Tianyi Chu

Oliver T. Schmidt

A stochastic SPOD-Galerkin model for broadband turbulent flows

Abstract The use of spectral proper orthogonal decomposition (SPOD) to construct low-order models for broadband turbulent flows is explored. The choice of SPOD modes as basis vectors is motivated by their opti-mality and space-time coherence properties for statistically stationary flows. This work follows the modeling paradigm that complex nonlinear fluid dynamics can be approximated as stochastically forced linear systems. The proposed stochastic two-level SPOD-Galerkin model governs a compound state consisting of the modal expansion coefficients and forcing coefficients. In the first level, the modal expansion coefficients are advanced by the forced linearized Navier-Stokes operator under the linear time-invariant assumption. The second level governs the forcing coefficients, which compensate for the offset between the linear approximation and the true state. At this level, least squares regression is used to achieve closure by modeling nonlinear interactions between modes. The statistics of the remaining residue are used to construct a dewhitening filter that facilitates the use of white noise to drive the model. If the data residue is used as the sole input, the model accurately recovers the original flow trajectory for all times. If the residue is modeled as stochastic input, then the model generates surrogate data that accurately reproduces the second-order statistics and dynamics of the original data. The stochastic model uncertainty, predictability, and stability are quantified analytically and through Monte Carlo simulations. The model is demonstrated on large eddy simulation data of a turbulent jet at Mach number $M = 0.9$ and Reynolds number $\mathrm{Re}_D \approx 10^6$.

1 Introduction

High-fidelity numerical simulations of turbulent flows have become common practice in engineering and science. Their computational cost, however, is often prohibitive for optimization, and their time-to-solution hinders real-time control. Reduced-order models (ROMs) are routinely employed to reduce computational complexity in this context.

Among the most successful models in fluid dynamics are POD-Galerkin models that seek a low-order representation of the flow dynamics in the space spanned by a modal basis obtained from proper orthogonal

decomposition (POD). Most commonly, the so-called method of snapshots introduced by Sirovich [37] is used to obtain a basis of space-only POD modes and temporal expansion coefficients. This most popular flavor of POD is computed from the eigendecomposition of the spatial cross-correlation tensor. The resulting POD modes optimally represent the data in terms of a spatial inner product. POD-based Galerkin models leverage the orthogonality of POD in the spatial inner product to obtain low-order models that govern the temporal evolution of the POD expansion coefficients. Holmes et al. [11] and Rowley & Dawson [32] summarized and illustrate this approach. POD-Galerkin models have been successfully applied to model the wall region of incompressible turbulent boundary layers by Aubry et al. [1] and later, with a focus on low wave number phenomena of turbulence generation, by Holmes et al. [12]. Noack et al. [23] proposed a hierarchy of POD-Galerkin models for viscous cylinder wakes and introduced a shift mode that accounts for the mean field correction. For compressible flows governed by the compressible governing equations, POD-Galerkin modeling is significantly complicated by the additional energy equation and the occurrence of triple products of the state variables, which may be defined in either primitive or conservative form. It was observed by Rempfer [28] that POD-Galerkin models of general compressible flows exhibit nonphysical instability. There are several previous attempts to overcome these difficulties for moderate Reynolds number flows. In pioneering works by Rowley et al. [29,30], the authors implemented POD-Galerkin models for compressible cavity flows for control purposes. Later, Rowley et al. [31] applied simplified Navier-Stokes equations to obtain a quadratic POD-Galerkin ROM that is valid for isentropic, cold flows at moderate Mach numbers. For linearized compressible flows, Barone et al. [2] devised an inner symmetry product that yields stable models. Expanding on this work, Kalashnikova & Barone [15,16] proposed stability-preserving model reduction techniques for non-linear systems. All of the work discussed so far is based on standard space-only POD computed from the method of snapshots.

As an alternative to POD-Galerkin techniques, which require knowledge of the governing equations, purely data-driven models are commonly used in the atmospheric sciences. Inverse stochastic models, for example, approximate the temporal evolution of the POD expansion coefficients directly from the input data. The simplest type of inverse stochastic models is the so-called linear inverse model (LIM) introduced by Penland [24,25]. The underlying assumption of this model is that a nonlinear dynamical system can be modeled as a deterministic linear system that is driven by stochastic forcing. In practice, the linear operator which governs the evolution is obtained by linear regression, and the regression residue is estimated as white noise. Within the LIM framework, additive noise is used to model both the nonlinear flow physics and process noise. LIMs have successfully been applied in climate and weather prediction, for example of the El Niño-Southern Oscillation [14,27], and other sea surface temperature anomalies [26]. Later, as a generalization of LIMs, Kondrashov et al. [18], Kravtsov et al. [19] introduced the multiple polynomial regression (MPR), which expands on the purely linear system assumption of LIMs by assuming a polynomial form for the expansion coefficients to better approximate nonlinear dynamics. Expanding on this idea, the authors further proposed the use of a multi-level regression (MLR) model, which inflates the MPR model by a hierarchy of additional levels until the residue can be modeled as white noise. The additional levels are used to account for the nonlinearity of the system and serial correlations in the residue. More recently, Kondrashov et al. [17] generalized the MLR models by introducing additional implicit variables. All of the above inverse stochastic models are applied to approximate the nonlinear dynamics of the expansion coefficients of space-only POD modes. More recently, Chekroun & Kondrashov [5] proposed a model for the time-domain expansion coefficients of data-adaptive harmonic modes, which to the best understanding of the present authors correspond to spectral POD (SPOD) modes, by specifying the main-level polynomial form in MLRs as a Stuart-Landau oscillator.

SPOD, the frequency-domain variant of POD, was first conceptually introduced by Lumley [20,21], and is computed as the eigendecomposition of the cross-spectral density tensor under the assumption of ergodicity, that is, for statistically stationarity flows. SPOD decomposes the data set into energy-ranked, monochromatic modes [36,40]. The resulting modes evolve coherently in both space and time, and optimally represent the second-order space-time flow statistics. SPOD has been applied to a variety of turbulent flows, including wall-bounded shear flows such as boundary layers [41], as well as bluff-body flows such as the wake behind a disk Tutkun et al. [42]. For turbulent jets, the use of SPOD was pioneered by Glauser et al. [8]. Later, different authors linked coherent structures identified by SPOD to concepts from linear stability theory for both experimental data [4,9,10] and large eddy simulation (LES) data [34,35,39,40].

In this work, we propose an SPOD-Galerkin two-level model for stochastic modeling of fully turbulent flows under the linear-time invariant (LTI) approximation. The use of SPOD modes as the basis for the model is motivated by: (i) their optimality for statistically stationary data, and (ii) by the prospect that SPOD modes can be modeled as eigenvectors of the resolvent operator [40]. The latter property has two important implications. First, the choice of modal basis is consistent with the LTI approximation, that is, that of forced linear dynamics,

and second, the potential to enable future models that use resolvent modes as basis (note that classical POD modes are data-driven and cannot readily be predicted by theory).

The goal of this work is to devise a ROM that can accurately reproduce a known flow trajectory, or, when forced stochastically, generates surrogate data which reproduces the second-order statistics and dynamics of chaotic, high-Reynolds number flows. Applications of the latter capability include Monte Carlo-based prediction under uncertainty. The key idea of this framework is to approximate the nonlinear fluid dynamics obtained from SPOD-Galerkin projection by a stochastically forced LTI system. The resulting forced ROM is considered as the first level of the model, which advances the state coefficients in time by the linearized Navier-Stokes operator. Due to the convective nonlinearity of the Navier-Stokes equations, white-in-time noise is generally not adequate as the sole stochastic input to linearized flow models. Zare et al. [44], for example, presented a modeling framework for turbulent channel flow which obtains colored-in-time stochastic excitation to account for second-order flow statistics by solving a convex optimization problem. Here, we take a different approach by incorporating the forcing coefficients as part of the state to offset the difference between the linear approximation and the true, nonlinear system dynamics. Finally, linear dependence between the evolution of the state and forcing coefficients is established by means of least squares regression. Closure of the model is achieved and can be tested *a priori*, if the second-level residue takes on the characteristics of white noise.

Following brief introductions of Galerkin projection and SPOD in Sect. 2, the stochastic two-level SPOD-Galerkin ROM is derived in Sect. 3, including residue modeling and uncertainty quantification. The performance of the SPOD-Galerkin two-level model is demonstrated in Sect. 4 using the LES data of the axisymmetric component of a turbulent jet at Mach number $M = 0.9$ and jet-diameter Reynolds number $\text{Re}_D \approx 10^6$. By verifying that the second-level residues are well approximated by white noise, model closure is confirmed. We show that the surrogate data obtained from our model accurately reproduces the second-order statistics and dynamics of the input data. The model uncertainty and stability are quantified by analytical means and Monte Carlo simulation. Finally, Sect. 5 summarizes and concludes the paper.

2 Background

2.1 Galerkin projection-based model-order reduction for linearized Navier-Stokes equations

We start by taking the Reynolds decomposition of the flow state $q \in \mathbb{C}^n$ into the temporal mean and fluctuating components $\overline{(\cdot)}$ and $(\cdot)'$, respectively, as

$$q(t) = \bar{q} + q'(t). \tag{1}$$

As an integral measure of energy, we define the inner product and associated norm

$$\langle q_1, q_2 \rangle_E = q_1^* W q_2, \quad \text{and} \tag{2}$$

$$\|q\|_E^2 = \langle q, q \rangle_E, \tag{3}$$

respectively, where W is a diagonal positive-definite weight matrix, and $(\cdot)^*$ denotes the Hermitian transpose.

As motivated above, we choose as the basis of the Galerkin ROM the first M SPOD modes (at all N_f frequencies),

$$V = \left[\underbrace{\psi_{\omega_1}^{(1)} \cdots \psi_{\omega_1}^{(M)}}_{\omega_1} \; \underbrace{\psi_{\omega_2}^{(1)} \cdots \psi_{\omega_2}^{(M)}}_{\omega_2} \; \cdots \; \underbrace{\psi_{\omega_{N_f}}^{(1)} \cdots \psi_{\omega_{N_f}}^{(M)}}_{\omega_{N_f}} \right]. \tag{4}$$

Here, $\psi_{\omega_k}^{(i)}$ are the ith SPOD modes of frequency ω_k and energy $\lambda_{\omega_k}^{(i)}$. Refer to Schmidt et al. [36] and Towne et al. [40] for more details. This rank-$M \times N_f$ (assuming all ψ are linearly independent) SPOD basis contains

$$E_M = \frac{\sum_{i=1}^{M} \sum_{k=1}^{N_f} \lambda_{\omega_k}^{(i)}}{\sum_{i=1}^{N_b} \sum_{k=1}^{N_f} \lambda_{\omega_k}^{(i)}} \cdot 100\% \tag{5}$$

of the total energy density.

Within the subspace $\mathcal{S}$ spanned by the columns of $\boldsymbol{V}$, the vector $\widetilde{\boldsymbol{q}}(t)$ that best (in a least-squared sense) approximates $\boldsymbol{q}'(t)$ in terms of $\| \cdot \|_E$ is found from the oblique projection

$$\widetilde{\boldsymbol{q}}(t) = \boldsymbol{P}\boldsymbol{q}'(t), \quad \text{where} \tag{6}$$

$$\boldsymbol{P} = \boldsymbol{V} \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \tag{7}$$

is the oblique projection matrix for non-singular $\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}$. Alternatively, we may directly express $\widetilde{\boldsymbol{q}}(t)$ as a linear combination

$$\widetilde{\boldsymbol{q}}(t) = \sum_{i=1}^{M} \sum_{k=1}^{N_f} a_k^{(i)}(t) \boldsymbol{\psi}_{\omega_k}^{(i)} = \boldsymbol{V} \boldsymbol{a}(t) \tag{8}$$

of the modes $\boldsymbol{\psi}_{\omega_k}^{(i)}$ by defining

$$\boldsymbol{a}(t) \equiv \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \boldsymbol{q}'(t) \tag{9}$$

as the vector of temporal expansion coefficients $a_k^{(i)}(t)$. The reader is referred to Nekkanti & Schmidt [22] for more details on the flow field reconstruction. The error vector between the true state and its approximation at any given time is defined as

$$\boldsymbol{e}(t) \equiv \boldsymbol{q}'(t) - \widetilde{\boldsymbol{q}}(t) = (\boldsymbol{I} - \boldsymbol{P}) \boldsymbol{q}'(t). \tag{10}$$

In the following, we will omit the explicit time dependence of $\boldsymbol{q}'$ and $\boldsymbol{a}$ for brevity. Next, we consider the linearized compressible Navier-Stokes equations under the linear time-invariant (LTI) assumption,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}' = \boldsymbol{L}_{\bar{q}}\boldsymbol{q}' + \boldsymbol{f}. \tag{11}$$

Here, the matrix $\boldsymbol{L}_{\bar{q}}$ is the discretized linearized Navier-Stokes operator.

