Towards robust data-driven reduced-order modelling for turbulent flows: application to vortex-induced vibrations

Received: 21 July 2021 / Accepted: 6 April 2022 / Published online: 23 May 2022
© The Author(s) 2022

Abstract This work presents a robust method that minimises the impact of user-selected parameter on the identification of generic models to study the coherent dynamics in turbulent flows. The objective is to gain insight into the flow dynamics from a data-driven reduced order model (ROM) that is developed from measurement data of the respective flow. For an efficient separation of the coherent dynamics, spectral proper orthogonal decomposition (SPOD) is used, projecting the flow field onto a low-dimensional subspace, so that the dominating dynamics can be represented with a minimal number of modes. A function library is defined using polynomial combinations of the temporal modal coefficients to describe the flow dynamics with a system of nonlinear ordinary differential equations. The most important library functions are identified in a two-stage cross-validation procedure (conservative and restrictive sparsification) and combined in the final model. In the first stage, the process uses a simple approximation of the derivative to match the model with the data. This stage delivers a reduced set of possible library function candidates for the model. In the second, more complex stage, the model of the entire flow is integrated over a short time and compared with the progression of the measured data. This restrictive stage allows a robust identification of nonlinearities and modal interactions in the data and their representation in the model. The method is demonstrated using data from particle image velocimetry (PIV) measurements of a circular cylinder undergoing vortex-induced vibration (VIV) at Re = 4000. It delivers a reduced order model that reproduces the average dynamics of the flow and reveals the interaction of coexisting flow dynamics by the model structure.

Keywords Reduced-order model · Vortex-induced vibration · Circular cylinder wakes · Spectral proper orthogonal decomposition · Nonlinear ODE system · Sparse systems

1 Introduction

In many engineering applications, the prediction and control of highly energetic coherent flow structures are important in designing, monitoring and optimising fluid systems. Most flows of practical interest are turbulent, and many strategies aim to gain insight into the flow dynamics by approximating only the coherent dynamics of the system with a reduced-order model (ROM). Over the last decades, refinements in experimental and numerical simulation methods have enabled access to very large amounts of high-fidelity data, making data-driven approaches an increasingly interesting choice for deriving ROMs [10]. However, the stochastic
nature of turbulent flows poses two inherent challenges for data-driven approaches. First, the posterior system identification requires extracting deterministic dynamics, which are to be represented by low-order models, from direct observations of the stochastic flow process. Second, the calibration procedure must account for random perturbations of the flow that are extraneous to the candidate model and may lead to overfitting. This work presents a generic approach, which uses a two-stage cross-validation procedure for identifying a low-dimensional system to represent deterministic flow dynamics, to obtain a robust and accurate ROM. This is different from related investigations that use a statistical approach to separate the deterministic and stochastic dynamics in the data [17, 36]. The latter relies on the evaluation of probability density functions of the state variables to infer a stochastic state space model. The main benefit of the proposed cross-validation procedure over statistical approaches is that it requires significantly less data, especially for increasing dimensions of the model.

Many approaches to separate fluid dynamics efficiently are based on modal decomposition techniques. Generally, spatial flow structures are represented by modes and their temporal variations are captured by the corresponding modal coefficients. This allows a representation of the flow dynamics by a reduced set of modes and a simpler description of the dynamics through the temporal evolution of the modal coefficients [10, 24]. Prevalent examples are the decomposition into Fourier modes, dynamic mode decomposition (DMD) [31], proper orthogonal decomposition (POD) [1] and many variations and modified versions of these techniques [39]. Most commonly the modal base of low-order models is constructed from an energy-ranked POD of the flow data [40]. However, a common shortcoming of POD is the lack of a clear separation of individual dynamics into distinct modes, which makes it difficult to build efficient models. SPOD has been developed with the intent to distinguish different dynamics [35] and mitigate this shortcoming of classical POD. The SPOD seeks a short-time temporal coherence in addition to the spatial coherence that is imposed by the modal decomposition. The SPOD used here is based in the time domain and is different from the frequency-domain method with the same name described by Towne et al. [42]. Both SPOD variants share the inclusion of the short-time temporal correlation as detailed by Sieber [34]. However, the SPOD used in the current investigation is implemented completely in the time domain, whereas the other SPOD employs a Fourier transform of the temporal evolution, seeking correlations in the frequency domain. The time-domain SPOD provides a natural way to describe the time dependence of the flow through the expansion coefficients without the need for an ambiguous inverse transformation required in the frequency-domain SPOD [23]. These time-dependent expansion coefficients constitute the backbone for the data-driven reduced order model. In the following, SPOD refers to the time-domain version of SPOD.

In contrast to classical POD, SPOD requires the specification of the temporal correlation domain defined by a filter size (window function). The filter size can be selected based on the governing time scales of the flow, commonly the period time of a dominant oscillation of interest. This decomposition has proven to be a good choice in the following applications. SPOD has shown to be very beneficial for building ROMs based on deep neural networks [22]. It was used also to study sparse models of modal interactions in a cavity flow [28]. Furthermore, through SPOD it was possible to identify and explain the interaction of an acoustic and a hydrodynamics instability in a combustor from experimental data [38]. In a separate study of a swirl stabilised combustor, two coexisting dominant modes could be separated and attributed to coexisting global hydrodynamic instabilities [43]. Applications of SPOD enabled the identification of coherent structures in biofluid dynamics [7]. From the applications, it is clear that SPOD allows the coherent dynamics in turbulent flows to be partitioned into modes that have interpretable dynamics, mostly related to specific hydrodynamic phenomena. This is advantageous for the construction of ROMs, since the correct separation of coexisting phenomena is essential for the construction of efficient and interpretable ROMs.

Recently, machine learning approaches have been introduced in fluid dynamics such as sparse identification of nonlinear dynamics (SINDy) [5]. These use modal decomposition as an input to span a possible function space and sparse regression to identify relevant terms, with the objective of achieving a parsimonious model. This technique showed that efficient representations of canonical fluid dynamic problems can be derived with a sparse system of nonlinear, ordinary differential equations (nODE). However, without precise tuning, adaption and prior knowledge of the flow, the applicability to more complex problems and noisy measurement data has proven challenging. Alternative approaches which are gaining increasing interest are deep neural networks (DNN) [15]. These computational models are inspired by the neural structure of brains. The DNNs are composed of multiple layers through which input features are transformed to capture complex interactions [19]. In contrast to the SINDy algorithm, a trained neural network is used for the representation of the system dynamics [22]. Due to the complexity of these networks, the interpretation of physical relationships becomes more challenging and the extrapolation and robustness of the model need to be ensured. While DNNs show
better performance for some flows [22], it comes at the expense of extensive training using high-fidelity data, higher computational efforts and the tacit risk of overfitting.

The data-driven identification of simplified equations from measurement or simulation data provides an efficient alternative to simulations based on Navier–Stokes equations. The present methodology is especially designed for an initial investigation of turbulent flows, as it does not require a priori knowledge of underlying models or dynamics. Further, it uses as few exogenous parameters as possible without resulting in trivial solutions. Projecting flow data onto a finite-dimensional subspace and using a reduced number of modes as a set of basis functions allows finding a formulation for the deterministic flow dynamics as a system of nODE. Then, with comparatively low computational effort, the implemented two-stage cross-validation (conservative and restrictive sparsification) results in a robust ROM. The ROM retains the capability of reproducing not only the flow statistics, but also provides insight of the physical interactions that can be deduced from the model structure and parameters. This approach is motivated by the potential benefits to engineering applications approximating real-time solutions, which could be used in, e.g. active flow control [14].

The proposed method is broadly comparable to the framework of the SINDy method. It similarly relies on a modal decomposition and the representation of nonlinear dynamics by an appropriate function library. The use of a polynomial library was motivated before by Noack’s xAMC method (now called xROM toolkit [33]), relying on the POD-based Galerkin projection of the Navier–Stokes equations [24]. The SINDy method augments xROM by introducing sparsification to the model. However, depending on the study, the way the sparsity of the system in enforced varies. Originally, a sequential thresholding procedure was proposed that does not balance the model sparsity against fidelity [5]. The resulting model sparsity directly depends on an arbitrarily selected iteration counter to prevent the algorithm to converge on the trivial limit of a stationary model. Further refinements of the SINDy approach suggest regularisation [20] that balance model complexity against fidelity similar to the LASSO method [29,41]. This balance relies on regularisation parameters and further cause a convex optimisation problem that is harder to solve. Furthermore, the user-selection of the regularisation parameters directly controls the model complexity and needs to be varied and adjusted from case to case [9]. The method proposed in this work tries to avoid the direct dependency of the model complexity on user selected parameters. This is achieved by a sequential evaluation of parameters in a brute-force manner and relies on cross-validation to evaluate the model fidelity. In summary, the method proposed here replaces the regularised minimisation problem used in the SINDy method. Note that the utilised SPOD-based ROM is related to the HAVOK approach [4] and can therefore be considered as convolutional coordinates [12], as explained in more detail in [34].

