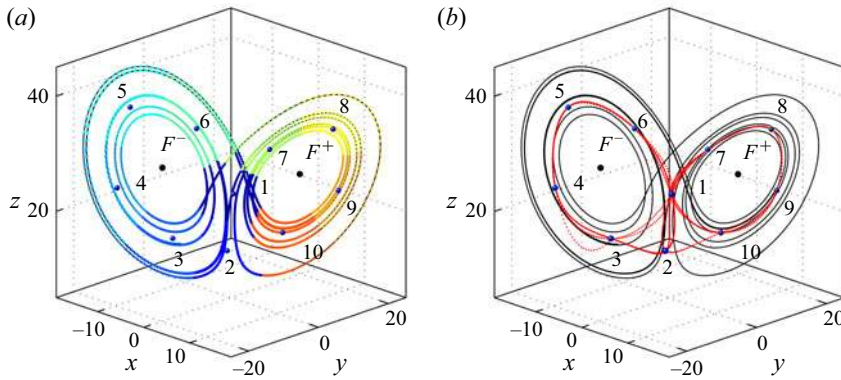
FIGURE 6. Cluster-based network model for the Lorenz attractor. (*a*) Cluster partitioning: the centroids are displayed as coloured solid circles. A trajectory is illustrated by a black curve. The dots on this trajectory are coloured according to their cluster affiliation. The clusters $k = 3, 4, 5, 6$ oscillate around the fixed point F^- and clusters $k = 7, 8, 9, 10$ around F^+. The clusters $k = 1, 2$ connect both 'ears' of the Lorenz attractor. (*b*) The trajectory of the CNM (red dashed line). The centroids represent the network nodes and edges represent possible transitions. Here, trajectory of CNM is obtained by a spline interpolation through the visited centroids.

with $\mathbf{x} = [x, y, z]^T$ are collected at a sampling time step $\Delta t = 0.005$ corresponding roughly to one thousandth of a typical oscillation period. The k-means++ algorithm partitions $M = 1\,000\,000$ snapshots into $K = 10$ clusters. Figure 6(*a*) displays a phase portrait of the corresponding clusters. The snapshots associated with one cluster are highlighted by the same colour. The 10 clusters feature three different subsets: the transition clusters $k = 1, 2$ between two butterfly wings, the F^- wing related cluster $k = 3, 4, 5, 6$ and clusters $k = 7, 8, 9, 10$ associated with the F^+ wing. The wing-related cluster groups represent approximately $90°$ phase bins and do not resolve the amplitude. Evidently, the 10 clusters are coarse representations of the state.

In the following, the dynamics is resolved by the network model of § 2.3. The 10 centroids are considered as nodes in the network. The transition between these centroids define directed edges characterized by direct transition matrix $\mathbf{Q}$ and the flight times $\mathbf{T}$. The connectivity is described by the adjacency matrix $H(\mathbf{Q})$ where H denotes the Heaviside function: non-vanishing elements of $\mathbf{Q}$ are replaced by unity (Newman 2010). Figure 7, displays the DTM $\mathbf{Q}$ (figure 7*a*) and associated transition time matrix $\mathbf{T}$ (figure 7*b*). The matrices reveal three distinct cluster groups consistent with the phase diagram of figure 6. Clusters 1 and 2 allow transitions to 3 and 10, i.e. the F^- and F^+ wing, respectively, and have been called flipper clusters by Kaiser *et al.* (2014). The cluster transition sequence $3 \to 4 \to 5 \to 6 \to 1 \to 2 \to 3$ is the dominant cyclic group associated with the F^- wing. Another cyclic group skips cluster 2: $3 \to 4 \to 5 \to 6 \to 1 \to 3$. A cyclic group through the F^+ wing reads $10 \to 9 \to 8 \to 7 \to 1 \to 10$. A longer sequence includes the second cluster: $10 \to 9 \to 8 \to 7 \to 1 \to 2 \to 10$. The transition times in the wing centroids are noticeably smaller than the passage through the flipper clusters.

Figure 8 compares the asymptotic population $\mathbf{p}^\infty$ predicted by CNM with the population from a long-term simulation. The CNM statistics are based on 20 000 transitions while the integration of the Lorenz equations is performed over 5000 time units. Both statistics correspond to the roughly 800 periods found to be sufficient for an accurate statistics. The relative error of the CNM is no more than approximately 10 %. This error does not decrease with much larger integration times, but is linked to the coarse-graining of the
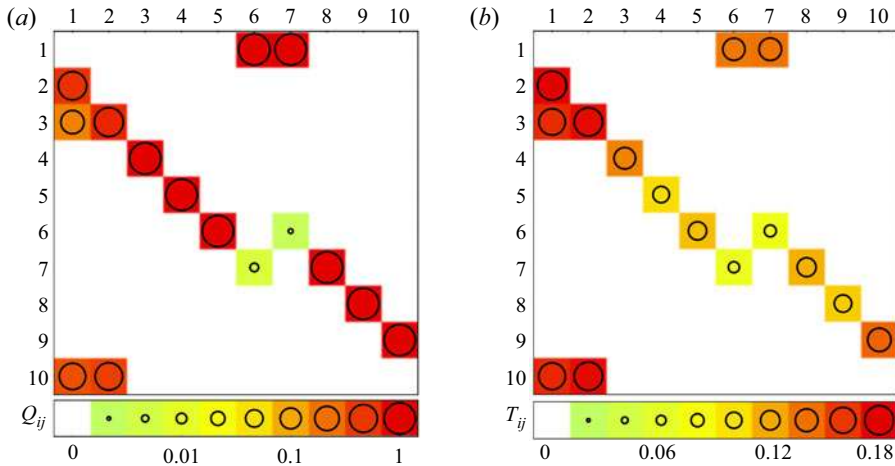
FIGURE 7. Cluster-based network model of the Lorenz attractor. (*a*) Direct transition matrix (Q_{ij}). (*b*) Transition time (T_{ij}).
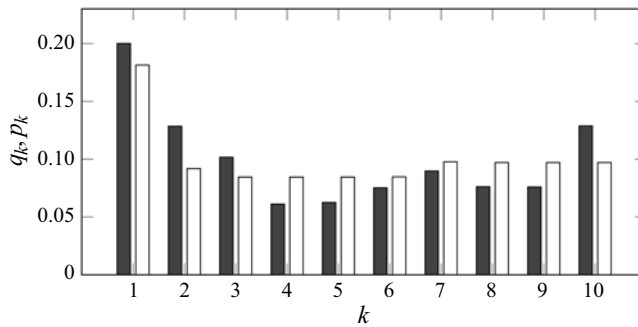

FIGURE 8. Cluster probability distribution of the Lorenz system (solid rectangles) and the corresponding cluster-based network model (open rectangles). The model results are based on 20 000 transitions.

state to Voronoi cells around the centroids. The assumed constant transition time τ_{ij} for all trajectories from cluster j to cluster i is a crude assumption. In fact, the transition times can vary by a large factor and can thus give rise to significant systematic errors. A more accurate transition model may, for instance, include earlier transitions for a more realistic representation of the trajectory. Intriguingly, the CMM has an error of only 0.5 % which is one order of magnitude lower. Due to the stroboscopic monitoring of the CMM states, no estimates of the transition times are required and one source of systematic errors is excluded by construction.

A distinguishing feature of the CNM is the resolution of the temporal dynamics illustrated in figure 9. The evolution of the model-based trajectory is hardly distinguishable from the one obtained by numerical integration. Smoothness of the CNM trajectory has been achieved by splines connecting the states between two consecutive centroid visits. Yet, the oscillatory amplitude growth in both wings cannot be resolved with this low cluster-based resolution. The CNM can only follow the simulations for a short time period, as nearby trajectories exponentially diverge with Lyapunov exponent 2.16 (Wolf *et al.* 1985) and the initial separation in each cluster is already large. Yet, the fluctuation
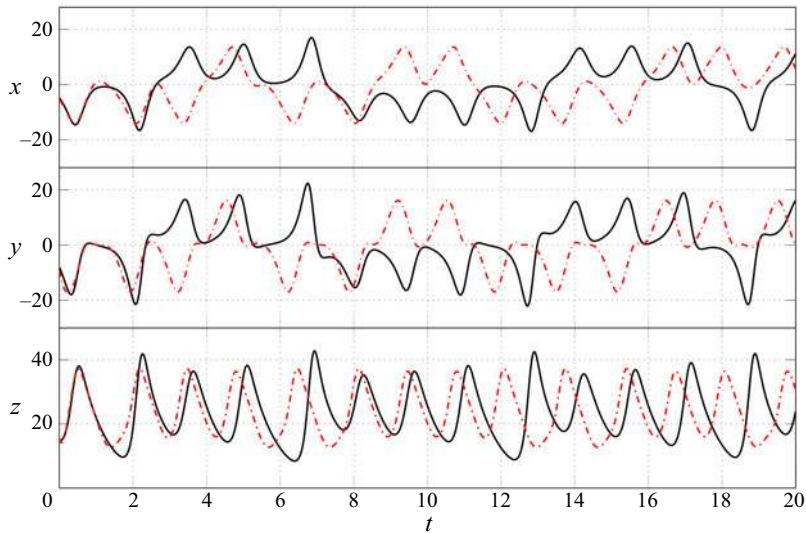
FIGURE 9. Evolution of the Lorenz system x, y, z, $t \in [0, 20]$ from integrating the dynamical system (black solid line) and from the prediction of the cluster-based network model (red dashed line).
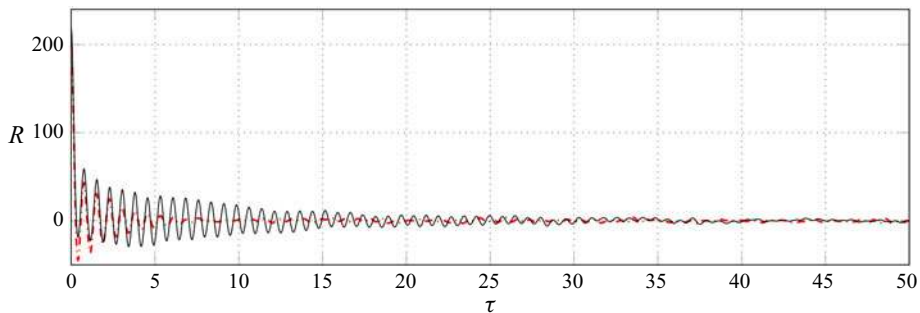

FIGURE 10. Autocorrelation function of the Lorenz system (black solid line) and the cluster-based network model (red dashed line).

amplitude, frequency content and bi-modality are well reproduced. A CNM with $K = 100$ clusters yields more realistic dynamics but require orders of magnitude more simulation data. At the least-order extreme, a CNM with two or three clusters only coarsely resolves the transitions between both ears of the Lorenz attractor, not the growing oscillations in each ear.

Finally, the autocorrelation of the simulation (black solid curve) and the CNM (red dashed curve) are presented for aggregate comparison in figure 10. CNM roughly reproduces the fast oscillatory decay of the autocorrelation function in the first five periods.

3. Cluster-based reduced-order modelling of the mixing layer

In this section, the cluster-based models are applied to a two-dimensional incompressible mixing layer with Kelvin–Helmholtz vortices undergoing vortex pairing. The flow configuration of the mixing layer and the employed direct Navier–Stokes

solver is presented in § 3.1. In § 3.2, the dominant flow features of the mixing layer are presented. Then (§ 3.3), the snapshots of incompressible mixing layer are coarse-grained into centroids. Following Kaiser *et al.* (2014), a cluster-based Markov model (CMM, § 3.4) is developed as a benchmark for the proposed network model (CNM, § 3.5).

3.1. *Flow configuration and direct numerical simulation*

The two-dimensional incompressible mixing layer and a velocity ratio of $3:1$ is considered as the test plant in this paper. The velocity ratio is a common choice in the literature (Comte, Silvestrini & Bégou 1998; Noack *et al.* 2005; Kaiser *et al.* 2014). The high- and low-speed streams have velocities U_1 and U_2, respectively. The convection velocity U_c of coherent structure is well approximated by the average velocity (Monkewitz 1988)

$$U_c = \frac{U_1 + U_2}{2}. \tag{3.1}$$

The initial vorticity thickness is denoted by δ_0. The Newtonian fluid is characterized by the density ρ and kinematic viscosity ν. The flow characteristics are described by the Reynolds number based on the convection velocity $Re = U_c\delta_0/\nu$ and velocity ratio. In the sequel, all quantities are assumed to be non-dimensionalized with the initial vorticity thickness δ_0, the high-speed velocity U_1 and the density ρ. The Reynolds number is set to 200.

The flow is described in a Cartesian coordinate system (x, y) with the origin at the maximum gradient location of the inlet profile. The x-axis points in the streamwise direction and the y-axis points in the direction of the high-speed stream. The velocity components in the x- and y-directions are denoted by u and v, respectively.