Different from standard Galerkin ROMs, we model linear dynamics and use the forcing as the compensation for the linearization. First, we obtain the forcing

$$\boldsymbol{f} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}' - \boldsymbol{L}_{\bar{q}}\boldsymbol{q}' \tag{12}$$

as the offset between the linear approximation and the true state. This procedure guarantees that the forcing is consistent with the discretization. The resulting Galerkin ROM of the LTI system

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} = \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \left(\boldsymbol{L}_{\bar{q}}\boldsymbol{q}' + \boldsymbol{f}\right) = \boldsymbol{L}_{\mathrm{Gal}}\boldsymbol{a} + \boldsymbol{b} \tag{13}$$

governs the evolution of the forced state. Both the state and the forcing are expressed in terms of their expansion coefficients, $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively. In Eq. (13),

$$\boldsymbol{L}_{\mathrm{Gal}} = \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \boldsymbol{L}_{\bar{q}} \boldsymbol{V} \tag{14}$$

is the system dynamics matrix, and

$$\boldsymbol{b} \equiv \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \boldsymbol{f} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} - \boldsymbol{L}_{\mathrm{Gal}}\boldsymbol{a} \tag{15}$$

the vector of expansion coefficients of the forcing field $\boldsymbol{f}$. Equation (13) is an m-dimensional first-order differential equation. If a particular trajectory, for example starting from $\boldsymbol{q}'(0)$ is of interest, the corresponding initial condition can be found as $\boldsymbol{a}(0) = \left(\boldsymbol{V}^* \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^* \boldsymbol{W} \boldsymbol{q}'(0)$. If $\boldsymbol{a}$ is statistically stationary, then

$$\bar{\boldsymbol{b}} = \frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{a}} - \boldsymbol{L}_{\mathrm{Gal}}\bar{\boldsymbol{a}} = 0 \tag{16}$$

implies that $\boldsymbol{b}$ has zero mean. This property is important in the context of inverse stochastic models, described next.

2.2 Inverse stochastic models

Inverse stochastic models are data-driven models that do not require knowledge of the linear operator $\boldsymbol{L}_{\text{Gal}}$. Instead, they approximate its action on the model coefficients from the data, most commonly using a least squares approximation and model the stochasticity of the original process as additive noise.

2.2.1 Linear inverse model (LIM)

The simplest inverse stochastic model is the linear inverse model (LIM)

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} = \widetilde{\boldsymbol{T}}\boldsymbol{a} + \boldsymbol{w}, \tag{17}$$

proposed by Penland [24,25]. It assumes a deterministic linear operator that is forced by white noise, $\boldsymbol{w}$. The operator itself is approximated from the data in a least squares sense as

$$\widetilde{\boldsymbol{T}} = \arg\min_{\widetilde{\boldsymbol{T}}} \sum_{j=1}^{N} \left(\left\| \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a}^{(j)} - \widetilde{\boldsymbol{T}}\boldsymbol{a}^{(j)} \right\|^2 \right). \tag{18}$$

Note that knowledge of $\boldsymbol{L}_{\text{Gal}}$ is not required for this procedure. The underlying assumption of the LIM is that the residue of the linear regression,

$$\boldsymbol{r} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} - \widetilde{\boldsymbol{T}}\boldsymbol{a}, \tag{19}$$

can be approximated as white noise. If the governing linear operator $\boldsymbol{L}_{\text{Gal}}$ is known, we may express $\widetilde{\boldsymbol{T}}$ as the sum of two matrices,

$$\widetilde{\boldsymbol{T}} = \boldsymbol{L}_{\text{Gal}} + \boldsymbol{T}. \tag{20}$$

Combining Eqs. (13), (17) and (20) yields the relation

$$\boldsymbol{b} = \boldsymbol{T}\boldsymbol{a} + \boldsymbol{w} \tag{21}$$

between the forcing coefficients, $\boldsymbol{b}$, and the model coefficients, $\boldsymbol{a}$ in terms of $\boldsymbol{T}$. If the feedback matrix $\boldsymbol{T}$ is known, a random realization of the forcing coefficients can be generated to drive the stochastic model. Observe that realizations of $\boldsymbol{b}$ are generated from white noise forcing. This implies that $\boldsymbol{T}$ accounts for the correlations between the state and forcing.

2.2.2 Linear multi-level regression (MLR) models

The LIM discussed above assumes that the residue in the model's definition, Eq. (19), is white in time. In practice, however, $\boldsymbol{r}$ is computed from the data using Eqs. (18) and (19). Hence, there is no guarantee that the white-noise assumption holds for any given nonlinear process. To address this problem, Kondrashov et al. [18]; Kravtsov et al. [19] introduced the so-called linear multi-level regression (MLR) model. The idea behind the MLR model is to inflate the original model by a hierarchy of additional levels. Each level describes the dynamics of the residue of the previous level and is found by linear regression. Closure of the model is archived once the white-noise assumption holds. Denoting by $\boldsymbol{r}_1$ the first-order residue, i.e., the residue of the original model obtained using Eq. (19), the first-level model takes the form $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} = \widetilde{\boldsymbol{T}}\boldsymbol{a} + \boldsymbol{r}_1$. The second-level system is then obtained by inflating the state vector by the first-order residue, $\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{r}_1 \end{bmatrix}$, and so on. Assuming that the first $(L-1)$ residues are differentiable, the linear MRL model takes the form

$$\text{Level 1:} \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} = \widetilde{\boldsymbol{T}}\boldsymbol{a} + \boldsymbol{r}_1,$$

$$\text{Level 2:} \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_1 = \boldsymbol{M}_1\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{r}_1\end{bmatrix} + \boldsymbol{r}_2,$$

$$\vdots \tag{22}$$

$$\text{Level } L: \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_L = \boldsymbol{M}_L\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{r}_1\\\vdots\\\boldsymbol{r}_L\end{bmatrix} + \boldsymbol{w}.$$

Here, $\boldsymbol{M}_l$ is the system matrix of size $MN_f \times (l+1)MN_f$, and $\boldsymbol{r}_l$ the l-th level residue. Alternatively, Eq. (22) can be written in the matrix form as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{r}_1\\\vdots\\\boldsymbol{r}_{L-1}\\\boldsymbol{r}_L\end{bmatrix} = \begin{bmatrix}\widetilde{\boldsymbol{T}} & \boldsymbol{I} & \boldsymbol{0} & \cdots & & \boldsymbol{0}\\ & -\boldsymbol{M}_1 - & \boldsymbol{I} & \ddots & & \vdots\\ & & \ddots & \ddots & & \boldsymbol{0}\\ & -\boldsymbol{M}_{L-1} & - & & & \boldsymbol{I}\\ & -\boldsymbol{M}_L & - & & &\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{r}_1\\\vdots\\\boldsymbol{r}_{L-1}\\\boldsymbol{r}_L\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}\\\vdots\\\vdots\\\boldsymbol{0}\\\boldsymbol{w}\end{bmatrix}. \tag{23}$$

Equation (22) reduces to the classical LIM if $\boldsymbol{r}_1$ is white-in-time to start with.

3 The stochastic two-level SPOD-Galerkin model

The underlying idea of the proposed approach is to model a statistically stationary flow as the superposition of large-scale coherent structures that evolve on the mean flow under the influence of stochasticity. This idea is reflected, for example, in the well-known triple-decomposition introduced by Hussain & Reynolds [13]. SPOD modes optimally represent the second-order space-time statistics of the stationary flow field. We choose SPOD modes as the basis of our model for this reason. The triple-decomposition further requires the 'background turbulence' to be stochastic in nature. To accomplish this goal, we employ a multi-level linear regression model, as introduced in Sect. 2.2.2.

In the following, we will demonstrate that the standard Galerkin-projection approach shown in Sect. 2.1 is particularly well-suited for this purpose as it requires only one additional level for closure. Section 3.1 introduces the two-level SPOD-Galerkin model and discusses its closure via residue modeling. The final, stochastically driven two-level SPOD-Galerkin model is summarized in Sect. 3.2. Finally, an uncertainty quantification study of the new model is conducted in Sect. 3.4.

3.1 Two-level SPOD-Galerkin model

To obtain the most compact representation of the dynamics, it is desirable to truncate the multi-level model, Eq. (22), at the lowest possible level. Our natural starting point is therefore the two-level model

$$\text{Level 1:} \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a} = \boldsymbol{L}_{\mathrm{Gal}}\boldsymbol{a} + \boldsymbol{b}, \tag{24}$$

$$\text{Level 2:} \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{b} = \boldsymbol{M}\boldsymbol{y} + \boldsymbol{r}, \quad \text{where} \quad \boldsymbol{y} \equiv \begin{bmatrix}\boldsymbol{a}\\\boldsymbol{b}\end{bmatrix}. \tag{25}$$

Motivated by the goal of creating a physics-based model that relies on regression for closure only, we deviate from the standard linear inverse modeling approach and retain the physics-based operator $\boldsymbol{L}_{\mathrm{Gal}}$ at first level. Consistent with Eq. (13), the residue at first level is identified as the forcing vector $\boldsymbol{b}$, obtained from Eq. (15).

Following the standard linear inverse modeling paradigm, we seek closure at the second level by solving the least squares problem

$$M = \arg\min_{M} \sum_{j=1}^{N} \left(\left\| \frac{d}{dt} b^{(j)} - M y^{(j)} \right\|^2 \right) \tag{26}$$

to obtain M. This least squares problem is formally equivalent to a linear system problem

$$\frac{d}{dt} B = MY, \tag{27}$$

where $Y = \left[y^{(1)}, y^{(2)}, \ldots, y^{(N)} \right]$ and $B = \left[b^{(1)}, b^{(2)}, \ldots, b^{(N)} \right]$ are the matrices of extended vectors and forcing vectors, respectively. The solution of Eq. (27) is

$$M = \left(\frac{d}{dt} B \right) Y^+, \tag{28}$$

where Y^+ denotes the pseudo-inverse of the (possibly singular) extended state matrix. Once M is known, we can obtain the second-level residue r from Eq. (24) as

$$r = \frac{d}{dt} b - M y. \tag{29}$$

Form Eq. (16), we may deduce that the second-level residue r has zero mean,

$$\overline{r} = \overline{\frac{d}{dt} b} - M \overline{y} = 0.$$

The residue r will be approximated as white noise that can be correlated in the subspace, described next. Once M and r are determined, the final two-level model is assembled as

$$\frac{d}{dt} y = \underbrace{\begin{bmatrix} L_{\text{Gal}} & I \\ - & M & - \end{bmatrix}}_{L_{\text{2-lvl}}} y + \begin{bmatrix} 0 \\ r \end{bmatrix}. \tag{30}$$

in the form of Eq. (23). Equation (30) is a forced first-order linear time-invariant system for the extended state vector. The final system dynamic matrix $L_{\text{2-lvl}}$ contains both the operator governing the linear evolution of large-scale coherent structures—represented by the basis vectors—about the mean flow, L_{Gal}, and correlation information between the state and the forcing that is learned from the data in M. In particular, the left square matrix M_{ab} and the right square matrix M_{bb} that constitute

$$M = \begin{bmatrix} M_{ab} & M_{bb} \end{bmatrix} \tag{31}$$

contain correlations between the expansion coefficients and the forcing, and the correlation between the forcing components, respectively. If the residual is computed from the data by means of Eq. (29), the proposed model, Eq. (30), accurately reproduces the original flow field over all times. The remaining task is to find a suitable model for the residue r.

3.1.1 Residue modeling

In particular, we seek a stochastic forcing model $\widetilde{r}$ that has the same second-order statistics as the residue r. The autocorrelation function that describes the second-order statistics of the residue r at zero time-lag is readily obtained from the data as $R_{rr} = \overline{rr^*}$. Assuming that the residue r is normally distributed in time with

$$r \Delta t \sim \mathcal{N}(0, R_{rr} \cdot (\Delta t)^2), \tag{32}$$

we may randomly generate $\widetilde{r}$ from Gaussian white noise $w \sim \mathcal{N}(0, I)$ as

$$\widetilde{r} = Gw, \tag{33}$$

where $\boldsymbol{G}$ is the unknown input distribution matrix. The role of $\boldsymbol{G}$ in our model is to color randomly generated white noise with the statistics of the residual. *Vice versa*, $\boldsymbol{G}^{-1}$ can be interpreted as a whitening filter [43]. Gaussian white noise can be generated, by sampling from a Wiener process $\boldsymbol{\xi}$, as

$$\boldsymbol{w} \equiv \frac{\mathrm{d}\boldsymbol{\xi}}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\boldsymbol{\xi}(t + \Delta t) - \boldsymbol{\xi}(t)}{\Delta t}. \tag{34}$$

For any time increment Δt, $\Delta\boldsymbol{\xi}$ is normally distributed with zero mean and covariance matrix $\boldsymbol{I}\Delta t$. We may now express Eq. (33) in terms of $\Delta\boldsymbol{\xi}$ as

$$\widetilde{\boldsymbol{r}}\Delta t = \boldsymbol{G}\Delta\boldsymbol{\xi}, \tag{35}$$

where

$$\boldsymbol{G}\Delta\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{G}\boldsymbol{G}^* \Delta t) \tag{36}$$

follows from the properties of the Wiener process. Comparing Eqs. (36) and (32) allows us to relate the unknown input distribution matrix, $\boldsymbol{G}$, to the known covariance matrix of the residue, $\boldsymbol{R}_{rr}$, as

$$\boldsymbol{R}_{rr}\Delta t = \boldsymbol{G}\boldsymbol{G}^*. \tag{37}$$

Based on the form of Eq. (37), we obtain $\boldsymbol{G}$ from the Cholesky decomposition of $\boldsymbol{R}_{rr}\Delta t$. Owing to this procedure of computation, $\boldsymbol{G}^{-1}$ may be referred to as a Cholesky whitening filter.