The method is demonstrated for two-dimensional PIV data in the wake of a circular cylinder at Re = 4000 undergoing VIV. Exposing an elastically mounted cylinder to a fluid flow can induce a cylinder vibration due to vortex shedding. The vortex shedding causes oscillating pressure fluctuations which apply forces onto the cylinder inducing a motion. This modifies the dynamics of the vortices, leading to a nonlinear fluid-structure interaction. The response (oscillation amplitude) of the cylinder depends on the oncoming stream velocity, $U_\infty$, and the natural frequency of the cylinder, $f_n$, and is classified through four regimes characterised through the reduced velocity $U^* = U_\infty / D f_n$, with $D$ the cylinder diameter. These regimes are the: (i) initial excitation regime; (ii) "upper branch" with a very high amplitude response; (iii) "lower branch" with a moderate amplitude response; and (iv) a desynchronisation regime [13]. Within these regions, certain wake pattern can be observed [44]. When the normalised velocity is gradually increased within the "upper" and "lower branch", a "lock-in" phenomenon can be observed, where the vortex shedding frequency synchronises with the cylinder oscillation. Further increase in the velocity leads to a desynchronisation of the vortex shedding and cylinder motion. VIV problems have been intensively studied over the last decades, as it entails several engineering challenges and covers numerous of disciplines. A good review of applied methods is provided in [30].

This work is organised as follows: Sect. 2 provides a detailed description of the methodology, introducing the modal decomposition technique (SPOD) and proposing a quantifying method to identify coupled mode pairs with a DMD (Sect. 2.1). The new algorithms of the two-staged sparsification procedure (conservative and restrictive sparsification) are presented in Sect. 2.2. In Sect. 3, the method is demonstrated for PIV data of circular cylinder undergoing VIV in the desynchronisation regime. The results and findings are outlined in Sect. 4.

2 Methodology

The aim of this work is to identify a simplified system for efficiently representing the dominant flow dynamics. As schematically shown in Fig. 1, instead of directly describing the temporal evolution of the flow field $v(x, t)$
with the Navier–Stokes equations ($\mathcal{N}$), the basis of the problem is changed, using a modal decomposition to approximate the instantaneous velocity field with only a few basis functions. Then, the evolution of the temporal coefficients $a_i(t)$ is represented with a nonlinear operator $\mathbf{F}$. In a final step, the approximated field is projected back to the canonical basis.

The general process can be divided into two main steps: first, reducing the dimensionality of the input data by a modal decomposition and, second, building robust ROM from data to approximate the nonlinear operator $\mathbf{F}$ (sparsification procedure). A flow chart summarizing the steps of the two sparsification algorithms is shown in Fig. 2.

### 2.1 Flow field decomposition and mode coupling

In this section, the essential steps of the data decomposition (SPOD) and a method to find coupled modes are presented following Sieber et al. [35]. Data, either measured (e.g. PIV) or computed (CFD), are recorded as snapshots ($M$ spatial points over $N$ time steps).

For PIV and CFD data, usually $M \geq N$ and the initial decomposition steps are more suitably carried out with the snapshot POD [37]. The POD projects the data onto an orthogonal base that is optimal with respect to the representation of kinetic energy of fluctuations, so that the most energetic structures can be represented by a minimal number of basis functions. As a first step, the velocity field $\mathbf{v}(x, t)$ is decomposed into a mean velocity field $\overline{\mathbf{v}}(x)$ and fluctuations $\mathbf{v}'(x, t)$. For the VIV data, the flow shows a spatial symmetry. Following Holmes et al. [10], this property is exploited and the decomposition is performed for a symmetrised flow field using the procedure described in “Appendix A”. By reinforcing natural symmetries of the modes, the convergence rates are increased with respect to the number of employed snapshots [3]. Note that this procedure also optimises the spatial convergence of the later applied filter operation of the SPOD.

Thus, as shown in Fig. 1, a decomposition of $\mathbf{v}(x, t)$ in the form

$$\mathbf{v}(x, t) = \overline{\mathbf{v}}(x) + \sum_{i=1}^{N} \Phi_i(x)b_i(t),$$

(1)

is sought, where $\Phi_i(x)$ are the spatial modes and $b_i(t)$ the corresponding temporal coefficients. The modal basis is constructed from correlated observations of fluctuations in the flow. Therefore, the correlation between individual snapshots is computed using the $L^2$ inner product:

$$\langle \mathbf{u}(x, t_i), \mathbf{v}(x, t_j) \rangle = \int_{\Gamma} \mathbf{u}(x, t_i)^T \mathbf{v}(x, t_j) \, dS,$$

(2)

where $\Gamma$ specifies the spatial region (PIV window) of integration. The elements of the covariance matrix $\mathbf{C}$ represent the covariance over all snapshots and are given by

$$C_{ij} = \frac{1}{N} \langle \mathbf{v}'(x, t_i), \mathbf{v}'(x, t_j) \rangle.$$

(3)
At this point, the only additional step of SPOD is applied. Instead of directly deriving the temporal coefficients with an eigendecomposition of $C$, an additional filter operation is applied to the covariance matrix, which requires the selection of a filter. The filter is specified by the size of a discrete filter in terms of sampling points $N_f$, which is equivalent to a short time span $N_f \Delta t$ given the sampling interval of the signal $\Delta t$. The elements of the resulting filter covariance matrix $W_{ij}$ are given by

$$W_{ij} = \sum_{k=-N_f}^{N_f} g_k C_{i+k,j+k}. \quad (4)$$
Here, a Gaussian filter is applied, where the vector elements $g_k$ are:

$$g_k = e^{-(2.285k/N_f)^2}.$$  

There are three options to deal with boundary points: (i) Periodic boundary conditions are imposed, if the periodic property applies to the investigated flow. (ii) Invalid boundary points are removed. If the filter operation relies on samples outside the acquired time frame, the analysis window is reduced (the option applied to the VIV data). This requires much smaller filter sizes than the number of snapshots to retain sufficient data after the truncation. (iii) Adding zeros to the end of the signal (zero padding). The filter operation allows a continuous transition of the decomposition between the POD and the discrete Fourier transform (DFT). Depending on the filter size, the temporal coefficients of the SPOD modes are better behaved than POD coefficients, i.e. the time evolution is smoother and with reduced spectral bandwidth. Thus, these coefficients are more suitable for deriving a ROM, especially for turbulent flows. The filter size is the only additional parameter compared to POD and should be chosen with respect to a characteristic time scale of the flow. For the VIV measurement, the selected scale corresponds to two periods of the cylinder oscillation. With the eigendecomposition of the filtered covariance matrix, the temporal coefficients $b_i(t)$ and the corresponding eigenvalues $\mu_i$ (representing twice the average kinetic energy of each mode) can be obtained

$$\sum_{k=1}^{N} W_{jk} b_i(t_k) = \mu_i b_i(t_j), \quad \mu_1 \geq \mu_2 \geq \cdots \geq \mu_N.$$  

Since $W$ is a real symmetric positive-definite matrix, its eigenvectors $b_i(t)$ are orthogonal and scaled with the mode energy

$$\frac{1}{N} \sum_{k=1}^{N} b_i(t_k) b_i(t_k) = \mu_i.$$  

From the projection of the snapshots onto the temporal coefficients, the spatial modes can be derived

$$\Phi_i(x) = \frac{1}{N \mu_i} \sum_{j=1}^{N} b_i(t_j) v'(x, t_j).$$  

The identification of mode pairs, typically having a similar amount of kinetic energy, but shifted in phase, is essential to represent the dynamics of coherent structures. Most approaches for identification involve an inspection of the spatial modes shapes and Lissajous figures, which becomes more difficult for complex dynamics. Sieber et al. [35] introduce an unbiased method to quantify coupled dynamics of modes, using DMD. The idea is to find temporal coefficients with the same spectral content, considering a $\pi/2$ phase lag to calculate a spectral proximity measure. Therefore, it is assumed that the temporal evolution of the mode coefficients is captured by a linear operator $T$:

$$b(t + \Delta t) = Tb(t).$$  

With two snapshot matrices $V_1$ and $V_2$ defined as follows

$$V_1 := [b(0) \ b(\Delta t) \ \ldots \ b((N-2)\Delta t)],$$  

$$V_2 := [b(\Delta t) \ b(2\Delta t) \ \ldots \ b((N-1)\Delta t)],$$  

the operator $T$ can be described by

$$V_2 = TV_1,$$

$$T = V_2 V_1^+,$$

where $V_1^+$ is the Moore–Penrose pseudo-inverse of $V_1$. Note, that the explicit computation of $V_1^+$ can be problematic for ill-condition snapshot matrices and Eq. (13) is only used for illustration. The problem can
also be described and solved by a least-square regression. Hence, the latter option is used in the implementation of the procedure using the MATLAB backslash operator. The DMD modes can then be derived by an eigendecomposition of the operator $T$: 

$$ T h_i = v_i h_i. $$

(14)

The modal representation of $b(t)$ can be written as

$$ b_j(t) = \sum_{i=1}^{N_c} h_{ij} e^{\ln(v_i) \Delta t}, $$

(15)

where $h_{ij}$ is the spectral content of the single mode coefficient $b_j$ with respect to the eigenvalue $v_i$. Note, that this notation is only exact for $N_c = N$; however, for the described process $N_c < N$. To mitigate measurement noise in the identification procedure, only modes with an acceptable signal-to-noise ratio should be considered for the calculation of $T$. Modes with very low energy are associated with a low signal-to-noise ratio. To remove these modes, the number of retained modes (sorted by energy in descending order) is truncated after $N_c$ modes.