Figure 11 illustrates the rectangular computational domain

$$\Omega := \{(x, y) \in \mathcal{R}^2 : 0 \le x \le 80 \wedge |y| \le 15\} \tag{3.2}$$

with 10 237 nodes and 2248 triangular elements. The location vector is denoted by $\boldsymbol{x} = (x, y)$. Similarly, the velocity vector is denoted by $\boldsymbol{u} = (u, v)$. The inlet velocity profile reads

$$u = 2 + \tanh\left(\frac{2y}{\delta_0}\right), \quad v = 0, \quad \text{where } \delta_0 = 1. \tag{3.3}$$

The Kelvin–Helmholtz vortices are triggered at the inlet by a stochastic perturbation of the u-component for $y \in [-2, 2]$ with a standard deviation of $0.01U_c$.

The streamwise extent of the domain is 80. This corresponds to a downwash time of 40 given the convection velocity of 2. This is the minimum time for the transient time as all initial interior vortices will leave the domain. A simulation over 400 convective units corresponds to 10 downwash times. This period is found to be sufficient for a good statistics of the mean value and fluctuation level. One simulation yields $M = 10\,000$ velocity snapshots $\boldsymbol{u}^m(\boldsymbol{x}) = \boldsymbol{u}(\boldsymbol{x}, t^m)$, where the sampling times $t^m = 0.04$ m start with $t = 0$ in the converged post-transient phase. The sampling frequency 25 is two orders of magnitude larger than the dominant shear-layer frequency of $f = 0.1075$ in the most active downstream region.

An in-house direct numerical simulation solver was employed to simulate the incompressible mixing layer. This solver is based on the finite-element method with third-order Taylor–Hood elements and implicit third-order time integration. The solver has been used for numerous configurations, like the cylinder wake (Noack *et al.* 2016), the mixing layer (Shaqarin, Noack & Morzyński 2018) and the fluidic pinball (Ishar *et al.* 2019), to name only a few.
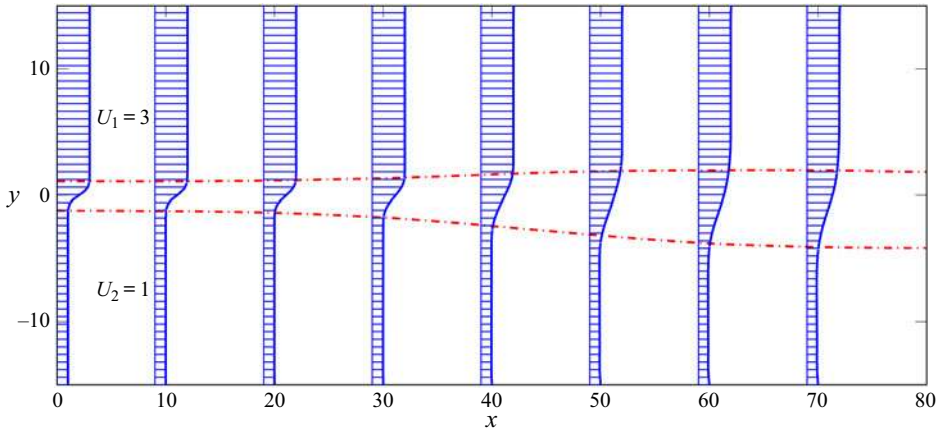
FIGURE 11. Numerical simulation sketch of incompressible mixing layer. An unperturbed tanh velocity profile $u(y) = 2 + \tanh(2y)$ is patched in the inlet of a rectangular domain. Time-averaged streamwise velocity profiles separated by $\Delta x = 10$ are visualized by blue lines. The red dashed curves mark the mixing-layer thickness. We chose the 90 % thickness of the profile starting with the average velocity, i.e. $u = 2.9$ and $u = 1.1$.
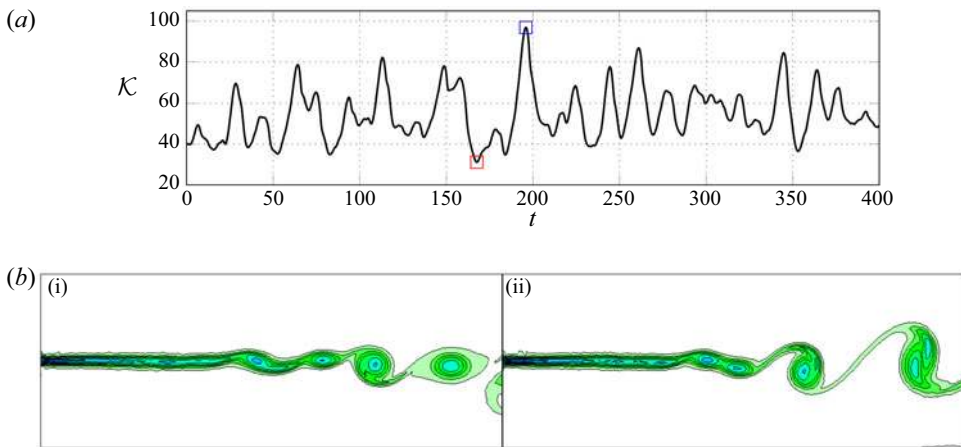

FIGURE 12. Mixing-layer simulation. (*a*) Energy fluctuation over time with the maximum marked by a blue square and minimum by a red square. (*b*) Vorticity fields associated with the maximum and minimum fluctuation energies. The minimum (*b* i) corresponds to a K–H vortex at $t = 167.60$, The maximum (*b* ii) features vortex pairing at $t = 195.92$. The curves represent the isolines of vorticity. Higher values corresponds to darker green areas.

3.2. *Flow features*

The incompressible mixing layer exhibits two typical behaviours. First, the initial dynamics is characterized by the roll-up of vorticity originating from the Kelvin–Helmholtz (K–H) instability (see figure 12*b* i). Second, these vortices pair further downstream, as can be seen at the outlet region of figure 12(*b* ii). This vortex pairing contributes to the mixing layer growth. The location of vortex pairing may change significantly in time. Upstream (downstream) vortex pairing is associated with high (low) fluctuation energy.
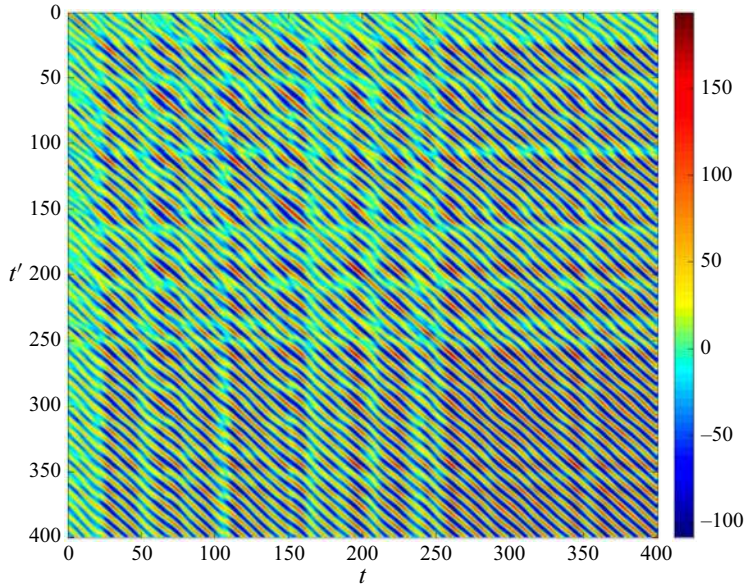
FIGURE 13. Autocorrelation matrix (3.4) of the mixing layer for $t, t' \in [0, 400]$. The value is presented in the colour bar. The plot is based on 401 snapshots collected at uniform time steps $\Delta t = 1$.

The time-averaged velocity field in figure 11 shows the mixing layer growth. The velocity thickness is visualized by a red dashed line and is defined as the distance between transverse locations where the mean streamwise velocity was equal to $U_1 - 0.05\Delta U$ and $U_2 + 0.05\Delta U$. The mixing-layer thickness increases significantly between $x = 30$ and $x = 60$. Here, vortex pairing leads to this thickness increase.

The temporal dynamics may be inferred from the evolution of the fluctuation energy in figure 12(*a*). The fluctuations indicate narrow bandwidth oscillatory behaviour. More refined insights may be gained from the correlation function between the flows at time t and t',

$$C(t, t') = \int_{\Omega} \boldsymbol{u}'(\boldsymbol{x}, t) \cdot \boldsymbol{u}'(\boldsymbol{x}, t') \, \mathrm{d}\boldsymbol{x}. \tag{3.4}$$

Figure 13 illustrates the autocorrelation matrix for $t, t' \in [0, 400]$. The fluctuation energy of figure 12(*a*) is quantified in the diagonal, $\mathcal{K}(t) = C(t, t)/2$. The wavy pattern indicates oscillatory coherent structures. The changes from pure periodicity are caused by vortex pairing at a large range of streamwise locations.

3.3. *Clustering*

Both considered reduced-order models are based on the direct numerical simulation of the two-dimensional incompressible mixing layer described in § 3.1; $M = 10\,000$ velocity field snapshots of the post-transient phase are sampled with a time step $\Delta t = 0.04$.

The computational load of clustering is significantly reduced by an effectively lossless POD compression detailed in appendix A. In fact, all operations are performed on the POD amplitude vector $\boldsymbol{a} = [a_1, a_2, \ldots, a_N]^{\mathrm{T}}$ instead of the snapshots.

The M snapshots are clustered with the k-means++ algorithm into $K = 10$ centroids. This number is small enough to allow for the physical interpretation of all centroids
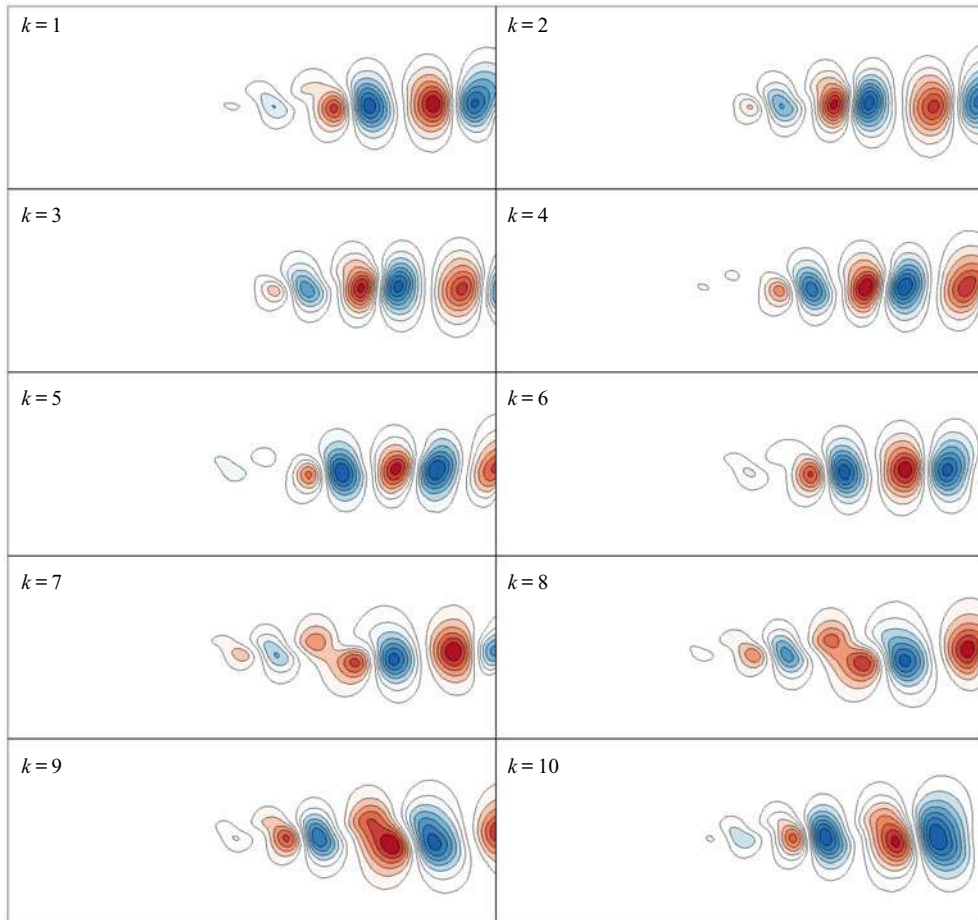
FIGURE 14. The cluster centroids c_k, $k = 1, \ldots, 10$, of the mixing layer. The transverse velocity fluctuation is depicted with contour lines. Red and blue regions mark positive and negative values.

and all transitions but large enough for a meaningful reduced-order model. Figure 14 illustrates the transverse velocity fluctuation of the centroids. The first six centroids show the streamwise convection of K–H vortices, while the next four centroids resolve a vortex pairing (VP) event. In centroid 7, two vortices merge at the beginning of the vortex chain. In the following three centroids, the merging is completed and leads to a large vortex. Note that the VP centroids $k = 7, 8, 9, 10$ have pronounced vortices at a similar position as the K–H centroids $k = 4, 5, 6$, respectively. The structures of the K–H and VP centroids are noticeably different. The main vortices of the K–H centroids are elliptical and the major axis is rotated in clockwise direction, i.e. the upper part of the vortices follow the faster stream. In contrast, the main elliptical vortices of VP centroids are rotated in mathematically positive direction, i.e. the upper part of these vortices move upstream with respect to their centre.

