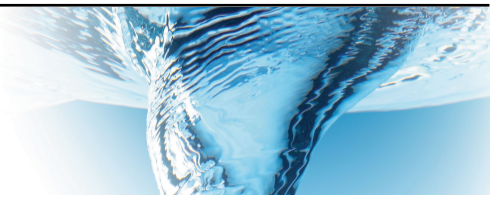
J. Fluid Mech. (2025), *vol.* 1010, R4, doi:10.1017/jfm.2025.383



Observable-augmented manifold learning for multi-source turbulent flow data

Kai Fukami[1] **and Kunihiko Taira**[2]

[1]Department of Aerospace Engineering, Tohoku University, Sendai 980-8579, Japan

[2]Department of Mechanical and Aerospace Engineering, University of California, Los Angeles, CA 90095, USA

Corresponding author: Kai Fukami, kfukami1@tohoku.ac.jp

(Received 29 January 2025; revised 2 April 2025; accepted 8 April 2025)

This study seeks a low-rank representation of turbulent flow data obtained from multiple sources. To uncover such a representation, we consider finding a finite-dimensional manifold that captures underlying turbulent flow structures and characteristics. While nonlinear machine-learning techniques can be considered to seek a low-order manifold from flow field data, there exists an infinite number of transformations between data-driven low-order representations, causing difficulty in understanding turbulent flows on a manifold. Finding a manifold that captures turbulence characteristics becomes further challenging when considering multi-source data together due to the presence of inherent noise or uncertainties and the difference in the spatiotemporal length scale resolved in flow snapshots, which depends on approaches in collecting data. With an example of numerical and experimental data sets of transitional turbulent boundary layers, this study considers an observable-augmented nonlinear autoencoder-based compression, enabling data-driven feature extraction with prior knowledge of turbulence. We show that it is possible to find a low-rank subspace that not only captures structural features of flows across the Reynolds number but also distinguishes the data source. Along with machine-learning-based super-resolution, we further argue that the present manifold can be used to validate the outcome of modern data-driven techniques when training and evaluating across data sets collected through different techniques. The current approach could serve as a foundation for a range of analyses including reduced-complexity modelling and state estimation with multi-source turbulent flow data.

Key words: machine learning, low-dimensional models, turbulent boundary layers

1. Introduction

Fluid mechanicians have developed approaches to measure turbulent flows in a high-fidelity manner. High-fidelity data sets collected from numerical simulations and experiments have been shared in the community (Li *et al.* 2008; Towne *et al.* 2023), supporting the development of approaches for fluid flow modelling and control. Such high-fidelity turbulent flow data involve their rich physics such as vortex dynamics, energy cascade processes, and anisotropic structures that evolve across a range of spatiotemporal length scales. While modern machine-learning techniques analyse high-fidelity turbulent flow data, it is challenging to evaluate data-driven models with data sets collected through different approaches since data characteristics such as measurement uncertainties and spatiotemporal length scales covered by each data source are different. Aiming to perform data-driven assessments of turbulent flows across different sources and scales, this paper proposes a nonlinear machine-learning technique that seeks a low-order representation of multi-source turbulent flow data.

To find such a low-rank representation, this study considers the concept of a manifold of turbulent flows. The long-term dynamics of infinite-dimensional turbulence systems often converge to a low-order manifold surface (Graham & Floryan 2021). This low-rank nature can be used to facilitate understanding and modelling of turbulent flows (Noack *et al.* 2003; Luchtenburg *et al.* 2009). Proper orthogonal decomposition (POD; Lumley 1967) has been considered in estimating a low-order manifold from turbulent flow data. However, finding a minimal representation of turbulence with such a linear technique becomes challenging as the variance of data increases because it linearly projects data onto a flat manifold. Increasing the Reynolds number of flows of interest falls under such a case since the difference between the minimum and maximum length scales is enlarged (Alfonsi & Primavera 2007). In dealing with multi-source turbulent flow data, including sensor readings, numerically simulated data, and experimental measurements, extracting features from data is further demanding due to measurement noise, uncertainties, the difference in degree of freedom and length scales captured by each data source.

In response, we consider a nonlinear autoencoder-based compression (Hinton & Salakhutdinov 2006) to seek a manifold capturing turbulence characteristics from various data sources. This technique achieves greater order reduction of turbulence compared with linear techniques, demonstrated with channel flow (Yousif *et al.* 2022), Kolmogorov turbulence (Racca *et al.* 2023) and a flow through an urban environment (Eivazi *et al.* 2022). In a reduced-order space, there often exists a submanifold forming a geometrical structure that compactly represents the characteristics and patterns underlying the flow data (Graham & Floryan 2021). Furthermore, such low-order representations can be used for modelling (Constante-Amores & Graham 2024) and controlling unsteady flows (Fukami *et al.* 2024*b*).