3.2 Stochastic two-level SPOD-Galerkin model

Closure of the model is achieved by introducing the stochastic forcing model $\widetilde{\boldsymbol{r}}$ into Eq. (30). The resulting stochastic two-level SPOD-Galerkin model takes the form of the stochastic differential equation

$$\mathrm{d}\boldsymbol{y} = \boldsymbol{L}_{\text{2-lvl}}\boldsymbol{y}\,\mathrm{d}t + \int_t^{t+\mathrm{d}t} \widetilde{\boldsymbol{w}}\,\mathrm{d}t, \quad \text{where} \quad \widetilde{\boldsymbol{w}} \equiv \begin{bmatrix} \mathbf{0} \\ \boldsymbol{G}\boldsymbol{w} \end{bmatrix} \tag{38}$$

is the process noise. Equation (38) is a stochastic differential equation (SDE) that can be solved numerically by approximating the time derivative using a forward difference as

$$\boldsymbol{y}^{(j+1)} = \boldsymbol{H}\boldsymbol{y}^{(j)} + \Delta\widetilde{\boldsymbol{\xi}}^{(j)}, \tag{39}$$

where j is the time index, Δt the time step, and

$$\boldsymbol{H} = \boldsymbol{I} + \boldsymbol{L}_{\text{2-lvl}}\Delta t \quad \text{and} \quad \Delta\widetilde{\boldsymbol{\xi}} \equiv \begin{bmatrix} \mathbf{0} \\ \boldsymbol{G}\Delta\boldsymbol{\xi} \end{bmatrix} \tag{40}$$

are the state transition matrix and the process noise, respectively. By construction, the process noise $\Delta\widetilde{\boldsymbol{\xi}}$ is a zero-mean Gaussian random sequence with covariance matrix

$$\widetilde{\boldsymbol{R}}_{rr} \equiv \overline{\Delta\widetilde{\boldsymbol{\xi}}\Delta\widetilde{\boldsymbol{\xi}}^*} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_{rr}\Delta t \end{bmatrix}. \tag{41}$$

This completes the model. The discrete-in-time algorithm is outlined in the following.

3.3 Algorithm: stochastic two-level SPOD-Galerkin

Algorithm Stochastic two-level SPOD-Galerkin model
Input: Fluctuating data matrix $Q' = [q^{(1)} - \bar{q}, q^{(2)} - \bar{q}, \ldots, q^{(N)} - \bar{q}]$ consisting N snapshots, discretized linearized Navier-Stokes operator $L_{\bar{q}}$, time step Δt.
Output: Matrices $L_{\text{2-lvl}}$ and G of the stochastic two-level SPOD-Galerkin model $\frac{d}{dt} y = L_{\text{2-lvl}} y + \widetilde{w}$, where $\widetilde{w} = \begin{bmatrix} 0 \\ Gw \end{bmatrix}$.

1. Compute the SPOD of Q' and store the SPOD modes to be retained in the model in the column (basis) matrix V.
2. Determine the expansion coefficients and the Galerkin system dynamics matrix as

$$A = \left(V^* W V\right)^{-1} V^* W Q' \quad \text{and} \quad L_{\text{Gal}} = \left(V^* W V\right)^{-1} V^* W L_{\bar{q}} V,$$

   respectively.
3. Following Eq. (15), calculate the forcing coefficients as

$$B = \frac{A_2^N - A_1^{N-1}}{\Delta t} - L_{\text{Gal}} A_1^{N-1}.$$

   where $A_2^N = [a^{(2)}, a^{(3)}, \ldots, a^{(N)}]$ and $A_1^{N-1} = [a^{(1)}, a^{(2)}, \ldots, a^{(N-1)}]$.
4. Let $Y \leftarrow \begin{bmatrix} A_1^{N-1} \\ B \end{bmatrix}$. Solving the linear system problem in Eq. (28) yields

$$M = \frac{B_2^{N-1} - B_1^{N-2}}{\Delta t} Y^* \left(Y Y^*\right)^{-1}.$$

5. Assemble the matrices to obtain

$$L_{\text{2-lvl}} \leftarrow \begin{bmatrix} L_{\text{Gal}} & I \\ -M & - \end{bmatrix}.$$

   .
6. Calculate the linear regression residue as

$$R = \frac{B_2^{N-1} - B_1^{N-2}}{\Delta t} - M Y_1^{N-2}.$$

7. Determine the matrix G by solving the Cholesky decomposition

$$G G^* = \frac{\Delta t}{N - 2} R R^*.$$

   The resulting stochastic two-level SPOD-Galerkin model is propagated in time as

$$y^{(j+1)} = [I + L_{\text{2-lvl}} \Delta t] y^{(j)} + \begin{bmatrix} 0 \\ G\sqrt{\Delta t} w \end{bmatrix},$$

where j is the time index, see Eqs. (38)–(40). The zero mean, unit variance Gaussian white noise w which drives the system is obtained from a random number generator.

3.4 Uncertainty quantification and spectral analysis

Next, we leverage the resemblance of Eq. (39) to first-order vector autoregression (VAR) processes to conduct an uncertainty quantification analysis of the model. Following Stengel [38], we analyze the propagation of the uncertainty in terms of the expected value of the extended state vector, $\langle \boldsymbol{y}^{(j)} \rangle$, where $\langle \cdot \rangle$ denotes the average over a large number of realizations of $\Delta \widetilde{\boldsymbol{\xi}}^{(j)}$. Since the forcing has zero-mean, the equation

$$\langle \boldsymbol{y}^{(j)} \rangle = \langle \boldsymbol{H} \boldsymbol{y}^{(j-1)} + \Delta \widetilde{\boldsymbol{\xi}}^{(j-1)} \rangle = \boldsymbol{H} \langle \boldsymbol{y}^{(j-1)} \rangle \tag{42}$$

shows that $\boldsymbol{H}$ functions as the propagator of the expected value of the state. If $\boldsymbol{H}$ is stable with $\max\{|\lambda(\boldsymbol{H})|\} < 1$, then the average of $\boldsymbol{y}^{(j)}$ at large time, denoted by $\lim_{j \to \infty} \langle \boldsymbol{y}^{(j)} \rangle$, is zero. Denoting by

$$\boldsymbol{y}'^{(j)} = \boldsymbol{y}^{(j)} - \left\langle \boldsymbol{y}^{(j)} \right\rangle \tag{43}$$

the fluctuation of the state, we may compute the auto-covariance matrix $\boldsymbol{P}^{(j)}$ of the state as

$$\begin{aligned} \boldsymbol{P}^{(j)} &= \left\langle \boldsymbol{y}'^{(j)} \boldsymbol{y}'^{(j)*} \right\rangle \\ &= \boldsymbol{H} \boldsymbol{P}^{(j-1)} \boldsymbol{H}^* + \widetilde{\boldsymbol{R}}_{rr}. \end{aligned} \tag{44}$$

Equation (44) is the propagation equation for auto-covariance of the state, which is readily obtained by combining Eqs. (39), (42), and (43). Here, we made use of the fact that the fluctuating state is uncorrelated to the process noise, that is, $\langle \boldsymbol{y}'^{(j-1)} \Delta \widetilde{\boldsymbol{\xi}}^{(j-1)*} \rangle = 0$. The reason is that the stochastic component of $\boldsymbol{y}'^{(j-1)}$ is computed from the evaluation of the stochastic process at the previous time instant, $\Delta \widetilde{\boldsymbol{\xi}}^{(j-2)*}$, see Eq. (39).

If the true state $\boldsymbol{y}^{(1)}$ is used as the initial condition, then $\boldsymbol{P}^{(0)} = 0$ from Eq. 44, and it can be shown that the matrix sequence $\{\boldsymbol{P}^{(j)}\}$ converges to the true covariance

$$\boldsymbol{P} = \lim_{j \to \infty} \boldsymbol{P}^{(j)} = \sum_{n=0}^{\infty} \boldsymbol{H}^n \widetilde{\boldsymbol{R}}_{rr} \left(\boldsymbol{H}^*\right)^n, \tag{45}$$

which solves the discrete-time Lyapunov equation

$$\boldsymbol{H} \boldsymbol{P} \boldsymbol{H}^* - \boldsymbol{P} + \widetilde{\boldsymbol{R}}_{rr} = 0. \tag{46}$$

Since $\widetilde{\boldsymbol{R}}_{rr}$ is hermitian and positive-definite by construction, the existence and uniqueness of the solution $\boldsymbol{P}$ are guaranteed by the Lyapunov theorem if $\boldsymbol{H}$ is stable. Since the process noise $\Delta \widetilde{\boldsymbol{\xi}}^{(j)}$ is Gaussian, the distribution of the realizations of the state, $\boldsymbol{y}$, is also Gaussian for the LTI system at hand. Therefore, the 95% confidence interval of $\boldsymbol{y}$ at t_j can be determined as

$$\text{CI} = \left(\langle \boldsymbol{y}^{(j)} \rangle - 2\sqrt{\text{diag}(\boldsymbol{P}_j)}, \ \langle \boldsymbol{y}^{(j)} \rangle + 2\sqrt{\text{diag}(\boldsymbol{P}_j)} \right). \tag{47}$$

In the limit of large times with $j \to \infty$, this interval converges to the bounded interval $\left(-2\sqrt{\text{diag}(\boldsymbol{P})}, 2\sqrt{\text{diag}(\boldsymbol{P})}\right)$. Under the assumption that the realizations generated by the discrete SDE, Eq. (39), are weakly stationary with zero mean and covariance matrix $\boldsymbol{P}$, the time-lagged autocorrelation function can be determined analytically in terms of the state transition matrix as

$$\boldsymbol{R}_{yy}(n) \equiv \langle \boldsymbol{y}^{(j)} \boldsymbol{y}^{(j+n)*} \rangle = \boldsymbol{P} \left(\boldsymbol{H}^*\right)^{|n|}. \tag{48}$$

By way of the Wiener–Khinchin theorem, the analytical expression for the spectral density function

$$\begin{aligned} \boldsymbol{S}_{yy}(\omega) &= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \boldsymbol{R}_{yy}(n) e^{-i\omega n} \\ &= \frac{1}{\sqrt{2\pi}} \boldsymbol{P} \left[\left(\boldsymbol{I} - e^{-i\omega} \boldsymbol{H}^*\right)^{-1} + \left(\boldsymbol{I} - e^{i\omega} \boldsymbol{H}^*\right)^{-1} - \boldsymbol{I} \right]. \end{aligned} \tag{49}$$

is found by means of the discrete-time Fourier transform. The diagonal of $\boldsymbol{S}_{yy}(\omega)$ then contains the power spectral density of $\boldsymbol{y}$.

4 Example of a turbulent jet

Take as an example of a statistically stationary flow a turbulent, iso-thermal jet at Mach number, based on the jet velocity and the far-field speed of sound, of $M = 0.9$ and Reynolds number, based on the nozzle diameter and the jet velocity, of $\text{Re} \approx 10^6$. The state vector

$$\boldsymbol{q} = [\rho, u_x, u_r, u_\theta, T]^T, \tag{50}$$

comprises the density ρ, temperature T, and cylindrical velocity components u_x, u_r and u_θ in the streamwise, x, radial, r, and circumferential, θ, directions, respectively. For more details, the reader is referred to Brès et al. [3]. Owing to the rotational symmetry of the jet, we may decompose the data, without loss of generality, into azimuthal Fourier components, m.

In the same way, we may decouple the governing equations and construct the linear operator $\boldsymbol{L}_{\bar{q}}$, defined in Eq. (11), for each azimuthal wavenumber independently. Upon linearization of the compressible Navier-Stokes equations, we obtain the linearized Navier-Stokes operator $\boldsymbol{L}_{\bar{q}}$ that governs the fluctuating state, $\boldsymbol{q}' = [\rho', u'_x, u'_r, u'_\theta, T']^T$, here defined in terms of primitive variables, as

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{q}' = \boldsymbol{L}_{\bar{q}} \boldsymbol{q}'. \tag{51}$$

Consequently, $\bar{\boldsymbol{q}} = [\bar{\rho}, \bar{u}_x, \bar{u}_r, 0, \bar{T}]^T$ is the long-time mean of the primitive state, whose azimuthal velocity component is zero for the round jet. For more details, the reader is referred to Schmidt et al. [35]. In this example, we construct a stochastic two-level SPOD-Galerkin model for the symmetric component of the jet with $m = 0$. We interpolate the data on a 950×195 Cartesian mesh that includes the physical domain $x, r \in [0, 30] \times [0, 6]$.