This number is calculated based on the energy resolved by the SPOD:

$$ E(N_c) = \frac{\sum_{k=1}^{N_c} \mu_k}{\sum_{k=1}^{N} \mu_k}. $$

(16)

For the VIV measurement, the modes are truncated around $E(N_c) = 0.95$. The amplification rates $\sigma_i$ and frequencies $\omega_i$ of the operator $T$ are related to eigenvalue $v_i$ as follows:

$$ \frac{\ln (v_i)}{\Delta t} = \sigma_i + i\omega_i, $$

(17)

where $i := \sqrt{-1}$. The measure of the proximity for ranking the spectral similarity of temporal coefficients $b(t)$ is given by

$$ H_{ij} = \Re \left\{ \sum_{k=1}^{N_c} h_{ik} h_{kj}^* \text{sgn}(\text{Im}(v_k)) \right\} = -H_{ji}. $$

(18)

The elements of the skew-symmetric matrix $H$ are sorted by decreasing magnitude. The first SPOD mode corresponds to the largest element $H_{mn}$ identifying the coupled modes $m$ and $n$. The next lower magnitude of $H$, excluding the row and column of the first element, defines the second coupled SPOD modes, and the process is repeated until all modes are paired. The coupled modes $b_m$ and $b_n$ form the real and imaginary parts of a complex SPOD mode $a_i(t)$ with an amplitude $A_i(t)$ and a phase $\psi_i(t)$

$$ a_i(t) = b_m(t) + ib_n(t) = A_i(t)e^{i\psi_i(t)}, $$

(19)

for fixed index pairs $(m, n)$. The subscript $i$ corresponds to the sorting order. The spatial modes are paired analogously:

$$ \Psi_i(t) = \Phi_m(t) - i\Phi_n(t), $$

(20)

for fixed index pairs $(m, n)$. The entire time-dependence is captured by the state vector $a(t)$. Using the complex definition of the modes, the reconstruction equivalent to Eq. (1) now reads:

$$ v(x, t) = \overline{v(x)} + \sum_{i=1}^{N} \Re(\Psi_i(x) a_i(t)), $$

(21)

Thus, the temporal evolution of the flow can be represented by deriving a ROM for the nonlinear operator $F$ to describe the deterministic dynamics of the flow:

$$ \frac{d}{dt} a(t) = F(a(t)). $$

(22)
2.2 Sparsification

While the decomposition of the velocity fluctuations into SPOD modes in Eq. (1) is exact (when all modes are used), it is desired to select a function basis using the least number of modes \( N_m \), while describing the dominant coherent dynamics sufficiently well.

Mode selection

The mode selection depends on the flow configuration, and the following points need consideration. The SPOD procedure already ranks modes in two ways. The kinetic energy ranking Eq. (6) considers the larger energetic contributions, and the harmonic correlation ranking Eq. (18) provides insight into the possible dynamic significance of the lower energy interactions. The highest ranked modes of both sortings should be selected. Note, that the highest ranked modes in both cases do not have to differ remarkably. The final model prediction can also provide insights, if sufficient modes have been selected to cover the most important interactions. For the VIV data, the first three SPOD modes with the highest energy are also those with the highest harmonic correlation. These represent the vortex shedding synchronised with the cylinder oscillation, the natural vortex shedding, and their dominating nonlinear interaction.

Computation of the derivative

To describe the deterministic dynamics, the derivative of the selected modes provides the basis for the two-stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage sparsification. In order to accurately estimate the numerical derivative from data, while avoiding error amplification due to measurement noise, a sixth-order compact scheme [18] was chosen. For each temporal stage ...
Here, \( A^{P_2}, A^{P_1}, \ldots \), denote higher order polynomials, where \( A^{P_2} \) denotes quadratic nonlinearities of the state vector \( a(t) \) and \((A\overline{A})^{P_2}\) denotes quadratic nonlinearities of the state vector \( a(t) \) and its complex conjugated \( a^{\ast}(t) \):

\[
A^{P_2} = \begin{bmatrix}
    a_1(t_1)^2 & a_1(t_1)a_2(t_1) & \ldots & a_2(t_1)^2 & a_{N_m}^2(t_1) \\
    a_1(t_2)^2 & a_1(t_2)a_2(t_2) & \ldots & a_2(t_2)^2 & a_{N_m}^2(t_2) \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    a_1(t_N)^2 & a_1(t_N)a_2(t_N) & \ldots & a_2(t_N)^2 & a_{N_m}^2(t_N)
\end{bmatrix}
\]

\( (A\overline{A})^{P_2} \) =

\[
\begin{bmatrix}
    a_1(t_1)a_1^\ast(t_1) & a_1(t_1)a_2^\ast(t_1) & \ldots & a_1(t_1)a_{N_m}^\ast(t_1) & a_{N_m}(t_1)a_{N_m}^\ast(t_1) \\
    a_1(t_2)a_1^\ast(t_2) & a_1(t_2)a_2^\ast(t_2) & \ldots & a_1(t_2)a_{N_m}^\ast(t_2) & a_{N_m}(t_2)a_{N_m}^\ast(t_2) \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    a_1(t_N)a_1^\ast(t_N) & a_1(t_N)a_2^\ast(t_N) & \ldots & a_1(t_N)a_{N_m}^\ast(t_N) & a_{N_m}(t_N)a_{N_m}^\ast(t_N)
\end{bmatrix}.
\]

Generally, the library may be composed of all possible forms of nonlinear functions and polynomial combinations of the state vector to account for nonlinear dynamics. In the next step, linear-dependent library functions are removed from the library, using a \( O-R \)-decomposition.

**Sparsification procedure**

Assuming that only a few of the library functions are required for the representation of \( F \), a sparse regression problem can be formulated. The idea is to determine the important library coefficients, i.e., finding the sparse vectors of library coefficients \( \Xi = [\xi_1 \xi_2 \ldots \xi_{N_m}] \) for each mode. Hence, Equation (22) can be formulated as

\[
\hat{A} = \Theta(A)\Xi.
\]

Nonzero entries (active coefficients) of \( \xi_k \) define which of the library functions are contributing to compute mode coefficient \( a_k(t) \). Initially, all coefficients are considered possible candidates.

The sparsification procedure seeks to identify the important, i.e. statistically significant library coefficients and thereby reduces the model complexity. This avoids the risk of overfitting and increases the robustness of the ROM, while maintaining its accuracy. The procedure consists of two cross-validation stages, which allow to evaluate the statistical importance of individual coefficients and minimise computational costs. A simplified conceptual schematic of the process for both stages is depicted in Fig. 3.

The first stage (conservative sparsification) focuses on the approximation of the derivative and the second stage (restrictive sparsification) on the model prediction. In both stages, the selected library coefficients are trained with a subset of the data based on the optimal approximation of the derivative and validated against the remaining data. For this, the measurement data are split into continuous data segments of length \( N_s \). Here, the length is based on the SPOD filtering operation \( (N_s = N_f) \). Then, these segments are \( N_{\text{train}} \) times
randomly assigned to training $a^T(t_{\text{train}})$ and validation $a^T(t_{\text{val}})$ data. Based on exploratory trials using different distribution ratios, results using 20% data segments for training and 80% for validation yielded consistently robust models. The ratio should be chosen with respect to the amount of available data maximising the validation data while keeping enough data for training to reduce the risk of overfitting. For the VIV data, $N_{\text{train}} = 100$ showed good convergence for both stages with respect to the final selection of the library coefficient.

Conservative sparsification

The calculation procedure for the conservative sparsification is presented in Algorithm 1. The principle of this cross-validation step is a bottom-up approach of a greedy algorithm to approximate the derivative for each mode, i.e. finding the optimal choice of a library coefficient in each iteration to improve the approximation of the derivative. Initially, all library coefficients are set passive, i.e. set to zero and excluded from training, and the individual coefficients are gradually activated, trained and evaluated. The training of the library coefficients for the $k$th individual mode for the $i$th training data set is computed as follows:

$$
\xi^\text{min}_k = \arg\min_{\xi} \| \dot{\hat{a}}^T_k(t_{\text{train}}) - \Theta (a^T_k(t_{\text{train}})) \xi \odot d^i_k \|
$$

(31)

$$
\xi^\text{train}_k = \xi^\text{min}_k \odot d^i_k.
$$

(32)

Here, the operator $\odot$ stands for an elementwise product and $d^i_k$ is a Boolean vector controlling which coefficients are active and passive. Note, that this notation is chosen to indicate dependencies of the variables and to be in line with Algorithm 1, but the equations are correct even without the index for the training set variation. With the trained coefficients, the approximation residual $e_{jk}$ for the $j$th library function of the $k$th mode of the derivative can be derived:

$$
e_{jk} = \| \dot{\hat{a}}^T_k(t_{\text{val}}) - \Theta (a^T_k(t_{\text{val}})) \xi^\text{min}_k \|^2.
$$

(33)

When all passive coefficients are individually tested, the library coefficient that shows the highest improvement of the residual in an iteration is kept (permanently) activated. This procedure is repeated until further activation of coefficients no longer result in improvements with respect to the residuals. This results in a set of possible candidates of library coefficients for each random training and validation data set, which are merged into a final list of possible candidates.