While a nonlinear autoencoder achieves significant order reduction of turbulent flows, the mapping from the input high-dimensional space to the latent subspace can involve various transformations even if their decoding leads to the same flow reconstruction. This study discusses the importance of incorporating prior knowledge of turbulent flow physics in the learning formulation of a nonlinear autoencoder to identify a submanifold that best describes the multi-source turbulent flow data in a low-order manner. As an example, the data sets of the transitional turbulent boundary layer collected by direct numerical simulations and particle image velocimetry are considered. We find that a nonlinear autoencoder with physics embedding provides a manifold that captures structural features of flows across the Reynolds number while distinguishing the data source. With machine-learning-based super-resolution, we further discuss that the present manifold can facilitate

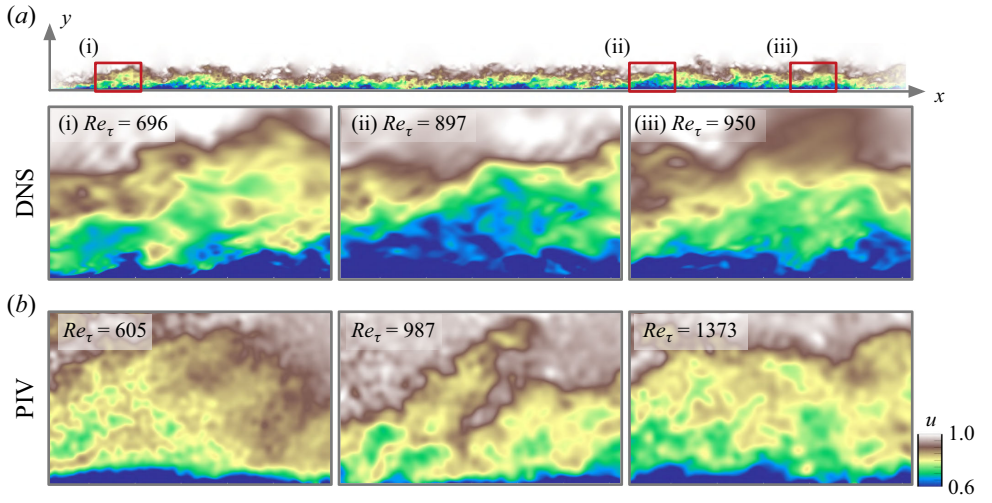


Figure 1. Transitional turbulent boundarylayer data sets. (*a*) Direct numerical simulation: part of the $x-y$ sectional domain and subdomains (i – iii) are shown. (*b*) Particle image velocimetry: three different Reynolds numbers are considered.

analysis, comparison and interpretation of turbulent flow data measured through different approaches.

This paper is organised as follows. The approach is described in § 2. Results of the manifold discovery for the transitional turbulent boundary-layer data sets are presented in § 3. Conclusions are offered in § 4.

2. Approach

This study develops a machine-learning approach for identifying a low-order submanifold that presents geometric structures capturing the essential features of turbulent flows, in hopes of facilitating data-driven analyses performed across different data sources. We consider the numerical and experimental data sets of a transitional turbulent boundary layer made available by the database of Towne *et al.* (2023). These data sets include time-resolved snapshots of a high-fidelity direct numerical simulation (DNS) across a wide domain in the streamwise direction and planar particle image velocimetry (PIV) measurements at different Reynolds numbers. As there is a variation of noise, measurement uncertainties, resolved length scales and Reynolds number across the data sets, the present set of DNS and PIV measurements calls for a comprehensive analysis for manifold discovery of multi-source turbulent flow data.

For the DNS, the zero-pressure-gradient flat-plate turbulent boundary layer is simulated by numerically solving the incompressible Navier–Stokes equations. The streamwise velocity field u is visualised in figure 1(*a*). The size of the computational domain is $(L_x, L_y, L_z)/\theta_{avg} = (469, 53, 79)$, where θ_{avg} is the streamwise-averaged momentum thickness, and L_x. L_y, and L_y are the streamwise, wall-normal, and spanwise extent of the computational domain. The recycling scheme by Lund *et al.* (1998) is used to generate the turbulence fluctuations at the inlet in which the recycling plane x_{ref}/θ_{avg} is set to 375, where x_{ref} is the location of the recycling plane for the inflow boundary condition. The friction Reynolds number $Re_\tau = u_\tau \delta/\nu$ spans from 481 to 1024 between the inlet and outlet, covering a range of turbulence characteristics such as nonlinear interactions and momentum exchange across the flow. All the variables are normalised by the kinematic

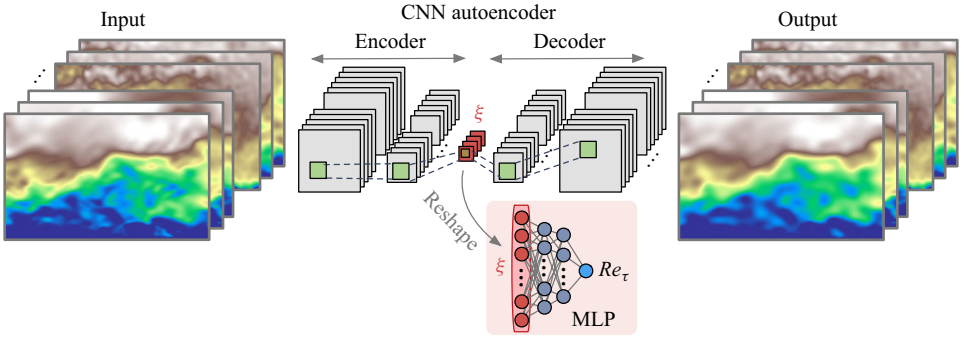


Figure 2. Observable-augmented autoencoder composed of convolutional neural network (CNN) and multi-layer perceptron (MLP) (Fukami & Taira 2023).

viscosity ν, the boundary-layer thickness δ at 99 % of the free-stream velocity, and the friction velocity at the wall $y=0$, $u_\tau=(\nu\partial\overline{u}^{z,t}/\partial y|_{y=0})^{1/2}$, where $\overline{(\cdot)}^{z,t}$ represents the average over the spanwise direction and time. Details of the simulation set-up are provided in Towne *et al.* (2023).