The centroids represent characteristic stages in the mixing-layer dynamics as can be elucidated in a proximity map. This map reflects the configuration matrix $\boldsymbol{D} = (D_{ij}) \in$
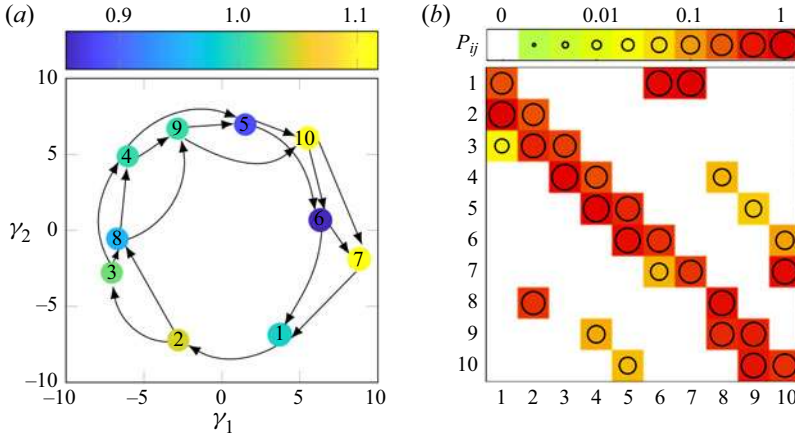
FIGURE 15. Cluster-based Markov model of the mixing layer. (*a*) Proximity map of centroids. Each centroid is marked by a solid coloured circle. The colour denotes the relative energy content (see colour bar on top). Unity corresponds to the average value. (*b*) Transition matrix. The probability value is displayed by the background colour and the radius of the corresponding circle.

$\mathcal{R}^{K \times K}$ comprising the distance between two centroids

$$D_{ij}^c := \left\| c_i - c_j \right\|_\Omega, \quad i,j = 1, 2, \ldots, K. \tag{3.5}$$

Following Kaiser *et al.* (2014), the proximity map is used to represent the configuration matrix **D** in a two-dimensional feature space $\gamma \in \mathcal{R}^2$ optimally preserving the relative distances. The proximity map employs classical multidimensional scaling (Mardia, Kent & Bibby 1979). Figure 15(*a*) displays centroids close to a circle which is characteristic of vortex shedding.

3.4. *Markov model*

The temporal mixing-layer evolution is characterized by the cluster transition matrix **P** illustrated in figure 15(*b*); P_{ij} represents the probability of moving from cluster j to i in one forward time step. Here, we choose a time step $\Delta t^c = T/10 = 1$ where the $T = 10$ is the dominant period of the evolved mixing layer.

The cluster transition matrix reveals two cyclic groups. The first group $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 1$ is consistent with the convection process of the K–H vortex shedding observed in the centroid visualization. This periodic process corresponds to a nearly uniform clockwise rotation in the proximity map. The second cyclic group $8 \rightarrow 9 \rightarrow 10 \rightarrow 7 \rightarrow 1 \rightarrow 2 \rightarrow 8$ comprises VP centroids $k = 8, 9, 10$ and shares two centroids with the K–H regime. This dynamics also leads to a nearly uniform clockwise rotation in the feature space. There are also transitions from the VP to K–H regime, e.g. $8 \rightarrow 4$, $9 \rightarrow 5$, $10 \rightarrow 6$ and $10 \rightarrow 7$ and in the opposite direction. All these transitions are between similar centroids of both groups. From the cluster index the orientation of the main elliptical vortices can be inferred. For $k \leq 6$ ($k \geq 7$), the upper part of the vortices are displaced in (against) the direction of the flow with respect to their centres.

The evolution of the cluster population vector p^l at $t = l\Delta t^c$ is investigated by iterating equation (2.8). Figure 16 compares the probability distribution of DNS data and the model-based asymptotic vector p^∞. The agreement is astonishingly good for such a
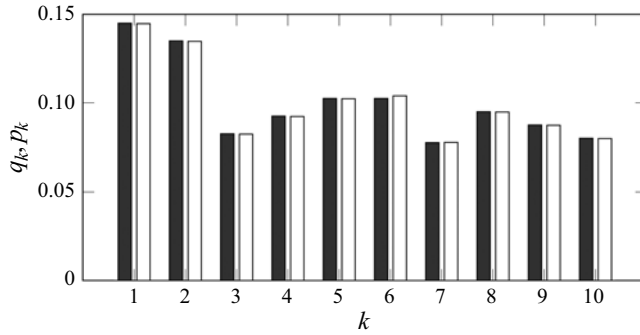
FIGURE 16. Cluster probability distribution of the mixing layer from the DNS data and the cluster-based Markov model. Solid rectangles denote the probability q_k from the DNS data. Open rectangles represent asymptotic values from the CMM after $l = 35$ iterations.

low-order model. The probability vector converges quickly to a unique, stationary probability distribution near $t = 20$.

In figure 17, the dynamics of CMM is illustrated for the first cluster probability p_1 and the first POD mode amplitude a_1 inferred from the flow state (2.19). Starting point is direct numerical simulation starting at $t = 0$ close to the first cluster c_1 which corresponds to the probability vector $p = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$. The probability and POD mode amplitude of CMM show a convergence after around $l = 35$ iterations or, equivalently, $t \approx 35$. The solid horizon line denotes q_1, i.e. the population of the first cluster from DNS data. The POD mode amplitude a_1 performs three strongly damped oscillations before vanishing.

Figure 17(c) shows an oscillating quickly decaying autocorrelation function of CMM which is consistent with the observations for a_1 and p_1. In contrast, the autocorrelation function associated with the DNS keeps oscillating around with an amplitude around 50 % of the average fluctuation level. This level indicates that half of the fluctuation energy resides in repeating oscillatory flow structures while the other half is of non-repeating stochastic nature.

3.5. *Network model*

In this section, a CNM is developed using the same snapshot data and same centroids. Starting point for the dynamic network is the cluster-affiliation function $k(t)$. Following § 2.3, the direct cluster transition matrix Q with associated average transition times T are derived. Figure 18 illustrates both matrices. These matrices have the almost same structure as the Markov model except for the diagonal elements which are vanishing by design. In other words,

$$Q_{ii} = T_{ii} = 0, \quad \forall i \in \{1, \ldots, K\}, \tag{3.6a}$$

$$H[P_{ij}] = H[Q_{ij}] = H[T_{ij}], \quad \forall i, j \in \{1, \ldots, K\} \wedge i \neq j, \tag{3.6b}$$

with H being again the Heaviside function. Vanishing diagonal elements (3.6a) arise from the requirement of non-trivial transitions. Theoretically, the trajectory may terminate in a cluster, like in a stable fixed point of a linear dynamical system. This case is not compatible with the goal to model a well-resolved non-trivial attractor and shall be ignored in this study. Equation (3.6b) requires a sufficiently small time step of the CMM. Otherwise, the
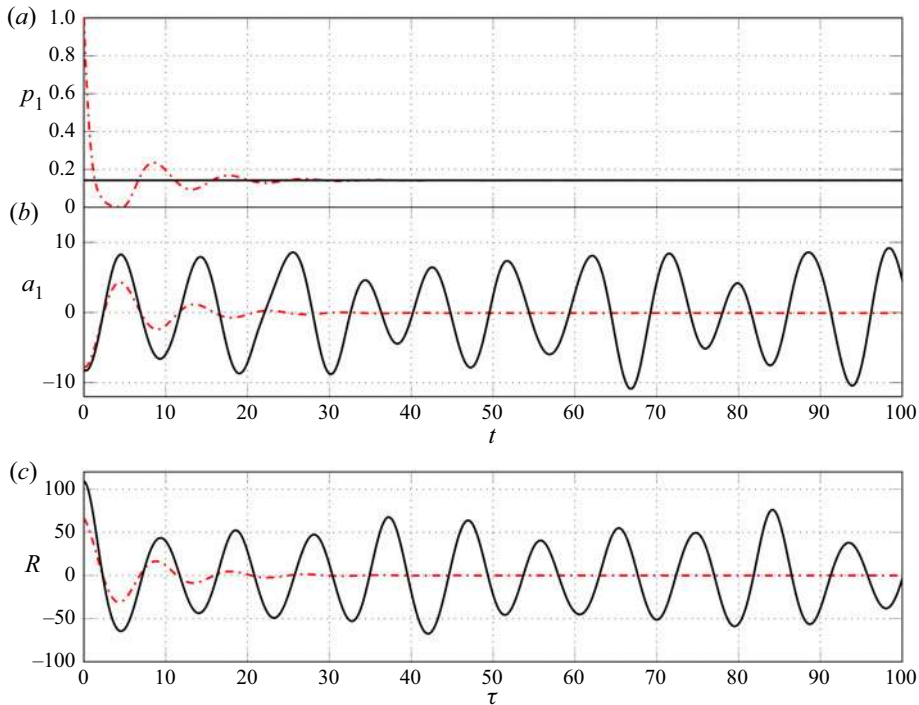
FIGURE 17. Dynamics of the mixing layer from DNS and the cluster-based Markov model. (*a*) Probability evolution for DNS (black solid line) and for CMM (red dashed line). The probability of the first cluster p_1 quickly converges to p_i^∞ around $t = 35$. The corresponding DNS value is 0.1325 and is represented by a horizontal line. (*b*) The evolution of the first POD mode amplitude a_1 for DNS (black solid line) and CMM (red dashed line). (*c*) The autocorrelation function for DNS (black solid line) and for CMM (red dashed line).

stroboscopic view on the trajectory may miss a crossing of an intermediate cluster. This happens with the transition from $1 \to 2 \to 3$ in one CMM time step Δt^c. Hence, $P_{31} \neq 0$ while $Q_{31} = 0$. However, this is a rare event, as indicated by the small value of P_{31}.

An inspection of $\boldsymbol{T}$ reveals that the transition time between K–H and VP centroids is relatively small. This is consistent with the closeness of the corresponding centroids in the proximity map (figure 15*a*). An exception is the transition between K–H centroid 2 to VP centroid 8 which are well separated in the proximity map. Intriguingly, the transitions within the K–H and VP regimes are also strongly correlated with the distances depicted in the proximity map. For instance, the smallest (largest) inner-regime transition from centroid 6 to 7 (8 to 9) is associated with a small (large) distance in the proximity map. The physical interpretation of the cycle-to-cycle variations of the CMM persist for CNM.

In the following, the temporal dynamics of CNM is investigated based on the identified centroids c_k, the description of their connectivity DTM $\boldsymbol{Q}$ and their flight times $\boldsymbol{T}$. Like a POD model, CNM is a grey-box model resolving the temporal dynamics and the associated coherent structures. We choose cluster $k = 1$ as initial condition for DNS and for the CNM and integrate over $l = 20\,000$ transitions. In figure 19, the asymptotic cluster population $\boldsymbol{p}^\infty$ from (2.17) is compared with $\boldsymbol{q}$ from the DNS. The discrepancies of few per cent seem expectable and tolerable for a 10-cluster model. This difference is not cured by increasing the amount of transition data in CNM. Intriguingly, the probability distribution of the
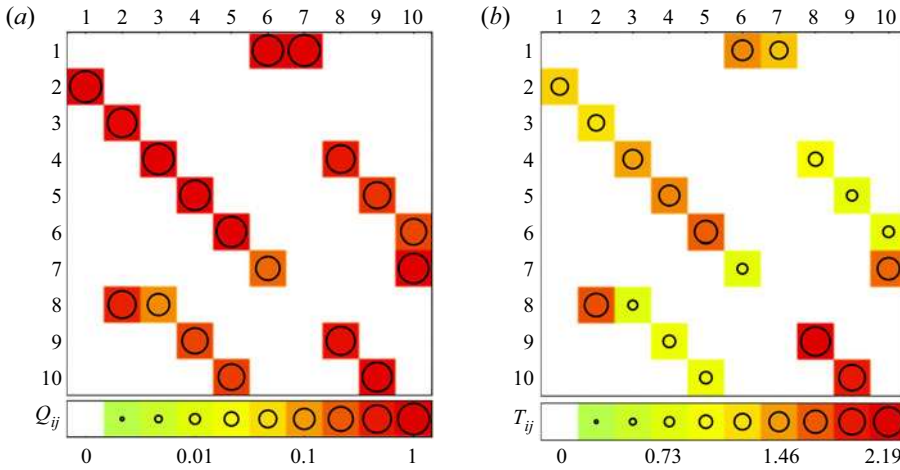
FIGURE 18. Dynamics of the cluster-based network model for the mixing layer. (*a*) Direct transition matrix. (*b*) Averaged transition time. The non-vanishing values are denoted by the circle radius and the colour code from the bottom caption.
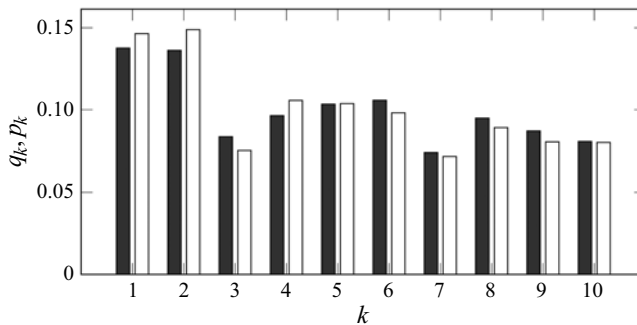

FIGURE 19. Cluster probability distribution of the mixing layer from DNS (solid rectangles) and cluster-based network model (open rectangles). The modelled values are obtained from simulating 20 000 clusters transitions.