Restrictive sparsification

The calculation procedure of the restrictive sparsification is given in Algorithm 2. This procedure is a top-down approach, that starts with the candidate list of the conservative sparsification. In contrast to the first stage, the idea is to optimise the coefficient selection based on the model prediction on a system level and not for each mode individually. In an iterative procedure, models are built excluding one of the library coefficients candidates, trained and then, a short time integration is performed. As a reference, an additional integration with all candidates included is computed. The training of the library coefficients follows the same procedure as in the conservative sparsification algorithm, but for the entire system simultaneously:

$$
\Xi^\text{min} = \arg\min_{\Xi} \| \hat{a}^T(t_{\text{train}}) - \Theta (a^T(t_{\text{train}})) \Xi \odot D \|
$$

(34)

$$
\Xi^\text{train} = \Xi^\text{min} \odot D.
$$

(35)

where $D$ is a Boolean matrix that contains the information of currently active and passive library coefficients. Data points at the beginning of every non-training segment are used as initial values. This ensures an equal distribution and weighting of measurement data used as initial values, i.e. avoiding initial values taken from consecutive data points. As the equal-length segmentation is used on a periodic flow, the start of the segment is randomly shifted within the length of one oscillation period to prevent an over-weighting of a specific phase.
To determine the deviation of the model prediction from the true state, the following initial value problem is \( N_p \) times solved (for the \( n \)th initial value \( \tilde{f}^0 \)) and the residual is computed:

\[
    r_n = \left\| \int_{\tilde{f}^0,n}^{t_n} a(t) - a(t_{0,n}) - \int_{\tilde{f}^0,n}^{t} F(a(s)) \, ds \right\|^2 dt. \tag{36}
\]

The number of simulated time steps \( N_t \) is based on two periods of the cylinder oscillation. A solver based on an explicit Runge–Kutta of fourth- and fifth-order (Dormand–Prince method) is used to integrate the model. If the integration is aborted (unstable model), the residual is penalised with \( r_0 k \). Simulations showed that this mainly occurs when the most important coefficients are excluded. With difference of the residuals (active and passive coefficient), a quality measure \( Q \) for each coefficient candidate can be computed (see Algorithm 2).

The quality measure allows to determine, if a coefficient is statistically significant. Hence, all coefficients that are not statistically significant (negative values of \( Q \)) are removed. With the new list (sparse matrix \( D \)) of candidates, the procedure is repeated, until the list of candidates remains unchanged. In a final step, the resulting library coefficients (active coefficients in matrix \( \Xi_{\text{final}} \)) are trained using the entire measurement data as training data. Then, with a simulation with resulting ROM of the temporal coefficients \( a^{\text{sim}}(t) \), the flow field can be reconstructed.

**Algorithm 1: Conservative sparsification**

- **Input**: Number of different training and validation data sets \( N_{\text{train}} \)
- **Output**: List of possible library coefficient candidates \( \Xi \)

\[
\begin{align*}
N_{\text{lib}} & = \text{number of different library functions in } \Theta; \\
& \text{repeat for different test sets} \\
\text{for } i = 1 \text{ to } N_{\text{train}} \text{ do} \\
& \quad \text{Randomly select training data } t_{\text{train}} \text{ and corresponding validation data } t_{\text{val}}; \\
& \quad \text{Initialise } D_i \text{ with all } 0; \quad \text{set all coefficients to passive} \\
& \quad \text{for } k = 1 \text{ to } N_{\text{m}} \text{ do} \\
& \quad \quad \text{// test all library coefficients} \\
& \quad \quad \text{if } d_{ik} = 0 \text{ then} \quad \text{activate coefficient} \\
& \quad \quad \quad \text{train the coefficients } \xi_{\text{train}}^k \text{ with Eq. (31) and Eq. (32)}; \\
& \quad \quad \quad \text{set } d_{ik} \text{ to } 1; \quad \text{deactivate coefficient} \\
& \quad \text{end} \\
& \quad \text{end} \\
& \quad \text{for } j = 1 \text{ to } N_{\text{lib}} \text{ do} \\
& \quad \quad \text{// if coefficient is not used} \\
& \quad \quad \text{if } e_{jk} < r_{c-1,k} \text{ then} \quad \text{add coefficient permanently} \\
& \quad \quad \quad j^* = \arg \min_j (e_{jk}); \quad \text{pick most significant coefficient} \\
& \quad \quad \quad d_{j^*k} = 1; \\
& \quad \quad \text{end} \\
& \quad \text{end} \\
& \quad c = c + 1; \\
& \text{end} \\
& \text{D} = D_1 \cup \ldots \cup D_{N_{\text{train}}}; \quad \text{final matrix with a reduced number of active coefficients}
\end{align*}
\]
Algorithm 2: Restrictive sparsification

Input: Number of different training sets $N_{\text{train}}$, Number of randomly selected initial values $N_p$, Number of integrated time steps $N_t$, $D$ matrix with initial selection of active library derived with conservative sparsification

Output: Sparse matrix of trained of library coefficients $\Xi^\text{final}$

1. $c = 0$; // iteration counter
2. $N_{\text{lib}} = \text{number of different library functions in } \Theta$; // initialise indices for coefficient deactivation
3. $(j^*, k^*) = \emptyset$; // repeat if coefficients were deactivated
4. while $c = 0 \text{ OR } (j^*, k^*) = \emptyset$ do // repeat for different test sets
5. set $Q^c$ with all zeros of size $\Xi$; // initialise quality function
6. for $i = 1 \text{ to } N_{\text{train}}$ do // repeat for different test sets
7. randomly select training data segments $t_{\text{train}}$;
8. train all library coefficients $\Xi_{\text{train}}$ with Eq. (34) and Eq. (35); // define function
9. for $n = 1 \text{ to } N_p$ do // iterate all initial values
10. randomly select initial value $t_0, n$;
11. integrate the model & compute deviation $r_n$ from the true state with Eq. (36);
12. end
13. for $k = 1 \text{ to } N_m$ do // iterate all modes
14. for $j = 1 \text{ to } N_{\text{lib}}$ do // iterate all library coefficients
15. if $d_{jk} = 1$ then // if coefficient is used
16. set $d_{jk}$ to 0; // deactivate current coefficient
17. train the coefficients $\Xi_{\text{train}}$ with Eq. (34) and Eq. (35);
18. $F(a) = \Theta(a) \Xi_{\text{train}} \circ D$; // update model function
19. for $n = 1 \text{ to } N_p$ do // iterate all initial values
20. integrate the model & compute the deviation $r_n$ from the true state with Eq. (36) $\Delta r_n = r_n^\text{all} - r_n$;
21. $q_{jk}^c = q_{jk}^c + \Delta r_n$; // update quality function
22. end
23. set $d_{jk}$ to 1; // activate current coefficient
24. end
25. end
26. $Q^c = \frac{1}{N_{\text{train}}} \frac{1}{N_p} Q^c$; // averaged quality function
27. end
28. find set of indices $(j^*, k^*) \subseteq \{(1, 1), \ldots, (N_m, N_{\text{lib}})\}$ where $Q^c < 0$;
29. for all sets: $d_{j^*, k^*} = 0$; // deactivate library coefficients that decrease model quality
30. $c = c + 1$;
31. end
32. select all measurement data as training data;
33. train the coefficients $\Xi_{\text{final}}$ with Eq. (34) and Eq. (35);

3 Application to experimental data

In this section, the presented method is applied to experimental data of the VIV of a circular cylinder. The datasets analysed during the current study are available from the authors on reasonable request.

3.1 Experimental setup

The experiments were carried out at the University of Calgary in a free-surface water channel facility. Details of the experimental setup can be found in “Appendix D” and Riches et al. [26]. This work focuses on the desynchronisation regime, as it involves complex dynamical interactions of both, the motion induced by the cylinder oscillation and the vortex shedding. In this regime, the spectral signature of the velocity fluctuations and the independently measured lift force show peaks at both the shedding frequency $f_{sh}$ and cylinder vibration frequency, $f_c$, allowing a good demonstration of the method. The experiments were conducted for a reduced velocity of $U^* = 8.75$. The system has one degree of freedom with the cylinder moving transversely to the flow.
Fig. 4 Basic representation of the VIV system: circular cylinder (black circle) with its maximal deflection (dashed lines), snapshot of the PIV window with the displayed $v$-component of the velocity field (blue and red) and streamlines (black) on the left and the reconstructed snapshot of the velocity field using the three most energetic SPOD modes on the right.

Fig. 5 The power spectra (By Parseval’s identity, the integration of the power spectral density (PSD) over the frequency (in Hz) returns the variance of the signal.) of the measured velocity fluctuations at a point $x/D = 1.19, y/D = -0.04$ (where $u'_i/u'_i$ is maximum): $u$-component in blue and $v$-component in orange. The dashed, black line shows the theoretical decay according to Kolmogorov’s -5/3 power law.