Time-resolved planar velocity fields measured by PIV in a wind tunnel are collected at three different Reynolds numbers of $Re_\tau=605$, 987 and 1373, as shown in figure 1(*b*). The friction velocity for the PIV cases is computed by the Clauser method (Clauser 1956). The free-stream turbulence intensity is approximately 0.5 % and the boundary layer is developed in a nominally zero-pressure-gradient environment. Details on camera set-ups and test sections are available in Towne *et al.* (2023).

In the present analysis, we consider the streamwise velocity u for both DNS and PIV data sets. This is intended to examine whether the dominant feature of the turbulent boundary layer mainly driven by the streamwise velocity can be extracted across the different data sources by machine learning. While the DNS is performed in a three-dimensional domain, subdomains sampled from an $x-y$ sectional field are used to align the data set-up with the PIV data, as shown in figure 1(*a*). The subdomain set-up will be explained later.

To seek a manifold that captures the physical features from these numerical and experimental data of the turbulent boundary layer, a nonlinear autoencoder-based data compression (Hinton & Salakhutdinov 2006) is performed. An autoencoder is trained to replicate the data between the input and output while possessing the bottleneck in the model architecture, as illustrated in figure 2. The data dimension at the bottleneck referred to as a latent vector $\boldsymbol{\xi}$ is generally much smaller than that of the input or output data. Hence, the latent vector $\boldsymbol{\xi}$ can be regarded as a compressed representation of the given data $\boldsymbol{q}$ if the autoencoder $\mathcal{F}_{AE}$ successfully reconstructs the data. With an encoder $\mathcal{F}_e$ and a decoder $\mathcal{F}_d$, the aforementioned process is expressed as

$$\boldsymbol{q}\approx\mathcal{F}_{AE}(\boldsymbol{q})=\mathcal{F}_d(\mathcal{F}_e(\boldsymbol{q})),\quad \boldsymbol{\xi}=\mathcal{F}_e(\boldsymbol{q}),\quad \boldsymbol{q}\approx\widehat{\boldsymbol{q}}=\mathcal{F}_d(\boldsymbol{\xi}),\tag{2.1}$$

where $\widehat{(\cdot)}$ denotes a reconstructed variable.

To promote the manifold identification capturing the development of turbulent boundary layers, we consider a friction Reynolds-number-based augmentation in performing the present nonlinear autoencoder-based compression, as illustrated in figure 2. In the formulation of a variable-based augmentation (Fukami & Taira 2023), an autoencoder is constructed to estimate an observable variable (Re_τ in the present study) from the latent vector $\boldsymbol{\xi}$ through a subnetwork depicted with a pink box in figure 2 while compressing the

data such that

$$\boldsymbol{w}^* = \mathrm{argmin}_{\boldsymbol{w}}[||\boldsymbol{q} - \hat{\boldsymbol{q}}||_2 + \beta||Re_\tau - \widehat{Re_\tau}||_2], \tag{2.2}$$

where $\boldsymbol{w}$ denotes the weights inside the autoencoder and β balances the flow field and Reynolds number loss terms. Here, the weights inside the main autoencoder and the subnetwork are simultaneously optimised. As an observable variable needs to be accurately estimated to minimise the above cost function, the optimal weights $\boldsymbol{w}^*$ are found to capture the coherent relationship between the given data $\boldsymbol{q}$ and an observable in the latent space, which has been used in a range of aerodynamic examples (Fukami & Taira 2023; Liu *et al.* 2024; Tran *et al.* 2024; Mousavi & Eldredge 2025; Eldredge & Mousavi 2025). This study uses $\beta = 0.05$ based on the L-curve analysis (Hansen & O'Leary 1993), which finds an appropriate regularisation parameter of the cost function. The case with $\beta = 0.01$ is also considered to observe how the latent description is modified through the observable augmentation. The current formulation can take not only an observable but also parameters derived from observables for variable-based augmentation, as performed in this study that uses Re_τ measured from the friction velocity u_τ.

Choosing the friction Reynolds number Re_τ is natural in our case as it characterises flow features in the wall-normal direction such as the balance between the near-wall region dominated by viscous effects and the outer region where turbulence is fully developed across the boundary layer. In other words, the primary features of the turbulent boundary layer in both the x and y directions can be incorporated in identifying a manifold that represents the flow characteristics across multiple data sources. Although not considered in this study, the momentum thickness Reynolds number Re_θ may also provide a similar result in the manifold identification for the present case as it increases across the streamwise direction with the friction Reynolds number.