CMM displayed in figure 16 is significantly more accurate. This behaviour can be linked to the simple transition time estimate which employs one single average value for a large range of observed transition times. We have developed more refined and more accurate transition time estimates leading to much better agreements of the cluster probability distributions. The price is increased complexity of the CNM which we deemed not helpful for our first publication.

Figure 20 shows the evolution of the first two POD mode amplitudes (red dashed curve). The CNM tracks well the amplitude and phase of the DNS over 100 time units. As for the Lorenz system, the temporal evolution is smoothed by a spline and does not use the non-smooth uniform motion between two consecutive centroid visits.

Figure 21 compares the autocorrelation function of the CNM and the DNS. We intentionally do not normalize this function to reveal the resolved fluctuation level at vanishing time delay. As expected, the model-based fluctuation level is significantly lower than the DNS value. This difference is quantified by the unresolved inner-cluster variance. Intriguingly, CNM and DNS functions become already similar after half a period. The
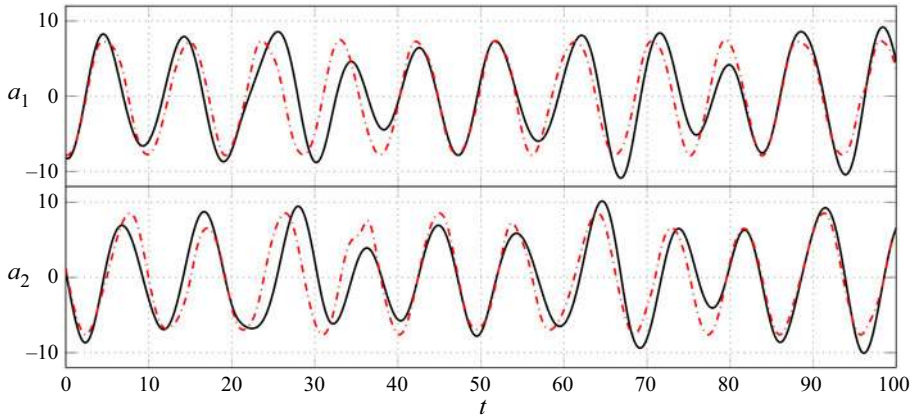
FIGURE 20. Evolution of mode amplitudes a_1, a_2 for the mixing layer $t \in [0, 100]$. The curves correspond to DNS (black solid line) and the cluster-based network model (red dashed line).
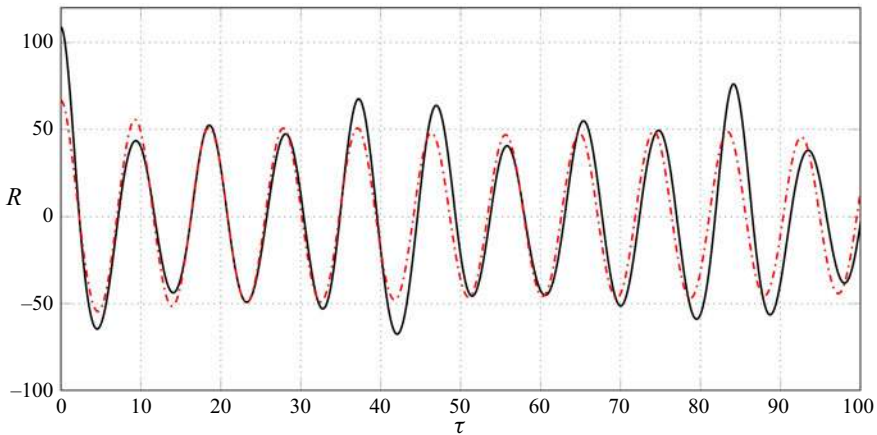

FIGURE 21. Autocorrelation function of the mixing layer for DNS (black solid line) and cluster-based network model (red dashed line).

asymptotic fluctuation level represents coherent structures which are well resolved by the chosen centroids and serve as coarse-grained recurrence points of the DNS. Due to the dominant oscillatory dynamics, the autocorrelation does not vanish with increasing time. The good reproduction autocorrelation function is *a posteriori* justification for the chosen cluster number.

4. Cluster-based network modelling of the actuated turbulent boundary layer

In this section, the cluster-based network modelling is implemented on a three-dimensional actuated turbulent boundary layer. First (§ 4.1), the flow configuration and the large-eddy simulation is described. The clustering results, which follow the same coarse-graining approach as for the shear layer, are presented in § 4.2. A cluster-based network model is developed and assessed in § 4.3.
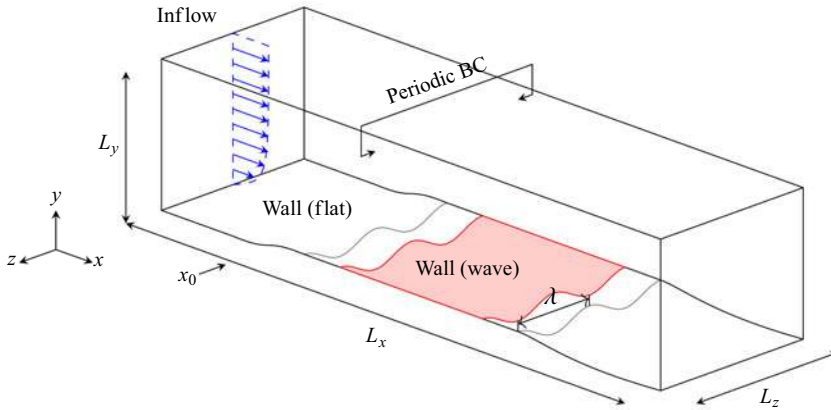
FIGURE 22. Overview of the physical domain of the actuated turbulent boundary layer flow, where L_x, L_y, and L_z are the domain dimensions in the Cartesian directions, λ is the wavelength of the spanwise travelling wave and x_0 marks the actuation onset. The shaded red surface A_{surf} marks the integration area of the wall-shear stress τ_w.

4.1. *Flow configuration and large-eddy simulation*

In this section, the actuated turbulent boundary layer configuration for skin-friction reduction is detailed. In particular, the actuation mechanism is presented, and the numerical set-up is described. For more details, the reader is referred to Albers *et al.* (2019) and Fernex *et al.* (2020). The fluid flow is described in a Cartesian frame of reference where the streamwise, wall-normal and spanwise coordinates are denoted by $x = (x, y, z)$ and the velocity components by $u = (u, v, w)$. The Mach number is set to $Ma = 0.1$ corresponding to a nearly incompressible flow. An illustration of the rectangular physical domain is shown in figure 22. A momentum thickness of $\theta = 1$ at x_0 is achieved such that the momentum thickness based Reynolds number is $Re_\theta = 1000$ at x_0. The domain length and height in the streamwise and wall-normal directions are $L_x = 190\,\theta$ and $L_y = 105\,\theta$. In the spanwise direction, different domain widths $L_z \in [21.65\,\theta, 108.25\,\theta]$ are used to simulate different actuation wavelengths.

At the domain inlet, a synthetic turbulence generation method is applied to generate a natural turbulent boundary layer flow after a transition length of 2–4 boundary layer thicknesses (Roidl, Meinke & Schröder 2013). Characteristic boundary conditions are used at the domain exit and a no-slip wall boundary condition is enforced at the lower domain boundary for the unactuated and actuated wall.

The actuation is performed by a transverse travelling wave on the surface. The corresponding wall motion is prescribed by the space- and time-dependent function

$$y_{wall}^+(z^+, t^+) = A^+ \cos\left(\frac{2\pi}{\lambda^+}z^+ - \frac{2\pi}{T^+}t^+\right), \tag{4.1}$$

in the interval $-5 \leq x/\theta \leq 140$. The quantities λ^+, T^+ and A^+ denote the wavelength, period and amplitude in inner coordinates, i.e. the parameters are scaled by the viscosity ν and the friction velocity of the unactuated reference case u_τ^n. In the area just upstream and downstream of the wave actuation region, a spatial transition is used from a flat plate to an actuated plate and *vice versa* (Albers *et al.* 2019). In total, 38 actuation configurations with wavelength $\lambda^+ \in [200, 500, 3000]$, period $T^+ \in [20, 120]$ and amplitude $A^+ \in [10, 78]$ are simulated. In the current study, we model one test case with $\lambda^+ = 1000$, $T^+ = 120$ and
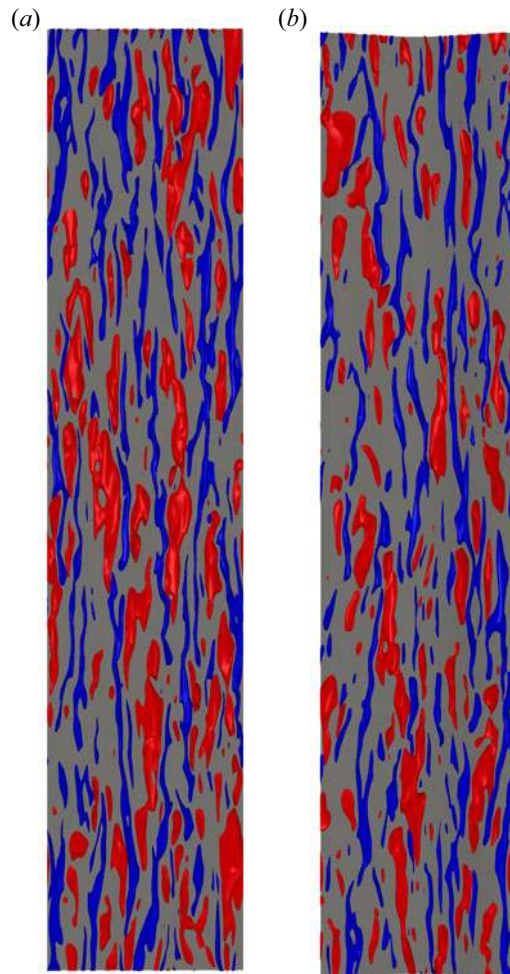
FIGURE 23. Contours of the random streamwise velocity fluctuations in the near-wall region of (*a*) a non-actuated reference case and (*b*) the actuated case. The actuation diminishes the near-wall streak intensity. The surface deformation is indicated at the bottom.

$A^+ = 40$ which yields the largest drag reduction of 3 % found at that wavelength. These actuation parameters correspond to case N36 in table 3 of Ishar *et al.* (2019) and in table 2 of Albers *et al.* (2020).

The physical domain is discretized by a structured block-type mesh with a resolution of $\Delta x^+ = 12.0$ in the streamwise and $\Delta z^+ = 4.0$ in the spanwise direction. In the wall-normal direction, a resolution of $\Delta_y^+|_{wall} = 1.0$ at the wall is used with gradual coarsening away from the wall. Depending on the domain width, the meshes consist of 24–120 million cells.

The actuated flat plate turbulent boundary layer flow is governed by the unsteady compressible Navier–Stokes equations in the arbitrary Lagrangian–Eulerian formulation for time-dependent domains. A second-order accurate finite-volume approximation of the governing equations is used in which the convective fluxes are computed by the advection upstream splitting method (AUSM) and time integration is performed via a 5-stage Runge–Kutta scheme. The smallest dissipative scales are implicitly modelled

through the numerical dissipation of the AUSM scheme. This monotonically integrated large-eddy simulation approach (Boris *et al.* 1992) is capable of accurately capturing all physics of the resolved scales (Meinke *et al.* 2002).

The actuated simulations are initialized by the solution from the unactuated reference case and the temporal transition from the flat plate to the actuated wall is initiated. When a converged state of the friction drag is obtained, statistics are collected for $tU_\infty/\theta = 1250$ convective times.

The actuation effects on the near-wall flow features are illustrated in figure 23, which shows contours of the streamwise velocity fluctuation of a reference natural (figure 23a) and the actuated case (figure 23b). The intensity of the near-wall streaks, which are known to contribute to skin friction, are observed to diminish with the actuation.

4.2. *Clustering*

Similar to the mixing layer, the clustering of the actuated boundary layer large-eddy simulation (LES) snapshots is performed using a lossless POD compression. Again, this compression dramatically reduces the computational load of clustering. Here, we perform the POD and the clustering on all 38 test cases simultaneously. Employing this enlarged set of POD modes yields a richer, and thus more accurate, dynamical representation of the individual test cases and allows for a direct comparison of different actuations (Ishar *et al.* 2019). Concatenating all configurations results in $M = 15\,873$ snapshots sampled at $\Delta t = 0.94$ time units.

