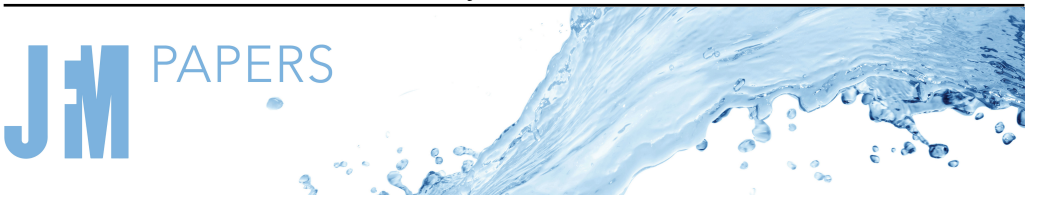
J. Fluid Mech. (2025), *vol.* 1020, A36, doi:10.1017/jfm.2025.10610



On deep-learning-based closures for algebraic surrogate models of turbulent flows

Benet Eiximeno[1,2], **Marcial Sanchis-Agudo**[3], **Arnau Miró**[1,2], **Ivette Rodriguez**[2], **Ricardo Vinuesa**[4] and **Oriol Lehmkuhl**[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain

[2]Universitat Politècnica de Catalunya, Terrassa, Spain

[3]FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, Sweden

[4]Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109, USA

Corresponding author: Benet Eiximeno, benet.eiximeno@bsc.es

(Received 5 December 2024; revised 16 July 2025; accepted 13 August 2025)

A deep-learning-based closure model to address energy loss in low-dimensional surrogate models based on proper-orthogonal-decomposition (POD) modes is introduced. Using a transformer-encoder block with an easy-attention mechanism, the model predicts the spatial probability density function of fluctuations not captured by the truncated POD modes. The methodology is demonstrated on the wake of the Windsor body at yaw angles of $\delta = [2.5°, 5°, 7.5°, 10°, 12.5°]$, with $\delta = 7.5°$ as a test case, and in a realistic urban environment at wind directions of $\delta = [-45°, -22.5°, 0°, 22.5°, 45°]$, with $\delta = 0°$ as a test case. Key coherent modes are identified by clustering them based on dominant frequency dynamics using Hotelling's T^2 on the spectral properties of temporal coefficients. These coherent modes account for nearly 60 % and 75 % of the total energy for the Windsor body and the urban environment, respectively. For each case, a common POD basis is created by concatenating coherent modes from training angles and orthonormalising the set without losing information. Transformers with different size on the attention layer, (64, 128 and 256), are trained to model the missing fluctuations in the Windsor body case. Larger attention sizes always improve predictions for the training set, but the transformer with an attention layer of size 256 slightly overshoots the fluctuation predictions in the Windsor body test set because they have lower intensity than in the training cases. A single transformer with an attention size of 256 is trained for the urban flow. In both cases, adding the predicted fluctuations close the energy gap between the reconstruction and the original flow field, improving predictions for energy, root-mean-square velocity fluctuations and instantaneous flow fields. For instance, in the Windsor

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

body case, the deepest architecture reduces the mean energy error from 37 % to 12 % and decreases the Kullback–Leibler divergence of velocity distributions from $\mathcal{D}_{\mathcal{KL}} = 0.2$ to below $\mathcal{D}_{\mathcal{KL}} = 0.026$.

Key words: machine learning, wakes

1. Introduction

Surrogate models are data-driven computational techniques used in various scientific and engineering fields to approximate complex systems or functions. These models serve as simpler substitutes for both experiments and computationally expensive simulations, thus providing quicker, yet sufficiently accurate results (Sun & Wang 2019). Surrogate models are mainly utilised to estimate the optimum-product solution or as instrumental tools to evaluate the performance in the initial stages of the vehicle development because they reduce the resource requirements for design exploration (Kuya *et al.* 2011; Yondo, Andrés & Valero 2018).

In the particular case of fluid dynamics applications, surrogates are typically built on a reduced space due to the complexity and high dimensionality of the original phenomenon (Yondo *et al.* 2018). The dimensionality reduction can be done either with algebraic methods, e.g. the proper-orthogonal decomposition (POD) (Lumley 1981), or employing deep-learning-based techniques. Proper-orthogonal decomposition was first introduced in fluid dynamics by Lumley (1981) to express the chaotic turbulent motions into modes representing some portion of the total fluctuating energy of the flow. Sirovich (1987) explored the relationship between POD and the dominant features of the flow, and showed that POD is a relevant tool for the study of vortex dynamics in all types of fluid flows. Recently, other modal decompositions have been introduced in order to obtain modes that are associated with a single frequency instead of the range of frequencies present in the time series of the temporal coefficients in POD. Among these new techniques, the most popular are dynamic-mode decomposition (DMD) (Schmid 2010) and spectral proper-orthogonal decomposition (SPOD) (Towne, Schmidt & Colonius 2018). Note that while POD and SPOD rank the modes in terms of their contribution to the reconstruction of the original flow, DMD obtains modes classified in terms of their dynamical importance to minimise errors in the reconstruction.

Alternatively, deep-learning methods for dimensionality reduction are based on unsupervised-learning methodologies such as autoencoders. There are application examples of several autoencoder architectures for dimensionality reduction in fluid dynamics, including vanilla (Eivazi *et al.* 2020), mode decomposing (Murata, Fukami & Fukagata 2020), hierarchical (Fukami, Nakamura & Fukagata 2020), physics-assimilated (Zhang 2023) and variational autoencoders (Akkari *et al.* 2022; Eivazi *et al.* 2022; Eiximeno *et al.* 2024*c*; Solera-Rico *et al.* 2024; Wang *et al.* 2024). All of them are able to capture the nonlinear behaviour of dynamical systems with a higher compression capacity than any POD-based methodology thanks to the excellent capabilities of spatial convolutions for nonlinear feature extraction (Brunton, Noack & Koumoutsakos 2020; Vinuesa & Brunton 2022).

It is particularly relevant to mention that β-variational autoencoders based on convolutional neural networks (CNN-βVAEs) have been used successfully to obtain a disentangled latent representation of turbulent fluid flows. For instance, Eivazi *et al.* (2022) compressed the turbulent flow around a simplified urban environment into five orthogonal latent variables containing more than 85 % of the flow energy. However, the

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

need of convolutional layers restricts the usage of this technique to geometries that can be represented on a regular grid. On the other hand, algebraic decompositions can be used on unstructured grids at the cost of losing a significant amount of the energy of the system. A good illustration of this is the aforementioned study from Eivazi *et al.* (2022), where five POD modes barely recover 30 % of the flow energy. Accurately capturing all the fluctuations in a turbulent flow would require selecting nearly all the modes of the system.

Couplet, Sagut & Basdevant (2003) proved that large-index POD modes drain energy from the more significant modes, yielding an energy-cascade structure. Such a modal-energy redistribution suggests that reduced-order models (ROMs) can be built on a small number of significant modes that represent the majority of flow features and the contribution of the rest of modes can be modelled as an additional term to the ROM. This conclusion has led to an intense research on closures for ROMs based on Galerkin and Petrov–Galerkin projections of the Navier–Stokes equations. These models constitute a fundamental pillar for the stability of the projection (Stabile & Rozza 2018; Kaptanoglu *et al.* 2021) and have been traditionally inspired by sub-grid scale models such as those used in large-eddy simulations (Wang *et al.* 2012; Hijazi *et al.* 2020; Imtiaz & Akhtar 2020). More recently, such closures have been modelled with data-driven techniques such as probabilistic neural networks (Maulik *et al.* 2020). A recent review and comparison of data-driven methods for ROM closures can be found in Prakash & Zhang (2024).

The main goal of this paper is to present a new data-driven model capable of recovering the energy loss due to modal truncation in POD. Instead of working in the reduced space as the aforementioned closures, this work is focused on learning the spatial probability density function (PDF) of the difference between the original field and the POD reconstruction using only the most significant modes with a transformer model (Vaswani *et al.* 2017). A transformer is a deep-neural-network architecture initially developed in the field of natural-language processing. Since then, it has revolutionised many areas of machine learning thanks to its attention mechanism, which enables identifying long-range dependencies in the data more effectively than traditional models (Yousif *et al.* 2023). A relevant requirement for the model is to be generalisable for flow conditions similar to those used in the training phase. Then, this closure will be a helpful technique to improve the accuracy of surrogate models. The rationale behind the approach proposed in this work is to reduce the dimensionality of the problem in order to build a cheaper surrogate model capable of predicting the most significant features of the flow, which are fully dependent on the geometry and initial conditions, and then use the present model as a separate correction for the smaller turbulent scales, which are lost during the model order reduction. It is important to note that the construction of the surrogate model for the prediction of the significant features of the flow is beyond the scope of this work, as it is fully focused on how to recover the energy lost after applying a ROM. Hence, all the test cases will be generated with the projection and truncation of the ground truth data. When used in practical engineering applications the large scales should be predicted using any other cheaper method as the parameterised DMD (Andreuzzi, Demo & Rozza 2023) or shallow recurrent decoder networks for ROM (SHRED-ROM) (Tomasetto *et al.* 2025). Such process ensures that the only source of error in the results is linked to the capability of the closure model to bridge the energy gap and not from the surrogate-modelling prediction.

The methodology is tested on the turbulent wake of the flow past the Windsor body (Littlewood & Passmore 2010), which is a simplified square-back vehicle, and on the flow in a realistic urban environment as the Zona Universitària neighbourhood, located in Barcelona. In particular, we have focused on the flow at pedestrian level around the headquarters of the Barcelona Supercomputing Center (BSC), where several authors of this paper are affiliated. Both datasets have been obtained by means of wall-modelled

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

large-eddy simulations (WMLES) under five different free-stream-velocity directions and are described in §§ 3.1 and 3.2, respectively. The final objective is showing how the proposed closure can recover the truncated fluctuations from the POD common basis. In other words, after adding the closure term the root-mean-square (r.m.s.) values of velocity fluctuations should be equivalent to those from the original simulations.

In the case of the Windsor body, the closure is trained to be valid for any free-stream-velocity direction in a yaw-angle range $2.5° \leqslant \delta \leqslant 12.5°$. This test case is highly relevant for the automotive industry because in any road vehicle the drag force increases linearly for yaw angles in the range of $0° \leqslant \delta \leqslant 15°$ (Howell 2015). This drag increase is completely independent of the zero-yaw drag, thereby making it impossible to extrapolate the performance in cross-flow conditions from the parallel-flow case (Howell 2015). Hence, car manufacturers need to evaluate the aerodynamic performance under yawed flows in the development loop of a new vehicle (D'Hooge *et al.* 2014). The development could be massively accelerated by using a surrogate model instead of re-running the simulations and wind-tunnel tests that are needed to characterise the aerodynamic performance of a road vehicle (Zhang, Toet & Zerihan 2006) at every angle of interest, and the closure presented in this work would play a key role in the accuracy of the turbulent kinetic energy (TKE) of the flow.

In the BSC building case, the dataset comprises high-fidelity simulations of the neutral atmospheric boundary layer (ABL) flow over a neighbourhood in Barcelona, with incoming wind directions ranging from $-45°$ to $+45°$, $0°$ being the wind coming from the south. Here, the variation in the wind direction aims to capture the influence of urban geometry on flow patterns such as channelling, stagnation and shear layers. Indeed, wind velocity plays a critical role in shaping the vertical and horizontal structure of the flow within the urban canopy layer, which is essential for understanding wind loading on buildings and urban ventilation dynamics.

The model performance is compared with the energy recovery given by a super-resolution generative adversarial neural network (SRGAN) (Ledig *et al.* 2017) trained to predict the original flow field when given the truncated POD reconstruction. Note that similar approaches have been used in numerous studies in recent years (Kim *et al.* 2021; Fukami *et al.* 2019, 2021). It is important to note that this methodology is based on convolutional layers. Hence, it is necessary to represent the flow field on a structured grid, and the method cannot handle any complex geometry that needs to be represented using an unstructured mesh. Thus, this comparison is only done on the Windsor body case as the flow around the BSC building includes the BSC geometry inside the domain.

The rest of this paper is organised as follows: § 2 describes how the closure is formulated, how the significant POD modes are selected and how the model is extended to multiple flow conditions; then, § 3 describes the datasets in which the methodology is tested, § 4 shows the accuracy of the closure in the wake behind the Windsor body and in the flow around the BSC building; and finally, § 5 summarises the main findings of the paper.

2. Methodology

This section describes the methods used in this paper, including a mathematical definition of POD, how the significant modes from which a surrogate model would be built are selected, and finally, the explanation of the model used to add the energy from the truncated modes.

2.1. *Proper-orthogonal decomposition (POD)*

Proper-orthogonal decomposition is used in this work as a dimensionality-reduction technique. It is an efficient way to capture an infinite-dimensional process with a reduced

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

number of modes (Holmes *et al.* 1997). This method is based on finding a set of deterministic functions that characterise the dominant features of the system given by the field $\boldsymbol{F}(\boldsymbol{X}, t)$. This decomposition can be written as

$$\boldsymbol{F}(\boldsymbol{X}, t) = \sum_{i=1}^{i=N} a_i(t)\boldsymbol{\Phi}_i(\boldsymbol{X}), \tag{2.1}$$

where N is the number of functions to decompose the field into. Proper-orthogonal decomposition requires that the basis for the spatial modes is orthonormal, i.e.

$$\int_{\boldsymbol{X}} \Phi_{i_1}(\boldsymbol{X})\Phi_{i_2}(\boldsymbol{X})\,\mathrm{d}x = \begin{cases} 1 \text{ if } i_1 = i_2, \\ 0 \text{ otherwise}, \end{cases} \tag{2.2}$$

and optimal, so that the the first N_r vectors are those that reconstruct the database with the minimum possible error.

In this work the chosen method to perform POD is the singular-value decomposition (SVD). The SVD decomposes the initial snapshot matrix, $\boldsymbol{\mathcal{X}}$, into the left singular vectors, $\boldsymbol{\Psi}$, the singular values, $\boldsymbol{S}$, and the right singular vectors, $\boldsymbol{V}$:

$$\mathcal{X} = \boldsymbol{\Psi}\boldsymbol{S}\boldsymbol{V}^T. \tag{2.3}$$

Each column of $\boldsymbol{\Psi}$ contains a spatial mode, $\boldsymbol{\Phi}_i(\boldsymbol{X})$ and each column of $\boldsymbol{V}$ gives the evolution of the time coefficient, $a_i(t)$, of the corresponding mode. The singular values are given in a diagonal matrix and are associated with the energy contribution of each mode in descending order. The higher the singular value, the more energy is contained in the mode. The POD analysis has been performed using pyLOM (Eiximeno *et al.* 2024*a*), a high-performance-computing reduced-order-modelling code that has a parallel and scalable algorithm for the SVD (Eiximeno *et al.* 2024*b*).

2.2. *On the significance of POD modes*

Turbulent flows are characterised by a flat tail of singular values, making it difficult to set an energy threshold to select the modes onto which the data has to be projected. This threshold is set arbitrarily and is decided based on a trade-off between accuracy of the model and evaluation cost of the future ROM. To overcome this issue, in this work the selection of the relevant modes is based on their frequency content. The objective is to select only the modes that contain relevant information on the frequency of the coherent structures of the flow. This step is necessary to ensure that any methodology used to predict a new flow field condition in the reduced space will take into account all the large scales of the case without the additional cost and inaccuracy given by non-coherent modes. Their contribution will be modelled by the probabilistic closure proposed in the present work.