A representative instant (PIV snapshot) of the wake flow is depicted in Fig. 4. Here, instantaneous streamlines overlay the iso-contours of the $v$-component of the velocity. The wake shows periodically shed vortices. The perturbations due to the cylinder displacement $s(t)$ interact with these structures. These interactions result in an asymmetrical displacement of vortices. To describe the dynamics of the perturbed vortex street, several modes in combination are required. The observed asymmetry is expected to be captured by at least one mode, which needs to be considered in the mode selection process. The turbulent nature of the wake flow can be inferred from Fig. 5. Shown are the power spectra of both velocity components for a representative point $(x/D = 1.19, y/D = -0.04)$ where $u'_i/u'_i$ is maximum. The dominant frequencies in the $v$-component can be associated with the cylinder oscillation ($f_c$) and the natural vortex shedding ($f_{sh}$). The energy associated with the mechanical cylinder oscillations is concentrated in a narrow frequency band, while that for the vortex shedding exhibits the typical spectral broadening associated with perturbations due to a stochastic process. In the $u$-component, dominant frequencies are the second harmonic of the cylinder oscillation ($2f_c$) and an inter-harmonic, $f_{sh} - f_c$. The excitation of the inter-harmonic is associated with interaction of the cylinder motion and the vortex shedding. Approximately 40% of the spectral energy content is associated with the broad-band stochastic fluctuations. The main challenge in the form of the ROM is its ability to represent the interactions relating these dominant frequencies. In the following, the resulting model and flow representation are discussed in greater detail.
3.2 SPOD results

Figure 6 illustrates the SPOD modes, which consist of coupled mode pairs using DMD as described in Sect. 2.1. Shown are the $v$-component of the spatial (paired) mode, the evolution of the corresponding temporal coefficients and the spectrum of $\Re(a_i)$, which is representative for the coupled mode as the spectrum of $\Im(a_i)$ is almost identical. The first mode in Fig. 6 corresponds to the cylinder oscillation, as $f_c$ matches the peak frequency of the independently measured lift force. The second mode is associated with the vortex shedding process, noting that its spectral energy is concentrated at the vortex shedding frequency $f_{sh}$ and the distribution of extrema in its spatial component matches the number of vortices depicted in Fig. 4. Similar to the measure power spectrum in Fig. 5, the peak of the vortex shedding frequency shows a broader bandwidth. This indicates some sort of variation or disturbance in the vortex shedding process. Generally, the envelope of the oscillations of the first mode tends to be inversely related to the amplitude of oscillations of the second SPOD mode suggesting an energy exchange, or interference, between these modes. The frequency of the third mode corresponds to the difference of the vortex shedding and cylinder oscillation frequencies ($f_t = f_{sh} - f_c$ from Fig. 6). Hence, the slow varying third mode is considered to represent an interaction between the first and second SPOD mode. This interaction is expected to increase the cycle-to-cycle variation in the vortex shedding frequency and lead to a broader frequency peak in the power spectra around the shedding frequency as can be observed in the measurement (Fig. 5). The spatial component of this mode is, in contrast to the previous two, asymmetrical. This mode can account for an asymmetry shift in the flow field observed in Fig. 4. The

Fig. 6 (left two columns) the spatial modes of the $v$-component of the four most energetic coupled SPOD modes of the velocity field, (mid) a representative time interval showing the evolution of the temporal coefficients $a_i$ and (right) power spectra of $\Re(a_i)$.
fourth SPOD mode corresponds to the second harmonic, $2 f_c$, of the cylinder oscillation. Its energetic content is relatively low and the amplitude of the oscillation, or envelope, appears directly related to that of the first mode. For the SPOD, a filter length $N_f = 32$ was used, which corresponds to two periods of the cylinder oscillation. With this choice, a good separation of the dynamics is already visible, as the mode spectra show very distinct and narrow peaks, while representing a significant amount of total fluctuation energy. Thus, the temporal coefficients are better behaved than the coefficients of a classical POD and therefore more suitable for the construction of a ROM. The influence of the SPOD filter lengths is discussed in “Appendix B”. Within the scope of illustrating the methodology, only the three most energetic coupled modes are selected to derive the ROM, as this is the minimal group of modes for a meaningful representation of the process and they already describe a complex interaction. As indicated with the fourth mode, additional coherent motions are captured in modes with significantly less energy content (low impact on the dominating dynamics), so that with the chosen degree of complexity these modes are not represented by the model.

3.3 Final reduced-order model

The sparsification procedure starts with building a polynomial function library of cubic order, as illustrated in Eq. (27) using the three most energetic modes depicted in Fig. 6. The conservative and restrictive sparsification algorithms (Algorithms 1, 2) are applied to retain only the statistically significant library coefficients and corresponding functions. With this step, the number of possible library coefficients is reduced from 252 to only 15. In a final step, the model is calibrated against all available data. The resulting selection is presented in Table 1 and the corresponding system of ODEs in Eq. (38). Table 1 shows the coefficients, the calibrated values, the corresponding library functions and the system residual change $Q^{\text{sys}}$ of the coefficient. The system residual change represents the statistical relevance of an individual coefficient for the system description. The determination of $Q^{\text{sys}}$ follows the idea of the restrictive sparsification procedure, i.e. evaluating the relevance of the coefficients by comparing the model performance including and excluding the coefficient. For this, the measurement data are split into $N_{\text{segs}}$ equal-length segments for which the model is simulated. To avoid an over weighting of a specific cylinder oscillation phases, the initial values of the integration are shifted randomly within a period. The system residual change for each coefficient is determined as:

$$Q^{\text{sys}} = \frac{1}{N_{\text{segs}}} \sum_{n=1}^{N_{\text{segs}}} R_n^{\text{coeff}} - R_n^{\text{all}}$$

where $R_n^{\text{all}}$ is the residual of the model using all selected coefficients, $R_n^{\text{coeff}}$ is the residual excluding the analysed coefficient and $R_n^0$ is the variance of the measured signal. With the coefficients in Table 1, the final system of ODEs then reads:

| No. | Mode | Coeff. | Library function | $Q^{\text{sys}}$ in % | Coeff. value |
|-----|------|--------|------------------|-----------------------|--------------|
| 1.  | $a_1$ | $\alpha_1$ | $a_1$ | 46.565 | $0.00518 + 0.13177i$ |
| 2.  | $a_2$ | $\beta_1$ | $a_2$ | 35.383 | $0.00028 + 0.20098i$ |
| 3.  | $a_3$ | $\gamma_1$ | $a_3$ | 21.71 | $-0.00450 + 0.07941i$ |
| 4.  | $a_3$ | $\gamma_2$ | $a_2 a_1^*$ | 1.281 | $0.00194 + 0.00021i$ |
| 5.  | $a_3$ | $\gamma_3$ | $a_3 [a_1]^2$ | 1.259 | $-0.00010 + 0.00030i$ |
| 6.  | $a_2$ | $\beta_2$ | $a_2 [a_1]^2$ | 1.053 | $-0.00011 - 0.00036i$ |
| 7.  | $a_1$ | $a_2$ | $a_2 [a_1]^2$ | 0.99 | $0.00003 + 0.00015i$ |
| 8.  | $a_2$ | $\beta_3$ | $a_2 a_3^*$ | 0.793 | $0.00247 + 0.00058i$ |
| 9.  | $a_2$ | $\beta_4$ | $a_1 [a_1]^2$ | 0.612 | $-0.00003 - 0.00008i$ |
| 10. | $a_1$ | $\alpha_3$ | $a_1 [a_1]^2$ | 0.232 | $-0.00005 - 0.00004i$ |
| 11. | $a_1$ | $\alpha_4$ | $a_1 [a_3]^2$ | 0.218 | $-0.00015 - 0.00006i$ |
| 12. | $a_1$ | $\alpha_5$ | $a_2 [a_1]^2$ | 0.095 | $0.00002 - 0.00004i$ |
| 13. | $a_2$ | $\beta_5$ | $a_1 [a_3]^2$ | 0.076 | $-0.00017 + 0.00026i$ |
| 14. | $a_3$ | $\gamma_4$ | $|a_1|^2$ | 0.071 | $0.00004 + 0.00005i$ |
| 15. | $a_1$ | $\alpha_6$ | $a_1 [a_1]^2$ | 0.059 | $0.00002 + 0.00004i$ |
\[
\frac{da_1}{dt} = \alpha_1 \cdot a_1 + \alpha_2 \cdot a_2 |a_2|^2 + \alpha_3 \cdot a_1 |a_1|^2 + \alpha_4 \cdot a_1 |a_2|^2 + \alpha_5 \cdot a_2 a_3 a_3^* + \alpha_6 \cdot a_1 a_1 a_3^* \quad (38a)
\]
\[
\frac{da_2}{dt} = \beta_1 \cdot a_2 + \beta_2 \cdot a_2 a_2 a_1^* + \beta_3 \cdot a_2 a_3 + \beta_4 \cdot a_1 a_1 a_3^* + \beta_5 \cdot a_1 |a_3|^2 \quad (38b)
\]
\[
\frac{da_3}{dt} = \gamma_1 \cdot a_3 + \gamma_2 \cdot a_2 a_1^* + \gamma_3 \cdot a_3 |a_1|^2 + \gamma_4 \cdot |a_1|^2 \quad (38c)
\]

The usage of specific library functions in combinations with the library coefficients provides insights into the flow physics. Here, the final selection (Table 1) can be categorised into different groups. The first and most important group with respect to the impact on the system are linear terms, which describe in this context a complex oscillator. The pairing of two real-valued coefficients to one complex-valued coefficient is introduced to link modes in the model, which are physically connected. On the level of the model ODE, this drastically simplifies how the oscillatory mode pair is represented. The description of linear oscillator by a pair of real-valued coefficients reads as follows [21,24]:

\[
\frac{d}{dt} \begin{bmatrix} a_r \\ a_i \end{bmatrix} = \begin{bmatrix} \sigma - \omega \\ \omega \sigma \end{bmatrix} \begin{bmatrix} a_r \\ a_i \end{bmatrix}, \quad (39)
\]

where \(a_r\) and \(a_i\) constitute the real and imaginary part of the coupled mode-pair \(a = a_r + ia_i\). The corresponding linear amplification rate \(\sigma\) and oscillation frequency \(\omega\) can be similarly composed to a complex coefficient \(\alpha = \sigma + i\omega\). This results in the simple representation of the complex oscillator as

\[
\frac{da}{dt} = \alpha a. \quad (40)
\]

Thus, for the linear terms, the imaginary part of the library coefficient determines the frequency of the corresponding mode. The frequencies of the modes derived by the calibrated library coefficients presented in Table 1 deviate less than 3\% from the peak frequencies of the spectra in Fig. 6.