Following Ishar *et al.* (2019), the M snapshots are clustered with the k-means++ algorithm into 50 centroids, corresponding to $K = 10$ centroids populated by the investigated actuation. It is worth noting that increasing K significantly, say $K = 100$, uncovers centroids with smaller length scale features associated with broadband turbulence of the boundary layer. In this study, we purposely choose to focus on the main energy-containing dynamics and thus limit the number of centroids to $K = 10$. Figure 24 presents four centroid distributions of the test case with $\lambda^+ = 1000$, $T^+ = 120$ and $A^+ = 40$. As the figure shows, the centroids have similar spatial distributions and are phase shifted with respect to one another. Such behaviour is consistent with a limit-cycle dynamics, indicating partial lock-on of the boundary layer dynamics to the periodic surface actuation. This lock-on phenomenon is sometimes associated with aerodynamic gains or losses depending on the targeted flow instability. It is synonymous with synchronization, and has been repeatedly investigated for drag reduction problems (Barros *et al.* 2016; Taira & Nakao 2018; Herrmann *et al.* 2020). Similar to these studies, a lower actuation threshold with sufficient authority is required to synchronize the flow.

The dynamics is well represented in the state space illustrated in figure 25, which is spanned by the first three POD mode coefficients. The cluster centroids are displayed as black solid circles and their indices are labelled. The snapshots are coloured according to their cluster affiliation. Similar to the shear layer, the dynamics of the actuated boundary layer appears to be driven by two physical phenomena: a cyclic behaviour synchronized with the surface actuation, and a quasi-stochastic component that forces the limit cycle to experience cycle-to-cycle variations (Cao *et al.* 2014). The latter phenomenon is associated with broadband turbulence of the boundary layer.

4.3. *Network model*

The CNM is generated based on the direct transition matrix and the averaged transition time matrix, which are illustrated in figure 26. We reiterate the vanishing diagonal
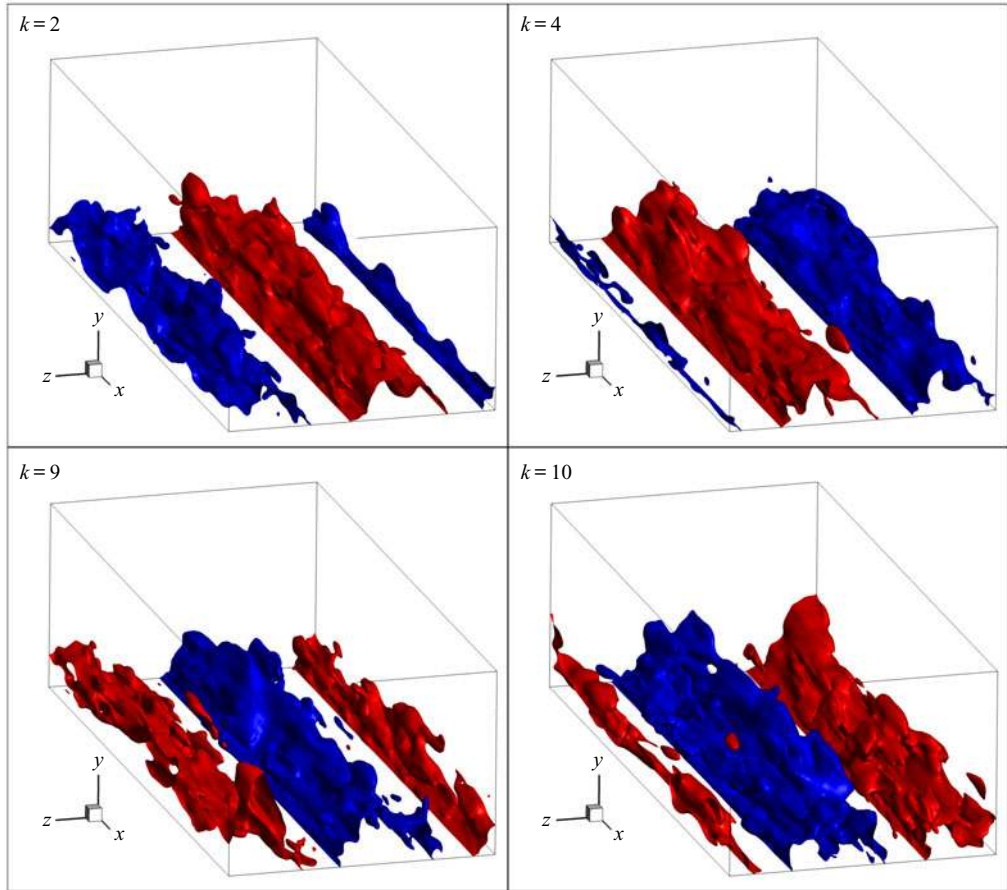
FIGURE 24. The cluster centroids c_k, $k = 2, 4, 9, 10$ of the actuated boundary layer. The cluster numbers are denoted in the state space of figure 25. The iso-surfaces correspond to constant wall-normal velocity of $V^+ = \pm 0.04$. Red and blue regions mark positive and negative values.

elements of both matrices, i.e. $Q_{ii} = T_{ii} = 0$, which are a result of enforcing non-trivial transitions. The direct transition matrix (cf. figure 26a) shows both the dominant transition probability of subsequent centroids being associated with the limit-cycle behaviour, and the wandering dynamics from the remaining transitions. The transition time matrix between the centroids (cf. figure 26b) reflects the same behaviour, and exhibits a quasi-constant transition time for limit-cycle 'flight times' and diverse transition times for the wandering effect.

Figure 27 compares the probability distribution of LES data and the model-based asymptotic vector p^∞. Again, we choose cluster $k = 1$ as the initial condition for LES and for the CNM and integrate over $l = 106$ transitions, which corresponds to a similar time range as that of the snapshots. The agreement between the two distributions is good.

The model performance is assessed against the reference LES results. Figure 28 shows the evolution of the first four POD mode amplitudes (red dashed curve). The dominance of the first two POD modes compared to the subsequent modes is expected for the current quasi-synchronous actuated flow. Similar to the previously presented results, the temporal evolution is smoothed with a spline. As the figure shows, CNM agrees very well with the
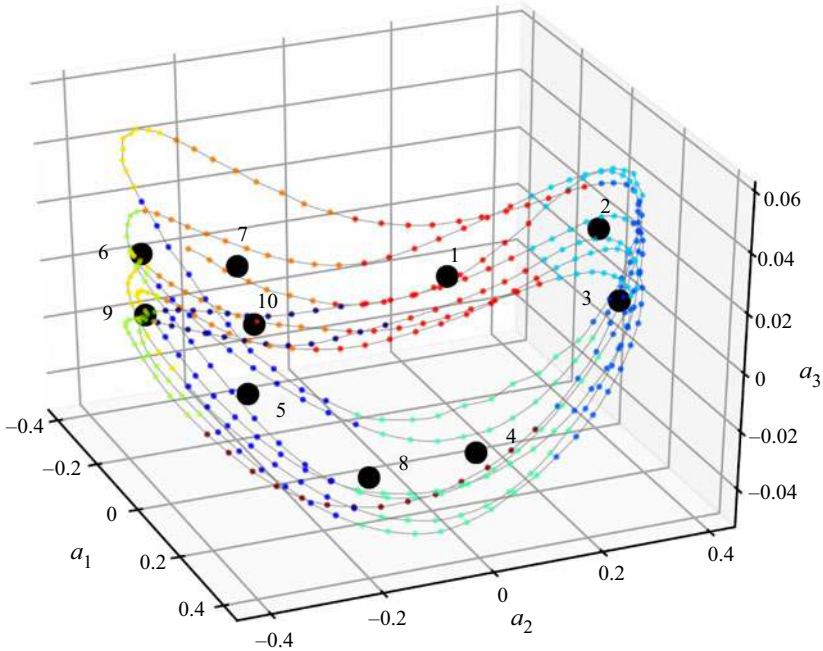
FIGURE 25. The state space is spanned by the first three mode coefficients of the lossless proper orthogonal decomposition of the LES data. The cluster centroids are displayed as black solid circles and the snapshots are coloured according to their cluster affiliation.
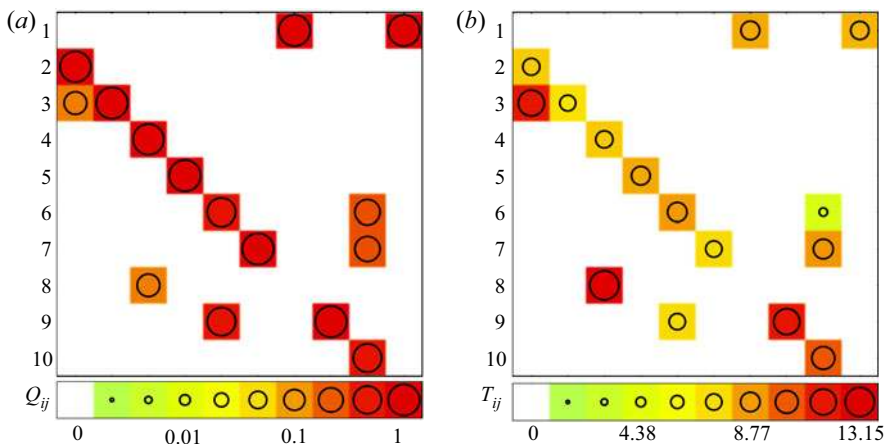

FIGURE 26. Dynamics of the cluster-based network model for the actuated boundary layer. (*a*) Direct transition matrix. (*b*) Averaged transition time. The non-vanishing values of the matrix elements are proportional to the circle radius and can be inferred from the colour code from the bottom caption.

amplitude and phase of the LES reference data over the entire approximately 400 time units.

The agreement between the model and the reference data is further corroborated by comparing the autocorrelation function. Figure 29 displays the autocorrelation function
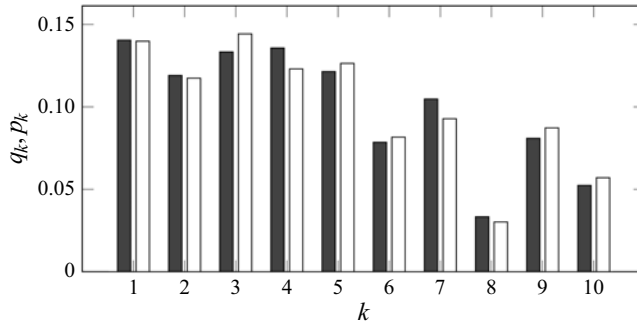
FIGURE 27. Probability distribution of the actuated boundary layer from large-eddy simulation (solid rectangle) and cluster-based network model (open rectangle). The CNM values are obtained from simulating $l = 106$ cluster transitions.
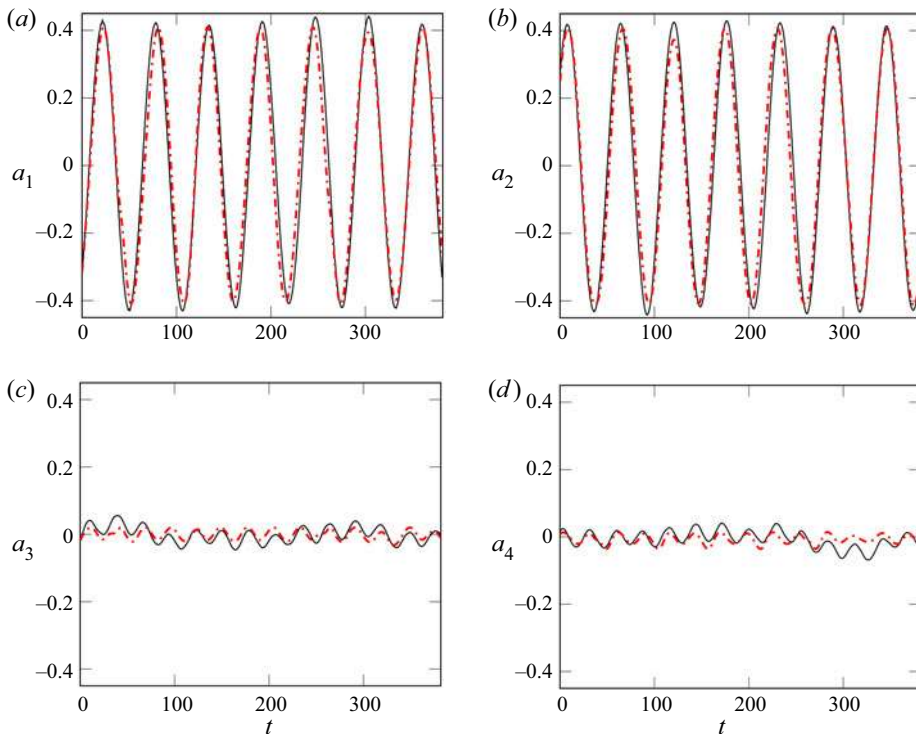

FIGURE 28. Evolution of mode amplitudes a_1–a_4 for the actuated boundary layer $t \in [0, 400]$. The curves correspond to LES (black solid line) and cluster-based network model (red dashed line).

of the CNM and the LES. As with the mixing layer, the model-based fluctuation level at vanishing time delay is lower than the LES value but becomes similar to oscillation level for an arbitrary larger time horizon. This large representation error at $\tau = 0$ relates to the unresolved inner-cluster variance. Yet, the centroids adequately resolve the periodic flow response of the flow to the periodic surface actuation.
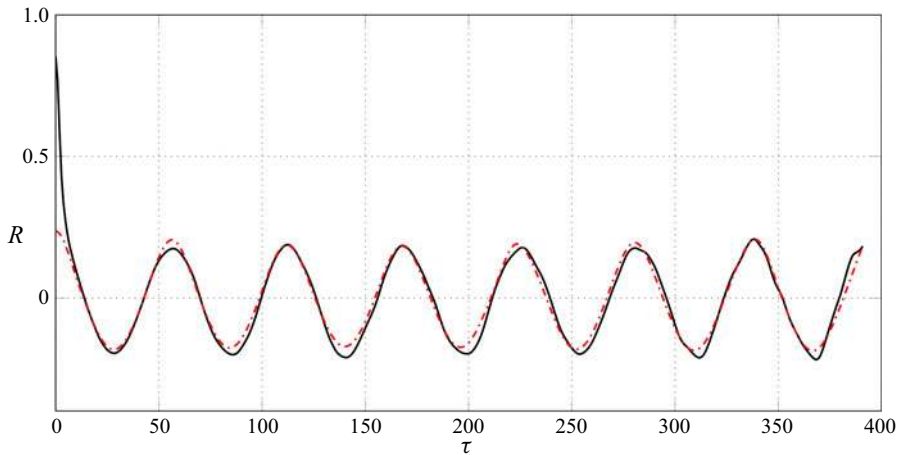
FIGURE 29. Autocorrelation function of the actuated turbulent boundary layer for LES (black solid line) and cluster-based network model (red dashed line).