The coherent modes are seen as the outlier modes in the power-spectral density (PSD) matrix of the temporal coefficients, $\boldsymbol{V}$. In other words, the selected modes are those that exhibit a frequency spectrum significantly different from the rest. The PSD matrix of $\boldsymbol{V}$ is computed by performing the Lomb–Scargle periodogram to the temporal coefficient of each mode of the system. Then, the outlier modes are identified with principal-component analysis (PCA). Principal-component analysis is analogous to POD once the data has been normalised with its variance and centred to its mean. Since the PCA model may contain numerous components, its information is summarised using Hotelling's T^2:

$$T^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a}\right)^2. \tag{2.4}$$

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

Here $t_{i,a}$ is the projection of the PSD of mode i into the PCA component a and s_a is the covariance of that component. Note that T^2 can be seen as the distance from the centre of the hyperplane formed by the components to the projection of the observation onto the hyperplane. The larger the T^2 value is, the more relevant frequency content the mode will have. Hence, the modes now can be selected with a T^2 threshold value that contains all the outliers. This threshold will be set after a qualitative analysis of the different clusters seen on each of the training angles for the two studied flows. The empirical value given to the threshold should be consistent at least for all flow conditions of the same case and ideally should be valid for the two configurations studied in the present work.

2.3. *The POD projection and reconstruction*

The POD basis for data projection is built using the spatial correlations of the N_r modes corresponding to the frequency outliers. When working with n different inlet conditions, one can find an optimal POD basis among them by concatenating the spatial correlations of the outlier modes from each case to create the following matrix $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \left[\boldsymbol{\Psi}_0\ \boldsymbol{\Psi}_1\ \ldots\ \boldsymbol{\Psi}_n\right]. \tag{2.5}$$

Then, POD is applied to matrix $\boldsymbol{Y}$ to find an orthonormal basis that contains the information of the selected modes for each of the inlet conditions:

$$\boldsymbol{Y} = \boldsymbol{\Psi}_Y \boldsymbol{S}_Y \boldsymbol{V}_Y^T. \tag{2.6}$$

The resulting basis can be truncated as long as there are no information losses, i.e. the selected modes are able to recover more than 99 % of the energy. The usage of common POD basis for dimensionality reduction of parameterised turbulent flows on complex geometries has already been used in other cases such as the flow around wind turbines (Céspedes Moreno *et al.* 2025).

The data matrix $\mathcal{X}$ can be projected now onto $\boldsymbol{U_Y}$ as

$$\hat{\mathcal{X}} = \boldsymbol{\Psi}_Y^T \cdot \mathcal{X}, \tag{2.7}$$

with the assurance that all coherent modes inside the inlet-conditions range are included in the ROM. This operation reduces the dimensionality of the numerical data and sets a latent space for any surrogate-modelling applications. Such a surrogate model can be used to perform temporal predictions of the system or to evaluate its response to any condition in the evaluated range.

When a prediction, $\hat{\mathcal{X}}_{\mathcal{P}}$, is reprojected back into the full-order space, i.e.

$$\mathcal{X}_{\mathcal{P}} = \boldsymbol{\Psi}_Y \cdot \hat{\mathcal{X}}_{\mathcal{P}}, \tag{2.8}$$

the main behaviour of the system is captured; however, the model lacks the energy from the modes that were discarded during its construction.

As stated in the introduction, the aim of this work is to model the error between the original data and the reconstruction from the truncated POD modes:

$$\mathcal{E} = \mathcal{X} - \mathcal{X}_P. \tag{2.9}$$

This error is responsible for the missing energy in the predicted flow field. Training a surrogate model to interpolate flow fields at different directions of the incident velocity is beyond the scope of this paper, hence, all the predictions in the reduced space, $\mathcal{X}_P$, will be obtained through the projection of the ground truth data to the common basis (2.7), regardless of whether the data was used when constructing the basis. This approach ensures that the only source of error in the results is linked to the capability of the

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

closure model to bridge the energy gap and not from the surrogate-modelling prediction. Such interpolation could be performed using parameterised DMD (Andreuzzi *et al.* 2023) or SHRED-ROM (Tomasetto *et al.* 2025). Both methodologies have been used successfully to predict POD coefficients in the prediction of different flow conditions on the bidimensional laminar flow around a circular cylinder; however, the extension to turbulent flows might deserve further studies.

2.4. *Closure model*

To build a closure for the missing scales in the POD projection and reconstruction process, it is essential to understand the spatial and temporal distribution of this error. In other words, it is necessary to determine where and when this error is more likely to occur. The strategy followed in this work involves learning the evolution of the error as a function of the recovered fluctuations, $\mathcal{X}_P$, since this field contains all relevant information about the system's state at all points in the domain for the studied time step. To achieve this, a transformer (Vaswani *et al.* 2017) encoder block is trained to minimise the difference between the actual error field, $\mathcal{E}$, and the predicted one, using the temporal series of $\mathcal{X}_P$ across all points in the domain. The training process employs a mean-squared-error loss function. Thus, if $\mathcal{X}_P$ is known for a given time step, the transformer can predict the corresponding error field, $\mathcal{E}$. From now on, the error predicted by the transformer is represented as $\mathcal{E}_T$ and the error of the model after considering the closure is defined as

$$\mathcal{E}_{\mathcal{M}} = X - (\mathcal{X}_{\mathcal{P}} + \mathcal{E}_{\mathcal{T}}). \tag{2.10}$$

The choice of using a transformer-based model is motivated by their ability to identify and predict the temporal dynamics of chaotic systems by capturing long-term dependencies in the data (Geneva & Zabaras 2022; Wu *et al.* 2022; Sanchis-Agudo *et al.* 2023). Additionally, transformers are well suited for forecasting time series based on other spatial variables (Wang 2023) through their variant known as visual transformers. Transformers can be seen as universal approximators to PDFs (Furuya *et al.* 2024). Hence, the proposed model actually learns the joint PDF of $\mathcal{E}$ given $\mathcal{X}_P$, $p(\mathcal{E} \mid \mathcal{X}_P)$. Furthermore, there exists an attention-only, transformer T with attention normalisation N such that, for any auto-regressive sequence, $(x_t)_{t\geq 1}$ converges exponentially fast as n goes to infinity, where n is the number of attention layers. Denoting

$$\mathcal{E}(x_{1:t}) := \lim_{n\to+\infty} \mathcal{E}_n(x_{1:t}), \tag{2.11}$$

one has

$$\lim_{t\to+\infty} (\mathcal{E}(x_{1:t}) - x_{t+1}) = 0. \tag{2.12}$$

For a more detailed study of the transformer's universality and the analytic intrinsics when approximating the theoretical measure, we refer the reader to Geshkovski, Rigollet & Ruiz-Balet (2024); Sander & Peyré (2024).

The latter definition ensures that the system modelled by the transformer is statistically equivalent to the original one and that the closure will be generalisable as long as the joint PDF $p(\mathcal{E} \mid \mathcal{X}_P)$ for a new set of data is similar to the original one. Such similarity is quantified using the Kullback–Leibler (KL) divergence, $\mathcal{D}_{\mathcal{KL}}$, i.e.

$$\mathcal{D}_{\mathcal{KL}}(P_i||Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P_i(x)}{Q(x)}\right) \mathrm{d}x, \tag{2.13}$$

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

Parameter	Windsor 1	Windsor 2	Windsor 3	BSC
Input dimension	48	48	48	48
Output dimension	1	1	1	1
Time projection	128	128	128	128
Attention heads	8	8	8	8
Attention size	64	128	256	256
Feed-forward layer	128	128	128	128
Activation function	tanh	tanh	tanh	tanh
Convolution layer	64	128	256	256
Fully connected layer	98 304	98 304	98 304	73 133
Number of parameters	19 141 505	38 057 345	75 938 177	57 061 294
Size of the model (Mb)	74	146	290	218

Table 1. Summary of the four architectures considered in the present work.

where $P_i(x)$ represents the joint PDF of the error for a single snapshot, while $Q(x)$ is the joint PDF for all snapshots included in the training.

In this study the input signal has a time-delay dimension of 48 steps, which means that the input to the transformer is a sequence of 48 consecutive time steps of the POD reconstruction. This choice is constrained by the number snapshots available per yaw angle (650 for each Windsor body case and around 1000 for the flow around the BSC). A further increase would reduce the amount of training samples to an extremely limited number. On the other hand, a smaller embedding size to increase the number of training samples would not give enough information on the time history to the transformer. The output is the error field between the POD reconstruction and the original data (2.9) of the first time instant of the input series (i.e. the oldest one). A time–space embedding module is added to each point time signal to incorporate temporal and spatial information before passing it to the transformer blocks, allowing the model to distinguish between the evolution of the velocity in different points at different time steps. An average pooling and a max-pooling layer are added to the time–space embedding. Both of them are one dimensional and have a stride of two steps.

A summary of the different transformer architectures trained in this work can be seen in table 1. All of them are based on a single transformer-encoder block (figure 1) with eight attention heads followed by a feed-forward layer. The Windsor body case is used to assess the effect of the transformer depth and its number of parameters on the closure accuracy. To do so, three different architectures with an increasing attention size are tested. The shallowest architecture has 64 attention layers, which are doubled to 128 for the second architecture and doubled again to 256 for the deepest architecture. Only the deepest transformer is trained for the closure model of the BSC building. The attention layers measure the importance of different parts of the input sequence when making predictions (Bahdanau, Cho & Bengio 2014). Note that, the dimension of the feed-forward layer is set to 128 in all three cases. This layer learns complex nonlinear relationships between the input and output sequences.

The choice of using multi-head attention is based on its outstanding performance over scaled dot product attention as it allows the model to jointly attend to information from different representation subspaces at different positions (Vaswani *et al.* 2017). In particular, the current architecture employs the easy-attention mechanism (Sanchis-Agudo *et al.* 2023), which has demonstrated promising performance in predicting the temporal dynamics of chaotic systems, significantly outperforming the self-attention transformer

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

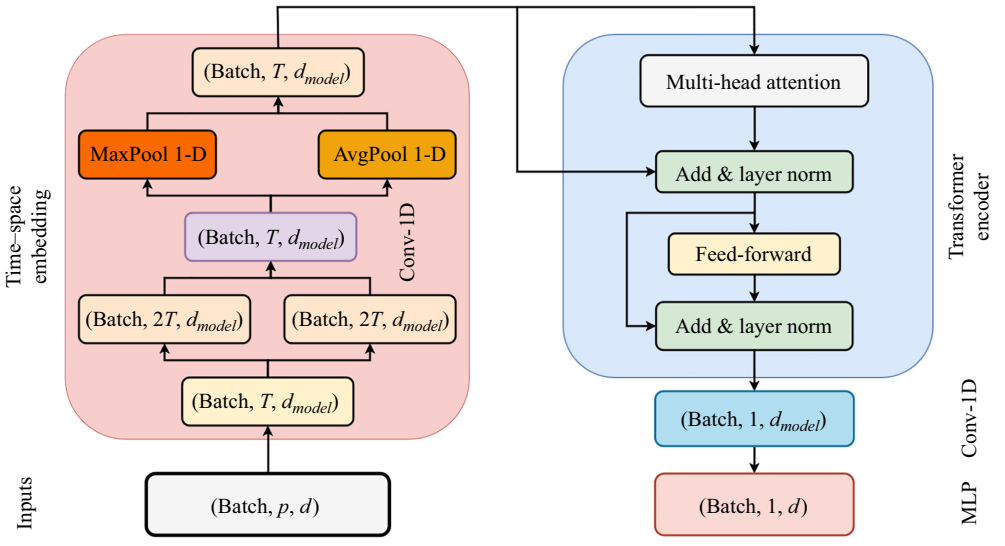


Figure 1. Easy-attention-based transformer with time–space embedding. Figure adapted from Sanchis-Agudo *et al.* (2023).

(Solera-Rico *et al.* 2024). The easy-attention mechanism originally presented by Sanchis-Agudo *et al.* (2023) is defined by the mapping $\mathbb{R}^{\mathrm{d}_T \times \mathrm{d}_S} \to \mathbb{R}^{\mathrm{d}_T \times \mathrm{d}_S}$, given by the equation $\boldsymbol{M} \to \hat{\boldsymbol{M}} = \alpha \boldsymbol{M} \boldsymbol{W}_V$, where the pseudo-input, the input after embedding and the output matrices have the same dimensions as the attention size (table 1). In this formulation, $\alpha \in \mathbb{R}^{\mathrm{d}_T \times d_T}$ and $\boldsymbol{W}_V \in \mathbb{R}^{\mathrm{d}_S \times \mathrm{d}_S}$ are matrices of trainable parameters, with d_T representing the temporal feature dimension and d_S the spatial feature dimension. Following the standard notation used in transformer architectures, $\boldsymbol{M} \cdot \boldsymbol{W}_V$ denotes the values, while the matrix α represents the attention weights. This mechanism, expressed as a kernel operation, can be formulated as

$$\hat{\boldsymbol{M}}(t, s) = \int_T \int_S \alpha(t, t') \boldsymbol{W}_V(s, s') \boldsymbol{M}(t', s') \, \mathrm{d}t' \mathrm{d}s'. \tag{2.14}$$

To extend the easy-attention mechanism to the multi-head attention strategy, we consider multiple attention heads so that each of them focus on different parts of the input space. In this case, the input $\boldsymbol{M}$ is projected into multiple subspaces, allowing the model to attend to different sources of information simultaneously. For each attention head, we perform the same kernel operation as defined in the original mechanism, but with distinct sets of trainable parameters for the attention weights and value projections.

The multi-head version of the kernel operation can be written as

$$\hat{\boldsymbol{M}}^{(h)}(t, s) = \int_T \int_S \alpha^{(h)}(t, t') \boldsymbol{W}_V{}^{(h)}(s, s') \boldsymbol{M}(t', s') \, \mathrm{d}t' \mathrm{d}s', \tag{2.15}$$

where $h \in \{1, 2, \ldots, H\}$ denotes the index of the attention head, H is the number of attention heads, and each $\alpha^{(h)}$ and $\boldsymbol{W}_V{}^{(h)}$ are distinct trainable parameters for the hth attention head. The final output of the multi-head attention is obtained by concatenating the outputs of each attention head:

$$\hat{\boldsymbol{M}} = \mathrm{Concat}\big(\hat{\boldsymbol{M}}^{(1)}, \hat{\boldsymbol{M}}^{(2)}, \ldots, \hat{\boldsymbol{M}}^{(H)}\big). \tag{2.16}$$

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

After the transformer block, a one-dimensional convolutional network of the same size as the attention layer and a fully connected layer of the size of the number of points of the spatial domain (98 304 for the Windsor body case and 73 133 for the flow around the BSC building) are added to decode the transformer output and form the final spatial prediction of the POD reconstruction error. The training of each architecture was conducted along 3500 epochs. An extensive discussion on the accuracy of each architecture is presented next in the results section.