For the present data-driven analyses, both DNS and PIV data are sampled such that the subdomain size in the streamwise and wall-normal directions becomes $L_{x,ML}/\delta \in [1, 3]$ and $L_{y,ML}/\delta \approx 1$, respectively. The DNS data are uniformly interpolated in the wall-normal direction to align the set-up of the PIV data. These collected data are then resized to be $N_{ML}^2 = 128^2$, where N_{ML} is the number of grid points for the resized data used in the data-driven analysis. As the current study aims to find a manifold assessing the data similarities across different sources, the present observable-augmented autoencoder is trained with both DNS and PIV snapshots together. We consider 19 000 snapshots for the present analysis, including 10 000 DNS samples composed of 10 subdomains per snapshot across 1000 time-resolved frames and 3000 PIV snapshots for each of three Reynolds numbers. We use 70 % of the snapshots for training and the remaining 30 % for validation with random splitting of the data set. The encoder and decoder parts of the autoencoder are constructed by convolutional neural networks (LeCun *et al.* 1998) while the subnetwork is composed of multi-layer perceptrons (Rumelhart *et al.* 1986), analogous to the original observable autoencoder network in Fukami & Taira (2023). A sample code (https://github.com/kfukami/Observable-CNN-AE) can be referred to for further details on the present autoencoder model.

3. Results

We apply the current autoencoder with the friction Reynolds-number-based augmentation to the DNS and PIV data sets of the turbulent boundary layer. The reconstructed fields through the observable-augmented autoencoder with $\beta = 0.05$ across different numbers of the latent dimension n_ξ are presented in figure 3. As representative cases, a DNS snapshot at $Re_\tau = 897$ and a PIV snapshot at $Re_\tau = 605$ are shown. The L_2 reconstruction error

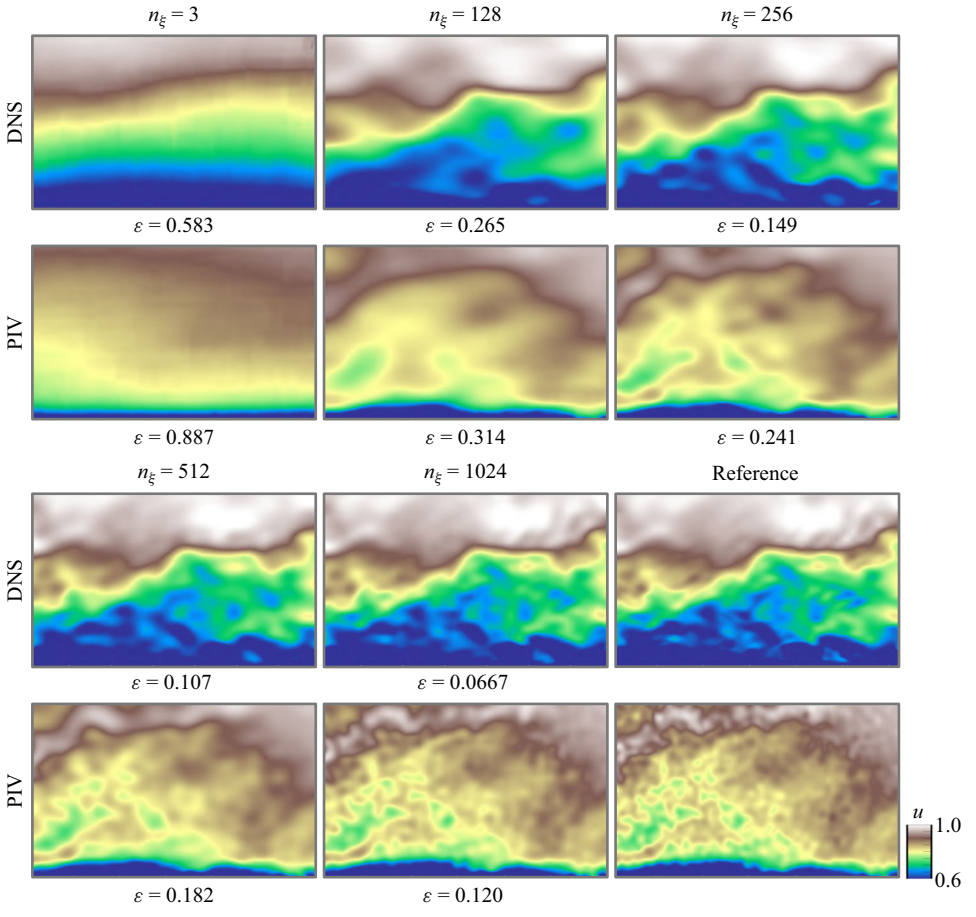


Figure 3. Representative reconstructed fields through nonlinear autoencoder compression across the latent dimension $n_{\boldsymbol{\xi}}$. The DNS ($Re_\tau = 897$) and PIV ($Re_\tau = 605$) fields are shown. The L_2 reconstruction error ε is reported underneath each field.

$\varepsilon = ||\boldsymbol{q} - \mathcal{F}_{AE}(\boldsymbol{q})||_2/||\boldsymbol{q}||_2$ is reported under each of the reconstructed velocity fields. The error decreases as the latent dimension increases, corresponding to the low compression of velocity data. While large-scale structures can be represented only with $n_{\boldsymbol{\xi}} \leqslant 256$, fine-scale features are captured well across the field with $n_{\boldsymbol{\xi}} \geqslant 512$.

As a reference, we note that the linear POD requires more than 3000 modes to achieve the same reconstruction level as the present nonlinear autoencoder with $n_{\boldsymbol{\xi}} = 512$. In other words, the use of nonlinear activation functions inside the model improves compression in seeking low-order representations of the turbulent boundary-layer data sets. Based on the L_2 error norm and the reconstruction of fine-scale structures, we hereafter choose the latent dimension of 512 for the discussions.