5. Conclusions

In the present study, we propose a new data-driven methodology for modelling nonlinear dynamical systems. We trade compatibility with first principles, as with a POD-based Galerkin model, with the simplicity and robustness of the modelling. Point of departure is the cluster-based Markov model (Kaiser *et al.* 2014) for time-resolved snapshot data. The snapshots are coarse-grained into a few representative centroids. The temporal evolution of the state is conceptualized as a straight constant velocity movement from one centroid to the next. The average flight time and the transition probabilities are inferred from the data. Thus, the dynamics is modelled by a deterministic–stochastic network model with the centroids as nodes, the straight trajectory segments as edges, the transition time as parameters of the edges and the transition probability characterizing the nodes.

The resulting cluster-based network model has several desirable features: (i) The methodology is simple and automatable. (ii) The off-line computational load is only slightly larger than a snapshot-based POD. After the computation of the POD, the clustering and network model requires a tiny fraction of the computational operation. If the CNM is computed with original flow data without POD compression, the computational costs are orders of magnitudes larger, as elaborated in appendix A. (iii) The CNM has the same recurrence properties as the original data: if one cluster is visited multiple times in the data, it will also be a recurrence point of the CNM. (iv) Long-term integration will never lead to a divergence – unlike POD models. (v) The framework is very flexible allowing us, for instance, to incorporate multiple operating conditions.

The simplicity and robustness have a price. On the kinematic side, the vanilla version of CNM does not have the possibility of extrapolating the data, e.g. resolving oscillations at higher amplitudes not contained in the data. On the dynamic side, we lose the relationship to first principles: the network model is purely inferred from the snapshot data, without links to the Navier–Stokes equations. In particular, cluster-based models are not natural frameworks for dynamic instabilities, as the notion of exponential growth and nonlinear saturation is intimately tied to Galerkin flow expansions. Subsequent generalizations need to overcome these restrictions.

Cluster-based network modelling is applied to the Lorenz attractor. A k-means++ algorithm yields 10 centroids from a long time-resolved solution. Four centroids represent

each ear of the attractor and two, the switching area. Despite the coarseness of the presentation, the temporal dynamics mimics well the oscillations in each ear and the switching between both ears. The agreement is mirrored by the similarity between the autocorrelation functions of the simulation and the CNM. Statistically, the cluster population is predicted with acceptable accuracy. The CNM dramatically outperforms the cluster-based Markov model (Kaiser *et al.* 2014) in terms of predicting the temporal evolution. In contrast, CMM is more accurate for the cluster population. The error source of the CNM can be traced back to the chosen simple model of transition times.

Two demonstrations of CNM are performed with a laminar two-dimensional mixing layer and with a periodically actuated turbulent boundary layer. The mixing layer features K–H vortices and occasional vortex pairing. The cycle-to-cycle variations are clearly distilled by the centroids and the proximity map shows the possible transitions. The transition probabilities and times are quantified in the CNM model parameters. The actuated turbulent boundary layer exhibits partial lock-on with a superimposed stochastic meandering. For both applications, the snapshots are coarse-grained into 10 centroids. For the mixing layer, one group of centroids can be associated with K–H vortices and a second group to vortex pairing – similar to Kaiser *et al.* (2014). In contrast, the centroid affiliations for the actuated turbulent boundary layer are less categorizable. The dominant periodic dynamics is superimposed with quasi-stochastic transitions associated with broadband turbulence. The CNM well resolves the temporal evolution of the main flow dynamics, the fluctuation level, the autocorrelation function and the cluster population. A noteworthy observation relates to the autocorrelation function. For vanishing time delay, this function displays the average representation which is significant both for the mixing layer and wall turbulence. Yet, the function is surprisingly well represented by the CNM after one characteristic period. This behaviour corroborates that the dominant periodic dynamics is well resolved by the CNM with 10 centroids and the local interpolation between them.

The CNM is found to have a distinct advantage over the departure point, CMM, namely the much longer prediction horizon as evidenced by the autocorrelation function. POD and DMD models may describe the same flow with a similar number of modes (Protas *et al.* 2015). We emphasize that the construction of the CNM could be fully automated in a software package. In contrast, data-driven nonlinear Galerkin models may be designed as insightful least-order representations with interpretable modes. Moreover, the Galerkin dynamics may reveal the interplay between linear and nonlinear terms, as beautifully displayed in mean-field theory (Stuart 1971), self-consistent models (Mantič-Lugo, Arratia & Gallaire 2014), resolvent operator approaches (Gomez *et al.* 2016), finite-time thermodynamics (Noack *et al.* 2008) and criteria for boundedness (Schlegel & Noack 2015). Yet, a functional model requires the careful choice of flow data, potentially shift and other non-standard modes, subscale closure models and calibration techniques. Thus, cluster-based and POD based models have different niche applications.

CNM opens a novel automatable avenue for nonlinear dynamical modelling. Moreover, CNM provides a framework for estimation and model-based control. This extension is elaborated in appendix C and complements model-free cluster-based control for open-loop actuation (Kaiser *et al.* 2017) and for feedback laws (Nair *et al.* 2019). The authors actively pursue this direction.

Acknowledgements

Declaration of interests

The authors report no conflict of interest.

Appendix A. Data compression for clustering

Clustering is a computationally expensive process based on a large number of area/volume integrals for the distance between snapshots and centroids. Let M and K be the number of snapshots and clusters, respectively, then a single k-means iteration requires the computation of $K \times M$ integrals. Let I be the number of k-means iterations and L be the number of repetitions then the total number of integrals is $L \times I \times K \times M$. Typical values are $K \sim 10$, $I \sim 10K$ and $L \sim 100$.

The computational load can be significantly reduced by pre-processing the snapshot data with a lossless POD. The most expensive step of a typical snapshot POD is the computation of the correlation matrix with $M \times (M + 1)/2$ area/volume integrals. Thus, the integral for the distance between two velocity fields transforms into the Euclidean norm with $(M - 1)$-dimensional vectors of POD mode amplitudes. The computational saving reads

$$\frac{M \times (M + 1)/2}{L \times I \times K \times M} = \frac{M + 1}{2L \times I \times K}. \tag{A 1}$$

With typical values, the savings are one or two orders of magnitudes.

For completeness and self-consistency, the snapshot POD algorithm is described. POD is performed with the whole computational domain Ω. The inner product between two velocity fields $\boldsymbol{v}(\boldsymbol{x})$, $\boldsymbol{w}(\boldsymbol{x})$ in the square-integrable Hilbert space $\mathcal{L}^2(\Omega)$ reads

$$(\boldsymbol{v}, \boldsymbol{w})_\Omega = \int_\Omega \mathrm{d}\boldsymbol{x}\, \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{w}(\boldsymbol{x}). \tag{A 2}$$

The corresponding norm is given by

$$\|\boldsymbol{v}\|_\Omega = \sqrt{(\boldsymbol{v}, \boldsymbol{v})_\Omega}. \tag{A 3}$$

The distance D between two velocity fields is based on this norm,

$$D(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\|_\Omega. \tag{A 4}$$

The inner product (A 2) uniquely defines the snapshot POD (see e.g. Holmes *et al.* 2012). The *m*th snapshot is represented by

$$\boldsymbol{u}^m(\boldsymbol{x}) := \boldsymbol{u}_0(\boldsymbol{x}) + \sum_{i=1}^{M-1} a_i^m \boldsymbol{u}_i(\boldsymbol{x}), \tag{A 5}$$

where $\boldsymbol{u}_0$ denotes the mean flow, $\boldsymbol{u}_i$ the *i*th POD mode and a_i^m the POD mode amplitude corresponding to the *m*th snapshot. It may be noted that the maximal number of POD modes is $M - 1$, e.g. two snapshots define a one-dimensional line, not a plane.

Let $\boldsymbol{v} = \boldsymbol{u}_0 + \sum_{i=1}^{M-1} b_i \boldsymbol{u}_i$ and $\boldsymbol{w} = \boldsymbol{u}_0 + \sum_{i=1}^{M-1} c_i \boldsymbol{u}_i$ be two velocity field representations, e.g. a snapshot and a centroid. Then, their distance is given by

$$D(\boldsymbol{v}, \boldsymbol{w})_\Omega = \sqrt{\sum_{i=1}^{M-1} (b_i - c_i)^2}. \tag{A 6}$$

Evidently, (A 6) is much quicker to compute than (A 4) assuming the typical case that the number of grid points is much larger than the number of snapshots.

Appendix B. On the optimal number of clusters

We investigate the prediction error of a CNM with direct transition matrix $\boldsymbol{Q}$ and transition time matrix $\boldsymbol{T}$ for K clusters from M snapshot data. The number of clusters K significantly influences the prediction error of the CNM. Coarse clustering (small K) means that the direct transition matrix $\boldsymbol{Q}$ can be inferred from a lot of transition data and is hence relatively accurate. Yet, the snapshots in each cluster have a large representation error. In contrast, a finely resolving clustering (large K) implies a more accurate representation of the true state. Yet, the transition matrix is larger and the error of the estimated transition probability increases. The extremes are $K = 1$ cluster with large representation error and $K = M$ with vanishing representation error, but large error of the transition matrix for new data. We can expect a sweet spot with optimal prediction error based on good representation error and an accurate estimate of the transition matrix.

In the following, we define the performance measure for the CNM. The starting point is the error between the model and true state $\delta \boldsymbol{u}(t) = \boldsymbol{u}^\circ(t) - \boldsymbol{u}^\bullet(t)$. The modelling error for a specified number of clusters K is defined as average error for all available snapshots with prediction horizon τ starting from the most accurate initial condition $\boldsymbol{u}^\circ(0) \approx \boldsymbol{u}^\bullet(0)$. The true initial state is taken from the snapshot data $\boldsymbol{u}^m$, while the modelled initial state is the closest centroid $\boldsymbol{c}_{k^m}$. The resulting error reads

$$C(\tau) := \overline{\|\boldsymbol{u}^\circ(t+\tau) - \boldsymbol{u}^\bullet(t+\tau)\|^2}_\Omega. \tag{B 1}$$

The overbar denotes the average over the prediction errors for all available snapshots $\boldsymbol{u}^\bullet(t)$ with data horizon until $t + \tau$; $C(0)$ corresponds to the representation error where the true state $\boldsymbol{u}^\bullet(t)$ is estimated by the modelled state $\boldsymbol{u}^\circ(t)$ as accurately as possible; $C(\tau)$ is the prediction error after time τ.

Figure 30 illustrates the temporal evolution of prediction error for $\tau \leq 50$, roughly corresponding to five Kelvin–Helmholtz shedding periods. This error expectedly increases with growing prediction horizon τ for all investigated numbers of clusters $K = 10, 20, 50, 100$. There is no uniformly superior prediction error for any number of clusters. Small (large) K correspond to large (small) representation and prediction error for a small
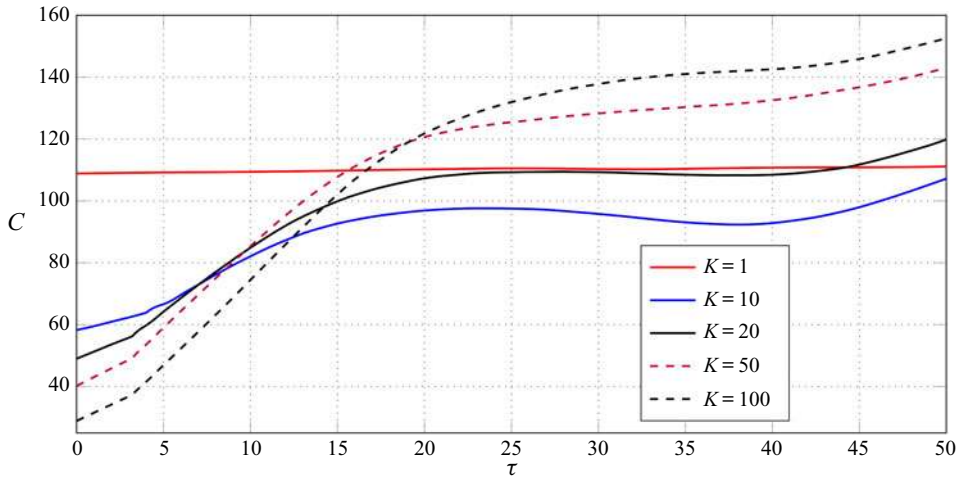
FIGURE 30. The prediction error $C(\tau)$ over time with selected number of clusters $K = 1$, 10, 20, 50, 100.

time horizon. However, the more inaccurate transition matrix leads to larger prediction errors in the long run. The CNM with $K = 10$ leads to the smallest prediction error for $\tau \in [12.5, 50]$ in comparison to all other investigated models. Hence, we conclude that $K = 10$ is a good choice for the cluster number for a prediction horizon with one to several shedding periods.