Similar to a nonlinear oscillating system described by the Stuart–Landau equation [16] and the generalised mean-field model [21,24], the ROM includes nonlinear terms, where the mode \(a_j\) is coupled with the amplitude but not the phase of a mode \(a_k\):

\[
\frac{da_j}{dt} = + \cdots a_j |a_k|^2 + \cdots. \quad (41)
\]

Within the model, these terms address nonlinear coupling that results in changes of the linear terms described above. Thus, they influence the energy transfer between the modes, which can lead to a saturating effect stabilising the system. From a perspective of the flow dynamics, these terms can be understood as a mean field correction [24].

Further investigations of the coefficients show, for example for mode \(a_3\) of Eq. (38c), a three mode interaction of the two complex coefficients \(a_1 = A_1 e^{i \omega_1 t}\) and \(a_2 = A_2 e^{i \omega_2 t}\), which results in the observed frequency mixing. This is clearly seen from the complex-valued multiplication

\[
a_2 a_1^* = A_2 e^{i \omega_2} A_1 e^{-i \omega_1} = A_1 A_2 e^{(i \omega_2 - \omega_1) t}, \quad (42)
\]

represented by coefficient \(\gamma_2\) of the model. It represents a forcing of mode \(a_3\) at the difference of the individual frequencies, which corresponds exactly to the observed composition of mode frequencies in Fig. 6c, where the difference of forced shedding \((a_1)\) at \(f_c = 0.128\) and natural shedding \((a_2)\) at \(f_{sh} = 0.203\) leads to the interaction mode frequency of \(f_i \approx 0.203 - 0.128 \approx 0.08\). The modelling process identifies mode \(a_3\) with a negative real part in the linear coefficient \(\gamma_1\). Therefore, this interaction mode is actually damped and only observed due to the forcing by the interaction of cylinder oscillation \(a_1\) and the natural vortex shedding \(a_2\).

To provide another example for the interpretation of the model coefficients, \(\beta_4 (a_1 a_1 a_3^*)\) is discussed. It describes a triadic interaction of the second harmonic of the cylinder oscillation \(a_1 a_1\) with the interaction mode \(a_3\) acting on the vortex shedding mode \(a_2\). It corresponds to a frequency of \(2 \times 0.128 - 0.08 = 0.176\), which will be further discussed with the dynamics of the model presented in the following section. Note that the physics discussed only represents correlations in the data and most likely, but not necessarily, represents actual fluid dynamic interactions.

In a comparable way, other terms can be analysed and their contributions assessed, which helps to conceptualise internal mechanisms and interaction of the underlying physics.
3.4 Model predictions

There are two aspects that need to be considered when assessing the performance of the ROM. First, the reduced number of modes used for the model only represents around 60\% of the turbulent kinetic energy (TKE) of the flow. Thus, only the most energetic and dominating flow structures are represented by the model and other coherent, and low-energy structures are not captured. Secondly, the model does not account for any random perturbation. Hence, the stochastic characteristics of the turbulence make an accurate long-term prediction with a deterministic model impossible. The deviations from constant amplitude (limit-cycle) oscillations are similarly observed for an isolated global instability in a turbulent flow [17,36]. There, the deviations from the limit-cycle are explained by turbulent perturbations and related to the turbulence intensity.

As a consequence, simulations of the final ROM often start to deviate from the actual temporal coefficients after a short period. In Fig. 7, observed and predicted evaluations of the real part of the temporal coefficients of the mode pairs are compared for representative time-series excerpts. Note, that the second part of the mode pair $\Im(a_i)$ shows the same behaviour. Generally, the higher-energy modes are better approximated over longer time intervals. For the lower-energy modes, for example $a_3$ in Fig. 7c, this deviation can be observed after two periods of the cylinder oscillation.

A more general statement of the performance can be obtained from the averaged simulation residual (for $N_{\text{init}} = 300$ randomly chosen initial values):

$$
\epsilon_S = \sqrt{\frac{1}{N_{\text{init}}} \sum_{n=1}^{N_{\text{init}}} |d_i - a_{\text{sim}}|^2} \sqrt{\frac{1}{N_{\text{init}}} \sum_{n=1}^{N_{\text{init}}} |a_i|^2},
$$

as displayed in Fig. 8. The gradient of the residual provides information on how fast the predictions deviate and to which extent the measurements are governed by stochastic influences. The convergence of the model to a certain error level indicates a general stability, as there is no unlimited growth independently of the initial values.
Fig. 8 Residual of various simulations as thin lines (different initial values) and the averaged simulation residual (thick lines) of the calibrated ROM for the three most energetic SPOD modes; a the mode related to the cylinder oscillation (blue), b the vortex shedding mode (orange), c the interaction mode (yellow) and d the averaged simulation residual of all three modes.

Fig. 9 Representative time interval showing the temporal evolution of the real part $\Re(a_i)$ of the three most energetic SPOD modes (blue), a concatenated prediction $a_{\text{sim}}^i$ with the final ROM (orange) and the initial values used for the simulations (red circles); modes related to a cylinder oscillation, b the vortex shedding c their interaction.

The limitation of the long-term prediction accuracy raises the question: How precise can the statistics of the flow dynamics be represented by the model? To address this, simulations were conducted over equal-duration segments of the time series. The simulations were reinitialised at the beginning of each segment. The concatenation $a_{\text{sim}}^i$ of the simulated segments reconstructs the entire time series of the temporal coefficients. Figure 9 depicts samples of the observed coefficients overlaid with the concatenated segments. With a variation of the simulation duration, i.e. increasing the length and reducing the number of segments, one can evaluate how well the deterministic model performs compared to the measured statistics. Figure 10 displays the variance...
Fig. 10 Variance of the original temporal coefficient (blue) and variance of reconstructed signals with different simulations lengths (orange, computed simulations lengths are indicated with a star); modes related to (a) cylinder oscillation, (b) the vortex shedding and (c) their interaction for each mode based on the measurement and the concatenated model prediction. The definition of variance is given by:

\[ \mathbb{E}(X) = \frac{1}{N} \sum_{i=1}^{N} X_i = \mu, \]  
\[ \text{Var}(X) = \frac{1}{N-1} \sum_{i=1}^{N} |X_i - \mu|^2, \]

where \( \mathbb{E}(X) \) is the expected value of the signal and \( \text{Var}(X) \) its variance. It is a representation of the average mode energy. Hence, the deviation of the predicted from the actual (measured) variance provides insight of the model energy balance. For short simulations \( tU_\infty/D \leq 50 \), Fig. 10 shows a very good agreement of the actual and simulated variance. This is in line with the expectation, since for these time intervals the influence of stochastic perturbations is considerably lower. However, for longer simulations the prediction of the energy balance within the model differs from that of experimental data. Here, the downward drift of the first mode Fig. 10a and the upward drift of the second mode Fig. 10b indicate that in the model more energy is transferred from the first to the second mode than observed in the data. For each segment of the concatenated coefficients, i.e. for each simulation of the temporal coefficient, a fast Fourier transform (FFT) of the measured data is computed. The average of these spectra (for \( tU_\infty/D \approx 240 \)) is shown in Fig. 11. These spectra are representative for the mode pair and correspond to relatively long simulations, i.e. long segments (30% of the realisation data) and show that the model still captures the general, spectral characteristics of the flow. The difference between the predicted and the measured spectra can be attributed to several factors. The deviation is most obvious in the decay of the flanks next to the peak. The model only explains a narrow spectral contribution at the peak frequency of each mode, whereas the observed (measured) mode coefficients include further dynamics. The previously illustrated triadic interaction represented by the coefficient \( \beta_4 \) of the model describes a forcing of the vortex shedding mode at a frequency of 0.176. This causes the hump to the left of the main peak of the simulated spectrum in Fig. 11b. However, such nonlinear mechanisms explain only part of the measured...
spectral broadening. Since the model only includes the first three SPOD mode pairs, which account for about 60% of the turbulent kinetic energy, the difference could be due to the neglected effect of higher modes on the leading modes. The model thus captures the interactions between the large amplitude motions, but does not explicitly account for lower energy perturbations. The discarded modes, representing further coherent dynamics or stochastic fluctuations, might thus be necessary to model the observed broader spectral peaks. Instead of adding further modes, the unresolved perturbations can be modelled by a generic stochastic forcing, which effectively emulates the observed broadening of the spectral peaks [25]. This approach does not improve the predictive capabilities of the model since it only adds random perturbations. However, it allows better reproduction of the average statistics [36]. In this context, it should be noted that, although the perturbations can be considered as additive random forcing for modelling purposes, in the experimental data they are not filtered out by the decomposition. The data were recorded from the entire system including the disturbances. Therefore, the unresolved disturbances are implicitly included as frequency fluctuations of the resolved data. While there could also be a broadening of the spectra due to measurement noise, this effect is expected to be secondary, as such contributions are generally poorly correlated in space and time and are filtered out by the modal decomposition [32]. Regardless of the reason for the discrepancy, obtaining directly a more accurate model would be at the cost of a significant increase in its complexity. The potential benefit of using just a few additional modes is usually small. Since the remaining modes contain much less energy, substantial differences in the model usually require many more modes. Furthermore, this mainly translates the problem of the neglected action of unresolved modes on resolved modes towards higher modes. The alternative representation by generic stochastic forcing requires further modelling of the forcing characteristics, which is not straightforward [2,6]. The choice of the right model complexity ultimately depends on the intended purpose of the ROM, balancing captured dynamics and predictive accuracy against complexity.