Building a closure for the truncated scales during dimensionality reduction can be seen as augmenting the resolution of the flow field. Thus, the results from the closure model presented in the current work are compared with those given by a SRGAN (Ledig *et al.* 2017), a model used for resolution augmentation in turbulent flows (Fukami, Fukagata & Taira 2023). The model is designed to generate a realistic original flow field given the POD reconstruction from the current instant. The architecture used is an adaptation from the one used by Güemes *et al.* (2021) to reconstruct turbulent-flow quantities from coarse wall measurements of wall-shear stress and wall pressure. In that case, the architecture is taken from the original paper introducing the SRGAN method (Ledig *et al.* 2017). The main changes in the present case are introduced to the generator. First of all, the number of residual blocks is reduced from 16 to 8. Then, the final sub-pix-convolution is eliminated because the output field has the same size as the input field.

This methodology is based on convolutional layers. Hence, it is necessary to represent the flow field on a structured grid. The input data has to be reshaped into an image-like snapshot. Thus, the 98 304 spatial points of the Windsor body slice are converted into an image with 384 points in the streamwise direction and 256 points in the cross-stream direction. For this case, the training is extended over 500 epochs using a batch size of 32 snapshots. The method is not tested for the flow around the BSC building as its geometry is embedded in the domain and cannot be represented accurately by a structured grid.

3. Datasets description

This section describes the two different flow datasets in which the methodology from the present work has been tested. These datasets are the WMLES of the flow around the Windsor body (§ 3.1) and the building of the BSC headquarters (§ 3.2) under five different directions of the free-stream velocity.

For the simulations, the spatially filtered incompressible Navier–Stokes equations,

$$\frac{\partial \overline{u}_i}{\partial x_i} = 0, \tag{3.1}$$

$$\frac{\partial \overline{u}_i}{\partial t} + \frac{\partial \overline{u}_i \overline{u}_j}{\partial x_j} - \nu \frac{\partial^2 \overline{u}_i}{\partial x_j \partial x_j} + \rho^{-1} \frac{\partial \overline{p}}{\partial x_i} = -\frac{\partial \mathcal{T}_{ij}}{\partial x_j}, \tag{3.2}$$

were numerically integrated using SOD2D (spectral high-order code 2 solve partial differential equations) (Gasparino *et al.* 2024*b*), a low-dissipation spectral-element-method code (Gasparino *et al.* 2024*a*). In (3.1) and (3.2), x_i are the spatial coordinates (or x, y and z), u_i (or u, v and w) stands for the velocity components and p is the pressure. Note that ρ is the density of the fluid. The filtered variables are represented by $(\overline{\cdot})$. The right-hand-side term in (3.2) represents the sub-grid stresses, and its anisotropic part is expressed as

$$\mathcal{T}_{ij} - \frac{1}{3}\mathcal{T}_{kk}\delta_{ij} = -2\nu_{sgs}\overline{\mathcal{S}}_{ij}, \tag{3.3}$$

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

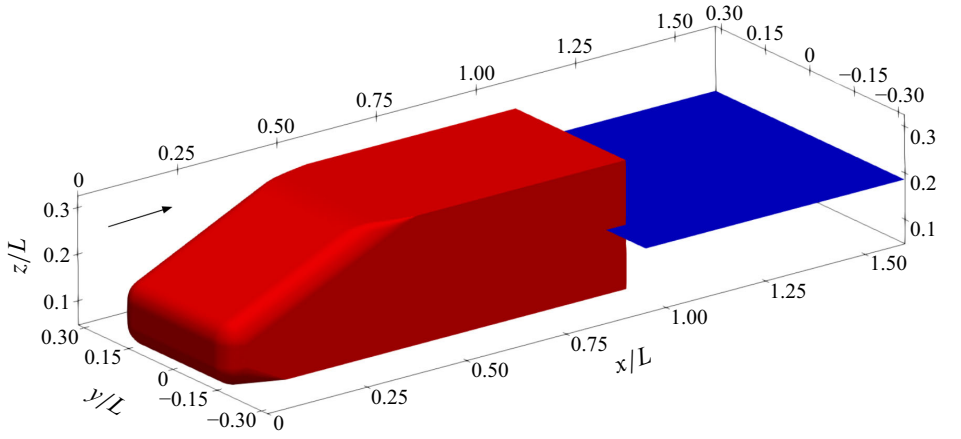


Figure 2. The geometry of the Windsor body (red) and working plane (blue) where the data are interpolated to develop the model. The plane is perpendicular to the vertical axis and is located at $z/L = 0.186$. The arrow indicates the flow direction.

where the large-scale rate-of-strain tensor $\overline{\mathcal{S}}_{ij}$ is evaluated as $\overline{\mathcal{S}}_{ij} = (1/2)(g_{ij} + g_{ji})$, with $g_{ij} = \partial \overline{u}_i / \partial x_j$ and δ_{ij} being the Kronecker delta. In the case of the Windsor body, the unresolved scales are modelled using the local formulation of the integral length-scale approximation (Lehmkuhl, Piomelli & Houzeaux 2019) and the Vreman model (Vreman 2004) is used in the simulations of the flow in the neighbourhood where the BSC is located. The near-wall region was modelled using the Reichardt wall law (Reichardt 1951) with an exchange location at the fifth node (Lehmkuhl *et al.* 2018).

3.1. *Windsor body*

The test dataset is the turbulent wake behind the Windsor body, the simplified square-back vehicle depicted in figure 2, at a Reynolds number of $Re_L = U_\infty L/\nu = 2.9 \times 10^6$, where U_∞ is the magnitude of the free-stream velocity, L is the length of the model and ν is the kinematic viscosity of the fluid. Its relative dimensions are similar to those of a sport utilitary vehicle (SUV): its width is 0.373 L and its height is 0.325 L. The data was generated by means of WMLES at yaw angles of $\delta = [2.5°, 5°, 7.5°, 10°$ and $12.5°]$.

After the initial transients had been washed out, all simulations were run for 60 additional convective time units, $t = 60L/U_\infty$, to collect 660 snapshots. The data for the model assessment was interpolated into the plane represented in figure 2. This plane is perpendicular to the vertical axis, therefore, it contains the dynamics of both the leeward and windward sides of the wake. It is located at $z/L = 0.186$, which is half of the vehicle height when measured from the bottom of the body.

In the present work only a brief comparison of the fluid flow at the different yaw angles is shown to illustrate the different conditions in which the closure needs to be valid. For more details on the numerical model, grid and simulations accuracy, we refer the reader to the previous work by Eiximeno *et al.* (2024*c*) on the development of a surrogate model for the base pressure of the Windsor body. The data in the assessed plane can be downloaded from the computational fluid dynamics (CFD) simulations section of the AC-1-12 case of the ERCOFTAC Knowledge Base Wiki (Eiximeno, Lehmkuhl & Rodriguez 2025), where additional details on the case are provided.

In terms of the averaged flow, the wake of square-back bluff bodies in a yawed free-stream flow is dominated by two vortices: one on the leeward side ($y/L > 0$) and one on the windward side ($y/L < 0$), as shown by Booysen, Das & Ghaemi (2022). In figure 3

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

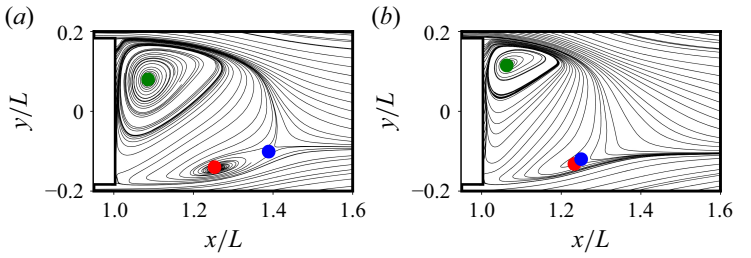


Figure 3. Time-averaged streamlines comparison between $\delta = 2.5°$ (*a*) and $\delta = 12.5°$ (*b*) at the plane $z/L = 0.186$. The green, red and blue dots represent the core of the leeward side vortex, the core of the windward side vortex and the saddle point, respectively.

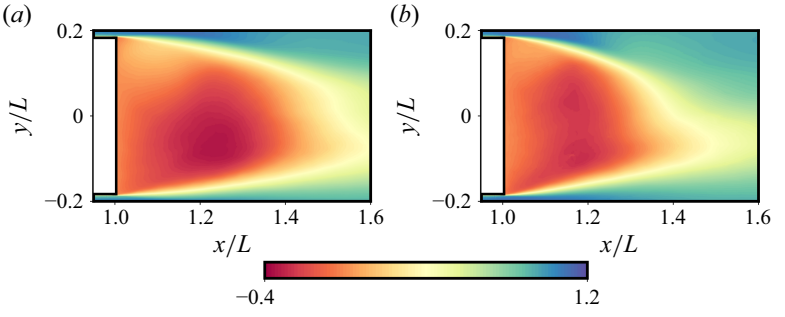


Figure 4. Mean streamwise velocity at $z/L = 0.186$ for $\delta = 2.5°$ (*a*) and $\delta = 12.5°$ (*b*).

the flow streamlines are plotted for $\delta = 2.5°$ and $\delta = 12.5°$. As reported by Booysen *et al.* (2022), the vortex on the leeward side dominates the recirculation and gains intensity over the windward vortex as the yaw angle increases. This effect moves the vortex centres and the saddle point to the leeward side of the vehicle and closer to the body.

The changes in the vortex intensity have an effect on the recirculation length. Lorite-Díez *et al.* (2020) identified, in a square-back Ahmed body, that this vortex interaction leads to a decrease of the recirculation length and to a deflection of the recirculation bubble towards the leeward side of the wake. Similar to the square-back Ahmed body, in the Windsor body, this trend is also observed. This can be seen in figure 4, where the mean streamwise velocity $\overline{u}$ for the cases at $\delta = 2.5°$ and $\delta = 12.5°$ is plotted. Here the recirculation length varies from 0.41 L to 0.28 L.

In figure 5 changes in the velocity fluctuations brought about with the yaw angle are illustrated by comparing the r.m.s of the streamwise velocity fluctuations, u_{rms}, at both $\delta = 2.5°$ and $\delta = 12.5°$. Figure 5 shows that a larger yaw angle increases the entrainment of the irrotational free stream into the near wake, resulting in a larger fluctuation intensity and a steeper shear layer angle on both sides of the vehicle. The latter leads to a narrower wake. This is in agreement with the findings from Li *et al.* (2019) on a square-back Ahmed body.

These changes in the mean flow with the yaw angle can also be observed when the mean streamwise velocity and its fluctuations are plotted along a streamwise line at $y/L = 0$ and over a cross-stream line at $x/L = 1.3$, respectively (see figure 6). For the objectives of the current work, it is relevant to remark that neither the fluctuations maxima nor their positions in the domain have a linear evolution with the yaw angle. Thus, it is not possible to formulate a linear model to predict them.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

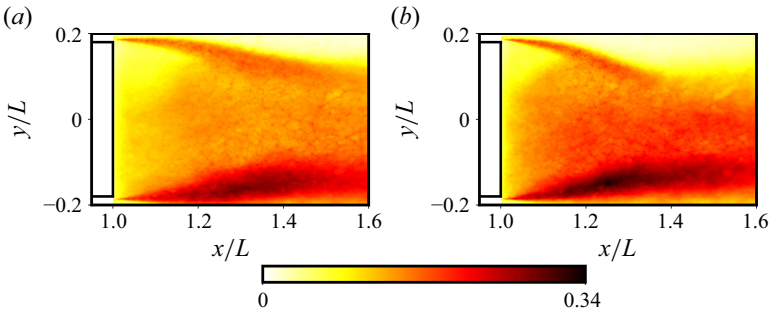


Figure 5. The r.m.s of the streamwise velocity fluctuations at $z/L = 0.186$ for $\delta = 2.5°$ (a) and $\delta = 12.5°$ (b).

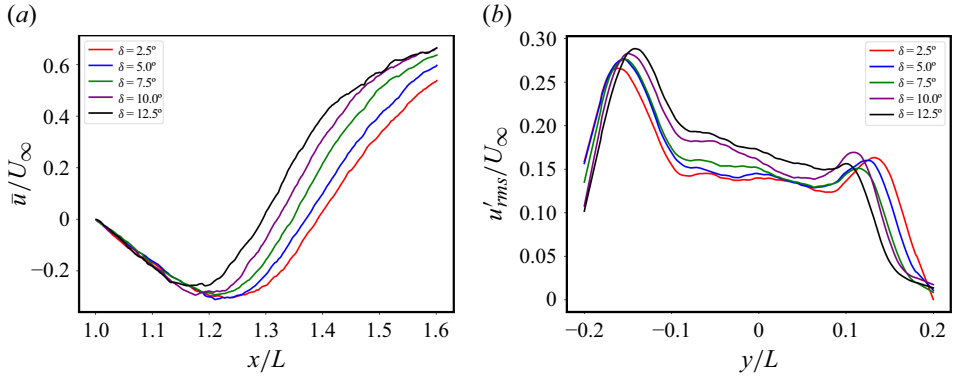


Figure 6. Mean streamwise velocity for all simulated angles at $y/L = 0$ (a) and the r.m.s. value of the velocity fluctuations at $x/L = 1.3$ (b).

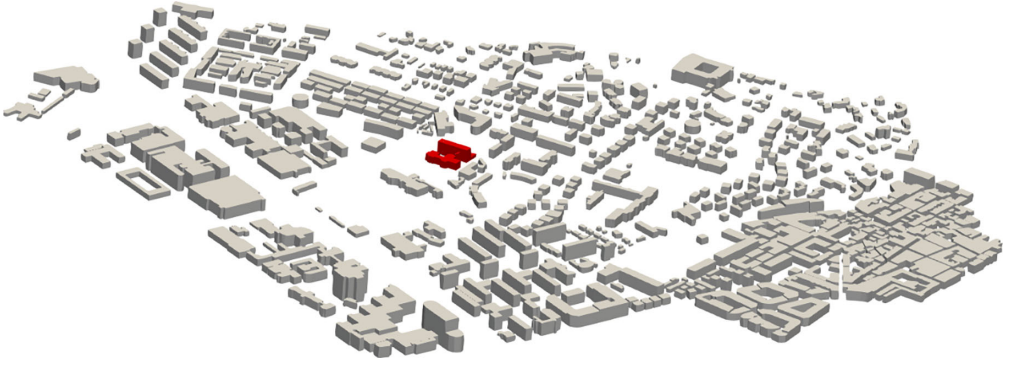


Figure 7. Geometry of the Zona Universitària neighbourhood in Barcelona. The building highlighted in red is the BSC headquarters.