We refrain from fine-tuning the optimal number of clusters as this number is a function of the prediction horizon τ. For $\tau = 0$, a CNM with $K = M$ reduces the representation error to zero, at least for the training data. For $\tau = \infty$, the trivial CNM with $K = 1$ yielding the mean flow outperforms all other models roughly by a factor 2. The average error between model and data can easily be shown to be larger than the average distance of the data to the mean. Let us consider the data $u^{\bullet} = \cos t$ and model $u^{\circ} = \cos 1.05t$ with small frequency difference, i.e. increasing phase error. Then, $\overline{(u^{\bullet} - u^{\circ})^2} = 1$ but $\overline{(u^{\circ} - 0)^2} = 1/2$. On average, u° stays closer to the mean 0 than to another harmonics which is occasionally out of phase.

Finally, we remark that the number of clusters K plays a similar role in cluster-based models than the number of POD modes N in Galerkin models. Human interpretability is easier for a low-dimensional flow representation while the accuracy increases with the model order. For instance, for periodic dynamics, the phase resolution of each centroid is approximately $360°/K$. However, there is a noticeable difference in robustness between CNM and POD models. POD models tend to become more fragile with increasing state-space dimension, as every new degree of freedom comes with many new coefficients and potential error amplifiers. In contrast, the robustness of cluster-based model does not suffer from increasing dimension. A second difference relates to the modes. Increasing the number of POD modes does not affect the lower-order modes by design. In contrast, all centroids change as K is just increased by 1. Similarly, all intervals of a one-dimensional finite-element discretization change as the number of elements increase by one.

Appendix C. POD versus cluster-based network modelling

The POD models and CNM belong to the family of data-driven dynamic grey-box models which resolve the evolution of coherent structures. Dynamic POD modelling was

pioneered by Aubry *et al.* (1988) and has enjoyed over three decades of rapid development on coherent-structure descriptions, dynamical systems, estimation and control. In contrast, networks have been recently introduced to reduced-order modelling of fluid flows (Nair & Taira 2015; Taira *et al.* 2016). In this section, we compare POD models and CNM with respect to kinematics (appendix C.1), dynamics (appendix C.2), estimation (appendix C.3) and control (appendix C.4) – foreshadowing promising future opportunities of CNM.

C.1. *Kinematics*

Starting point of most data-driven grey-box models are M flow snapshots $\{u(x)\}_{m=1}^{M}$ typically resolving first and second statistical moments. Like the dynamic mode decomposition (Rowley *et al.* 2009; Schmid 2010), the snapshots are assumed to resolve the coherent-structure evolution in time so that the temporal dynamics can be identified. The POD expands the fluctuations around the mean flow u_0 into a given number N of orthonormal modes u_i:

$$u(x, t) = u_0(x) + \sum_{i=1}^{N} a_i(t) u_i(x) + \epsilon(x, t). \tag{C 1}$$

By design, this expansion minimizes the averaged representation error $\sum_{m=1}^{M} \|\epsilon^m\|_{\Omega}^{2}/M$ with respect to all Galerkin expansions of N modes. The POD modes are linear combinations of the fluctuations $u^m - u_0$.

Clustering coarse-grains the snapshot data to a given number K of centroids $\{c_k\}_{k=1}^{K}$. Each snapshots with index m belongs to the closest centroid k_m. The centroids are requested to minimize the averaged representation error $\|u^m - c_{k^m}\|^2$. Similar to POD modes, centroids are linear combinations of the snapshots.

The representation error of centroids can be further reduced by allowing for interpolations:

$$u(x, t) = \sum_{k=1}^{K} w_k(t) c_k(x), \quad \sum_{k=1}^{K} w_k(t) = 1, \quad \forall k \colon w_k(t) \geq 0. \tag{C 2}$$

In case of the Markov model, the weights are the evolving probabilities $w_k(t) = p_k(t)$ and make the expansion (C 2) converge to the mean flow. In case of the network model, the weights characterize 'flights' with uniform velocity between two centroids, say from k to j, and typically re-visit all centroids in finite time. The Markov model might be compared with unsteady Reynolds-averaged Navier–Stokes equations converging to the mean flow while the network model is reminiscent of the large-eddy simulations.

We emphasize that (C 1) and (C 2) look similar but have quite different ranges of applications. The POD expansion is based on the superposition of modes with arbitrary mode amplitudes a_i. Neither the mean flow nor the POD modes are realizable states. POD could be considered a data-driven analogue of the Fourier expansion. In contrast, the cluster-based expansion is only meant to describe a local interpolation for CNM. The centroids are coarse-grained approximations of realizable states. The centroids may be conceptualized as collocation points for a finite-element inspired ansatz and the associated Voronoi cells serve as finite elements. As a corollary, POD expansions can describe new states which are far from the snapshot database, because the mode amplitudes are not confined. In contrast, cluster-based expansions are bound to stay

close to the training data by the non-negativity $w_k \geq 0$ and the normalization constraint $\sum_{k=1}^{K} w_k = 1$. By construction, global POD expansions have lower representation error as cluster-based expansions with the same number of modes. Some POD modes of simple dynamics may have a physical meaning as they resolve instability modes or harmonics. Typically, however, POD modes comprise a mix of frequencies and are difficult to interpret. In contrast, all centroids are human-interpretable coarse-grained flows which are representative for a certain state-space region. Summarizing, the choice between POD and clustering strongly depends on the intended applications.

C.2. *Dynamics*

The temporal evolution of the incompressible viscous flow can be derived from the Galerkin expansion and the Navier–Stokes equations (Fletcher 1984) for steady domains with stationary boundary conditions. The resulting Galerkin system for the mode amplitude vector $\boldsymbol{a} = [a_1, a_2, \ldots, a_N]^{\mathrm{T}}$ is the autonomous system

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{a}). \tag{C 3}$$

For turbulent flows, only a fraction of the fluctuation energy is resolved by the POD modes and the effect of the remaining unresolved fluctuations must be accounted for. Myriad of subgrid turbulence models and calibration techniques have been proposed and the identification of a robust realistic dynamical system constitutes a challenge. Even the basic physical requirement of a globally bounded dynamics is often not met (Schlegel & Noack 2015).

CNM might be conceptualized as flights between airports (centroids) from a discrete network of routes with destination probabilities (transition matrix) and flight times (transition times). In CNM, the chosen 'destination' j from 'airport' k at time t^m from centroid k at time t^m to centroid j during $t^{m+1} = t^m + T_{jk}$ is described by

$$j = \text{realization according to } Q_{jk}, \tag{C 4a}$$

$$\boldsymbol{u}(\boldsymbol{x}, t) = w_k(t)\boldsymbol{c}_k(\boldsymbol{x}) + w_j(t)\boldsymbol{c}_j(\boldsymbol{x}), \tag{C 4b}$$

$$w_j(t) = (t - t^m)/T_{jk}, \quad w_k(t) = 1 - w_j(t). \tag{C 4c}$$

At time t^{m+1}, a similar decision on the next destination is made, and so on. The CNM (C 4) describes a deterministic–stochastic dynamics in contrast to the deterministic (C 3).

In contrast to POD models, the CNM (C 4) contains no design parameter beyond the number of clusters and is fully automated. Moreover, the dynamics is robust and cannot diverge, unlike POD models. The price is the confinement to the neighbourhood of the training data. Again, the decision in favour of the POD model or CNM strongly depends on the goal. The POD models may allow deeper dynamics insights; CNM is much simpler and much more robust by design.

C.3. *Estimation*

In most experiments, only few signals, denoted by the vector $\boldsymbol{s}(t)$, can be recorded. Let $\boldsymbol{u}^m(\boldsymbol{x}, t^m)$, $m = 1, \ldots, M$, be the snapshots associated with the sensor readings $\boldsymbol{s}^m = \boldsymbol{s}(t^m)$.

The easiest realization of the estimator

$$\hat{\boldsymbol{u}}(\boldsymbol{x}, t) = \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{s}(t)) \tag{C5}$$

for sensor reading $\boldsymbol{s}$ is to find the closed sensor data $\boldsymbol{s}^m$ from the data base and to take the corresponding snapshot $\boldsymbol{u}^m$ as an estimator. This simplistic 1-nearest neighbour estimator can be refined in numerous ways. An interpolation with K data points can be performed with a K-nearest neighbour approach (Loiseau, Noack & Brunton 2018). The sensor signals may be lifted to a feature space without dynamic false neighbours, for instance with time-delay coordinates (Loiseau *et al.* 2018). Or the structure of $\boldsymbol{G}$ may be pre-assumed as in linear stochastic estimation.

The estimated flow field is canonically transcribed into POD mode amplitudes $\hat{\boldsymbol{a}}$ and permissible centroid weights $\hat{\boldsymbol{w}} = [\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_N]^T$. Summarizing, the estimation can easily be realized as add-ons in POD models and CNM. For completeness, we mention the possibility of dynamic observers exploiting the dynamical system.

C.4. *Control*

The POD models may be enriched with a forcing term. In a simple case, like a volume force, the forcing term is additive and linear in the actuation command $\boldsymbol{b} = [b_1, b_2, \ldots, b_{N_b}]^T$ with the gain matrix $\boldsymbol{B}$:

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{B}\boldsymbol{b}. \tag{C6}$$

From here on, stabilizing control laws may be derived from linearizations or other strategies (Brunton & Noack 2015).

The control design for CNM is more complex. The actuation command $\boldsymbol{b}$ affects the dynamics (C4) via changed transition probabilities $\boldsymbol{Q}(\boldsymbol{b})$ and changed transition times $\boldsymbol{T}(\boldsymbol{b})$. In a simple case, this dependency may be linearized:

$$\boldsymbol{Q} = \boldsymbol{Q}_0 + \sum_{l=1}^{N_b} b_l \boldsymbol{Q}_l, \quad \boldsymbol{T} = \boldsymbol{T}_0 + \sum_{l=1}^{N_b} b_l \boldsymbol{T}_l. \tag{C7a,b}$$

Here, the subscript '0' corresponds to the unforced state, while the subscript 'l' denotes changes caused by the actuation command b_l. The matrices may be identified from actuated flow data. After, the forced CNM (C4)–(C7) is identified, a regression solver can be employed to optimize the control law with respect to a cost function. Genetic programming has proven to be a powerful method for this method in dozens of turbulence control experiments (Noack 2019).

We remark that the deterministic–stochastic network dynamics rules out 'simple' control design based on local linearizations, but requires the numerical solution of a non-convect nonlinear optimization problem. Thus, the computational cost of this approach is significantly larger than the model-based linear control. Yet, cluster-based network model may enable nonlinear infinite-horizon control at a fraction of the computational cost of linear optimal control using the Navier–Stokes equations. The authors actively pursue this novel avenue of cluster-based network control for turbulence. Nair *et al.* (2019) and Kaiser *et al.* (2017) present a model-free cluster-based control as a prelude to these efforts.

REFERENCES

ALBERS, M., MEYSONNAT, P. S, FERNEX, D., SEMANN, R., NOACK, B. R. & SCHRÖDER, W. 2020 Drag reduction and energy savings by spanwise travelling transversal surface waves for flat plate flow. *Flow Turbul. Combust.* **105**, 125–157.

ALBERS, M., MEYSONNAT, P. S. & SCHRÖDER, W. 2019 Actively reduced airfoil drag by transversal surface waves. *Flow Turbul. Combust.* **102** (4), 865–886.

ASHURST, W. T. & MEIBURG, E. 1988 Three-dimensional shear layers via vortex dynamics. *J. Fluid Mech.* **189**, 87–116.

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.

BARROS, D., BORÉE, J., NOACK, B. R. & SPOHN, A. 2016 Resonances in the forced turbulent wake past a 3D blunt body. *Phys. Fluids* **28** (6), 065104.

BORIS, J. P., GRINSTEIN, F. F., ORAN, E. S. & KOLBE, R. L. 1992 New insights into large eddy simulation. *Fluid Dyn. Res.* **10** (4–6), 199–228.