To quantify the flow energy, we use the compressible energy inner product

$$\langle \boldsymbol{q}_1, \boldsymbol{q}_2 \rangle_E = \int_\Omega \boldsymbol{q}_1^* \mathrm{diag} \left(\frac{\overline{T}}{\gamma \bar{\rho} M^2}, \bar{\rho}, \bar{\rho}, \bar{\rho}, \frac{\bar{\rho}}{\gamma(\gamma - 1)\overline{T} M^2} \right) \boldsymbol{q}_2 \mathrm{d}\boldsymbol{x} = \boldsymbol{q}_1^* \boldsymbol{W} \boldsymbol{q}_2, \tag{52}$$

devised by Chu [6], in Eq. (2). $\boldsymbol{W}$ is the weight matrix containing both the numerical quadrature weights and weights associated with this inner product. The wide range of time and length scales of this fully developed turbulent flow becomes apparent from the velocity field.

4.1 SPOD of the turbulent jet

We compute the SPOD from the 10,000 snapshots of the turbulent jet by partitioning the data into $N_b = 77$ blocks of 256 snapshots with an overlap of 50%. These spectral estimation parameter are obtained following the best practices outlined in Towne et al. [40]; Schmidt & Colonius [33]. Owing to the rotational and temporal symmetry of the jet, it suffices to consider the $N_f = 129$ non-negative frequency components.

Figure 1(a) shows the instantaneous streamwise fluctuating velocity, u'_x, and its symmetric component (azimuthal wavenumber $m = 0$), $\hat{u}_x$, of the turbulent jet. Figure 1(b) shows the SPOD eigenvalue spectra of the axisymmetric component. The reader is referred to Schmidt et al. [36] for more details on the physical interpretation of the SPOD.

Figure 2 shows the first, the second and the 10th of the SPOD modes that constitute the basis of our models. Three representative frequencies are presented. Large-scale coherent structures associated with Kelvin–Helmholtz instability waves are observed in the leading modes.

4.2 Subspace modeling

Figure 3 shows the SPOD expansion coefficients, $\boldsymbol{a}_i(t)$, obtained from the oblique projection defined in Eq. (9) for the three modes shown in Fig. 2a–c. Results are shown for the rank 1×129 basis, that is, the basis consisting exclusively of the leading SPOD modes (one mode per frequency). Note again that each SPOD mode is, by construction, associated with a single frequency. To confirm that the oblique projection truthfully represents this property, we compute the periodograms (right column) to reveal the frequency content of the
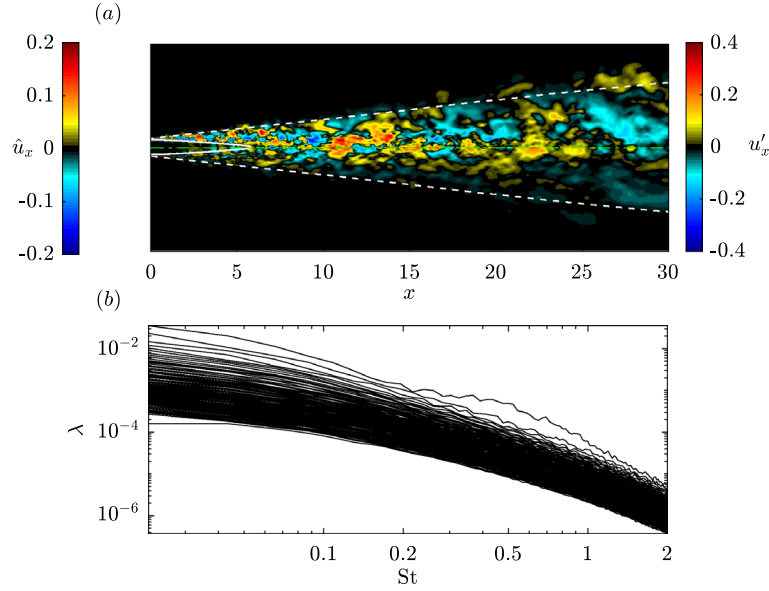
Fig. 1 Axisymmetric component of a $M = 0.9$ transonic jet: **a** streamwise cross section of the instantaneous streamwise fluctuating velocity (top half) and its symmetric component (bottom half); **b** SPOD eigenvalue spectra
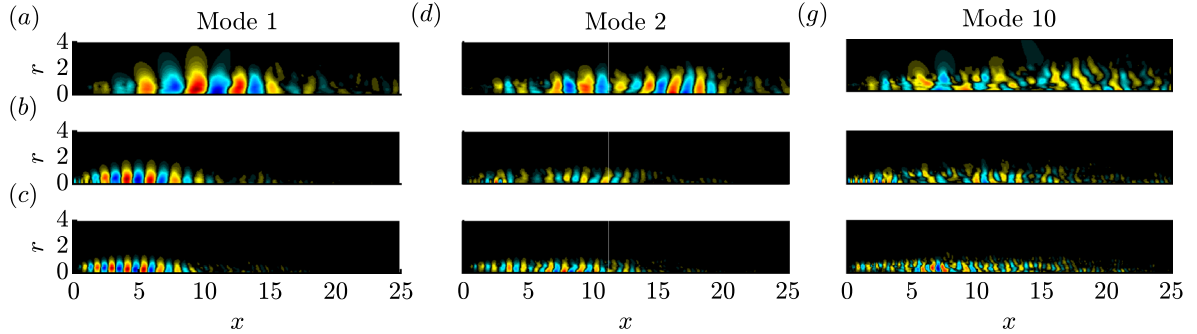

Fig. 2 Examples of SPOD modes that form the basis of the model: **a**, **d**, **g** $St = 0.2$; **b**, **e**, **h** $St = 0.41$; **c**, **f**, **i** $St = 0.63$. The normalized pressure components of the 1st, 2nd, 10th modes are shown in $x, r \in [0, 25] \times [0, 4]$

expansion coefficients. It is observed that the PSD indeed peaks at the respective SPOD frequencies. This confirms that the SPOD modes in fact predominantly represent the spectral content they optimally account for by construction. Even though the model does technically not depend on this property, we note that this observation can be interpreted as an *a posteriori* justification for the use of the oblique projection introduced in Eq. (9), and therefore also of the use of SPOD modes as a modal basis in the time domain.

Figure 4a compares the compressible energy norm, $\|q(x, t)\|_E$, of the full LES data to the energy of the partial reconstructions, $\widetilde{q}$, for modal bases of different sizes. It can be observed that all reconstructions follow the dynamics of the LES. The approximately constant offset between the energy of the LES and the different low-rank approximations is similarly reflected in the retained SPOD energy. In particular, the rank 1×129, 2×129, 3×129, and 10×129 approximations account for 13%, 21%, 27%, and 52% of the total energy, respectively. Figure 4b shows that the offset between the low-rank approximation and the true state is almost constant. This implies that the error stems, almost exclusively, from the energy contained in scales that are not important from a dynamics perspective.

In Fig. 5, the stability of the naive Galerkin ROM of the LTI system, Eq. (14), and the two-level model, Eq. (30), is addressed in terms of the eigenvalue spectra of the corresponding operators, L_{Gal} and $L_{\mathrm{2\text{-}lvl}}$, respectively. The dashed blue line in the spectra corresponds to the disc of radius $\frac{1}{\Delta t}$, centered about $-\frac{1}{\Delta t}$. It demarcates the region of stability; eigenvalues inside the circle are associated with temporal decay, whereas eigenvalues on the outside are associated with exponential amplification. Clearly, the simple Galerkin ROM is unstable, whereas the two-level model is stable. Furthermore, the eigenvalues of the 2-level model are mostly
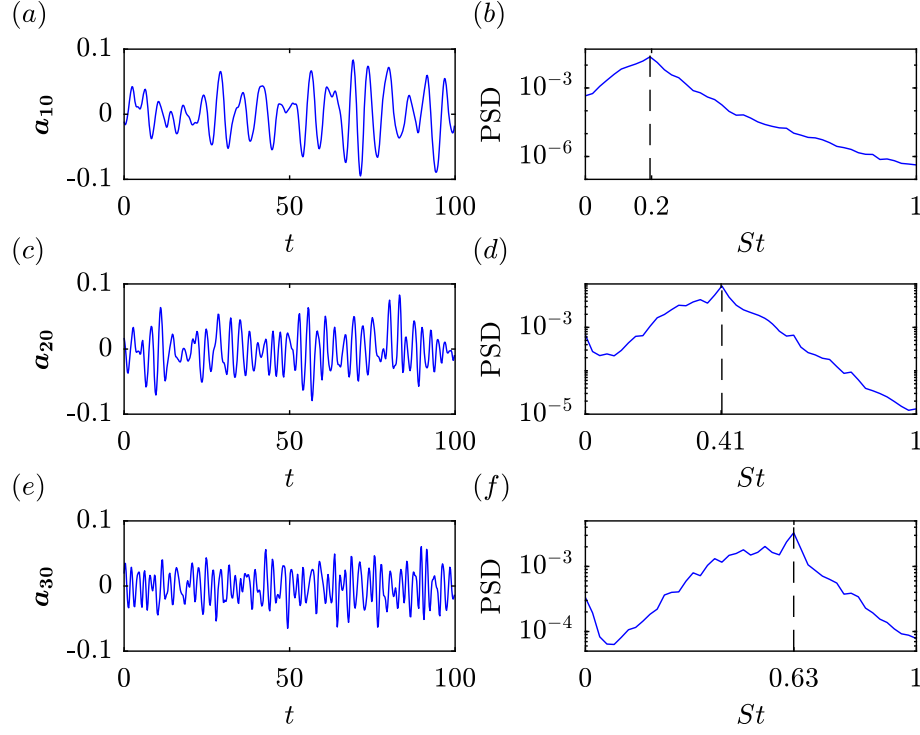
Fig. 3 SPOD expansion coefficients $a_i(t)$ for the three representative modes shown in Fig. 2a–c as part of the rank 1×129 basis: **a**, **b** $St = 0.2$; **c**, **d** $St = 0.41$; **e**, **f** $St = 0.63$. Time trace over 100 time units (left), and PSD (right). The frequency of the SPOD modes is marked by dashed lines in the spectra
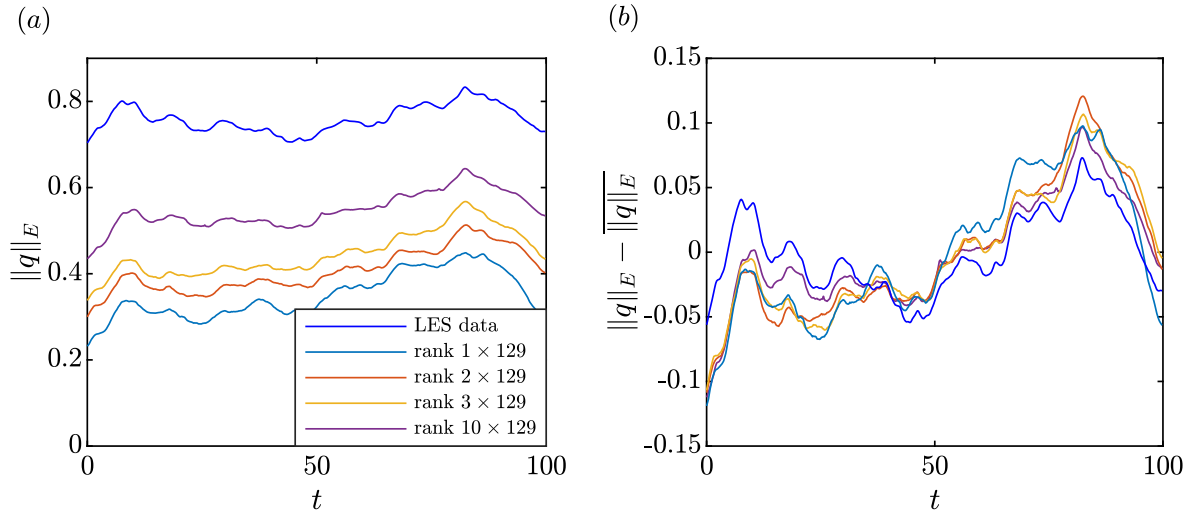