3.2. *Headquarters of the BSC*

The simulation models the Zona Universitària neighbourhood of Barcelona (figure 7), covering 1700 m in length and 1900 m in width. The computational domain is configured such that the urban area is located $20H_{max}$ upstream of the inlet, $30H_{max}$ from the lateral boundaries and $40H_{max}$ from the outlet, where $H_{max} = 67$ m corresponds to the height of the tallest building. The vertical extent of the domain is set to $20H_{max}$.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

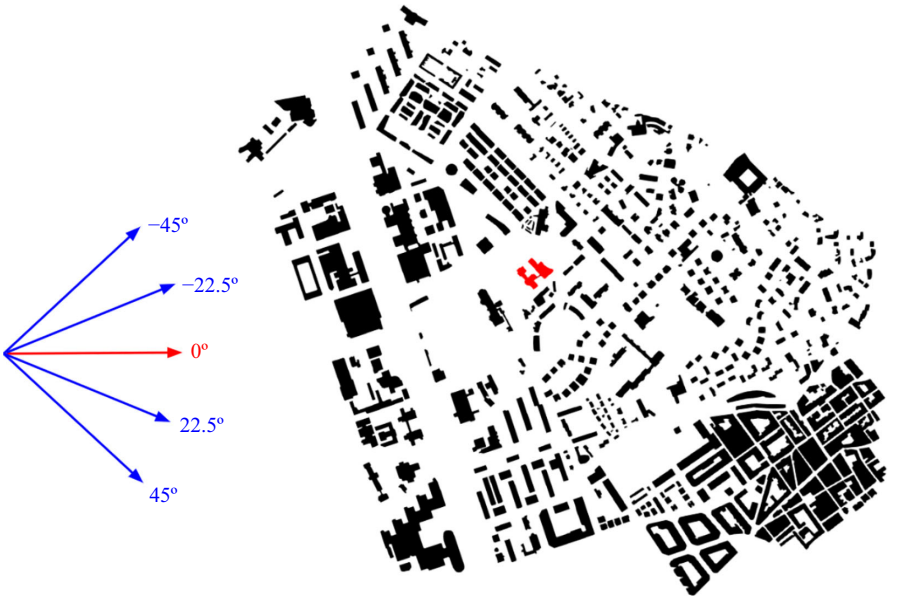


Figure 8. Simulated wind directions. Directions in blue are those used for training and the direction in red was used for the test. The red building is the BSC headquarters.

To impose a realistic turbulent inflow representative of a neutral ABL, an online periodic precursor simulation is coupled upstream of the main domain. This precursor enables the generation of a fully developed, statistically stationary turbulent flow at the inlet. The ABL is driven by a constant pressure gradient in the x direction, resulting in a logarithmic mean velocity profile characterised by a friction velocity $u^* \approx 0.597\,\mathrm{m\,s^{-1}}$ and an aerodynamic roughness length $z_0 = 1.53\,\mathrm{m}$. The reference velocity is set to $U_{ref} = 6.1\,\mathrm{m\,s^{-1}}$ at a height of $H_{ref} = 100$ m. The rest of the boundaries are set as follows. The top and lateral boundaries are treated as free-slip walls, truncating the boundary layer while enforcing zero normal gradients for all flow variables. A convective boundary condition with zero static pressure is applied at the outlet, while no-slip conditions are imposed on the ground and all surfaces of the urban geometry.

In order to resolve the complex turbulent motions around the urban geometry, an unstructured mesh composed of fourth-order hexahedral elements is used. The computational mesh is designed to achieve high spatial accuracy while efficiently capturing the broad range of turbulent scales, particularly within the urban canopy and near the ground. Several nested refinement regions are implemented, reaching a ground-level resolution of 0.75 m and leading to a final mesh consisting of approximately 520 million grid points.

The flow is computed for the five different wind directions represented in figure 8. All directions are equispaced 22.5° between themselves. The wind blowing from south to north is taken as the reference direction, i.e. the case at 0°. In the other four cases, the incident velocity is rotated 22.5° and 45° towards both the east and west directions.

Once the simulations were performed, the data around the headquarters of the BSC (figure 9), where several authors of this paper are affiliated, has been extracted. In particular, the work is focused on training the closure model for the flow at pedestrian level (1.50 m).

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

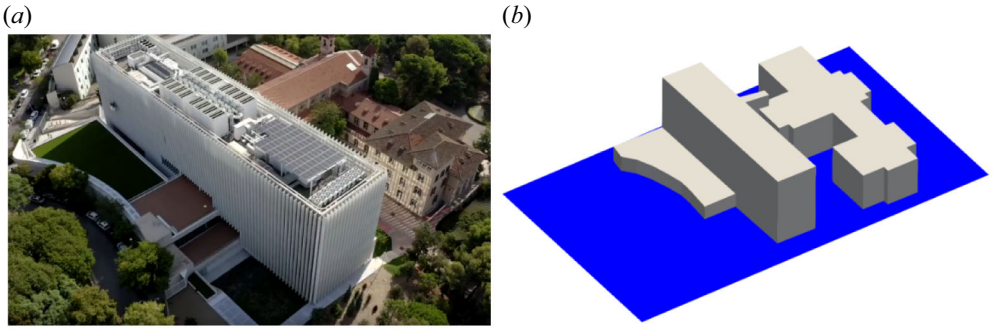


Figure 9. Picture of the BSC headquarters (*a*) and the simulated geometry (*b*). The simulated geometry includes the plane at $z = 1.50$ m, depicted in blue, in which the closure model has been tested.

4. Results

This section presents the performance of the probabilistic closure model for the POD reconstruction of the turbulent flow in the wake behind the Windsor body and around the building of the BSC headquarters. Both datasets are described in § 3. In each dataset, four out of the five flow conditions are used to train the models and the wind angle in between is used to test the model on an unseen condition. Then, the common basis for the Windsor body is built using the data at $\delta = [2.5°, 5°, 10°$ and $12.5°]$ and the basis for the BSC building is computed for $\delta = [-45°, -22.5°, 22.5°$ and $45°]$. All the snapshots collected at these flow conditions are projected through the common basis of their case and reconstructed to generate the training dataset for each closure model.

Once a transformer for each case is trained, the high-fidelity results at the test conditions are projected into their corresponding POD basis and reconstructed with the additional closure term from the transformer to assess its performance on unseen data. In the case of the Windsor body, the results are compared with those obtained by enhancing the POD reconstructions using a SRGAN. The training of the SRGAN is done following the same splitting of the flow conditions between training and the test.

As the physics of both flows is dominated by the streamwise component of the flow, all results are obtained for the streamwise velocity fluctuations, therefore, $\mathcal{X}$ in (2.3) is equivalent to u'.

4.1. *The POD common basis*

The first step to build the common basis is to perform the POD of each of the training angles individually. After that, the PSD-based mode-selection process described in § 2 is applied. Figure 10 shows the T^2 clustering results for one of the training angles of the Windsor body ($\delta = 2.5°$) and one of the training angles of the flow around the BSC headquarters ($\delta = 22.5°$). The clustering plots for the remaining angles of both cases can be seen in Appendix A. Figure 10 proves that such classification of POD modes yields consistent cluster shapes in different cases and the same threshold value can be set for all the flow conditions in both cases. The threshold for the coherent-mode selection is set to $T^2 = 3$ as it includes all the coherent modes, i.e. those which belong to the upper left region. Although a slight variation to the threshold value could change the number of retained modes, any value around $T^2 = 3$ ensures containing the large scales of the flow. A modification would only add or reduce some lower energy modes when building a future surrogate model.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

δ	Number of modes	Recovered energy
2.5°	39	57.5 %
5°	36	58.1 %
10°	35	56.0 %
12.5°	32	52.0 %

Table 2. Number of coherent modes and total amount of energy recovered by them for each training angle for the Windsor body case.

δ	Number of modes	Recovered energy
−45°	106	74.8 %
−22.5°	98	72.7 %
22.5°	101	73.1 %
45°	107	76.4 %

Table 3. Number of coherent modes and total amount of energy recovered by them for each training angle for the BSC building.

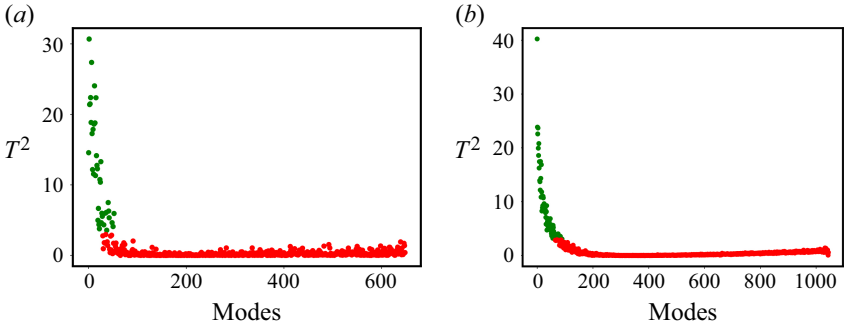


Figure 10. The T^2 clustering results for the wake of the Windsor body at $\delta = 2.5°$ (*a*) and for the BSC building at $\delta = 22.5°$ (*b*). Green dots represent the coherent modes and red dots the non-coherent modes.

Table 2 shows the number of selected modes in each angle of the Windsor body and the amount of energy recovered. It can be seen that the tendency is to have between 30 and 40 modes per angle containing coherent structures that represent a somewhat larger amount than half of the total energy. The case at $\delta = 12.5°$ is the one in which the coherent modes account for the smallest energy percentage, 52.0 %, while for $\delta = 5°$, they account for up to 58.1 %. It is important to note that the rest of the energy is shared among the remaining 600 non-coherent modes, therefore, each of them has a small individual contribution to the total energy of the system. In the case of the flow around the BSC (table 3), the clustering yields around 100 coherent modes that contain about 75 % of the energy.

Two modes of the Windsor body case at $\delta = 10°$ are used to illustrate the clustering process. Figure 11 compares the spatial correlation of a coherent mode with that of a non-coherent mode. Note that the chosen coherent mode is the fifth most energetic one and the non-coherent mode is the 450th most energetic one. The coherent mode is clearly dominated by four large correlated regions linked to the vortex shed from the windward side of the vehicle, whereas the non-coherent mode depicts multiple small scales.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

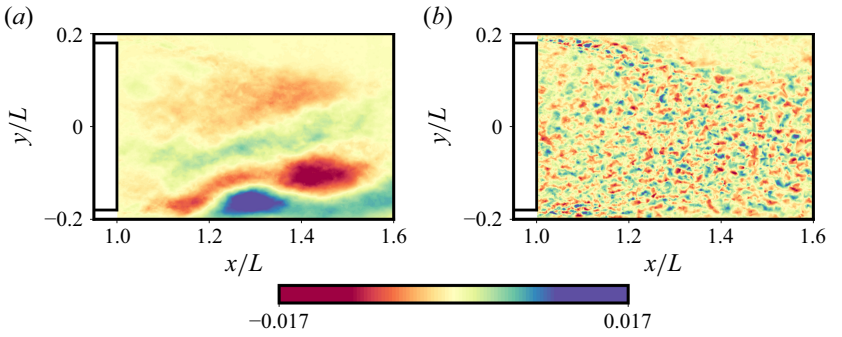


Figure 11. Spatial correlations of the 5th (*a*) and 450th (*b*) most energetic modes of the Windsor body at $\delta = 10°$, clustered as coherent and non-coherent, respectively.

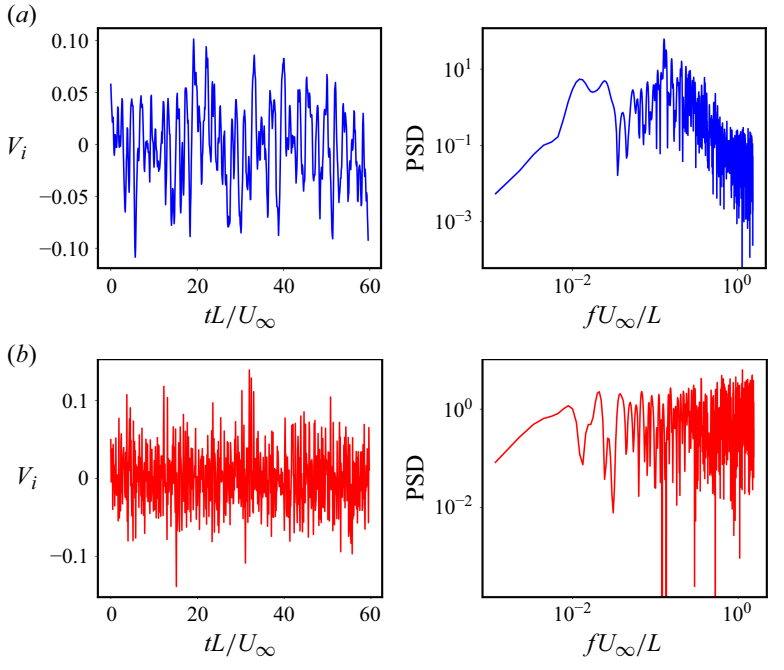


Figure 12. Temporal coefficient (left) and its power spectrum (right) of the 5th (*a*) and 450th (*b*) most energetic modes of the Windsor body at $\delta = 10°$, clustered as coherent and non-coherent, respectively.

Figure 12 compares the temporal coefficient and its spectrum for both modes. The spectrum of the coherent mode (right panel of figure 12*a*) exhibits a peak at the non-dimensional frequency of $fH/U_\infty = 0.13$. This peak corresponds to the windward vortex-shedding frequency (Booysen *et al.* 2022; Eiximeno *et al.* 2024*c*). Note that no dominant frequencies can be observed in the spectrum of the non-coherent mode (right panel of figure 12*b*), which is completely flat as in the case of pure white noise signals. These modes are seen as noise in the reduced system as their temporal coefficients are completely uncorrelated. The temporal coefficients of the non-coherent modes (left panel of figure 12*b*) suggest that the lack of correlation might come from an inadequate sampling frequency, this one being lower than the dominant frequency of these modes. The noisy and random evolution of the non-coherent modes, together with their small individual

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

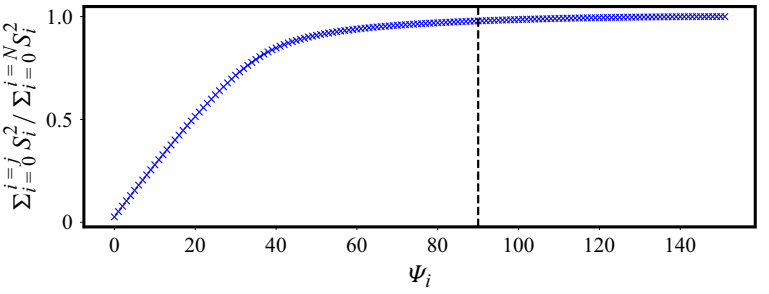


Figure 13. Cumulative energy for the POD to find the common basis between the selected modes for the Windsor body case.

energy contribution, would increase the cost of a surrogate model if they were included in the reduced system. However, they cannot be discarded without an efficient closure that accounts for the large energy percentage that they contain as a group.

The spatial correlations of the selected modes are then concatenated to create the matrix $\boldsymbol{Y}$ as in (2.5). As some coherent modes might be repeated in the yaw-angle range, POD is applied to matrix $\boldsymbol{Y}$ to find the optimal and orthonormal basis that contains the information of the selected modes for the four angles. Figure 13 shows the cumulative singular values to prove that for the case of the Windsor body, instead of using all the 142 coherent modes, 90 vectors are enough to represent the information of all the coherent modes in the yaw-angle range under study. In the case of the BSC building, 384 out of 412 modes are needed to build a basis without information loss. This can be explained by the larger difference between wind directions, which lead to a wider variety of features among the studied cases.