BRUNTON, S. L. & NOACK, B. R. 2015 Closed-loop turbulence control: Progress and challenges. *Appl. Mech. Rev.* **67** (5), 050801.

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.

CAO, Y., KAISER, E., BOREÉ, J., NOACK, B. R., THOMAS, L. & GUILAN, S. 2014 Cluster-based analysis of cycle-to-cycle variations: application to internal combustion engines. *Exp. Fluids* **55**, 1837.

COATS, C. M. 1997 Coherent structures in combustion. *Prog. Energy Combust. Sci.* **22**, 427–509.

COMTE, P., SILVESTRINI, J. H. & BÉGOU, P. 1998 Streamwise vortices in large–eddy simulations of mixing layer. *Eur. J. Mech.* B **17**, 615–637.

DELVILLE, J., UKEILEY, L., CORDIER, L., BONNET, J. P. & GLAUSER, M. 1999 Examination of large-scale structures in a turbulent plane mixing layer. Part 1. Proper orthogonal decomposition. *J. Fluid Mech.* **391**, 91–122.

DU, Y. & KARNIADAKIS, G. E. 2000 Suppressing wall turbulence by means of a transverse traveling wave. *Science* **288** (5469), 1230–1234.

FAN, B. & DONG, G. 2016 *Principles of Turbulence Control*. National Defense Industry Press; Wiley.

FERNEX, D., SEMANN, R., ALBERS, M., MEYSONNAT, P. S., SCHRÖDER, W., ISHAR, R., KAISER, E. & NOACK, B. R. 2019 Cluster-based network model for drag reduction mechanisms of an actuated turbulent boundary layer. *Proc. Appl. Maths Mech.* **19** (1), 1–2.

FERNEX, D., SEMAAN, R., ALBERS, M., MEYSONNAT, P.S., SCHRÖDER, W. & NOACK, B. R. 2020 Actuation response model from sparse data for wall turbulence drag reduction. *Phys. Rev. Fluids* **5** (7), 073901.

FIEDLER, H. & FERNHOLZ, H. H. 1990 On the management and control of turbulent shear flows. *Prog. Aeronaut. Sci.* **27**, 305–387.

FLETCHER, C. A. J. 1984 *Computational Galerkin Methods*, 1st edn. Springer.

GAD-EL-HAK, M. 2000 *Flow Control: Passive, Active and Reactive Flow Management*, 1st edn. Cambridge University Press.

GOMEZ, 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.

HAMA, F. R. 1962 Streaklines in a perturbed shear flow. *Phys. Fluids* **5**, 644–650.

HERRMANN, B., PHILIPP, O., RICHARD, S. & BRUNTON, S. L. 2020 Modeling synchronization in forced turbulent oscillator flows. *Nat. Commun. Phys.* (submitted).

HOLMES, P., LUMLEY, J. L., BERKOOZ, G. & ROWLEY, C. W. 2012 *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, 2nd edn. Cambridge University Press.

ISHAR, R., KAISER, E., MORZYNSKI, M., ALBERS, M., MEYSONNAT, P., SCHRÖDER, W. & NOACK, B. R. 2019 Metric for attractor overlap. *J. Fluid Mech.* **874**, 720–752.

JORDAN, P. & COLONIUS, T. 2013 Wave packets and turbulent jet noise. *Ann. Rev. Fluid Mech.* **45**, 173–195.

JUNG, W. J., MANGIAVACCHI, N. & AKHAVAN, R. 1992 Suppression of turbulence in wall-bounded flows by high-frequency spanwise oscillations. *Phys. Fluids* A **4** (8), 1605–1607.

KAISER, E., NOACK, B. R., CORDIER, L., SPOHN, A., SEGOND, M., ABEL, M. W., DAVILLER, G., ÖSTH, J., KRAJNOVIĆ, S. & NIVEN, R. K. 2014 Cluster-based reduced-order modelling of a mixing layer. *J. Fluid Mech.* **754**, 365–414.

KAISER, E., NOACK, B. R., SPOHN, A., CATTAFESTA, L. N. & MORZYŃSKI, M. 2017 Cluster-based control of nonlinear dynamics. *Theor. Comput. Fluid Dyn.* **31** (5–6), 1579–593.

KASTEN, J., REININGHAUS, J., HOTZ, I., HEGE, H. C., NOACK, B. R., DAVILLER, G., COMTE, P. & MORZYŃSKI, M. 2016 Acceleration feature points of unsteady shear flows. *Arch. Mech.* **68**, 55–80.

KLUMPP, S., MEINKE, M. & SCHRÖDER, W. 2011 Friction drag variation via spanwise transversal surface waves. *Flow Turbul. Combust.* **87** (1), 33–53.

LI, W., JESSEN, W., ROGGENKAMP, D., KLAAS, M., SILEX, W., SCHIEK, M. & SCHRÖDER, W. 2015 Turbulent drag reduction by spanwise traveling ribbed surface waves. *Eur. J. Mech. B/Fluids* **53**, 101–112.

LI, H. & TAN, J. 2020 Cluster-based Markov model to understand the transition dynamics of a supersonic mixing layer. *Phys. Fluids* **32** (5), 056104.

LIU, J. T. C. 1989 Coherent structures in transitional and turbulent free shear flows. *Ann. Rev. Fluid Mech.* **21**, 285–315.

LLOYD, S. 1982 Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** (2), 129–137.

LOISEAU, J.-C., NOACK, B. R. & BRUNTON, S. L. 2018 Sparse reduced-order modeling: sensor-based dynamics to full-state estimation. *J. Fluid Mech.* **844**, 459–490.

LORENZ, E. N. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

LUHAR, M., SHARMA, A. S. & MCKEON, B. J. 2016 On the design of optimal compliant walls for turbulence control. *J. Turbul.* **17** (8), 787–806.

LUMLEY, J. L., REMPFER, D. & BLOSSEY, P. N. 1999 Interaction between near-wall turbulent flows and compliant surfaces. *Tech. Rep.* AFRL-SR-BL-TR-99.

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. University of California Press.

MANTIČ-LUGO, V., ARRATIA, C. & GALLAIRE, F. 2014 Self-consistent mean flow description of the nonlinear saturation of the vortex shedding in the cylinder wake. *Phys. Rev. Lett.* **113** (8), 084501.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. 1979 *Multivariate Analysis*. Academic Press.

MEINKE, M., SCHRÖDER, W., KRAUSE, E. & RISTER, T. 2002 A comparison of second-and sixth-order methods for large-eddy simulations. *Comput. Fluids* **31** (4–7), 695–718.

MICHALKE, A. 1964 On the inviscid instability of the hyperbolic-tangent velocity profile. *J. Fluid Mech.* **19**, 543–556.

MONKEWITZ, P. A. 1988 Subharmonic resonance, pairing and shredding in the mixing layer. *J. Fluid Mech.* **188**, 223–252.

NAIR, A. G. & TAIRA, K. 2015 Network-theoretic approach to sparsified discrete vortex dynamics. *J. Fluid Mech.* **768**, 549–571.

NAIR, A., YEH, C. A., KAISER, E., NOACK, B. R., BRUNTON, S. L. & TIARA, K. 2019 Cluster-based feedback control of turbulent post-stall separated flows. *J. Fluid Mech.* **875**, 345–375.

NEWMAN, M. 2010 *Networks: An Introduction*. Oxford University Press.

NOACK, B. R. 2016 From snapshots to modal expansions – bridging low residuals and pure frequencies. *J. Fluid Mech.* **802**, 1–4.

NOACK, B. R. 2019 Closed-loop turbulence control-from human to machine learning (and retour). In *Proceedings of the 4th Symposium on Fluid Structure-Sound Interactions and Control (FSSIC), Tokyo, Japan* (ed. Y. Zhou, M. Kimura, G. Peng, A. D. Lucey & L. Hung), pp. 23–32. Springer.

NOACK, B. R., PAPAS, P. & MONKEWITZ, P. A. 2005 The need for a pressure-term representation in empirical Galerkin models of incompressible shear flows. *J. Fluid Mech.* **523**, 339–365.

NOACK, B. R., SCHLEGEL, M., AHLBORN, B., MUTSCHKE, G., MORZYŃSKI, M., COMTE, P. & TADMOR, G. 2008 A finite-time thermodynamics of unsteady fluid flows. *J. Non-Equilib. Thermodyn.* **33**, 103–148.

NOACK, B. R., STANKIEWICZ, W., MORZYŃSKI, M. & SCHMID, P. J. 2016 Recursive dynamic mode decomposition of transient and post-transient wake flows. *J. Fluid Mech.* **809**, 843–872.

PASTOOR, M., KING, R., NOACK, B. R., KING, R. & TADMOR, G. 2005 Observers and feedback control for shear layer vortices. In *44th IEEE Conference on Decision and Control and European Control Conference ECC 2005, Seville, Spain, 12–15 December 2005*.

PODVIN, B. 2009 A proper-orthogonal-decomposition based model for the wall layer of a turbulent channel flow. *Phys. Fluids* **21**, 015111.

PODVIN, B. & LUMLEY, J. 1998 A low-dimensional approach for the minimal flow unit. *J. Fluid Mech.* **363**, 121–155.

PRANDTL, L. 1904 Über Flüssigkeitsbewegung bei sehr kleiner Reibung. In *Verhandlungen des dritten internationalen Mathematiker-Kongresses in Heidelberg von 8. bis 13. August 1904* (ed. A. Krazer), pp. 484–491. B.G. Teubner. [Translated 1928 in *NACA memo no.* 452.]

PROTAS, B., NOACK, B. R. & ÖSTH, J. 2015 Optimal nonlinear eddy viscosity in Galerkin models of turbulent flows. *J. Fluid Mech.* **766**, 337–367.

QUADRIO, M., RICCO, P. & VIOTTI, C. 2009 Streamwise-travelling waves of spanwise wall velocity for turbulent drag reduction. *J. Fluid Mech.* **627**, 161.

ROIDL, B., MEINKE, M. & SCHRÖDER, W. 2013 A reformulated synthetic turbulence generation method for a zonal RANS–LES method and its application to zero-pressure gradient boundary layers. *Intl J. Heat Fluid Flow* **44**, 28–40.

ROWLEY, C. W., MEZIĆ, I., BAGHERI, S., SCHLATTER, P. & HENNINGSON, D. S. 2009 Spectral analysis of nonlinear flows. *J. Fluid Mech.* **645**, 115–127.

SASAKI, K., PIANTANIDA, S., CAVALIERI, A. V. G. & JORDAN, P. 2017 Real-time modelling of wavepackets in turbulent jets. *J. Fluid Mech.* **821**, 458–481.

SASAKI, K., TISSOT, G., CAVALIERI, A. V. G., SILVESTRE, F. J., JORDAN, P. & BIAU, D. 2018 Closed-loop control of a free shear flow: a framework using the parabolized stability equations. *Theor. Comput. Fluid Dyn.* **32** (6), 765–788.

SCHLEGEL, M. & NOACK, B. R. 2015 On long-term boundedness of Galerkin models. *J. Fluid Mech.* **765**, 325–352.

SCHMID, P. J. 2010 Dynamic mode decomposition for numerical and experimental data. *J. Fluid Mech.* **656**, 5–28.

SHAQARIN, T., NOACK, B. R. & MORZYŃSKI, M. 2018 The need for prediction in feedback control of a mixing layer. *Fluid Dyn. Res.* **50**, 1–22.

SPARROW, C. 1982 *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, 1st edn. Applied Mathematical Sciences, vol. 41. Springer.

STEINHAUS, H. 1956 Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.* **4** (12), 801–804.

STUART, J. T. 1971 Nonlinear stability theory. *Ann. Rev. Fluid Mech.* **3**, 347–370.

TAIRA, K., BRUNTON, S. L., DAWSON, S., ROWLEY, C. W., COLONIUS, T., MCKEON, B. J., SCHMIDT, O. T., GORDEYEW, S., THEOFILIS, V. & UKEILEY, L. S. 2018 Modal analysis of fluid flows: an overview. *AIAA J.* **55** (12), 4013–4041.

TAIRA, K., NAIR, A. G. & BRUNTON, S. L. 2016 Network structure of two-dimensional decaying isotropic turbulence. *J. Fluid Mech.* **795**, R2.

TAIRA, K. & NAKAO, H. 2018 Phase-response analysis of synchronization for periodic flows. *J. Fluid Mech.* **846**, R2.

UKEILEY, L., CORDIER, L., MANCEAU, R., DELVILLE, J., BONNET, J. P. & GLAUSER, M. 2001 Examination of large-scale structures in a turbulent plane mixing layer. Part 2. Dynamical systems model. *J. Fluid Mech.* **441**, 61–108.

WALSH, M. J. & LINDEMANN, M. 1984 Optimization and application of riblets for turbulent drag reduction. In *AIAA 22nd Aerospace Sciences Meeting*.

WEI, M. & ROWLEY, C. W. 2009 Low-dimensional models of a temporally evolving free shear layer. *J. Fluid Mech.* **618**, 113–134.

WOLF, A., SWIFT, J. B., SWINNEY, H. L. & VASTANO, J. A. 1985 Determining Lyapunov exponents from a time series. *Physica* D **16**, 285–317.