Fig. 4 Compressible energy norm in (**a**), and deviation from the mean energy in (**b**). The LES data and the low-rank approximations with 13%, 21%, 27%, 52% energy are reported. Part (b) emphasizes that the dynamics (deviation from the mean) are well captured even by the lowest-rank approximation

confined to the upper half of the stability region. This behavior is explained by the symmetry of the SPOD spectrum and the restriction of the modal basis to non-redundant positive frequency content. Figure 5d–f shows that the eigenvalues of $\boldsymbol{L}_{\text{2-lvl}}$ remain confined to a specific area within the region of stability when the model rank is increased. These results demonstrate that the transition matrices $\boldsymbol{H}$ computed using Eq. (40) are stable, that is, $\max\{|\lambda(\boldsymbol{H})|\} < 1$. Hence, when there is no stochastic input to the SDE, Eq. (39), the extended state vector $\boldsymbol{y}$ vanishes as $t \to \infty$ as its expected value is zero. In accordance with the modeling philosophy, this implies that the appropriate forcing input is necessary to sustain the turbulent flow.
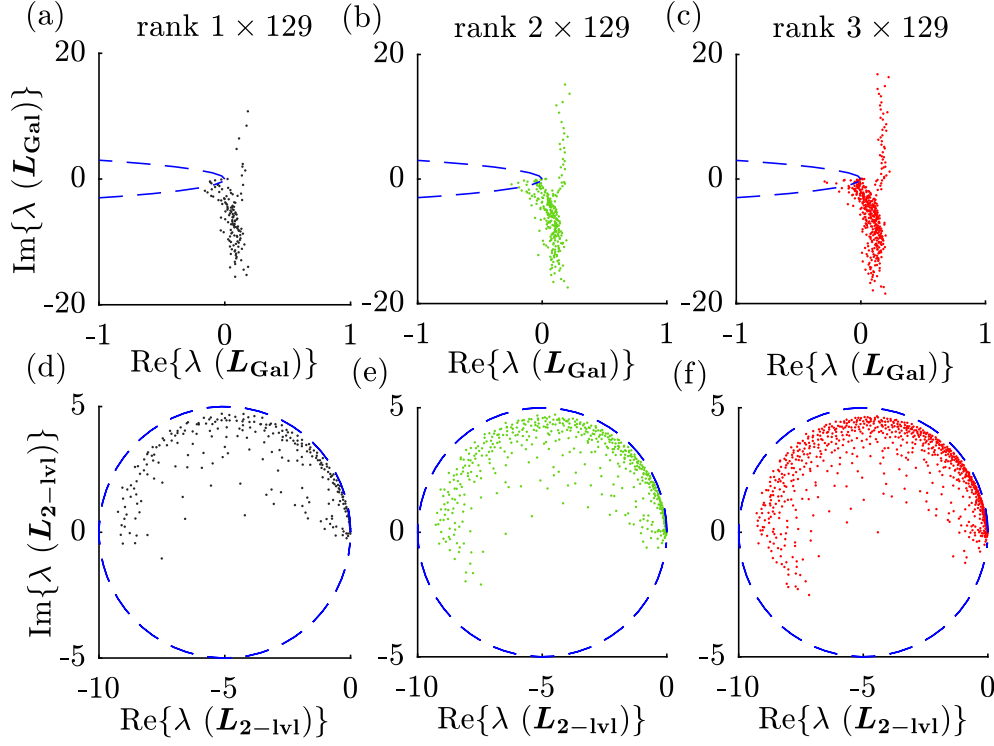
Fig. 5 Eigenvalues of the standard SPOD-Galerkin ROM, $\lambda(\boldsymbol{L}_{\text{Gal}})$, and the proposed two-level model, $\lambda(\boldsymbol{L}_{\text{2-lvl}})$: **a, d** rank 1×129 (black); **b, e** 2×129 (green); **c, f** 3×129 (red). The blue dashed circle represents the stability region. The eigenvalues outside of the unit disc in **a–c** indicate that Galerkin projection yields an unstable model. On the contrary, the two-level model in **d–f** is stable (colour figure online)

As the turbulence closure relies on linear regression, it has to be shown that $\boldsymbol{L}_{\text{2-lvl}}$ converges as more data are used for its construction. Consider the sequence of matrices, $\{\boldsymbol{L}_{\text{2-lvl}}^{(n)}\}$, where n is the number of snapshots used in the linear regression, Eq. (27). We define a normalized norm of these the matrix sequence, $\| \cdot \|_D$, as

$$\|\boldsymbol{L}_{\text{2-lvl}}\|_D \equiv \frac{\|\boldsymbol{L}_{\text{2-lvl}}^{(n)} - \boldsymbol{L}_{\text{2-lvl}}^{(N)}\|_F}{\|\boldsymbol{L}_{\text{2-lvl}}^{(N)}\|_F}, \tag{53}$$

where $\| \cdot \|_F$ denotes the Frobenius norm, and N is the total number of available snapshots. The norm $\| \cdot \|_D$ measures the normalized distance between the matrices constructed with n and N snapshots. The convergence of $\boldsymbol{L}_{\text{2-lvl}}$ as more and more data is added is apparent from Fig. 6.

Figure 7 compares the left, $\boldsymbol{M}_{ab}$ (left column), and the right, $\boldsymbol{M}_{bb}$ (right column), square matrices that together form $\boldsymbol{M}$, as defined in Eq. (31), for modal basis of three different sizes. The matrix $\boldsymbol{M}_{ab}$ represents the cross-correlation information between the expansion and the forcing coefficients. Compared to the off-diagonal entries, a notable positive correlation is observed on the diagonal of the matrix $\boldsymbol{M}_{ab}$. This suggests that the correlation between the expansion and the forcing coefficients at the same frequencies dominate. The matrix $\boldsymbol{M}_{bb}$ shows the cross-correlation between the forcing coefficients. This matrix, too, is diagonally-dominant; however, the autocorrelation of the forcing is negative. In the following discussion of Fig. 8, we will show that this procedure of computing $\boldsymbol{M}$ leads to an almost flat residue spectrum.

Following the steps outlined in Sect. 3.1.1, we proceed with the modeling of the second-level residue. Recall that the proposed model closure hinges on the assumption that the highest-level residue can be modeled as random noise. This assumption is tested in Fig. 8, which examines the temporal evolution and the spectra of the forcing coefficients, $\boldsymbol{b}(t)$, and the residue, $\boldsymbol{r}(t)$. The rank 1×129 case is shown as an example. It is observed that the residue is of significantly lower amplitude than the forcing. Similar to the trend observed for the SPOD expansion coefficient in Fig. 3, the PSD of the forcing coefficients, shown in the right column, attains its maximum value at the corresponding mode frequency. For the frequencies at hand, a separation of at least two orders of magnitude between the maximum and minimum values of the PSD is found. The PSD
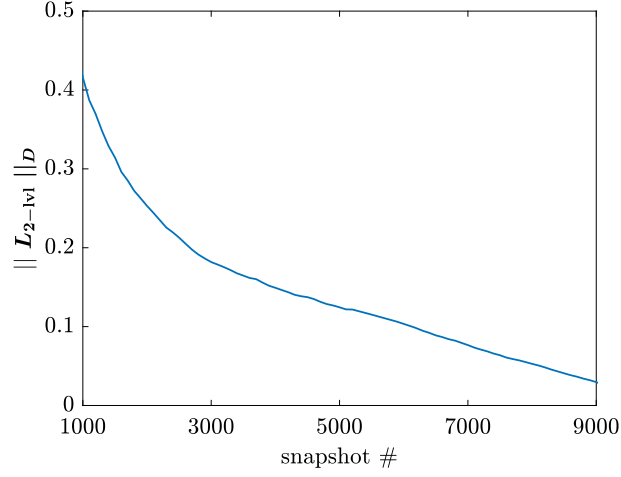
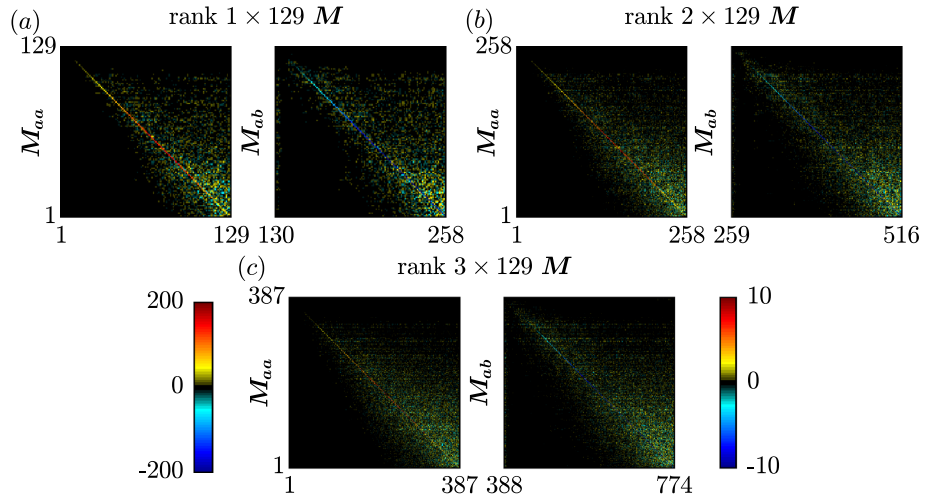**Fig. 6** Convergence of the matrix sequence $\{L_{2\text{-lvl}}^{(n)}\}$

Fig. 7 Matrices M_{ab} (left column) and M_{bb} (right column), computed from the least-squares problem (27) for models of different rank: **a** rank 1 × 129; **b** rank 2 × 129; **c** rank 3 × 129

of the residue, on the contrary, is significantly flatter. To further quantify the white noise assumption, Fig. 8c,f,i shows the corresponding (normalized) autocorrelations calculated as $R_{xx}(\tau) = \langle x(t)x(t+\tau)\rangle$, where the ensemble average $\langle\cdot\rangle$ is taken over the same blocks for the SPOD, see Sect. 4.1. The autocorrelations of the residue decay rapidly, whereas the forcing coefficients are correlated over 20 or more time units. This suggests that the residue can be modeled as white-in-time. Both the spectral flatness and the rapid drop of the autocorrelation motivate the truncation of the multi-level stochastic model at the second level.

To further motivate the proposed model truncation at the second level, we next examine the probability distributions of the rank-1×129 coefficients a and b, and the corresponding residue r in Fig. 9. As desired, the probability distributions of residue at different frequencies are nearly Gaussian, which is clearly not the case for the mode and forcing coefficients. In accordance with the observations made in the context of Fig. 8, this suggests that r may be modeled with components that are mutually correlated, but that are white-in-time. In the model, the correlation between the components of the residue is accounted for by the matrix G [see Eq. (37)] that filters the white-in-time input w to generate the process noise $\widetilde{w}$ that drives the final model.
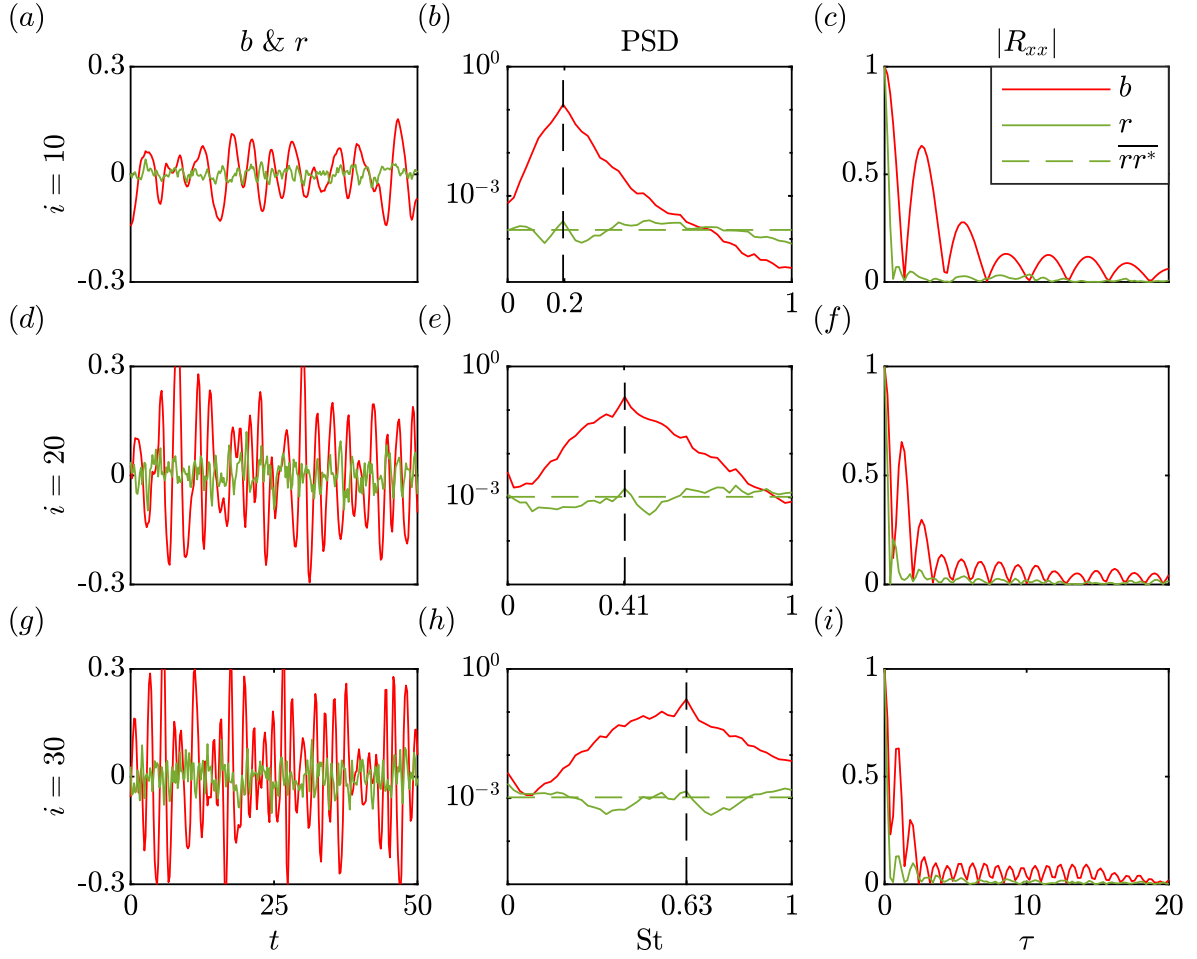
Fig. 8 Time traces (left column), spectra (middle column) and autocorrelation estimates (right column) of the rank 1×129 forcing coefficients, $\boldsymbol{b}$ (red), and residue, $\boldsymbol{r}$ (green), at three representative frequencies: $(a - c)$ $i = 10$, or $St = 0.2$; $(d - f)$ $i = 20$, or $St = 0.41$; $(g - i)$ $i = 30$, or $St = 0.63$. In **b**, **e**, **h**, vertical dashed lines indicate the SPOD mode frequencies and horizontal dashed green lines the mean PSD of the residue. The absolute values of the normalized autocorrelation is shown in **(c, f, i)** (colour figure online)

4.3 Subspace and flow field realizations

Consistent with Eq. (39), we use the forward Euler method to march the SDE, Eq. (38), forward in time. The time step of the original data, $\Delta t = 0.2$, is used to guarantee consistency with the modeling of the forcing and the residue.