Figure 14(a) presents the kernel density estimate of all the training snapshots of the Windsor body case for the original field, $\mathcal{X}$, its reconstruction after being projected into the common POD basis, $\mathcal{X}_P$, and the error between both of them, $\mathcal{E}$. The most likely situation is to have fluctuations close to zero in the original and reconstructed fields. This is explained by the large unperturbed area in the leeward ($y \geq 0$) side of the domain. The source of error is then the filtering performed by the POD reconstruction of the high-amplitude fluctuations. Such filtering yields a field that is more likely to have points with velocity fluctuations close to zero than in the original case. Figure 14(a) also confirms that this holds true for the test case at $\delta = 7.5°$, bringing evidence that the common basis is valid for any angle in the studied range. Figure 14(b) compares the joint PDF of the error given the reconstruction from the common basis, $p(\mathcal{E} \mid \mathcal{X}_P)$, for the training and tests fields. As stated in § 2, this is the PDF learned by the closure model as it ensures that the predicted error yields a statistically equivalent system to the original one. Consistent with the results shown in figure 14(a), the most likely case in both the training and test datasets is to have a state with the velocity reconstruction and its error with the original field being close to zero. The most probable values for the test set match those of the training set; however, the limits of $p(\mathcal{E} \mid \mathcal{X}_P)$ for the training set are wider than those at $\delta = 7.5°$. Figures 14(c) and 14(d) confirm that the filtering phenomenon given by the projection and reconstruction to the common POD basis described for the Windsor body case is consistent with the flow data around the BSC headquarters.

4.2. *Model size convergence*

The three different architectures described in table 1 for the Windsor body are tested in order to assess the correct size of the attention layer. In figure 15 the PDF, $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$, given by the transformer output with the original one, represented in figure 14(b), for both the

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

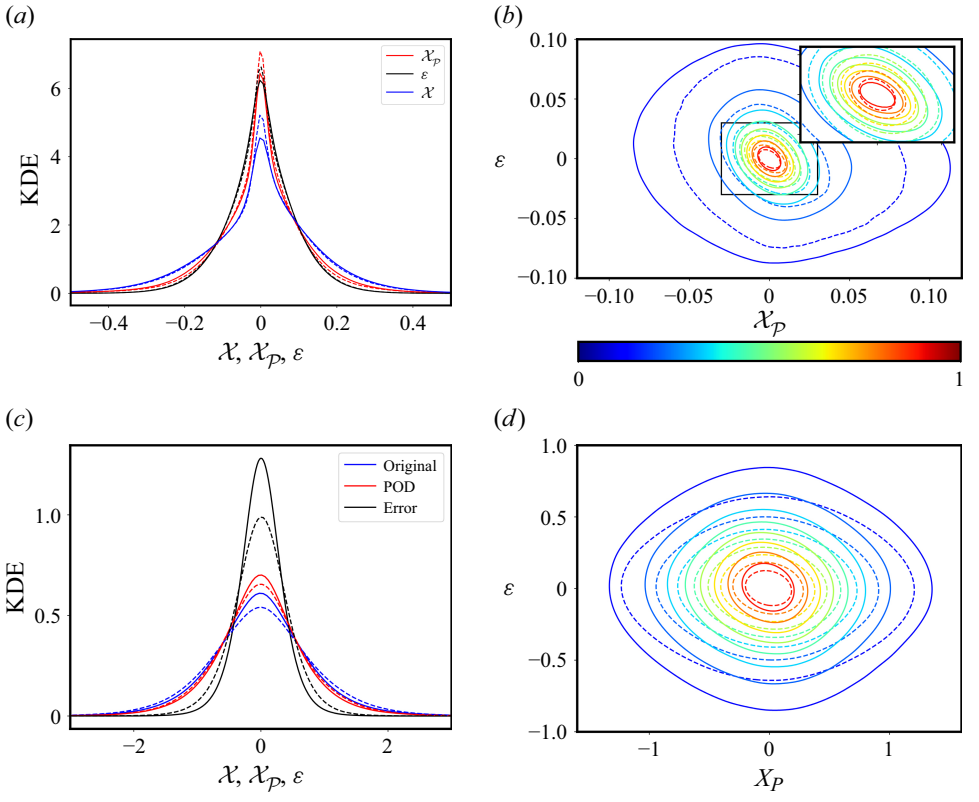


Figure 14. Kernel density estimate of the velocity fluctuations, its reconstruction from the common basis and the error between both for all snapshots in the training (solid) and test (dashed) datasets for (*a*) the Windsor body case and (*c*) the flow around the BSC. Joint PDF of the error depending on the reconstruction from the common basis, $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$, for all snapshots in the training (solid) and test (dashed) datasets for (*b*) the Windsor body case and (*d*) the flow around the BSC.

training and test datasets are compared. Architecture 1, with an attention layer of size $d_{model} = 64$, performs poorly in learning both the centre and the limits of the distribution. The KL divergence between the transformer prediction and the original probability for the training data is of $\mathcal{D}_{\mathcal{KL}} = 0.301$ and for the test data is $\mathcal{D}_{\mathcal{KL}} = 0.139$. Both values are the highest ones obtained during the architecture refinement process. In this case, the main source of error is that the PDF learned by the transformer is much narrower than the original one, meaning that the model fails to recover the fluctuations with larger amplitude.

Increasing the attention layer to $d_{model} = 128$, with its subsequent duplication of the number of parameters, allows the transformer to learn a wider area of $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$. This reduces the KL divergence with the original data to $\mathcal{D}_{\mathcal{KL}} = 0.135$ for the training snapshots and $\mathcal{D}_{\mathcal{KL}} = 0.023$ for the test snapshots. It is relevant to mention that this architecture nearly matches the output distribution for the test set as the limits of $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$ for $\delta = 7.5°$ are narrower than those found in the training set.

Duplicating the attention size to $d_{model} = 256$ leads to the best match of the training dataset of the three architectures. The learnt PDF expands for a wider area of fluctuations and the KL divergence is reduced to $\mathcal{D}_{\mathcal{KL}} = 0.031$. Now the KL divergence on the test set has raised again up to $\mathcal{D}_{\mathcal{KL}} = 0.031$, matching that obtained for the training dataset. The change in tendency of the KL divergence on the test set accounts for the larger fluctuations

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

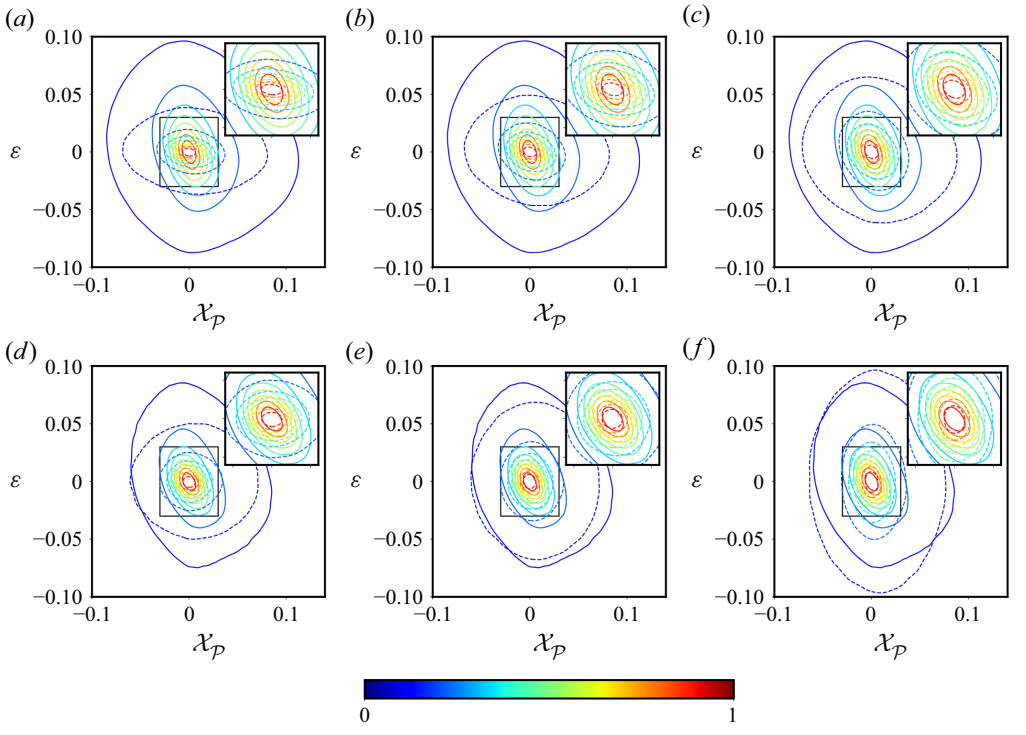


Figure 15. Joint PDF of the error depending on the reconstruction from the common basis, $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$, for the Windsor body case. Solid lines represent the reference values and dashed lines represent the values learned by the closure. Panels (*a*,*b*,*c*) show the accuracy on the training set ($\delta = [2.5°, 5°, 10°, 12.5°]$) and panels (*d*,*e*,*f*) show the accuracy for the validation set ($\delta = 7.5°$). (*a*,*d*) Architecture 1, (*b*,*e*) architecture 2 and (*c*,*f*) architecture 3.

from the training set that are not present in the case of $\delta = 7.5°$ and are already learned by the transformer. This is the first sign of overfitting to the training data as in this case the value of $\mathcal{D}_{\mathcal{KL}}$ is slightly larger than that found with architecture 2. This is the last step of architecture refinement because the evaluation of the test set has already crossed the ideal prediction in which the KL divergence would be null.

The wider area of $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$ learned by the architectures with a larger attention size can be linked to the amount of TKE,

$$k = \int_\Omega \frac{1}{2} u' u' \mathrm{d}\Omega, \tag{4.1}$$

recovered by the closure model. The TKE recovered in each case is quantified with the kernel density estimate among all the snapshots of the training and test sets separately. Figure 16 effectively showcases that the most likely energy value, $\overline{k}$, after the reconstruction from the POD common basis is significantly lower than that of the original flow. For the training snapshots, it is reduced from $\overline{k} = 0.0053$ to $\overline{k} = 0.0032$ and, for the test dataset, it decreases from $\overline{k} = 0.0050$ to $\overline{k} = 0.0029$.

Figure 16 also shows that the most likely energy value when adding the closure term increases with the attention layer size of the transformer used to model the missing fluctuations. In the training angles, $\overline{k}$ increases from $\overline{k} = 0.0038$ to $\overline{k} = 0.0041$ when the attention sizes change from $d_{model} = 64$ to $d_{model} = 128$. It finally reaches the value of $\overline{k} = 0.0049$ with the largest architecture of $d_{model} = 256$. A similar behaviour is observed

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

Attention size (d_{model})	$\mathcal{D}_{\mathcal{KL}}$ Training	$\mathcal{D}_{\mathcal{KL}}$ Test
64	0.301	0.139
128	0.135	0.023
256	0.031	0.031

Table 4. Küllback–Leibler divergence between the original $p(\mathcal{E}|\mathcal{X}_{\mathcal{P}})$ and the one learned by each transformer architecture for the Windsor body case.

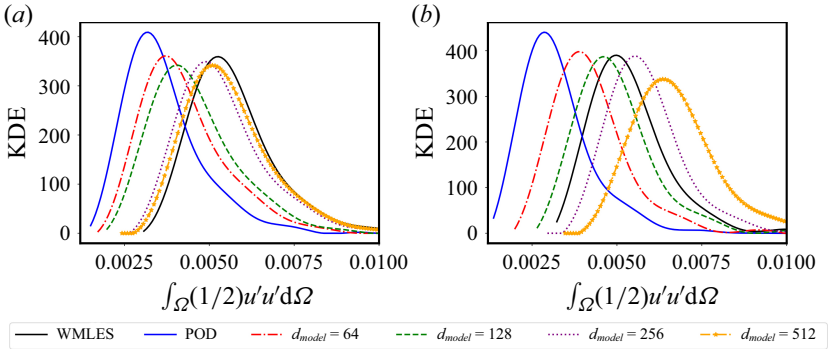


Figure 16. Kernel density estimate of the energy for the training (*a*) and test (*b*) datasets of the Windsor body case.

with the case at $\delta = 7.5°$ for the architectures with $d_{model} = 64$ and $d_{model} = 128$ as $\overline{k}$ goes up to $\overline{k} = 0.0039$ and $\overline{k} = 0.0046$, respectively. However, for the architecture with $d_{model} = 256$, the most likely energy value, $\overline{k} = 0.0053$, is slightly higher than in the original flow. This can be explained by the fact that the PDF of the fluctuations predicted by the transformer is wider than those of the real case (figure 15).

These results show the relevance of choosing the size of the transformer appropriately and that it is necessary to fine tune the depth of the model. Exceptionally, figure 16 includes the results obtained with a deeper transformer ($d_{model} = 512$). The performance of this architecture is not presented anywhere else in this paper due to its large energy overprediction for the case at $\delta = 7.5°$. This result shows that the deeper the transformer, the better the energy prediction of the training set. However, it demonstrates that a representative test set is needed to decide the model size in order to avoid overfitting of the model and its consequent energy overprediction.

This analysis also brings evidence that the closure model actually reduces the offset between the energy of the POD reconstruction and that of the original system. It is important to note that the accuracy of the energy prediction is directly linked to the KL divergence between the predicted $p(\mathcal{E}|\mathcal{X}_{\mathcal{P}})$ by the transformer and the ground truth. When the KL divergence tends to decrease, the energy added by the closure is still smaller than the gap between the original flow and the POD reconstruction. Zero KL divergence would mean a perfect match between the model and the ground truth with no energy deviation. For the last scenario, an increasing KL divergence indicates that the model is overshooting the predicted fluctuations, and with it the TKE.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

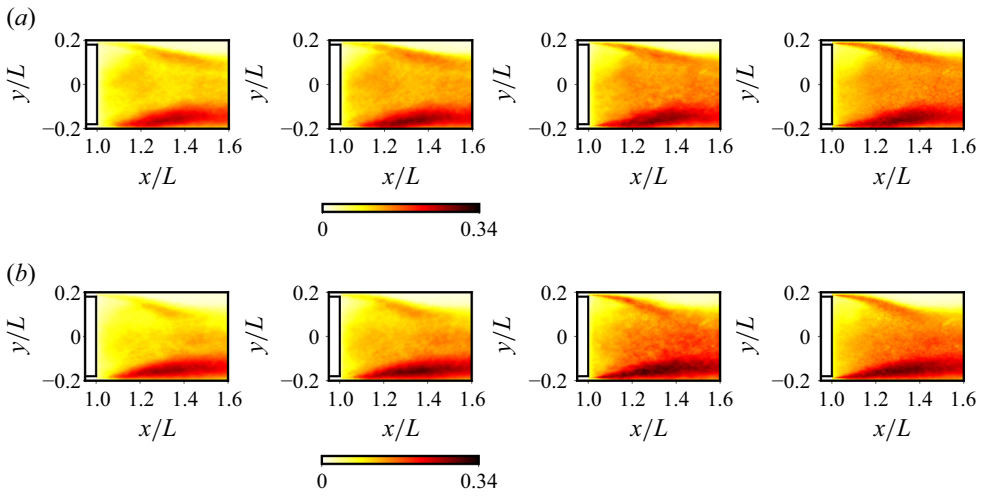


Figure 17. The r.m.s. value of the streamwise velocity fluctuations for the cases at $\delta = 2.5°$ (a) and $\delta = 7.5°$. From left to right, the pictures represent: reconstruction from the common POD basis, reconstruction enhanced with a SRGAN, reconstruction with the closure model with an attention size of $d_{model} = 256$ and the original flow for the Windsor body case.