Next, we are interested in what can be captured through the present nonlinear compression of multi-source turbulent boundary-layer data. Let us perform POD on the compressed representation $\boldsymbol{\xi}$ to examine the primary features extracted from the turbulent flows with the autoencoder. The coordinate composed of three dominant POD coefficients $a_1 - a_2 - a_3$ is shown in figure 4. In addition to the cases of the observable autoencoder with $\beta = 0.01$ and 0.05, a regular autoencoder without the Reynolds-number augmentation, i.e. $\beta = 0$, is also presented for comparison.

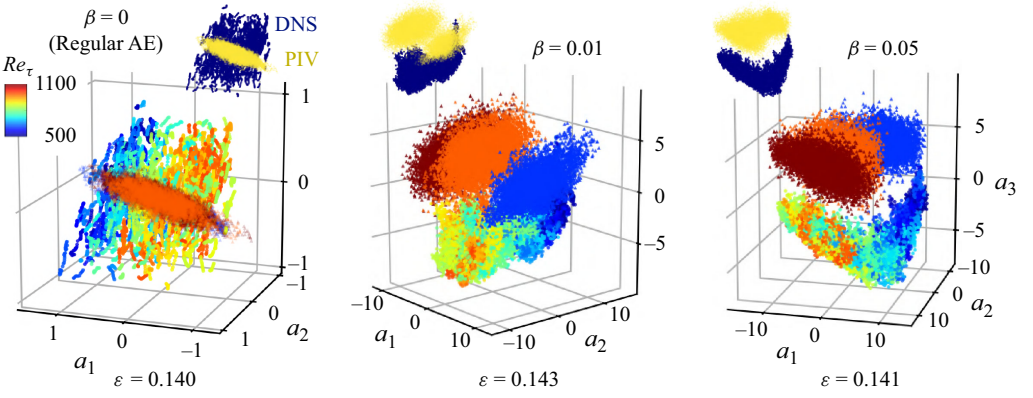


Figure 4. Cross-source manifold identified through the observable-augmented nonlinear autoencoder (AE). The coefficient space composed of the three dominant POD modes is shown. The L_2 error ε averaged over the samples is reported underneath each coefficient space.

All three spaces learn the friction Reynolds-number dependence for the DNS data. The clear trend with respect to Re_τ is observed in the a_1 axis with $\beta \leqslant 0.01$ and the diagonal direction on the $a_1 - a_2$ plane with $\beta = 0.05$. The main difference between the cases with and without the Reynolds-number augmentation is their low-order expression about the data source. In the space of a regular autoencoder, the PIV cases generally overlay the centre region of the DNS cases. In contrast, there is a clear distinction between the DNS and PIV cases by introducing the Reynolds-number-based augmentation. The a_3 axis with $\beta = 0.01$ and 0.05 captures the difference in the data source.

Noteworthy here is that the reconstruction performance over the three cases is almost the same, presenting the L_2 error of 14 % as reported in figure 4. In other words, while all nonlinear autoencoders achieve similar compression of the turbulent flows, what they learn in the latent space from the data becomes different from each other. With the friction Reynolds-number-based augmentation, the current submanifold compactly represents the Reynolds number and the difference in the data source as their characteristics of given data sets. In addition, the latent expression with $\beta = 0.05$ provides a continuous transition between the data sources of DNS and PIV, forming a V-shape submanifold of the DNS data. Since this submanifold extracts structural characteristics from high-resolution flow fields in addition to the Reynolds-number dependence, there are some variations of Re_τ at $-10 < a_1 < 0$ in the latent space with $\beta = 0.05$.

To further examine how the present manifold can be used for assessing data-driven techniques with multi-source turbulent flow data, we consider machine-learning-based super-resolution analysis of turbulent flows (Fukami *et al.* 2019). Super-resolution aims to reconstruct high-resolution data from the corresponding low-resolution signal, which has been examined for turbulent flow reconstruction (Fukami *et al.* 2023). The process of super-resolution reconstruction in the context of fluid flows is expressed as $\boldsymbol{q}_{HR} = \mathcal{F}_{SR}(\boldsymbol{q}_{LR})$, where the high-resolution flow field $\boldsymbol{q}_{HR}$ is reconstructed from the low-resolution flow data $\boldsymbol{q}_{LR}$ with a machine-learning model $\mathcal{F}_{SR}$.

To recognise turbulent flow structures that have been learned as the difference between the DNS and PIV data sets on the manifold, we use the machine-learning-based super-resolution reconstruction in the following manner. The present super-resolution model is first trained with the DNS data only and then evaluated with the PIV data. We then examine how the super-resolved flow fields from PIV data are described on the identified manifold through the nonlinear autoencoder.