Figure 10 shows the comparison between the rank 1×129 approximation and a single realization of the rank 1×129 model, as well as their power spectra at different frequencies. Starting from an initial condition taken from the data, this example of a realization of the stochastic model follows the initial transient dynamics of the data for approximately 3 time units. To demonstrate that the model accurately captures the known trajectory, we run the model without stochastic forcing and the true residue as the input $\boldsymbol{r}$. It can be seen that the model output (red dots) and the LES trajectory (black line) are indistinguishable. A more rigorous approach to quantify the predictability of the model is use of the analytical 95% confidence interval defined by Eq. (47), and Monte Carlo simulation. For the latter, 10^4 realizations of the stochastic model, all starting from the same initial condition, were computed. It can be seen that the envelope of uncertainty closely follows the reference in the vicinity of $t = 0$. As theoretically predicted in the context of Eq. (47), the region of uncertainty stays bounded for larger times, implying that the model is stable. A good agreement is also found between the power spectra of the coefficients of the model and the rank 1×129 approximation. This observation confirms that the model preserves the direct correspondence between modes and frequencies inherent to SPOD.
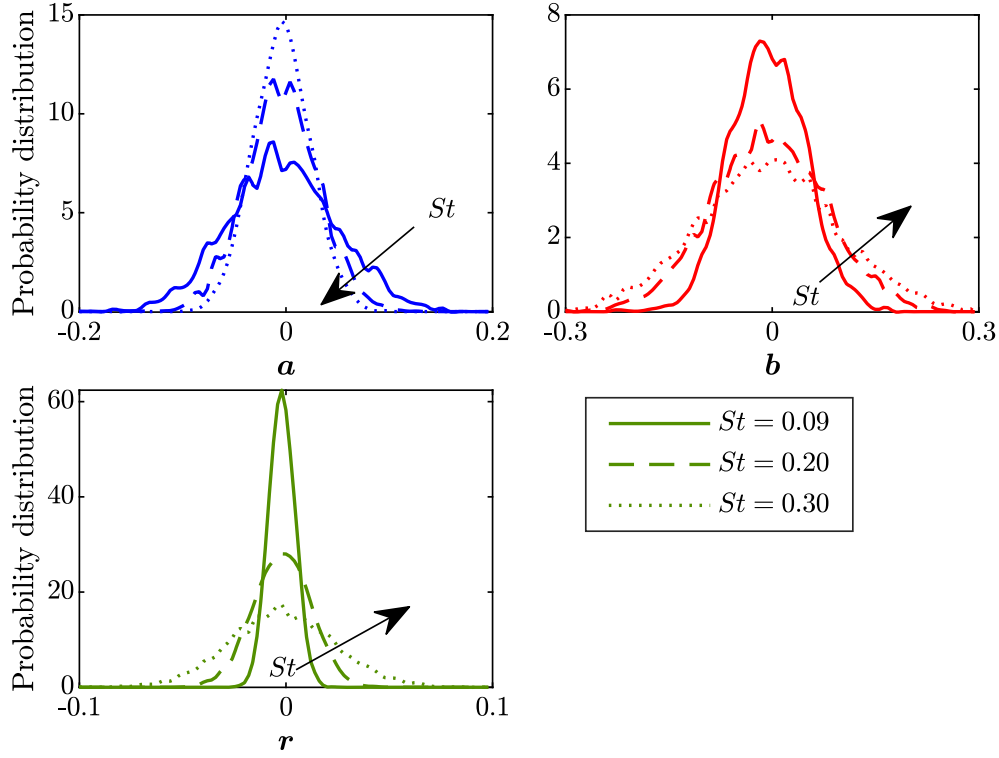
Fig. 9 The probability distribution of rank 1×129 state coefficients $\boldsymbol{a}$ (blue), forcing coefficients $\boldsymbol{b}$ (red) and the residue $\boldsymbol{r}$ (green) at different frequencies: $St = 0.09$ (solid), $St = 0.20$, (dashed), $St = 0.30$ (dotted) (colour figure online)

Figure 11 shows, analogously to Fig. 10b,d,f, the power spectra of the approximation and model coefficients at different frequencies for the rank 10×129 approximation. For all frequencies, the power spectra of the rank 10×129 approximation and model follow the order of the SPOD eigenvalues. As for the rank 1×129 model, a favorable agreement between approximation and model is observed. In summary, Figs. 10 and 11 demonstrate that both the dynamics and statistics of the state coefficients $\boldsymbol{a}$ are well described by two-level models of different fidelity. More results for the rank 2×129 and 3×129 models are provided in Appendix A.

Having found that the spectral content of individual projection coefficients is represented well by the model, we now focus on the entire fluctuating flow field $\boldsymbol{q}'$. Figure 12 compares the compressible energy norm of the LES data, the rank 10×129 , and a single realization of rank 10×129 model. Also indicated in gray is the region of uncertainty obtained from 2500 Monte Carlo realizations of the model that start from the same initial condition. Two shifted axes are used to account for the constant offset between the oblique projection and model, on one hand, and the full data, on the other hand (see discussion of Fig. 4). A favorable agreement of the general dynamics is observed between the original data a and the rank 10×129 model. Starting from the initial condition of the data, the random realization shown here (first Monte Carlo sample) follows the general trend of the data for, arguably, up to 100 time units. The model uncertainty region show that significantly larger variations are of course possible, but that the model stays bounded within an expected range. As desired, the ensemble average of the Monte Carlo simulations converges to the long-time mean of reduced-rank approximation.

After establishing that the approach yields a good model of the flow dynamics in terms of energy, we next compare the original, approximated and modeled flow fields. As examples, Fig. 13 compares the instantaneous pressure fields at $t = 5$ for rank 1×129 and rank 10×129. The same realization of the stochastic rank 10×129 model as in Fig. 12 (dashed line) is shown. From Fig. 12, we expect that the model closely follows the approximation (and therefore the data) at this short time after its initialization. This can be clearly seen for both the rank 1×129 and rank 10×129 approximations and models in Fig. 13b–e, respectively. As expected, the higher-rank approximation and model yield a more detailed picture of the flow. But even for rank 1×129, many of the eminent features of the LES data are accurately captured. Due to the stochastic nature of the
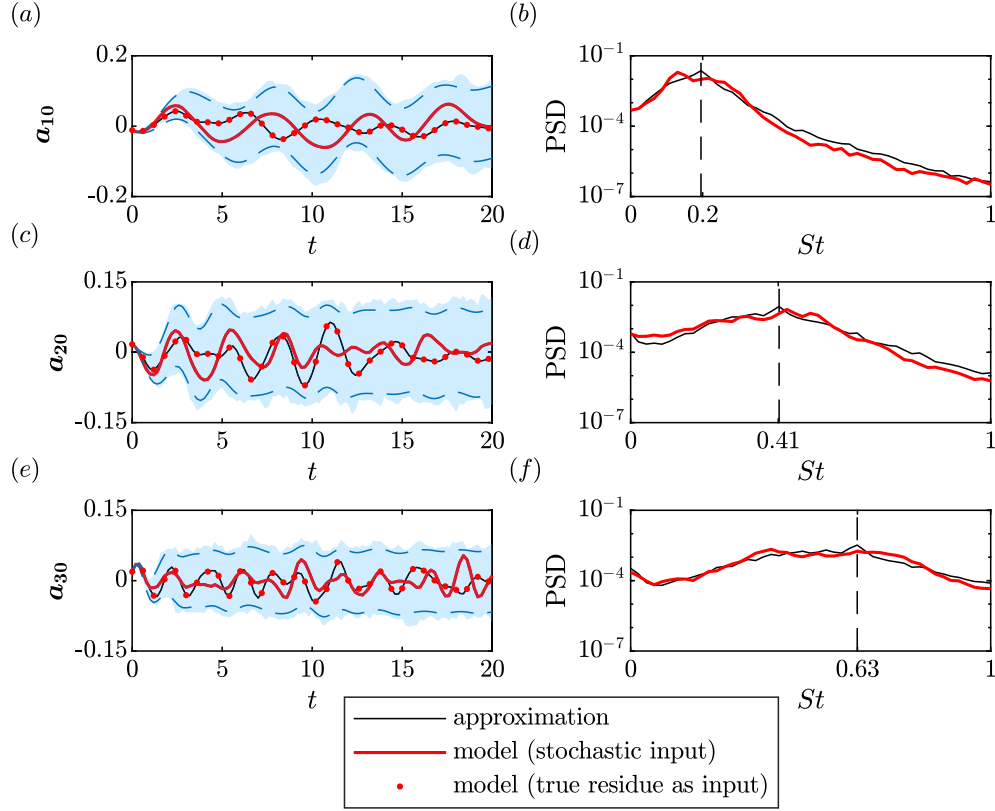
Fig. 10 Time series and spectra for the rank 1×129 approximation (black line, representing the original data) and a single realization of the rank 1×129 model (red line) for three representative frequencies: **a,b** $St = 0.2$; **c,d** $St = 0.41$; **e,f** $St = 0.63$. Red dots show the model output when the true residue taken from the data is used as input. It is indistinguishable from the data. Blue dashed lines mark the analytic 95% confidence interval. The blue shaded area outlines a Monte Carlo envelope based on 10^4 realizations of the stochastic model, for comparison (colour figure online)

problem, this similarity fades for larger times. Instead, the stochastic model will yield a unique flow trajectory that can be interpreted as surrogate data that accurately reproduces the second-order statistics and dynamics of the input data.

The temporal evolution of the fluctuating pressure field is investigated in terms of x-t diagrams at $r = 0.5$ (along the lipline) in Fig. 14. The convective nature of the flow field becomes apparent from the diagonal pattern corresponding to the advection of the wavepackets previously seen in figure 13. The original LES data shown in Fig. 14a is compared to its low-rank approximations (middle) and the model output (right). It can be seen that the rank 1×129 model in Fig. 14c exhibits dynamics that are highly reminiscent of the rank 1×129 approximation and the full data. Similar observations are made for the rank 10×129 approximation and model in Fig. 14d and e, respectively. We note that the vertical stripes in Fig. 14d stem from spatial aliasing, which occurs for subdominant modes at high frequencies. We emphasize that this issue is linked to the discretization of the data set, which was interpolated from the unstructured LES grid to a cylindrical grid for post-processing. The phenomenon is, however, accurately reproduced by the model, as can be seen in Fig. 14e.