With the simulated temporal coefficient and the corresponding spatial modes, a simulated velocity field can be reconstructed using Eq. (21). This allows to estimate the Reynolds stresses in the cylinder wake and gain some insight of predicted flow physics. Figure 12 displays the shear Reynolds stresses at the cross section with the maximum measured Reynolds stresses for: the measured field, a reconstructed field using the first three SPOD modes and the stresses reconstructed with the simulated temporal coefficients. The estimation of the shear stresses provides a measure to evaluate if the model can represent the influence of the large flow structures. Using the measured temporal coefficients to reconstruct the flow field provides a reference for comparison.

As both reconstructed Reynolds stresses account only for around \( \sim 60\% \) of the TKE, the extrema of these stresses are expected to be lower than the stresses derived from the measurement. However, the simulated stresses are already very close to the stresses described with the three SPOD modes. Slightly reducing the simulation length leads to an excellent match of the reconstructed stresses. Thus, with respect to the represented energy content, the deterministic ROM is capable of accurately reproducing the statistics associated with the flow physics. Further, for this flow configuration the influence of the stochastic perturbations is limited to the extent that the dominating dynamics can be captured with a few modes and represented by the model.
4 Concluding remarks

A robust method has been presented for deriving reduced order models from data to describe the underlying flow dynamics. The methodology is demonstrated for the turbulent wake of a cylinder undergoing vortex-induced vibration. Its robustness is achieved by establishing objective criteria in the decision process underpinning a two-step sparsification procedure. This approach aims to minimise user interaction and ensure an unambiguous procedure.

The separation of the flow dynamics is realised by a modal decomposition of the flow data. A reduced set of spatial modes represents dominating flow structures and their dynamics through their corresponding temporal coefficients. The evolution of these coefficients allows a simplified description of the dynamics. A system of ODEs is constructed using linear and nonlinear combination (here polynomial) of the coefficients to model the dynamics. These combinations define a library of possible functions. The critical step in the presented method is the optimised selection (sparsification) process to identify the minimal number of relevant library functions and corresponding coefficients implementing a two-stage cross-validation procedure. The first stage (conservative sparsification) works as a prefiltering of the library coefficients, where all of these coefficients are step-wise tested and relevant coefficients with respect to the approximation of the derivative are kept. In the second stage (restrictive sparsification), an iterative process is used to quantify the relevance of each library coefficient for the system and remove non-physical coefficients. Briefly, these coefficients are used to build models, validated them against the measured evolution of the temporal coefficients and perform short-time predictions. The final selection of library coefficients follows a statistical significance test by the successive exclusion of coefficients.

The major improvement gained through the sparsification procedure addresses three challenges in particular. Firstly, in contrast to previous approaches, it not only identifies coefficients that influence the approximation of the derivative, but also determines how sensitive the prediction is to a given coefficient. This enhances the reliability and robustness of the final ROM. Secondly, introducing cross-validation in both stages of the sparsification reduces the risk of overfitting. Within the conservative sparsification, this is accomplished by using randomly chosen segments of the computed derivative to train the coefficients and the non-training segments for validation. For the second stage, the training is performed in a similar way, but the validation is based on the prediction of the temporal evolution of the non-training segments and not its derivative. Using only 20% of the data for training and 80% for the validation further minimises the risk of overfitting. Finally, a central goal of the new sparsification algorithms is the definition of an objective criterion for the termination of the sparsification procedure. Thus, the system identified is not sensitive to user-selected thresholds and the complexity of the model is determined by the data.
The choice of decomposition technique is not considered essential to the process. However, the proprieties of the decomposition do influence the general performance of the ROM. Here, SPOD was chosen because it enables an efficient separation of the dynamics and leads to well-behaved temporal coefficients. This is beneficial for the construction of ROM and increases the probability, that the process results in an accurate and robust representation of dominating flow dynamics. Another optional step is the identification and coupling of modes pairs with a DMD and their representation as a complex mode. By minimising the number of possible library functions and forcing the process to handle modes as paired couples, the computational costs in the sparsification procedure are reduced and the description of the system simplified. The procedure can be applied using real valued modes and without prior coupling.

The number of selected modes is a balance between model complexity and the dynamics resolved by the model. Note, this applies to all approaches based on modal decomposition. Additional models using a larger number of modes have also been computed. The resulting models are more complex but benefit the representation by incorporating interactions between the higher energy modes and the added modes. However, it remains that the more complex models do not change the basic form and relationships between the modes presented in this work. Consequently, the simpler model is found to be more suitable for an initial demonstration of the method.

The results from the presented method show its capability to derive a robust ROM solely from measurement data. It can capture the deterministic part of the underlying dynamics of a turbulent wake. The only exogenous parameters that need to be defined within the process can either be derived from dynamical characteristics (e.g. vortex shedding frequency) of the flow or address procedural adjustments such as the number of repeated iterations on different parts of the data. Moreover, the process design exhibits a low sensitivity to control parameter variations. Hence, the process converges towards a final model, i.e. robustly identifying the same basis functions and library coefficients repeatedly. Additionally, the analytic structure of the utilised library functions resembles the terms of a Galerkin projection of the Navier–Stokes equations. Thus, it enables to gain insight of the model mechanics and interactions by analogy. Since specific terms of the ROM can be related to fluid mechanical phenomena, the model also allows an interpretation of flow physics. Due to its fast prediction capability, the model may potentially be suitable for applications where rapid approximations of the flow evolution are critical, such as for active flow control, process monitoring and assessing design performance.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

---

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

**A Symmetrisation**

In case the investigated flow field shows a symmetry, it is recommended to make use of symmetrical properties. This enables an improved separation of the dynamics with a modal decomposition. The wake of a circular cylinder undergoing VIV showed a symmetry with respect to the $y = 0$-plane. Thus, before performing a modal decomposition, the fluctuation is decomposed into a symmetric $v^{(s)}(x, t)$ and an asymmetric $v^{(a)}(x, t)$ part.

\begin{align}
  v(x, t) &= v(x) + v'(x, t) \\
  v'(x, t) &= v^{(s)}(x, t) + v^{(a)}(x, t)
\end{align}

(46)

(47)

The symmetric and asymmetric split is computed as follows:

\begin{align}
  u^{(s)}(x, y, z, t) &= \frac{1}{2} \left[ u'(x, y, z, t) + u'(x, -y, z, t) \right]
\end{align}

(48a)
Data-driven identification of low-order models for turbulent flows

\[ u''(x, y, z, t) = \frac{1}{2} \left[ u'(x, y, z, t) - u'(x, -y, z, t) \right] \]  \hspace{1cm} (48b)

\[ v''(x, y, z, t) = \frac{1}{2} \left[ v'(x, y, z, t) - v'(x, -y, z, t) \right] \]  \hspace{1cm} (48c)

\[ w''(x, y, z, t) = \frac{1}{2} \left[ w'(x, y, z, t) + w'(x, -y, z, t) \right] \]  \hspace{1cm} (48d)

With both, POD and SPOD, the symmetric and asymmetric fluctuations can be described by a sum of spatial modes \( \Phi_i(x) \) and corresponding temporal coefficients \( b_i(t) \): With Eqs. (46), (47) and (49), the symmetric and asymmetric part can be written in a combined sum:

\[ \mathbf{v}'(x, t) = \sum_{i=1}^{N} \Phi_{i}^{(s)}(x) b_{i}^{(s)}(t) + \sum_{i=1}^{N} \Phi_{i}^{(a)}(x) b_{i}^{(a)}(t). \]  \hspace{1cm} (49)

Then, Eq. (49) can be written as a combined sum:

\[ \mathbf{v}'(x, t) = \sum_{i=1}^{2N} \Phi_i(x) b_i(t), \]  \hspace{1cm} (50)

where the modes are ordered by energy. From here on, the procedure continues as described in Sect. 2.

B SPOD filter

SPOD results of the symmetrised VIV data for two filter lengths \( (N_{\ell} = 0, 32) \) are illustrated in Fig. 13. A limiting case, computed with the minimal filter length \( (N_{\ell} = 0) \), is equivalent to results of a classical POD. The case using \( N_{\ell} = 32 \) corresponds to the results discussed in Sect. 3. This filter length approximately corresponds to a characteristic time scale of the flow (two periods of the cylinder oscillation, \( N_{\ell} \cdot \Delta t = 1.33 \) s, \( T_p = 1/f_c \approx 0.67 \) s). For each case, so-called SPOD spectra [35] are depicted in Fig. 13 (left side), where a single dot represents a coupled mode pair. The size of the dot scales with the magnitude of its harmonic correlation \( H_{ij} \) determined according to Eq. (18). The colour also indicates the magnitude and is used to enhance visibility. The frequency \( f \) of the identified mode pair is computed by the weighted sum of the DMD eigenvalues \( \nu_k \) (Eq. 14) sorted in descending order [35]:

\[ f = \frac{\sum_{i=1}^{n} \ln(\nu_j) \tilde{h}_i}{2\pi \Delta t \sum_{i=1}^{n} \tilde{h}_i} \]  \hspace{1cm} (51)

\[ \tilde{h}_i = h_{ij}^2 + h_{ik}^2 \]  \hspace{1cm} (52)

Here, an averaged over three eigenvalues \( (n = 3) \) is used, to mitigate possible corruption by noise. The relative energy content \( K \), with respect to the total energy, of the mode pair is given by:

\[ K = \frac{\mu_j + \mu_j}{\sum_{k=1}^{N_e} \mu_k}. \]  \hspace{1cm} (53)

The subscripts \( i \) and \( j \) refer to the indices of the coupled SPOD modes. Three modes are selected (indicated by numbers), which are related to the cylinder oscillation (1), the vortex shedding (2) and their interaction (3). In Sect. 3.3, these modes are discussed in detail for \( N_{\ell} = 32 \). To the right of each SPOD spectrum (Fig. 13), the spatial mode shapes \( \Phi_{i,j}(x) \) (upper row) and the power spectral density (PSD) function of the real part of the corresponding temporal coefficients (lower row) are displayed for the three selected modes. The spatial mode shapes are visualised by the \( v \)-component of the velocity.