4.3. *Spatial field reconstruction*

Up to this point of the discussion, it has been proven that adding a field of fluctuations based on the $p(\mathcal{E}|\mathcal{X}_\mathcal{P})$ learned by the proposed transformer architectures is enough to close the energy gap of a POD reconstruction and the original flow field. However, it remains to be proven that the closure model can distribute these fluctuations adequately across the spatial and temporal domains.

Figure 17 compares the r.m.s. of the velocity fluctuations in the Windsor body case for the reconstruction from the common POD basis, the reconstruction enhanced with a SRGAN, the reconstruction plus the closure term given by the transformer with $d_{model} = 256$ and those of the original field. The r.m.s. of velocity fluctuations is closely related with the local contribution to the total TKE of the flow; hence, a field with the closure term matching the r.m.s. fluctuations of the original case could be considered accurate in space and statistically equivalent in time. In figure 17 the case at $\delta = 2.5°$ is used to illustrate the performance on the various training angles. The results at $\delta = 7.5°$ are also plotted to show the performance of the model on unseen data. Moreover, the four mentioned reconstructions together with those including the closures at $d_{model} = 64$ $d_{model} = 128$ are evaluated along the line at $x/L = 1.3$ in figure 18. We refer the reader to Appendix B for the equivalent figures to figures 17 and 18 corresponding to the training cases at $\delta = [5°, 10°, 12.5°]$.

Both figures illustrate that the common basis captures the positions of the fluctuation maxima and their correct distribution along the domain, ensuring that the main flow structures are preserved throughout the projection and reconstruction processes (2.7) and (2.8). This is also valid for the case at $\delta = 7.5°$, despite the fact that its features were not explicitly included in the basis.

In all analysed angles, the reconstruction from the common basis misses the actual value by an offset associated with the filtered fluctuations. Figure 18 shows that increasing the attention size helps to close the gap in the r.m.s. fluctuations in all areas of the domain. The larger range of fluctuations learned by the deeper architecture ($d_{model} = 256$) and its additional kinetic energy added to the flow, translates to a nearly perfect match of the

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

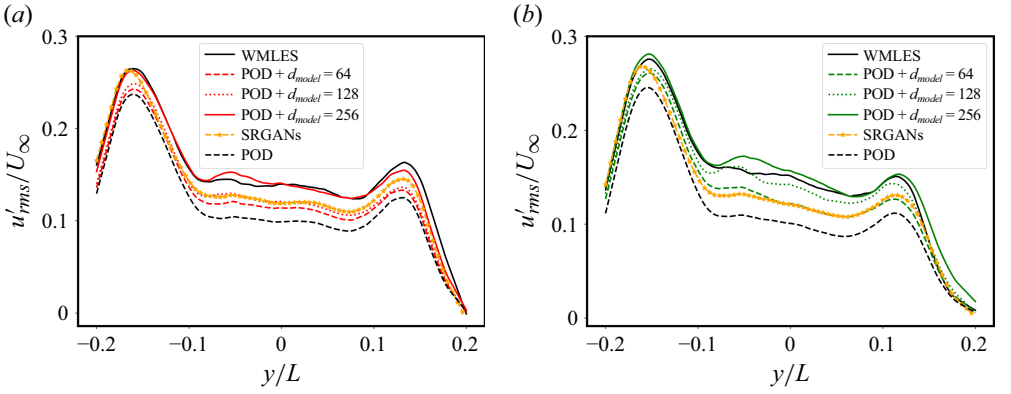


Figure 18. The r.m.s. of the streamwise velocity fluctuations for the cases at $\delta = 2.5°$ (*a*) and $\delta = 7.5°$ (*b*) on a cross-stream line at $x/L = 1.3$ for the Windsor body case.

r.m.s. of the velocity fluctuations. It is worth mentioning that in the training cases most of the differences between the original field and the closure prediction arise from the model underestimating the fluctuations; however, at $\delta = 7.5°$ all the error of the closure is attributed to a slight overprediction.

As the offset between the POD reconstruction and the original field is not constant throughout the whole domain, figures 17 and 18 are also the evidence that the closure learns how much energy the POD reconstruction missed depending on the domain region. This proves that the closure model does not only close the energy gap statistically over all the points of all snapshots, but also that it can give accurate predictions of what happens in every point in the domain.

Although, figures 17 and 18 prove that the inference of the POD reconstructions through the SRGAN closes the energy gap between the raw surrogate and the actual CFD results, figure 18 highlights that the SRGAN does not preserve the location of the fluctuation maximum. Such a maximum is the energy peak due to the windward shear layer and presents a displacement when compared with the original flow field and its POD reconstruction. This means that the SRGAN modifies the large scales predicted by the surrogate. An additional drawback from the SRGAN architecture is its lower energy recovery when compared with the deeper transformer closures. It also has to be acknowledged that the SRGANs have the limitation of being based on convolutional neural networks, which is a large drawback when building models for aerodynamics of industrial cases dealing with complex geometries. On the contrary, the transformer can deal with data from unstructured grids and handle any geometry representation.

To show an example of how the proposed model can learn a closure for data represented on an unstructured grid, a transformer architecture equivalent to the highest accuracy on the Windsor body wake has been trained to generate the truncated fluctuations when reducing the dimensionality of the flow around the BSC building.

Figure 19 compares the r.m.s. value of the velocity fluctuations of the flow around the BSC building for the reconstruction from the common basis, the reconstruction adding the fluctuations predicted by the closure model and the original case. This figure should be analysed together with the plot over a line at $x/H_{max} = 1.79$ (figure 20), as it effectively showcases the energy given by the closure and how the additional fluctuations improve the accuracy of a potential surrogate model to predict transient flow fields.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

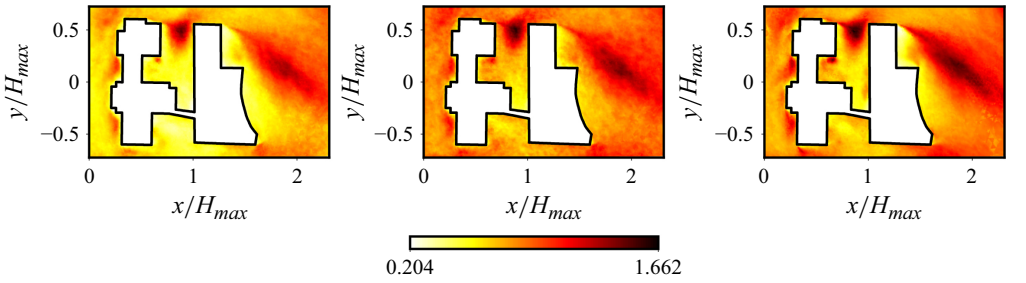


Figure 19. The r.m.s. value of the velocity fluctuations for the case at 0°. From left to right, the pictures represent: reconstruction from the common POD basis, reconstruction with the closure model and the original flow around the BSC building.

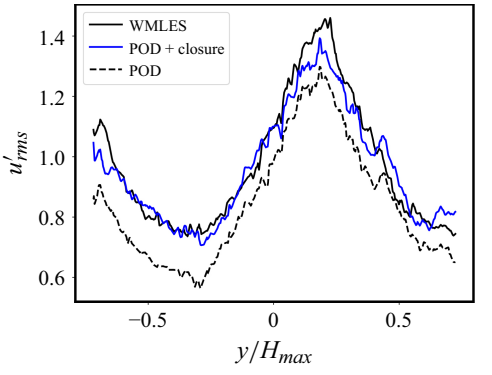


Figure 20. The r.m.s. value of the velocity fluctuations for the case at 0° on a cross-stream line at $x/H_{max} = 1.79$ for the flow around the BSC building.

4.4. *Instantaneous-field reconstruction*

After discussing how the closure model can emulate a field that is statistically equivalent to the original one in all points of the domain, it is time to discuss its impact on the reconstruction of the instantaneous fields. In this case, all comparisons are done with the deepest architecture ($d_{model} = 256$) for the Windsor body case.

The first step is proving that the closure learns the amount of energy missing in each time step. To do so, figure 21 shows the temporal evolution of the total TKE together with its POD reconstruction and the reconstruction corrected with the closure model for the case at $\delta = 7.5°$. This case is the evidence that common POD basis successfully captures the instants of all energy maxima and minima of the original field. Once again, this is still valid even if the flow condition was not included in the database.

Figure 21 also shows that the closure model represents the energy missing in each snapshot instead of adding the same energy to all of them. However, in the particular case of $\delta = 7.5°$, the actual energy predicted by the closure is consistently higher than that of the original flow in each snapshot, as discussed in the previous paragraphs. Table 5 links the better prediction of the energy temporal evolution with the mean relative error regarding the energy of the original field. In all angles this error has been reduced from over 37 % to a margin between 7 % and 12 %.

Closing the energy gap appropriately in each snapshot also comes with a better prediction of the instantaneous fluctuations. To exemplify this, figure 22 compares the

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

δ	POD reconstruction	Closure ($d_{model} = 256$)
2.5°	38.6 %	8.2 %
5°	37.0 %	8.8 %
7.5°	41.8 %	10.7 %
10°	38.5 %	7.1 %
12.5°	40.4 %	11.4 %

Table 5. Mean relative error between the energy of the original field, the energy recovered by the POD reconstruction and the POD reconstruction with the closure model for the Windsor body case.

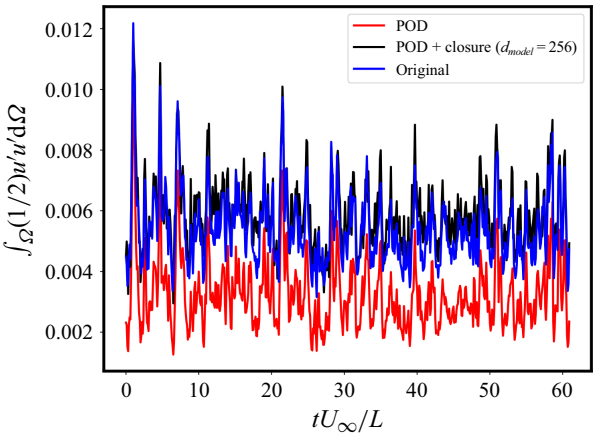


Figure 21. Temporal evolution of the energy of the system for the Windsor body case at $\delta = 7.5°$.

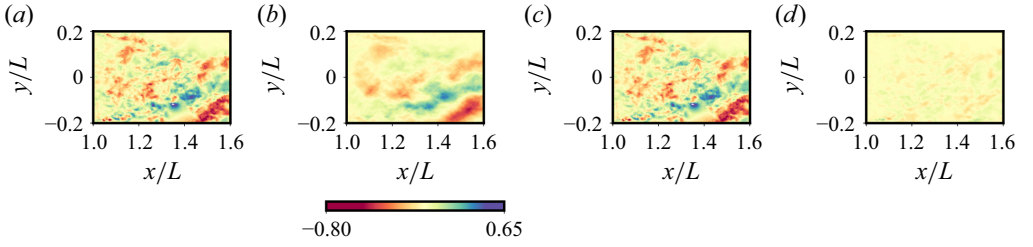


Figure 22. From (*a–d*): the Windsor body case: original streamwise velocity fluctuations, POD reconstruction, POD reconstruction with the closure term and the error between the closed reconstruction and the original field ($\mathcal{E}_M$) for a snapshot at $\delta = 7.5°$.

original instantaneous field and its reconstruction for a handpicked snapshot at $\delta = 7.5°$, effectively showcasing that the POD reconstruction exhibits large deviations from the original data. In fact, figure 23(*a*) shows that the reconstruction from the standard POD basis leads to a relative error larger than $|\mathcal{E}|/\mathcal{X} \geq 0.5$ in 57.8 % of the points in the domain. After adding the closure term, the accuracy of the reconstruction is increased so that only 13.7 % of the points have a relative error higher than $|\mathcal{E}_M|/X \geq 0.5$.

The comparison between the PDF of the fields, figure 23(*b*), agrees with figure 14(*a*) on showing that the POD reconstruction filters the high-amplitude fluctuations by increasing the points with fluctuations close to zero. When adding the closure term, the PDF of

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

δ	POD reconstruction	Closure ($d_{model} = 256$)
2.5°	0.2166	0.0240
5°	0.2085	0.0269
7.5°	0.2362	0.0253
10°	0.2111	0.0222
12.5°	0.2062	0.0245

Table 6. Kullback–Leibler divergence, $\mathcal{D}_{\mathcal{KL}}$, between the original field and the reconstruction with and without the closure term. The results are averaged over all snapshots for each angle of the Windsor body case.

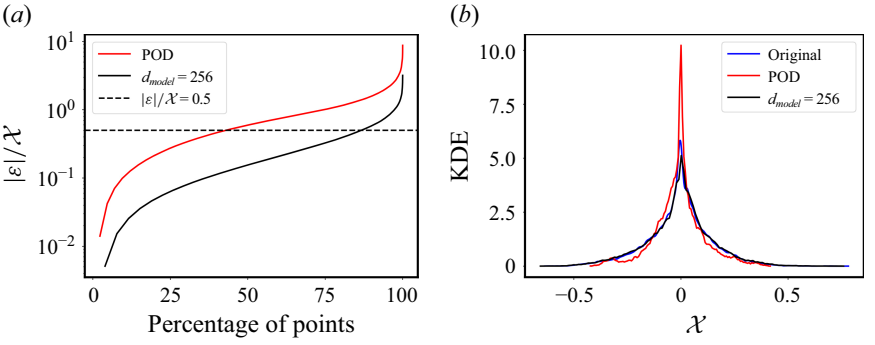


Figure 23. Cumulative density function of the relative error between the reconstructions and the original field (*a*) and PDF for the three fields (*b*) for the Windsor body case.

the velocity fluctuations is nearly identical to that of the original field. In fact, the KL divergence between the reconstructed PDFs and the original one is reduced from $\mathcal{D}_{\mathcal{KL}} = 0.2447$ to $\mathcal{D}_{\mathcal{KL}} = 0.0052$ after adding the term predicted by the transformer. Table 6 shows the mean $\mathcal{D}_{\mathcal{KL}}$ over all snapshots to show that adding the closure term reduces the KL divergence with the original field of instantaneous velocity fluctuations regardless of the yaw angle.

4.5. *Model cost*

The training of every transformer architecture shown in the present work for the Windsor body required around 8 h and 15 mins using an NVIDIA H100 GPU from the accelerated partition of the supercomputer MareNostrum 5 (Barcelona Supercomputing Center 2024). This time is slightly lower than that needed to fit the SRGAN architecture used as a benchmark for the methodology. In this case, the total time for training on the same machine is 8 h and 47 mins. An economic cost comparison between both trainings can be done by taking the reference hourly cost of using a NVIDIA H100 GPU on a cloud computing platform. According to Owen (2024) this cost is of 6.74€, and yields a total of 55.61€ for the transformer training and 58.98€ for the SRGANs.