After establishing that the model reproduces a surrogate flow field that is qualitatively very similar to the original data, we now focus on the flow statistics. An obvious choice is to use the SPOD for this purpose as well. In Fig. 15, we hence compare the SPOD eigenvalue spectra of LES data to both the rank 1×129 and 10×129 approximations and models, respectively. For clarity, only the leading two eigenvalues are shown for the LES, and the leading 3 eigenvalues for the rank 10×129 approximation and model. It can be seen that both models accurately reproduce the eigenvalue spectra of the LES data for all but the lowest frequencies. The rank 10×129 model accurately follows the 10×129 approximation at frequencies with $St > 0.2$, but somewhat under-predicts the 10×129 approximation at lower frequencies. For the 1×129 model, we observe an overall very good fit between approximation and model for the leading mode. In accordance with the conjectures

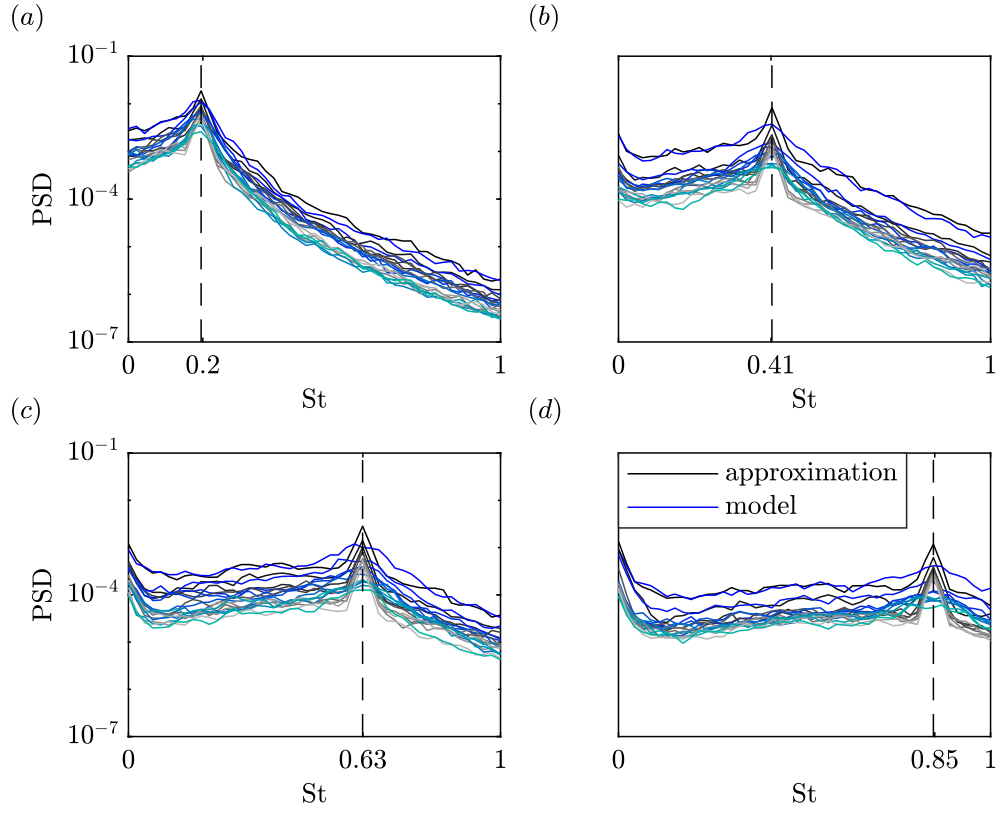**Fig. 11** The power spectra of the first three state coefficients $\boldsymbol{a}_i$ of rank 10×129 approximation (black) and model (blue) at four representative frequencies: **a** $St = 0.2$; **b** $St = 0.41$; **c** $St = 0.63$; **d** $St = 0.85$ (colour figure online)
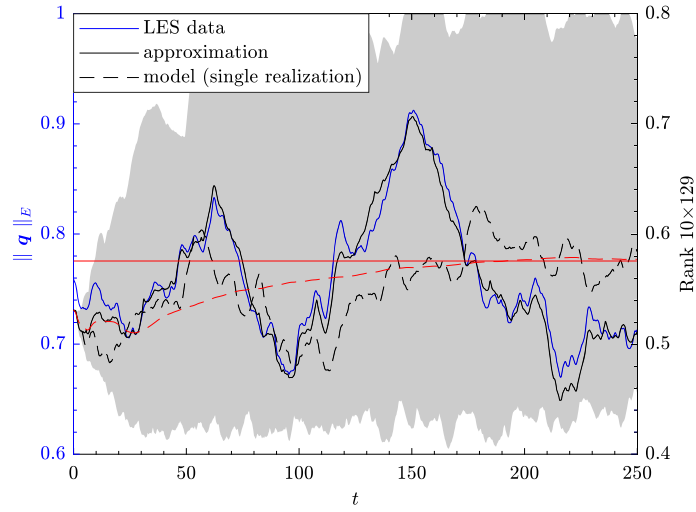

Fig. 12 Comparison between the compressible energy norms of LES data (blue), rank 10×129 approximation (black), and a single realization of the stochastic model (black dashed). The shaded area is a Monte Carlo envelope based on 2500 model evaluations. The ensemble average of Monte Carlo simulations (red dashed) and the long-time mean of rank 10×129 approximation (red solid) are shown for comparison (colour figure online)
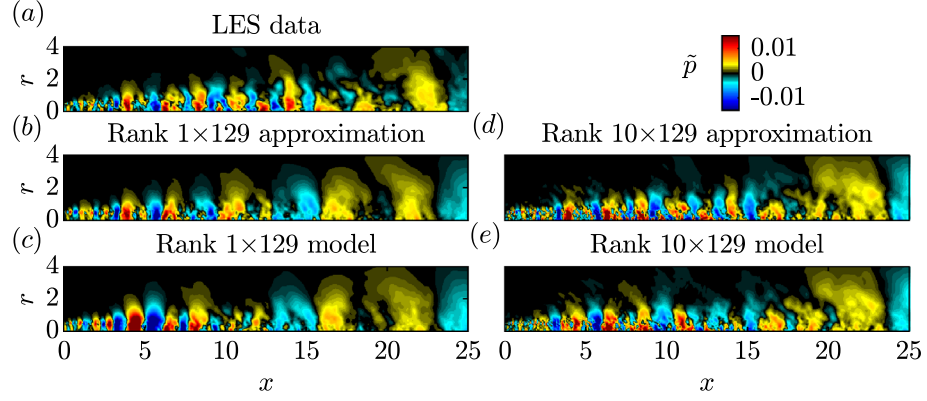
Fig. 13 Comparison between pressure fields of the LES data, its low-rank approximations, and random realizations of the two-level model at $t = 5$: **a** LES data; **b, c** rank 1×129 and model; **d, e** rank 10×129 and model. The pressure is normalized by its root mean square for comparability (see discussion of Fig. 4)
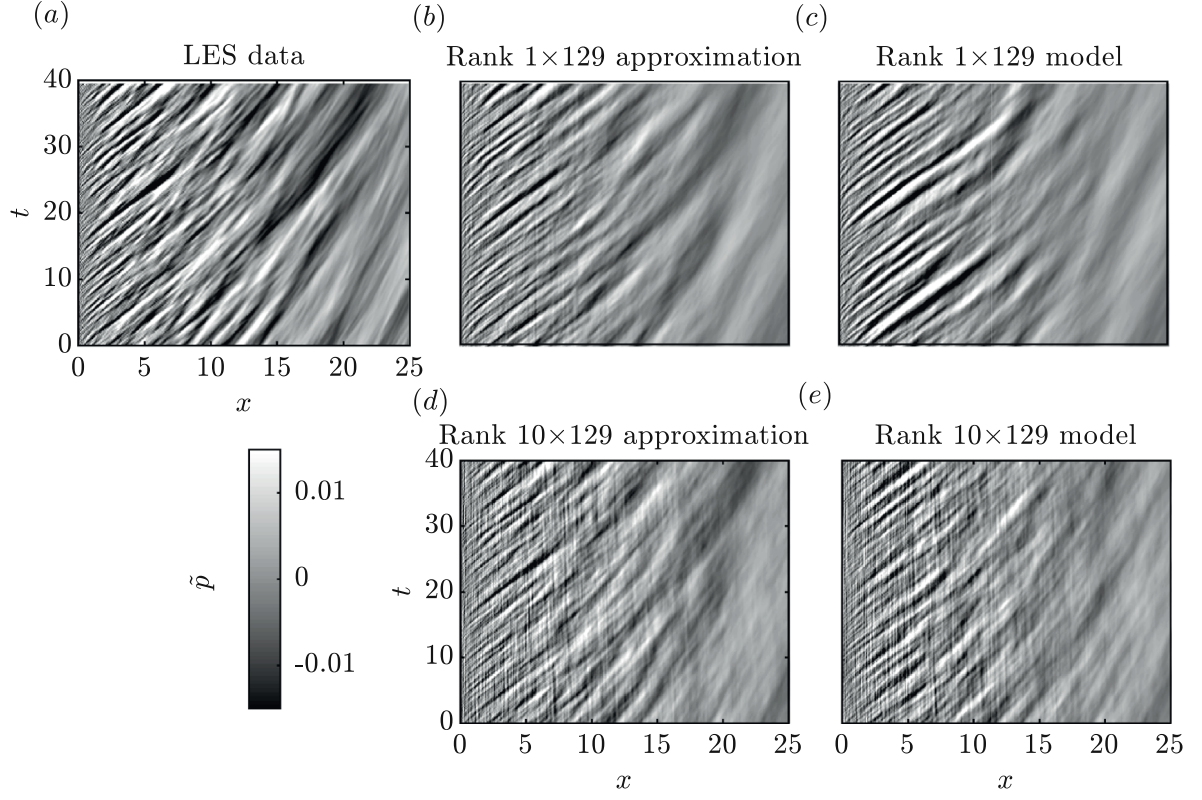

Fig. 14 Comparison between LES data **(a)**, rank 1×129 and model **(b,c)**, rank 1×129 and model **(d,e)** of normalized pressure component at $r = 0.5$ for different time
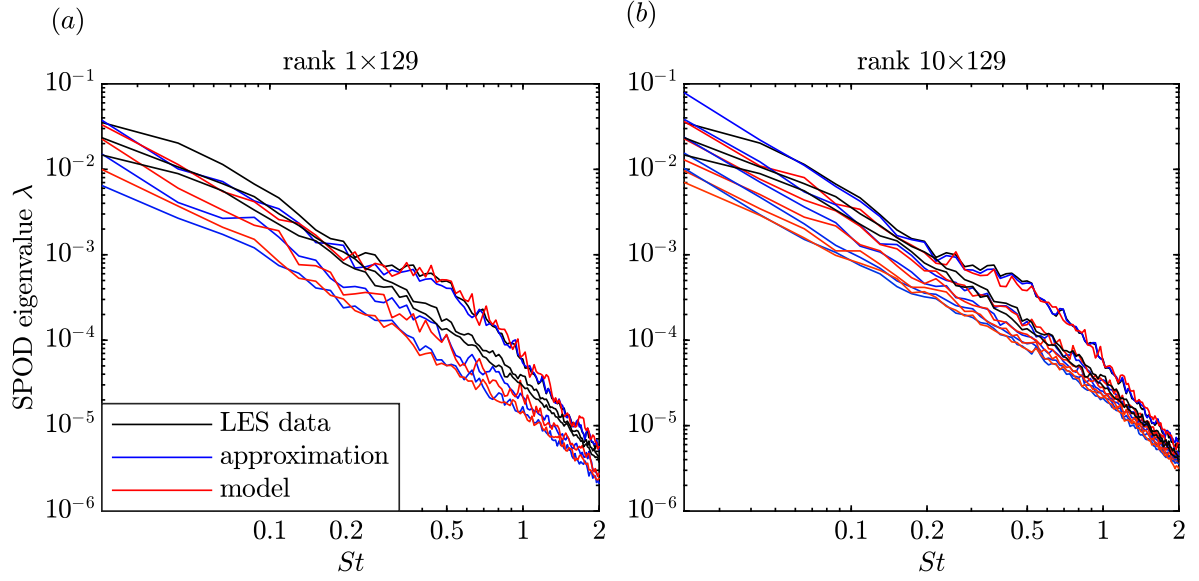
Fig. 15 Comparison of SPOD eigenvalue spectra of the LES data (black), low-rank approximations (blue) and 2–level models (red): **a** rank 1×129; **b** rank 10×129. The leading two eigenvalues are shown for the LES data (with shaded area between them), and the leading three eigenvalues for the approximations and models, respectively

drawn from Fig. 14, we conclude that the rank 1×129 model produces an accurate surrogate flow field, both qualitatively and statistically.

5 Conclusions

In this work, we propose stochastic two-level SPOD-Galerkin model for turbulent flows. The model is stable and produces surrogate data with accurate dynamics and statistics. The first level of the model is a forced Galerkin-ROM which propagates the SPOD-expansion coefficients in time. The SPOD-expansion coefficients are obtained by oblique projection of the data onto the SPOD basis. Consistent with this procedure, the system dynamics matrix of the ROM is obtained by projecting the discretized governing equations onto a low-dimensional SPOD basis. Here, we consider the linearized compressible Navier-Stokes equations. The offset between the linear approximation and the true state evolution is interpreted as a forcing of this linear system. The second level of the model is a stochastically forced system for this forcing. This closure is inspired by the purely data-driven linear inverse model by Penland [24,25] and its generalization, the linear multi-level regression model by Kravtsov et al. [19]; Kondrashov et al. [18]. The basic idea behind these models is to achieve closure by inflating the state with additional levels until the residue can be modeled as white noise. In the purely data-driven model of Kravtsov et al. [19], this criterion was met at the third level. For the turbulent jet governed by the compressible Navier–Stokes equations, we show that closure is achieved at the second level if a physical model is used describe the linear dynamics at the first level.

The two main limitations of our model are that it requires access to the discretized linear operator and a long time-series with sufficiently converged space-time statistics. The computationally most involved part of the model is the computation of the SPOD in the first place. The model itself is a system of $M \times N_f$ ordinary differential equations. It is therefore computationally cheap to integrate and lends itself to Monte Carlo simulations.