For a compact representation, the mode spectra and spatial mode shapes are given without axis labelling. The axis corresponds to analogous plots in Fig. 6.
Fig. 13 Resulting SPOD spectrum for different filter lengths: a $N_f = 0$ (POD) and b $N_f = 32$ (SPOD) displayed as a scatter plot, where a single dot indicates one couple SPOD mode pair and the size and colour refer to the harmonic correlation Eq. (18) of the modes. For both filter lengths, three mode pairs are selected (indicated by numbers in the SPOD spectrum) and their spatial mode of the $v$-component (upper row) and the PSD of the corresponding temporal $\Re(a_i)$ (lower row).

The POD ($N_f = 0$) spectrum in Fig. 13a shows few modes with a high harmonic correlation at lower Strouhal numbers $St = fD/U_\infty$ and a high energy content $K$. In the following, the effect of the filter size is primarily demonstrated for the selected modes. The spatial shapes of these modes are clearly symmetric (1) and (2) or anti-symmetric (3). The mode spectra of the corresponding temporal coefficients show broad bandwidth with several higher peak frequencies at different Strouhal numbers. Some of these frequencies are present in multiple spectra. For example, all three of the mode spectra show peaks at $St = 0.128$ and $St = 0.209$. This indicates that the dynamics are not fully separated.

In contrast, the SPOD ($N_f = 32$) spectrum yields similar modes with respect to the spatial mode shape, the peak frequency and energy content, but with a higher harmonic correlation, Fig. 13b. Only minor differences in the spatial shapes for the selected modes are observable, e.g., slight shifts in the distribution of the extrema. Note, that due to the previously applied symmetrisation, the effect of the filter operation on the spatial mode shapes is reduced. The mode spectra, however, are significantly changed. Instead of showing multiple high spectral peaks, Fig. 13a, each of the SPOD mode spectra is centred around a single frequency (Fig. 13b) with narrow bandwidth, so that the peaks, such as $St = 0.128$ and $St = 0.209$, are now associated with a specific SPOD mode. Briefly, when comparing SPOD to the POD modes, the spatial mode shapes and the energy content remain almost unchanged. However, the mode spectra indicate that with the SPOD a better separation of the dynamics can be achieved resulting in temporal coefficients with an increased harmonic correlation.

The filter size $N_f$ is an important parameter of the SPOD. Gradually increasing $N_f$ allows a continuous transition from POD to DFT. Accordingly, the spectral separation appears to be continuous, too. Selecting the filter size...
is a compromise between two competing effects. Increasing $N_f$ increases the spectral separation at the cost of reducing the energy content of each mode. Thus, the selection of the filter size is not critical in the sense of finding a single optimal parameter, but identifying the range, where the dynamics are separated and the modes represent the most energy. Selecting the filter size based on a characteristic time scales of the flow showed suitable results for different flow configurations [35].

**C Numerical derivative**

Assuming non-periodic boundary conditions, the coefficients for the sixth-order compact scheme at the boundary are defined as follows:

\[
\begin{bmatrix}
1 & 5 & 0 & 0 & \cdots \\
\frac{2}{11} & 1 & \frac{2}{11} & 0 & 0 & \cdots \\
0 & \frac{1}{5} & 1 & \frac{1}{5} & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots \\
0 & \frac{2}{5} & 1 & \frac{2}{11} & & \\
\end{bmatrix}
\begin{bmatrix}
\dot{a}_1 \\
\dot{a}_2 \\
\dot{a}_3 \\
\vdots \\
\dot{a}_{N-2} \\
\dot{a}_{N-1} \\
\dot{a}_N \\
\end{bmatrix} = \frac{1}{\Delta t}
\begin{bmatrix}
-197 & 5 & -5 & 5 & -1 & 0 & \cdots \\
-33 & -33 & 34 & -33 & 33 & -132 & 0 & \cdots \\
-36 & -18 & 0 & 14 & 18 & 36 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
-1 & 36 & 0 & 14 & 18 & 35 & 33 & 33 & -7 & 33 & -132 & 0 & \cdots \\
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_3 \\
\vdots \\
a_{N-2} \\
a_{N-1} \\
a_N \\
\end{bmatrix}
\]  

(54)

**D Experimental data**

Velocity and force data were obtained from a study conducted at the University of Calgary [27]. A cyber-physical apparatus [8,11,44], described in Riches et al. [26], was used to operate a real-time simulation of the mechanical system described in Eq. (55)

\[
m\ddot{y} + c\dot{y} + ky = F_y, \tag{55}
\]

where $m$, $c$, and $k$ are the structural mass, damping, and stiffness, respectively. $F_y$ is a time-dependent forcing from the wake dynamics in the transverse direction of the incoming free stream flow [13]. The measured forces are used as an input to calculate the kinematics, so that the structural mass, stiffness and damping can be modelled in software. With a motorised traverse, the cylinder is then forced to follow these kinematics. The experiment was conducted to analyse the wake of a hollow acrylic cylinder (diameter $D = 19.05$ mm and length $L \approx 20D$) with a non-dimensionalised mass ratio $m^*$ and a damping ratio $\xi$:

\[
m^* = \frac{4m}{\pi \rho D^2 L} = 8 \tag{56}
\]

\[
\xi = \frac{c}{2\sqrt{km}} = 0.005 \tag{57}
\]

at $Re = 4000$ and $U^* = 8.75$. Data of the time-resolved two-component planar PIV were collected in a plane of the cylinder midspan. The images were captured with a high-speed Phantom Miro M340 digital camera ($2560 \times 1600$). The flow was seeded with 11.7 $\mu$m hollow-glass spheres, and the plane was illuminated by a laser sheet from a Photonics 20 mJ Nd:YLF high repetition rate pulsed laser with a wavelength of $\lambda = 527$ nm. Both, camera and laser were configured to operate in dual-pulse mode at an effective data image per acquisition rate of 24Hz. For the processing of the PIV images, the LaVision DaVis 8.3 software suite was used. The time-resolved velocity vector fields were calculated using a multi-pass cross-correlation algorithm with a final interrogation window size of $24 \times 24$ pixels resulting in a vector spacing of 0.725 mm. More information and details of experimental setup and technical components are available in Riches et al. [26] and [27]. Within three realisations of the experiment, a total number of 4767 vector fields (1589 per realisation) were obtained at a sample frequency of 24Hz comprising approximately 298 periods of the cylinder oscillation.
34. Sieber, M.: Data-driven identification and modelling of coherent dynamics in turbulent flows. Ph.D. thesis, Technische Universität Berlin (2021). https://doi.org/10.14279/DEPOSITONCE-11749
35. Sieber, M., Paschereit, C.O., Oberleithner, K.: Spectral proper orthogonal decomposition. J. Fluid Mech. 792, 798–828 (2016). https://doi.org/10.1017/jfm.2016.103
36. Sieber, M., Paschereit, C.O., Oberleithner, K.: Stochastic modelling of a noise-driven global instability in a turbulent swirling jet. J. Fluid Mech. (2021). https://doi.org/10.1017/jfm.2021.133
37. Sirovich, L.: Turbulence and the dynamics of coherent structures. Part I: coherent structures. Part II: symmetries and transformations. Part III: dynamics and scaling. Q. Appl. Math. 45(3), 561–571 (1987)
38. Stöhr, M., Oberleithner, K., Sieber, M., Yin, Z., Meier, W.: Experimental study of transient mechanisms of bi-stable flame shape transitions in a swirl combustor. J. Eng. Gas Turbines Power (2017). https://doi.org/10.1115/1.4037724
39. Taira, K., Brunton, S.L., Dawson, S.T.M., Rowley, C.W., Colonius, T., McKeon, B.J., Schmidt, O.T., Gordeyev, S., Theofilis, V., Ukeiley, L.S.: Modal analysis of fluid flows: an overview. AIAA J. 1452(February), 1–29 (2017). https://doi.org/10.2514/1.J056060
40. Taira, K., Hemati, M.S., Brunton, S.L., Sun, Y., Duraisamy, K., Bagheri, S., Dawson, S.T., Yeh, C.A.: Modal analysis of fluid flows: applications and outlook. AIAA J. 58(3), 998–1022 (2020). https://doi.org/10.2514/1.J058462
41. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) 58(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
42. Towne, A., Schmidt, O.T., Colonius, T.: Spectral proper orthogonal decomposition and its relationship to dynamic mode decomposition and resolvent analysis. J. Fluid Mech. 847, 821–867 (2018). https://doi.org/10.1017/jfm.2018.283
43. Vanierschot, M., Müller, J.S., Sieber, M., Percin, M., van Oudheusden, B.W., Oberleithner, K.: Single- and double-helix vortex breakdown as two dominant global modes in turbulent swirling jet flow. J. Fluid Mech. (2020). https://doi.org/10.1017/jfm.2019.872
44. Williamson, C.H., Govardhan, R.: Vortex-induced vibrations. Annu. Rev. Fluid Mech. 36(1982), 413–455 (2004). https://doi.org/10.1146/annurev.fluid.36.050802.122128

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.