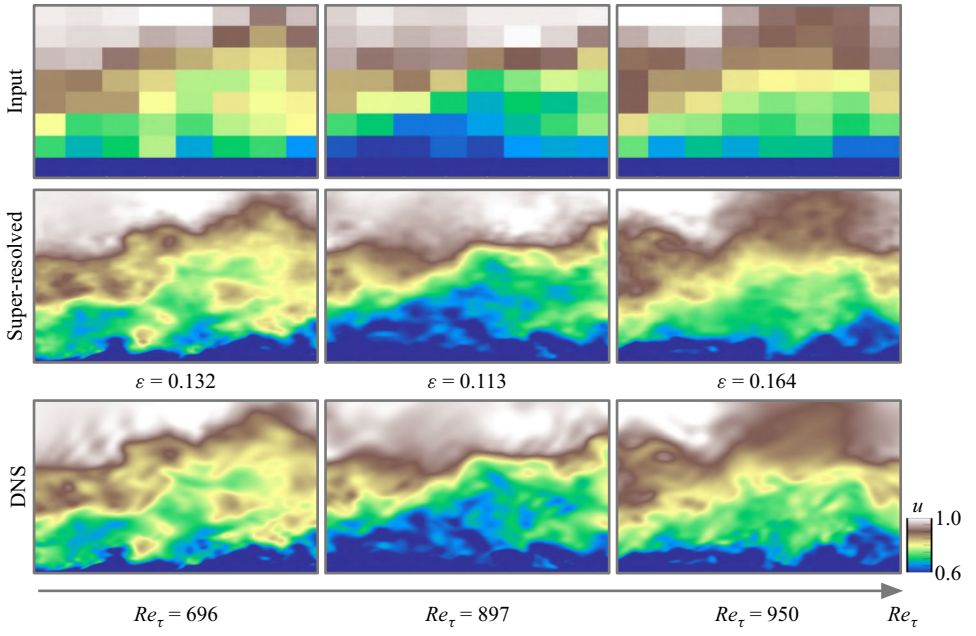


Figure 5. Machine-learning-based super-resolution reconstruction for DNS data of turbulent boundary layers. The L_2 reconstruction error ε is reported underneath each field.

For successful super-resolution reconstruction of turbulent flows, the machine-learning model $\mathcal{F}_{SR}$ should be carefully constructed to accommodate a range of spatial length scales while accounting for scale invariance of turbulent flow structures (Fukami *et al.* 2024*a*). We use the interconnected hybrid downsampled skip-connection/multi-scale (DSC/MS) model (Fukami *et al.* 2023) based on CNN (LeCun *et al.* 1998). This model is aimed at capturing rotational and translational invariance while incorporating multi-size filter operations that greatly assist in learning the correlation across a range of length scales in turbulent flows. We refer to Fukami *et al.* (2023) and a sample code (http://www.seas.ucla.edu/fluidflow/codes.html) for further details on the present super-resolution model. In this study, the model is trained to reconstruct the high-resolution velocity field of size 128^2 from the corresponding low-resolution data of size 8^2 generated by average pooling (Fukami *et al.* 2019).

Let us assess the present super-resolution model with the DNS data, as shown in figure 5. The super-resolution model $\mathcal{F}_{SR}$ is trained with 25 000 samples composed of 10 subdomains per snapshot across 2500 time-resolved DNS frames. The DSC/MS model accurately produces the fine-scale structures in the flow fields across the Reynolds number. The super-resolution model $\mathcal{F}_{SR}$ with the DNS data is then tested with the PIV data, as presented in figure 6. Here, the PIV fields are downsampled to be the size 8^2, and then these low-resolution PIV input data $\boldsymbol{q}_{LR,PIV}$ are given to the model $\mathcal{F}_{SR}$. While the reconstructed fields are generally similar to the reference PIV field $\boldsymbol{q}_{PIV}$ across the Reynolds number, their velocity profiles exhibit slight differences near the wall. This is further evident from the root mean square of streamwise velocity fluctuation u_{rms} across the wall-normal direction shown in figure 6. Here, the u_{rms} profile is plotted across $y^+ = yu_\tau/\nu$ obtained from the PIV data set-up. The profile of the PIV data presents overestimation near the wall due to the reflection of the laser sheet on the surface in