The inference of both methods for all the snapshots of the test angle is extremely quick. In this case the SRGAN is the cheapest architecture, as it takes only 3 s compared with 19 s for the transformer. The model has to be inferred 113 times to compensate for the time difference in training. While the cost of running a new CFD simulation should be compared with the inference of the models, a full comparison is unfair in our point of view until a full surrogate model is built in later stages of this research. Having said that, each

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

flow condition from the present Windsor body database was computed using 20 NVIDIA H100 GPUs for 80 h. Therefore, once the surrogate model for the interpolation of the POD coefficients inside the common basis is built, the potential savings for each inference of the model are of 10,784€.

5. Conclusions

This paper presents a deep-learning-based closure model for truncated POD modes. The main objective is to provide a methodology to recover the energy lost when reducing the dimensionality of the data. To do so, a transformer model is used to learn the spatial PDF of the difference between the original flow field and the POD reconstruction from the modes that would be included in a surrogate. This transformer learns and compresses the smaller scales of turbulent flows, enabling future strategies for more efficient data storage and potential surrogate models able to reproduce accurately the first-order statistics from the original field.

The methodology is tested to build two separate closures for the truncated scales of streamwise velocity fluctuations. The first model is trained on a slice in the wake of the Windsor body at the yaw angles of $\delta = [2.5°, 5°, 7.5°, 10°, 12.5°]$, and the second one learns the flow around the building of the BSC building at pedestrian level. In the latter case, five different wind directions are also considered. The wind blowing from south to north is taken as the reference direction, i.e. the case at 0°. In the other four cases, the incident velocity is rotated 22.5° and 45° towards both the east and west directions. As the model has to be generalisable for unseen data, two angles from each dataset are discarded during the training process and are used only for test purposes. In the case of the Windsor body, this angle is the one at $\delta = 7.5°$ and in the flow around the BSC is the one at 0°.

Before working on the transformer model, a set of POD modes at each of the training angles is selected. These modes have to be the most meaningful ones in the system as they would constitute the core of a future ROM. Truncating these modes would imply information loss on the large scales of the flow and should also be kept in case of working in a data compression paradigm. The selection process is based on performing PCA on the PSD of the temporal coefficients. Then the modes with an outstanding frequency behaviour are clustered with Hotelling's T^2. These modes are named as coherent modes and this selection process ensures that they are the only ones that present relevant frequency dynamics.

Despite the differences in the energy distribution, the clustering regions are fairly similar between all training angles of the two studied cases. In the particular case of the Windsor body, less than 10 % of the modes are coherent, however, they account for nearly 60 % of the energy. The remaining 40 % is distributed along the more than 600 non-coherent modes and is the one that needs to be modelled by the closure. In the case of the BSC building, around 100 coherent modes containing around 75 % of the energy are identified per wind direction. The clustered modes from each training angle are concatenated to build a common basis for each of the two cases that preserves the coherent structures inside the studied yaw-angle range. Proper-orthogonal decomposition is applied to the concatenated modes to ensure that all vectors from the basis are orthonormal between themselves and that they are the optimal representation of the coherent modes in that range. On the one hand, in the Windsor body case, this operation reduced the number of basis vectors from 142 to 90 without any additional information loss. On the other hand, nearly all the modes, 384 out of 412, are needed to build a basis without information loss.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

Projecting any snapshot (regardless of whether it was included in the training set or not) into the common POD basis, filters the high-amplitude fluctuations. A transformer-encoder block with an easy-attention mechanism is used to learn the PDF of the missing fluctuations depending on the reconstructed value from the POD common basis. Three different transformer architectures are trained for the Windsor body flow in order to assess its effect on the recovered fluctuations. The main difference between the architectures is the change in the attention size. In the shallowest architecture it takes the value of $d_{model} = 64$. Then it is doubled twice to get an attention size of $d_{model} = 128$ and $d_{model} = 256$. The closure for the flow around the BSC is only trained with an attention size of $d_{model} = 256$.

The Windsor body case shows that the larger the attention size, the more fluctuations from the training set are recovered. The accuracy of the prediction is quantified with the KL divergence between the transformer output and the original field. For the training set, it reduces from $\mathcal{D}_{\mathcal{KL}} = 0.301$ to $\mathcal{D}_{\mathcal{KL}} = 0.0031$ when the attention size changes from $d_{model} = 64$ to $d_{model} = 256$. In the case of the test set, there is also an accuracy improvement when doubling the attention size up to $d_{model} = 128$, but then, the first signs of overfitting to the training data are seen with the deepest architecture. The evaluation of the closure at $\delta = 7.5°$ for the architecture with $d_{model} = 256$ is the only case in which the KL divergence increases when compared with a shallower architecture. Note that a change of tendency in the KL is a sign of overfitting, which in this case is due to the transformer learning larger fluctuation amplitudes from the training set that are not present in the test set.

Adding the fluctuations field predicted by the transformer reduces the energy gap between the POD reconstruction and the original field. In the case of the Windsor body, the architectures with a larger attention size recover more energy than the shallower transformers. For instance, in the training set, the most likely energy value after adding the closure with $d_{model} = 64$ is $\overline{k} = 0.0038$, and it rises to $\overline{k} = 0.0049$ for $d_{model} = 256$. In this case, since the fluctuations from the training set are larger than those of the test set, the overfitting observed when comparing the PDF leads to an overshoot in the energy prediction. The evaluation of the test set with $d_{model} = 256$ yields $\overline{k} = 0.0053$, but the most likely energy value in the original flow is $\overline{k} = 0.0050$.

These fluctuations also leads to an improvement in the prediction of the r.m.s. value of the velocity fluctuations. The reconstruction from the common POD basis is able to capture the distribution of all local maxima and minima, but it falls short when matching the correct value. Then, the closure model helps to recover the missing fluctuations in the correct part of the domain. Once again, an increase in the attention size leads to a better closure of the offset. The evaluation of the deepest architecture at $\delta = 7.5°$ is the only case in which the r.m.s. prediction is larger than the original flow value. This is related to the energy overshoot discussed in the previous paragraph. It is relevant to mention that using the present methodology is more accurate than enhancing the POD reconstruction through a SRGAN. The main drawback from the SRGAN is that the inference from the truncated fields does not preserve the location of the fluctuation maximum, meaning that the network modifies the large scales predicted by the surrogate. Moreover, the energy recovery is lower than that given by the transformer closure. Another advantage of the current method when compared with SRGAN is the possibility to handle complex geometries represented on unstructured grids, such as the flow around the BSC building. In this case, adding the fluctuations recovered by the transformer also give a significant improvement to the truncated model.

Finally, this work proves that the energy added via the predicted fluctuations also reduces the error in the instantaneous flow field prediction. This is particularly true for

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

the architecture with $d_{model} = 256$. In the Windsor body case, the temporal mean of the energy prediction error is reduced on all angles from more than 37 % to less than 12 %. Moreover, the KL divergence between the velocity distribution reconstructed by POD and that of the original field is consistently larger than $\mathcal{D}_{\mathcal{KL}} = 0.2$, but the closure reduces it to less than $\mathcal{D}_{\mathcal{KL}} = 0.026$.

To conclude this paper, we would like to remark that this contribution is a first step towards a more general deep-learning-based closure model for data that has been truncated during dimensionality-reduction processes. Future research is mainly encouraged by the promising results in the BSC building case. These results lead to the conclusion that transfer learning of the model could be done between different cases or a different set of parameters. This would be particularly beneficial if the probabilistic distribution of the missing fluctuations is similar to those seen in the training phase and would increase the practical advantages of the model from the point of view of engineering performance, as the amount of training data during the reusability of the model would be drastically reduced. Therefore, it is an encouraging research path to learn more on how to quantify the similarity of such distribution, explain if and why it is possible to do transfer learning on cases with different geometries and how to reuse a trained model for a different input size.

There is an additional follow-up on how to optimise the amount of training data needed. Although the results presented in this paper are focused on recovering the fluctuations in a single slice of the domain, future studies could focus on how to select the spatial points for training. In other words, identifying which are the most informative points in the domain for the model and what is the minimum amount of spatial points needed to train so as to get a valid closure for the rest of the points.

Last but not least, the training of the closure can be coupled with the actual surrogate model for the large-scales interpolation. Such interpolation could be performed using the parameterised DMD from Andreuzzi *et al.* (2023) or the SHRED-ROM methodology developed by Tomasetto *et al.* (2025). When building the full surrogate, it will be necessary to do a deeper study of the models sensitivity to the amount of training data needed for turbulent flows and on how to choose the optimal training data points, i.e. which flow conditions are the most informative to use as training cases. This choice could be made based on low-fidelity simulations of the cases or using resolvent analysis.

Acknowledgements. The authors acknowledge the Barcelona Supercomputing Center for the usage of MareNostrum 5 during the development of this paper. The authors also acknowledge the insightful conversations with Fermín Mallor during the development of the initial ideas giving rise to this work. The authors acknowledge Josep Maria Duró for his help on extracting the data used for the realistic test case of the flow around the BSC building.

Funding. The research leading to this work has been partially funded by the project TIFON with reference PLEC2023-010251/ AEI/10.13039/501100011033. B. Eiximeno's work was funded by a contract from the Subprograma de Ayudas Predoctorales given by the Ministerio de Ciencia e Innovación (PRE2021-096927). O. Lehmkuhl has been partially supported by a Ramon y Cajal postdoctoral contract (Ref: RYC2018-025949-I). The authors acknowledge the support of Departament de Recerca i Universitats de la Generalitat de Catalunya to the Research Group Large-scale Computational Fluid Dynamics (Code: 2021 SGR 00902) and the Turbulence and Aerodynamics Research Group (Code: 2021 SGR 01051).

M. Sanchis-Agudo and R. Vinuesa would like to acknowledge the support from Marie Sklodowska-Curie Actions project MODELAIR, funded by the European Commission under the Horizon Europe program through grant agreement number 101072559.

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

Data availability statement. The data that support the findings of this study is available upon request. See JFM's research transparency policy for more information

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

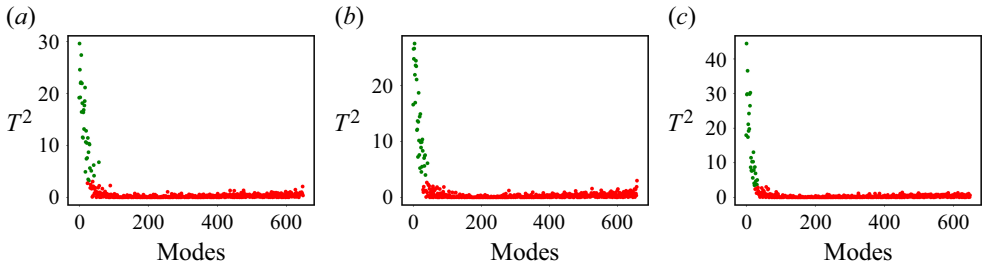


Figure 24. The T^2 clustering results for the wake of the Windsor body at $\delta = 5°$ (*a*), $\delta = 10°$ (*b*) and $\delta = 12.5°$ (*c*). Green dots represent the coherent modes and red dots the non-coherent modes.

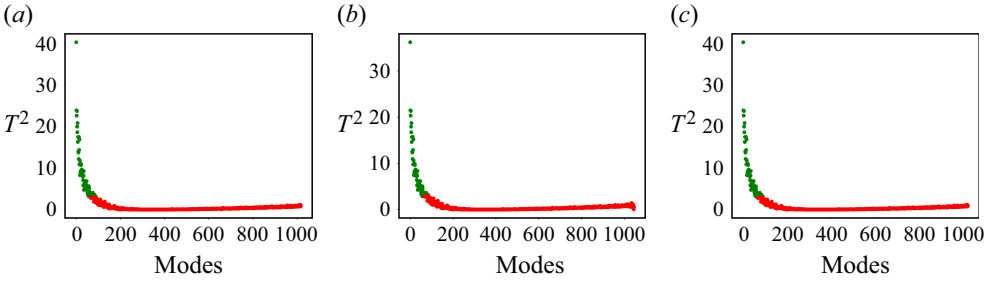


Figure 25. The T^2 clustering results for the flow around the BSC building at $\delta = -45°$ (*a*), $\delta = -22.5°$ (*b*) and $\delta = 45°$ (*c*). Green dots represent the coherent modes and red dots the non-coherent modes.

Author contributions. **B. E.:** writing – review & editing, writing – original draft, visualisation, validation, software, methodology, investigation, formal analysis, data curation, conceptualisation. **M. S-A.:** writing – review & editing, writing – original draft, software, methodology, investigation, conceptualisation. **A. M.:** writing – review & editing, supervision, software. **I. R.:** writing – review & editing, supervision, methodology, funding acquisition and project administration. **R. V:** writing – review & editing, supervision, resources, project administration, funding acquisition. **O. L.:** writing – review & editing, validation, supervision, software, resources, project administration, methodology, investigation, funding acquisition.

Appendix A. Clustering of coherent modes

This appendix presents the T^2 clustering plots for the angles $\delta = [5°, 10°, 12.5°]$ of the Windsor body case (figure 24) and the cases of wind blowing at $-45°$, $-22.5°$ and $45°$ of the flow around the BSC building (figure 25). These figures are analogous to figure 10 presented in § 4.

Appendix B. Accuracy of the closure on the training angles

This appendix presents the comparison of the r.m.s. value of the streamwise velocity fluctuations for the angles of $\delta = [5°, 10°, 12.5°]$. These angles were also included in the common basis as the case of $\delta = 2.5°$. Moreover, the error of their reconstruction was also included in the dataset used for the training of the transformer. Figure 26 complements figure 17 and figure 27 complements figure 18 – in these figures only the case at $\delta = 2.5°$ was used to illustrate the effect of the closure on the cases used during training.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

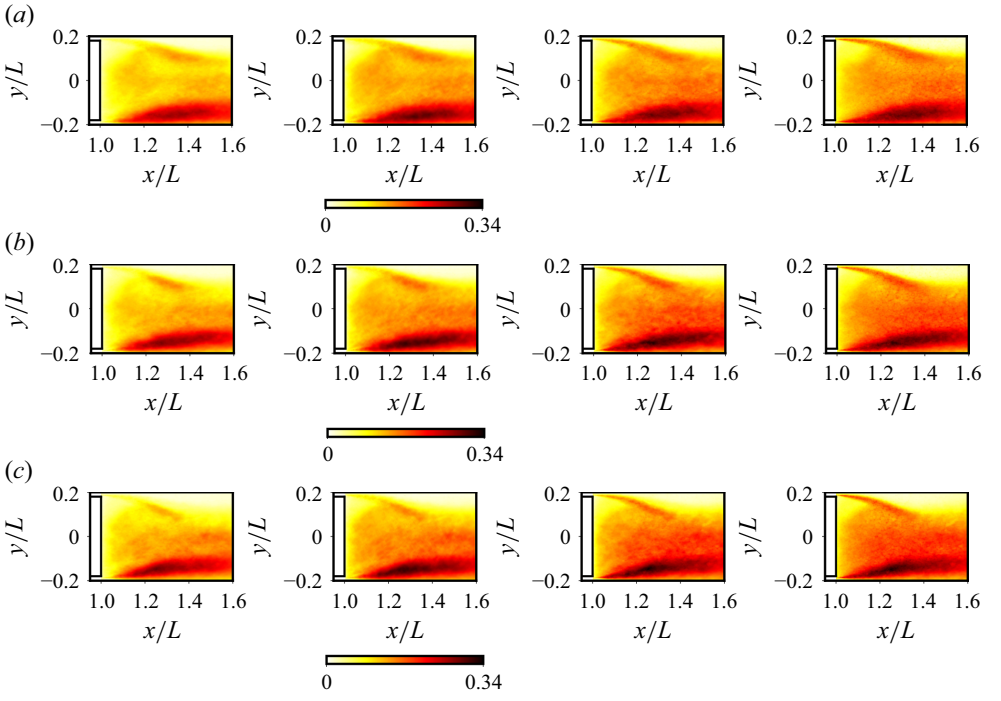


Figure 26. The r.m.s. of the streamwise velocity fluctuations for the Windsor body cases at $\delta = 5°$ (*a*), $\delta = 10°$ (*b*) and $\delta = 12.5°$ (*c*). From left to right, the pictures represent: reconstruction from the common POD basis, reconstruction enhanced with the SRGAN method, reconstruction adding the closure term predicted with an attention size of $d_{model} = 256$ and the original flow.