The proposed model is inspired by recent progress in flow modeling based on the resolvent operator, and the correspondence between resolvent and SPOD modes [40] under the assumption of white noise forcing. A possible extension of this work is hence to substitute the SPOD basis by a resolvent basis, which only requires knowledge of mean flow. Another potential venue to improve the model is to further offset differences between the reduced-order representation and the true data, for example by using the ROM Error Surrogates (ROMES) method by Drohmann and Carlberg [7], which implements Gaussian-process regression to construct error surrogates.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable reques

A Stochastic SPOD-Galerkin two-level model for rank 2×129 and rank 3×129 cases

This appendix reports the additional results for the rank 2×129 and rank 3×129 cases, which were omitted in Sect. 4.3 for brevity. Figure 16a–d show the comparison between the power spectra of the state coefficients a_i and the rank 2×129 expansion coefficients at different frequencies. Figure 16e–h report the corresponding results for the rank 3×129 case. Good agreements between the approximations and models are observed in both cases. The corresponding SPOD eigenvalue spectra shown in Fig. 16i,j show that both models accurately reproduce the eigenvalue spectra of the LES data for a wide range of frequencies. From these observations and the favorable results obtained for the 1×129 baseline model, as previously reported in Figs. 10 and 15, it concluded that subdominant SPOD modes are not required in the modal expansion.
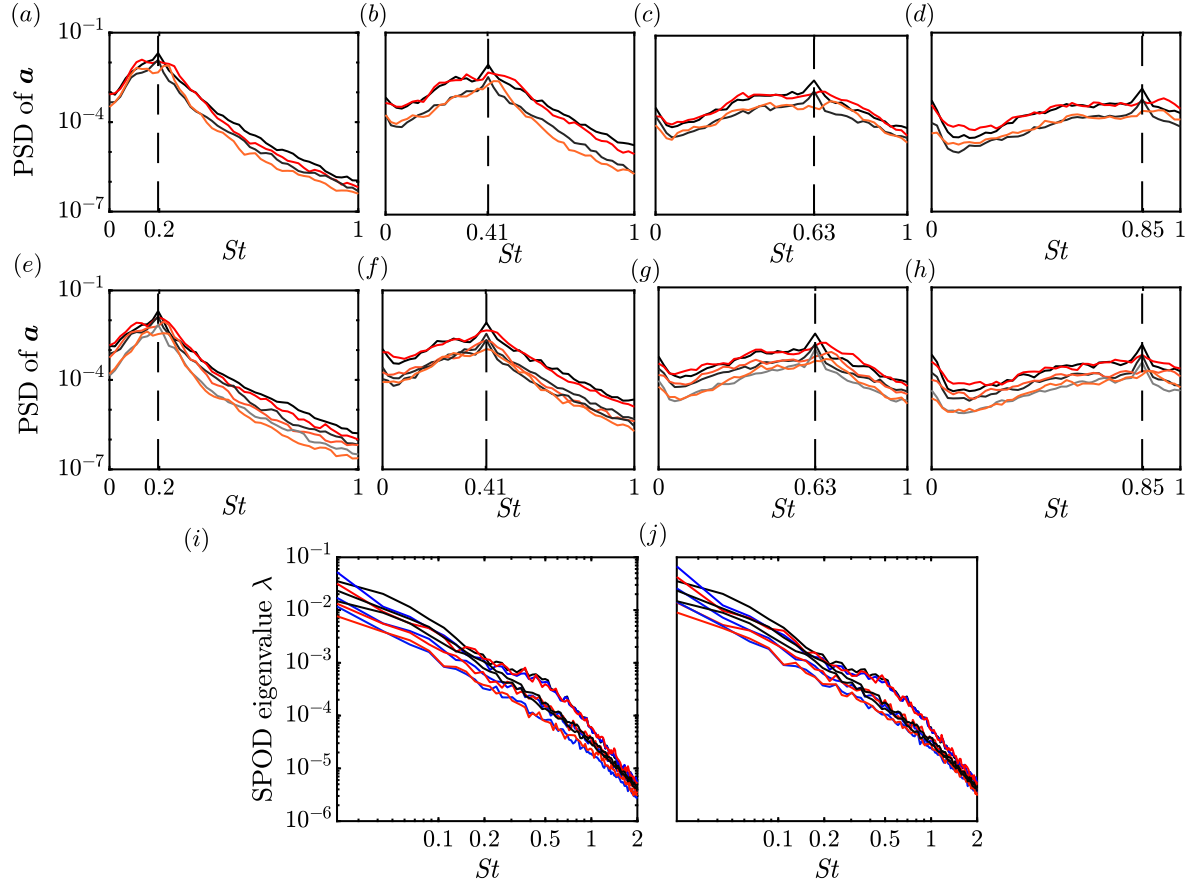

Fig. 16 Power spectra of the state coefficients a_i of low-rank approximations (red) and 2–level models (black) of rank 2×129 (first row) and rank 3×129 (second row) at four representative frequencies: **a,e** $St = 0.2$; **b,f** $St = 0.41$; **c,g** $St = 0.63$; **d,h** $St = 0.85$. Comparison of SPOD eigenvalue spectra of the LES data (black), low-rank approximations (blue) and 2–level models (red): **i** rank 2×129; **j** rank 3×129. Compare Figs. 10, 12 and 15

References

1. Aubry, N., Holmes, P., Lumley, J.L., Stone, E.: The dynamics of coherent structures in the wall region of a turbulent boundary layer. J. Fluid Mech. **192**, 115–173 (1988)
2. Barone, M.F., Kalashnikova, I., Segalman, D.J., Thornquist, H.K.: Stable galerkin reduced order models for linearized compressible flow. J. Compt. Phys. **228**(6), 1932–1946 (2009)
3. Brès, G.A., Jordan, P., Jaunet, V., Le Rallic, M., Cavalieri, A.V.G., Towne, A., Lele, S.K., Colonius, T., Schmidt, O.T.: Importance of the nozzle-exit boundary-layer state in subsonic turbulent jets. J. Fluid Mech. **851**, 83–124 (2018)
4. Cavalieri, André VG., Rodríguez, Daniel, Jordan, Peter, Colonius, Tim, Gervais, Yves: Wavepackets in the velocity field of turbulent jets. J. Fluid Mech. **730**, 559–592 (2013)
5. Chekroun, M.D., Kondrashov, D.: Data-adaptive harmonic spectra and multilayer stuart-landau models. Chaos: Interdiscip. J. Nonlinear Sci. **27**(9), 093110 (2017)
6. Chu, B.-T.: On the energy transfer to small disturbances in fluid flow (part 1). Acta Mech. **1**(3), 215–234 (1965)
7. Drohmann, M., Carlberg, K.: The romes method for statistical modeling of reduced-order-model error. SIAM/ASA J. Uncertain. Quantif. **3**(1), 116–145 (2015)
8. Glauser, M, N., Leib, S. J., Geroge, W. K.: Coherent structures in the axisymmetric turbulent jet mixing layer. In: Turbulent Shear Flows 5, (ed. F. Durst et al.), pp. 134–145. Springer. (1987)
9. Gordeyev, S.V., Thomas, F.O.: Coherent structure in the turbulent planar jet. Part 1. Extraction of proper orthogonal decomposition eigenmodes and their self-similarity. J. Fluid Mech. **414**, 145–194 (2000)
10. Gudmundsson, K., Colonius, T.: Instability wave models for the near-field fluctuations of turbulent jets. J. Fluid Mech. **689**, 97–128 (2011)
11. Holmes, P., Lumley, J.L., Berkooz, G.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge University Press, Cambridge (1996)
12. Holmes, P., Lumley, J.L., Berkooz, G., Mattingly, J.C., Wittenberg, R.W.: Low-dimensional models of coherent structures in turbulence. Phys. Rep. **287**(4), 337–384 (1997)
13. Hussain, A.K.M.F., Reynolds, W.C.: The mechanics of an organized wave in turbulent shear flow. J. Fluid Mech. **41**(2), 241–258 (1970)
14. Johnson, S.D., Battisti, D.S., Sarachik, E.S.: Empirically derived markov models and prediction of tropical pacific sea surface temperature anomalies. J. Clim. **13**(1), 3–17 (2000)
15. Kalashnikova, I., Barone, M.F.: Stable and efficient galerkin reduced order models for non-linear fluid flow. In: 6th AIAA Theoretical Fluid Mechanics Conference, p. 3110 (2011)
16. Kalashnikova, I., Barone, M.F.: Efficient non-linear proper orthogonal decomposition/galerkin reduced order models with stable penalty enforcement of boundary conditions. Int. J. Numer. Meth. Eng. **90**(11), 1337–1362 (2012)
17. Kondrashov, D., Chekroun, M.D., Ghil, M.: Data-driven non-markovian closure models. Phys. D **297**, 33–55 (2015)
18. Kondrashov, D., Kravtsov, S., Robertson, A.W., Ghil, M.: A hierarchy of data-based enso models. J. Clim. **18**(21), 4425–4444 (2005)
19. Kravtsov, S., Kondrashov, D., Ghil, M.: Multilevel regression modeling of nonlinear processes: derivation and applications to climatic variability. J. Clim. **18**(21), 4404–4424 (2005)
20. Lumley, J. L. (1967) The structure of inhomogeneous turbulent flows. In Atmospheric turbulence and radio wave propagation (ed. A. M. Yaglom & V. I. Tatarski), pp. 166–178. Nauka
21. Lumley, J.L.: Stochastic Tools in Turbulence. Academic Press, Cambridge (1970)
22. Nekkanti, A., Schmidt, O. T.: Frequency-time analysis, low-rank reconstruction and denoising of turbulent flows using spod. arXiv preprint arXiv:2011.03644 (2020)
23. Noack, B.R., Afanasiev, K., Morzyński, M., Tadmor, G., Thiele, F.: A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. J. Fluid Mech. **497**, 335–363 (2003)
24. Penland, C.: Random forcing and forecasting using principal oscillation pattern analysis. Mon. Weather Rev. **117**(10), 2165–2185 (1989)
25. Penland, C.: A stochastic model of indopacific sea surface temperature anomalies. Phys. D **98**(2–4), 534–558 (1996)
26. Penland, C., Matrosova, L.: Prediction of tropical atlantic sea surface temperatures using linear inverse modeling. J. Clim. **11**(3), 483–496 (1998)
27. Penland, C., Sardeshmukh, P.D.: The optimal growth of tropical sea surface temperature anomalies. J. Clim. **8**(8), 1999–2024 (1995)
28. Rempfer, D.: On low-dimensional galerkin models for fluid flow. Theor. Comput. Fluid Dyn. **14**(2), 75–88 (2000)
29. Rowley, C. W.: Modeling, simulation, and control of cavity flow oscillations. PhD thesis, California Institute of Technology (2002)
30. Rowley, C. W., Colonius, T., Murray, R.: Dynamical models for control of cavity oscillations. In: AIAA Pap., p. 2126 (2001)
31. Rowley, C.W., Colonius, T., Murray, R.M.: Model reduction for compressible flows using pod and galerkin projection. Phys. D **189**(1–2), 115–129 (2004)
32. Rowley, C.W., Dawson, S.T.: Model reduction for flow analysis and control. Annu. Rev. Fluid Mech. **49**, 387–417 (2017)
33. Schmidt, O.T., Colonius, T.: Guide to spectral proper orthogonal decomposition. AIAA J. **58**(3), 1023–1033 (2020)
34. Schmidt, O. T., Colonius, T., Bres, G. A.: Wavepacket intermittency and its role in turbulent jet noise. In: 55th AIAA Aerospace Sciences Meeting, p. 0686 (2017a)
35. Schmidt, O.T., Towne, A., Colonius, T., Cavalieri, A.V., Jordan, P., Brès, G.A.: Wavepackets and trapped acoustic modes in a turbulent jet: coherent structure eduction and global stability. J. Fluid Mech. **825**, 1153–1181 (2017)
36. Schmidt, O.T., Towne, A., Rigas, G., Colonius, T., Brès, G.A.: Spectral analysis of jet turbulence. J. Fluid Mech. **855**, 953–982 (2018)
37. Sirovich, L.: Turbulence and the dynamics of coherent structures. I. Coherent structures. Quart. Appl. Math. **45**(3), 561–571 (1987)
38. Stengel, R.F.: Stochastic Optimal Control: Theory and Application. Wiley, New York (1986)

39. Towne, A., Cavalieri, A.V., Jordan, P., Colonius, T., Schmidt, O., Jaunet, V., Brès, G.A.: Acoustic resonance in the potential core of subsonic jets. J. Fluid Mech. **825**, 1113–1152 (2017)
40. 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)
41. Tutkun, M., George, W.K.: Lumley decomposition of turbulent boundary layer at high reynolds numbers. Phys. Fluids **29**(2), 020707 (2017)
42. Tutkun, M., Johansson, P.B.V., George, W.K.: Three-component vectorial proper orthogonal decomposition of axisymmetric wake behind a disk. AIAA J. **46**(5), 1118–1134 (2008)
43. Van Trees, H.L.: Detection, Estimation, and Modulation Theory, Part I. Wiley, New York (1968)
44. Zare, A., Jovanović, M.R., Georgiou, T.G.: Colour of turbulence. J. Fluid Mech. **812**, 636–680 (2017)