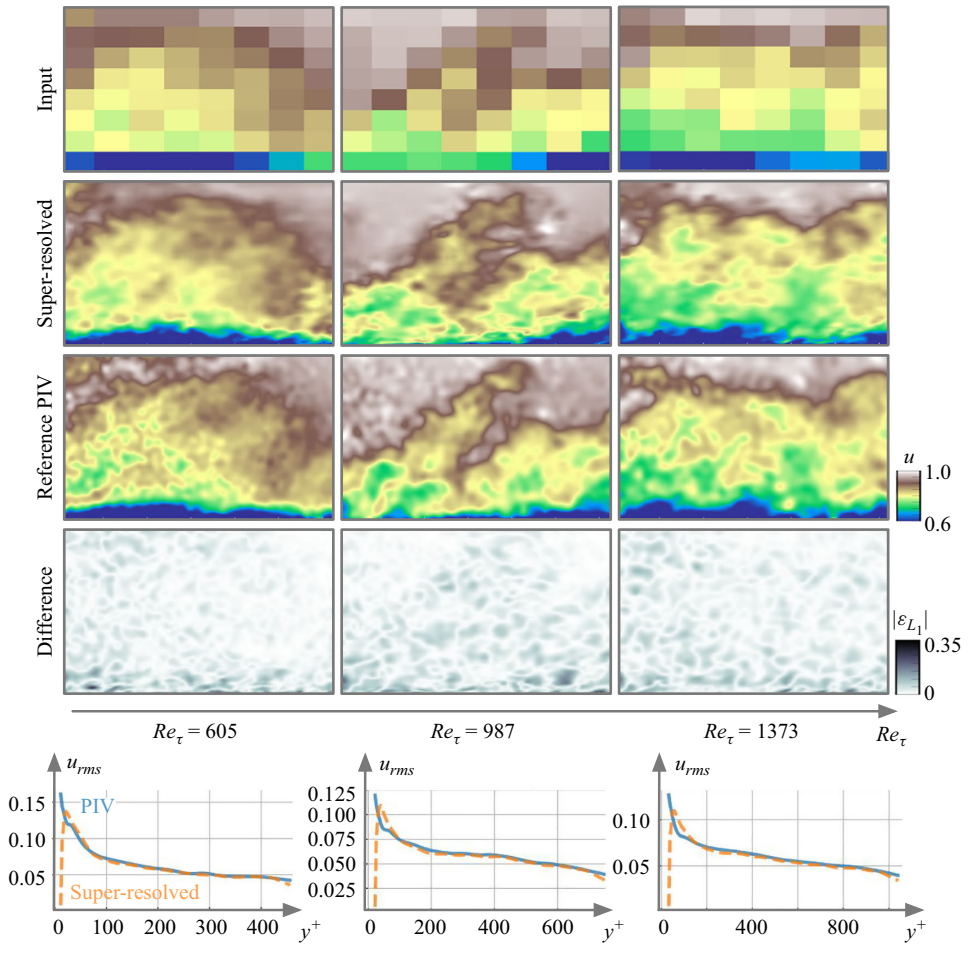


Figure 6. Application of machine-learning-based super-resolution model $\mathcal{F}_{SR}$ trained with the DNS data to the PIV data sets. The reconstructed fields, the L_1 difference field between the reconstructed and original PIV $|\varepsilon_{L_1}| = |\boldsymbol{q}_{PIV} - \mathcal{F}_{SR}(\boldsymbol{q}_{LR,PIV})|$, and the root mean square of streamwise velocity fluctuation u_{rms} across the wall-normal direction are shown.

measuring the data (Towne *et al.* 2023). This overestimation is corrected by the super-resolution model trained with the DNS data possessing a finer spatial resolution near the wall. The PIV cases of $Re_\tau = 987$ and 1373 do not resolve the near-wall region of $y^+ \approx 15$, where the inner peak of u_{rms} resides, even at the first grid point from the wall. In other words, the super-resolution model performs some nonlinear stretching of data in the wall-normal direction such that the inner peak at $y^+ \approx 15$ can be captured, despite not being resolved with PIV.

At last, we provide these super-resolved PIV data to the encoder $\mathcal{F}_e$ of the observable autoencoder with $\beta = 0.05$ to examine the behaviour on the identified low-rank coordinate such that $\boldsymbol{\xi}_{SRPIV} = \mathcal{F}_e(\mathcal{F}_{SR}(\boldsymbol{q}_{LR,PIV}))$. The coefficient space composed of the three dominant POD modes for the latent vector $\boldsymbol{\xi}_{SRPIV}$ is shown in figure 7. Due to the correction near the wall, the super-resolved PIV data are projected between the DNS and PIV data points in the coefficient space. This suggests that the nonlinear autoencoder $\mathcal{F}_{AE}$ learns the difference in the resolved length scales near the wall across the a_3 direction to distinguish the data sources. Although the decoder does not receive Re_τ

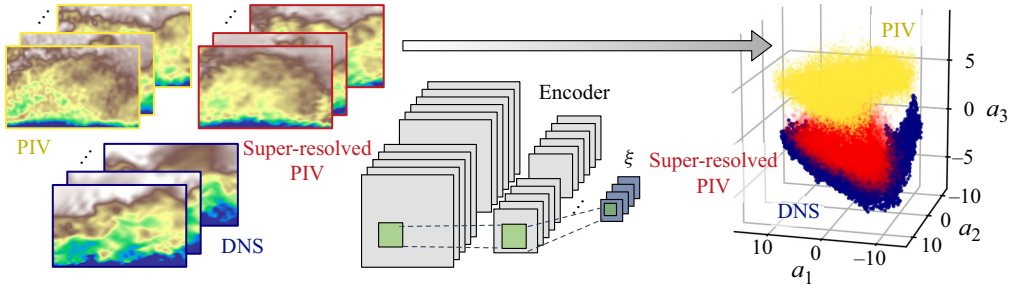


Figure 7. Assessment of super-resolved PIV fields on the identified coordinate. The coefficient space composed of the three dominant POD modes is shown.

as an input, the latent space of the observable-augmented autoencoder includes the information of Re_τ as presented in figure 4, thereby reflecting the reconstructed inner peak from the data in the latent space as the projection between DNS and PIV. In other words, the identified manifold gains the feature of given turbulence data in a compact manner, enabling standardised evaluation of machine-learning analysis when considering multi-source turbulent flow data together.

We emphasise that the present finding of the manifold is achieved by the observable augmentation with Re_τ that characterises the balance of turbulent flow features between the near-wall and outer regions while providing information on the development of turbulent boundary layers. This selection of observable variables capturing the flow features in both the x and y directions is based on our foundational understanding of turbulent boundary-layer flows. Rather than naively applying data-science techniques without consideration, appropriately incorporating prior knowledge into the model design is essential to learn multi-source turbulent flow data with nonlinear machine learning.