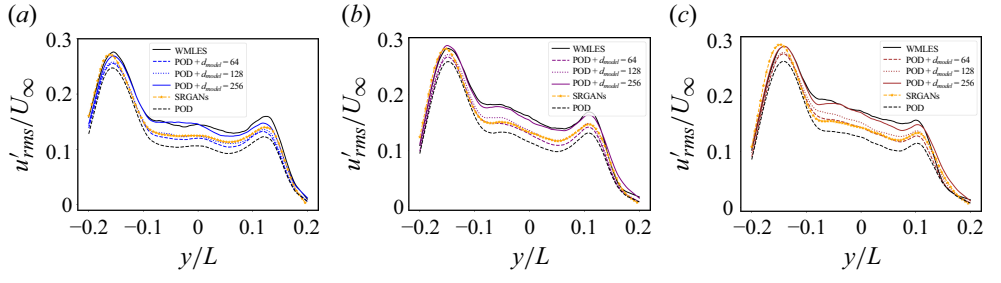


Figure 27. The r.m.s. of the streamwise velocity fluctuations for the Windsor body cases at $\delta = 5°$ (*a*), $\delta = 10°$ (*b*) and $\delta = 12.5°$ (*c*) at $x/L = 1.3$.

References

Akkari, N., Casenave, F., Hachem, E. & Ryckelynck, D. 2022 A Bayesian nonlinear reduced order modeling using variational autoencoders. *Fluids* **7** (10), 334–358.

Andreuzzi, F., Demo, N. & Rozza, G. 2023 A dynamic mode decomposition extension for the forecasting of parametric dynamical systems. *SIAM J. Appl. Dyn. Syst.* **22** (3), 2432–2458.

Bahdanau, D., Cho, K. & Bengio, Y. 2014 Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barcelona Supercomputing Center Marenostrum 5 guide. Available at: https://www.bsc.es/supportkc/docs/MareNostrum5/intro/. Accessed: 2024-11-06.

Booysen, A., Das, P. & Ghaemi, S. 2022 Large-scale 3D-PTV measurement of Ahmed-body wake in crossflow. *Expl Therm. Fluid Sci.* **132**, 110562.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

Brunton, S.L., Noack, B.R. & Koumoutsakos, P. 2020 Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52** (1), 477–508.

Céspedes Moreno, J.F., Murcia León, J.P. & Andersen, S.J. 2025 Convergence and efficiency of global bases using proper orthogonal decomposition for capturing wind turbine wake aerodynamics. *Wind Energy Sci.* **10** (3), 597–611.

Couplet, M., Sagut, P. & Basdevant, C. 2003 Intermodal energy transfers in a proper orthogonal decomposition–Galerkin representation of a turbulent separated flow. *J. Fluid Mech.* **491**, 275–284.

D'Hooge, A., Palin, R., Rebbeck, L., Gargoloff, J. & Bradley, D. 2014 Alternative simulation methods for assessing aerodynamic drag in realistic crosswind. *SAE Intl J. Passenger Cars - Mech. Syst.* **7** (2), 617–625.

Eivazi, H., Le Clainche, S., Hoyas, S. & Vinuesa, R. 2022 Towards extraction of orthogonal and parsimonious non-linear modes from turbulent flows. *Expert Syst. Appl.* **202**, 117038.

Eivazi, H., Veisi, H., Naderi, M.H. & Esfahanian, V. 2020 Deep neural networks for nonlinear model order reduction of unsteady flows. *Phys. Fluids* **32** (10), 105104.

Eiximeno, B., Begiashvili, B., Miro, A., Valero, E. & Lehmkuhl, O. 2024*a* PyLOM: low order modelling in python. Available at https://github.com/ArnauMiro/pyLowOrder.

Eiximeno, B., Lehmkuhl, O. & Rodriguez, I. 2025 Database on flow past the Windsor body at different yaw angles. Accessed: 2025-05-13.

Eiximeno, B., Miró, A., Begiashvili, B., Valero, E., Rodriguez, I. & Lehmkuhl, O. 2024*b* PyLOM: a HPC open source reduced order model suite for fluid dynamics applications. arXiv preprint arXiv:2405.15529.

Eiximeno, B., Miró, A., Rodriguez, I. & Lehmkuhl, O. 2024*c* Toward the usage of deep learning surrogate models in ground vehicle aerodynamics. *Mathematics* **12** (7), 998.

Fukami, K., Fukagata, K. & Taira, K. 2019 Super-resolution reconstruction of turbulent flows with machine learning. *J. Fluid Mech.* **870**, 106–120.

Fukami, K., Fukagata, K. & Taira, K. 2021 Machine-learning-based spatio-temporal super resolution reconstruction of turbulent flows. *J. Fluid Mech.* **909**, A9.

Fukami, K., Fukagata, K. & Taira, K. 2023 Super-resolution analysis via machine learning: a survey for fluid flows. *Theor. Comput. Fluid Dyn.* **37**, 421–444.

Fukami, K., Nakamura, T. & Fukagata, K. 2020 Convolutional neural network based hierarchical autoencoder for nonlinear mode decomposition of fluid field data. *Phys. Fluids* **32** (9), 095110.

Furuya, T., de Hoop, M.V. & Peyré, G. 2024 Transformers are universal in-context learners. arXiv:2408.01367.

Gasparino, L., Muela, J. & Lehmkuhl, O. 2024*a* SOD2D repository. Available at https://gitlab.com/bsc_sod2d/sod2d_gitlab.

Gasparino, L., Spiga, F. & Lehmkuhl, O. 2024*b* SOD2D: a GPU-enabled spectral finite elements method for compressible scale-resolving simulations. *Comput. Phys. Commun.* **297**, 109067.

Geneva, N. & Zabaras, N. 2022 Transformers for modeling physical systems. *Neural Networks* **146**, 272–289.

Geshkovski, B., Rigollet, P. & Ruiz-Balet, D. 2024 Measure-to-measure interpolation using transformers. arXiv:2411.04551.

Güemes, A., Discetti, S., Ianiro, A., Sirmacek, B., Azizpour, H. & Vinuesa, R. 2021 From coarse wall measurements to turbulent velocity fields through deep learning. *Phys. Fluids* **33** (7), 075121.

Hijazi, S., Stabile, G., Mola, A. & Rozza, G. 2020 Data-driven POD-Galerkin reduced order model for turbulent flows. *J. Comput. Phys.* **416**, 109513.

Holmes, P.J., Lumley, J.L., Berkooz, G., Mattingly, J.C. & Wittenberg, R.W. 1997 Low-dimensional models of coherent structures in turbulence. *Phys. Rep.* **287** (4), 337–384.

Howell, J. 2015 Aerodynamic drag of passenger cars at yaw. *SAE Intl J. Passenger Cars - Mech. Syst.* **8** (1), 306–316.

Imtiaz, H. & Akhtar, I. 2020 Nonlinear closure modeling in reduced order models for turbulent flows: a dynamical system approach. *Nonlinear Dyn.* **99** (1), 479–494.

Kaptanoglu, A.A., Callaham, J.L., Aravkin, A., Hansen, C.J. & Brunton, S.L. 2021 Promoting global stability in data-driven models of quadratic nonlinear dynamics. *Phys. Rev. Fluids* **6**, 094401.

Kim, H., Kim, J., Won, S. & Lee, C. 2021 Unsupervised deep learning for super-resolution reconstruction of turbulence. *J. Fluid Mech.* **910**, A29.

Kuya, Y., Takeda, K., Zhang, X. & Forrester, A.I.J. 2011 Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. *AIAA J.* **49** (2), 289–298.

Ledig, C. *et al.* 2017 Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

LEHMKUHL, O., PARK, G.I., BOSE, S.T. & MOIN, P. 2018 Large-eddy simulation of practical aeronautical flows at stall conditions. In *Proceedings of the 2018 Summer Program*, Center for Turbulence Research, Stanford University. pp. 87–96.

LEHMKUHL, O., PIOMELLI, U. & HOUZEAUX, G. 2019 On the extension of the integral length-scale approximation model to complex geometries. *Intl J. Heat Fluid Flow* **78**, 108422.

LI, R., BORÉE, J., NOACK, B.R., CORDIER, L. & HARAMBAT, F. 2019 Drag reduction mechanisms of a car model at moderate yaw by bi-frequency forcing. *Phys. Rev. Fluids* **4**, 034604.

LITTLEWOOD, R. & PASSMORE, M. 2010 The optimization of roof trailing edge geometry of a simple square-back. *Tech. Rep.* 2010-01-0510, SAE International, USA.

LORITE-DÍEZ, M., JIMÉNEZ-GONZÁLEZ, J.I., PASTUR, L., CADOT, O. & MARTÍNEZ-BAZÁN, C. 2020 Drag reduction on a three-dimensional blunt body with different rear cavities under cross-wind conditions. *J. Wind Engng Ind. Aerodyn.* **200**, 104145.

LUMLEY, J.L. 1981 Rational approach to relations between motions of differing scales in turbulent flows. *Phys. Fluids* **10** (7), 1405.

MAULIK, R., FUKAMI, K., RAMACHANDRA, N., FUKAGATA, K. & TAIRA, K. 2020 Probabilistic neural networks for fluid flow surrogate modeling and data recovery. *Phys. Rev. Fluids* **5**, 104401.

MURATA, T., FUKAMI, K. & FUKAGATA, K. 2020 Nonlinear mode decomposition with convolutional neural networks for fluid dynamics. *J. Fluid Mech.* **822**, A13-1–A13-15.

OWEN, H. 2024 *HPC technology focus group summary. In, presentations of autoCFD, 4*. Barcelona Supercomputing Center, Belfast.

PRAKASH, A. & ZHANG, Y.J. 2024 Projection-based reduced order modeling and data-driven artificial viscosity closures for incompressible fluid flows. *Comput. Meth. Appl. Mech.* **425**, 116930.

REICHARDT, H. 1951 Vollständige darstellung der turbulenten geschwindigkeitsverteilung in glatten leitungen. *Z. Angew. Math. Mech.* **31** (7), 208–219.

SANCHIS-AGUDO, M., WANG, Y., DURAISAMY, K. & VINUESA, R. 2023 Easy attention: a simple self-attention mechanism for transformers, arXiv preprint arXiv:2308.12874.

SANDER, M.E. & PEYRÉ, G. 2024 Towards understanding the universality of transformers for next-token prediction. arXiv:2410.03011.

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

SIROVICH, L. 1987 Turbulence and the dynamics of coherent structures. Part I: coherent structures. *Q. Appl. Maths* **14**, 15–27.

SOLERA-RICO, A., VILA, S., CARLOS, G.-L., MIGUEL, W., YUNING, A., ABDULRAHMAN, D., SCOTT, T.M. & VINUESA, R. 2024 β-variational autoencoders and transformers for reduced-order modelling of fluid flows. *Nat. Commun.* **15** (1), 1361.

STABILE, G.& ROZZA, G. 2018 Finite volume POD-Galerkin stabilised reduced order methods for the parametrised incompressible Navier–Stokes equations. *Comput. Fluids* **173**, 273–284.

SUN, G. & WANG, S. 2019 A review of the artificial neural network surrogate modeling in aerodynamic design. *Proc. Inst. Mech. Engnrs G: J. Aerosp. Engng* **233** (16), 5863–5872.

TOMASETTO, M., WILLIAMS, J.P., BRAGHIN, F., MANZONI, A. & KUTZ, J.N. 2025 Reduced order modeling with shallow recurrent decoder networks. arXiv:2502.10930.

TOWNE, A., SCHMIDT, O.T. & COLONIUS, T. 2018 Spectral proper orthogonal decomposition and its relationship to dynamic mode decomposition and resolvent analysis. *J. Fluid Mech.* **847**, 821–867.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, L. & POLOSUKHIN, I. 2017 Attention is all you need. arXiv preprint arXiv:1706.03762.

VINUESA, R. & BRUNTON, S.L. 2022 Enhancing computational fluid dynamics with machine learning. *Nat. Comput. Sci.* **2** (6), 358–366.

VREMAN, A.W. 2004 An eddy-viscosity subgrid-scale model for turbulent shear flow: algebraic theory and applications. *Phys. Fluids* **16** (10), 3670–3681.

WANG, Y. 2023 Convolution-compacted vision transformers for prediction of local wall heat flux at multiple Prandtl numbers in turbulent channel flow. Master's thesis, KTH, Stockholm, Sweden.

WANG, Y., SOLERA-RICO, A., SANMIGUEL VILA, C. & VINUESA, R. 2024 Towards optimal eta-variational autoencoders combined with transformers for reduced-order modelling of turbulent flows. *Intl J. Heat Fluid Flow* **105**, 109254.

WANG, Z., AKHTAR, I., BORGGAARD, J. & ILIESCU, T. 2012 Proper orthogonal decomposition closure models for turbulent flows: a numerical comparison. *Comput. Meth. Appl. Mech.* **237–240**, 10–26.

WU, P., QIU, F., FENG, W., FANG, F. & PAIN, C. 2022 A non-intrusive reduced order model with transformer neural network and its application. *Phys. Fluids* **34** (11), 115130.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press

YONDO, R., ANDRÉS, E. & VALERO, E. 2018 A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses. *Prog. Aerosp. Sci.* **96**, 23–61.

YOUSIF, M.Z., ZHANG, M., YU, L., VINUESA, R. & LIM, H.C. 2023 A transformer-based synthetic-inflow generator for spatially developing turbulent boundary layers. *J. Fluid Mech.* **957**, A6.

ZHANG, B. 2023 Nonlinear mode decomposition via physics-assimilated convolutional autoencoder for unsteady flows over an airfoil. *Phys. Fluids* **5** (0164250), 95–115.

ZHANG, X., TOET, W. & ZERIHAN, J. 2006 Ground effect aerodynamics of race cars. *Appl. Mech. Rev.* **59** (1), 33–49.

https://doi.org/10.1017/jfm.2025.10610 Published online by Cambridge University Press