4. Concluding remarks

This study considered nonlinear machine learning to seek a low-rank manifold that captures the characteristics of multi-source turbulent flow data. With an example of transitional turbulent boundary-layer data from both numerical and experimental sources, we found that a low-order subspace is identified through nonlinear autoencoder-based compression with an observable augmentation. In the identified space, the Reynolds-number dependence and the difference in the resolved length scale in a flow snapshot are compactly represented. The observable-augmented autoencoder provides a continuous transition between numerical and experimental data in the latent subspace while extracting structural features of turbulent flows in a low-order manner, which differs from what can be achieved through a standard classifier. Furthermore, the present coordinate enables the comparison of the data even with different spatial resolutions from multiple sources, which is challenging with a traditional norm. We showed that the super-resolved experimental flow fields that have no comparable solutions can be assessed in the identified subspace through projection. The current findings further support the importance of considering prior knowledge for data-driven studies in learning multi-source turbulent flow data.

The present manifold learning across multi-source turbulent flow data may be considered for analysis at a range of Reynolds numbers by carefully preparing the data sets with respect to their spatiotemporal resolution and Reynolds numbers. To make the identified subspace more robust, multi-fidelity data at different Reynolds numbers obtained from numerical simulations such as large-eddy simulations and detached eddy

simulations could be incorporated. As experiments can typically achieve higher Reynolds numbers compared with numerical simulations, one can consider manifold learning with the data sets composed of DNS data at low Reynolds numbers and PIV data at high Reynolds numbers. The resulting manifold with such data pairs would learn the characteristics of spatially high-resolved DNS at low Reynolds numbers and that of spatially low-resolved PIV at high Reynolds numbers. Leveraging such a latent space, super-resolution reconstruction with respect to not only spatial resolution but also Reynolds number could be performed through the decoder part of the observable-augmented autoencoder.

While this study primarily used flow field data obtained by DNS and PIV, sensor measurements could be further integrated into the coordinate identification process as either input/output data or variables used for the subnetwork-based augmentation. The denoising capability of autoencoder-based techniques also encourages multi-source data analysis of turbulent flows in seeking low-rank representations (Smith *et al.* 2024). Although the current analysis used POD to identify the primary characteristics in the latent subspace, sensitivity analysis (Mo *et al.* 2024) and geometric analysis with magnification factors (Kelshaw & Magri 2024) may also be helpful to examine the latent space characteristics. Depending on the physics and flow of interest, a careful choice of turbulent flow data given into nonlinear machine-learning models is necessary as well as preparing appropriate learning formulations to embed prior knowledge of turbulence. With this in mind, we can study multi-source turbulent flow data with nonlinear machine-learning techniques. The proposed approach to seeking a manifold of multi-source turbulent flow data offers guidance on how to examine turbulence characteristics across data sets with varying spatiotemporal data resolutions, measurement techniques and length scales, enhancing the reliability of cross-method comparisons and comprehensive data-fusion analyses of turbulence.

Funding. K.T. thanks the support from the US Air Force Office of Scientific Research (FA9550-21-1-0178) and the US Department of Defense Vannevar Bush Faculty Fellowship (N00014-22-1-2798). Part of the machine-learning analysis was performed on Delta GPU at the National Center for Supercomputing Applications (NCSA) through the ACCESS program (Allocation PHY230125).

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

References

ALFONSI, G. & PRIMAVERA, L. 2007 The structure of turbulent boundary layers in the wall region of plane channel flow. *Proc. R. Soc. Lond. A* **463** (2078), 593–612.

CLAUSER, F.H. 1956 The turbulent boundary layer. *Adv. Appl. Mech.* **4**, 1–51.

CONSTANTE-AMORES, C.R. & GRAHAM, M.D. 2024 Data-driven state-space and Koopman operator models of coherent state dynamics on invariant manifolds. *J. Fluid Mech.* **984**, R9.

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.

ELDREDGE, J.D. & MOUSAVI, H. 2025 A review of Bayesian sensor-based estimation and uncertainty quantification of aerodynamic flows. arXiv: 2502.20280.

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. 2023 Super-resolution analysis via machine learning: a survey for fluid flows. *Theor. Comput. Fluid Dyn.* **37** (4), 421–444.

FUKAMI, K., GOTO, S. & TAIRA, K. 2024*a* Data-driven nonlinear turbulent flow scaling with Buckingham Pi variables. *J. Fluid Mech.* **984**, R4.

FUKAMI, K., NAKAO, H. & TAIRA, K. 2024*b* Data-driven transient lift attenuation for extreme vortex gust–airfoil interactions. *J. Fluid Mech.* **992**, A17.

FUKAMI, K. & TAIRA, K. 2023 Grasping extreme aerodynamics on a low-dimensional manifold. *Nat. Commun.* **14** (1), 6480.

GRAHAM, M.D. & FLORYAN, D. 2021 Exact coherent states and the nonlinear dynamics of wall-bounded turbulent flows. *Annu. Rev. Fluid Mech.* **53** (1), 227–253.

HANSEN, P.C. & O'LEARY, D.P. 1993 The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* **14** (6), 1487–1503.

HINTON, G.E. & SALAKHUTDINOV, R.R. 2006 Reducing the dimensionality of data with neural networks. *Science* **313** (5786), 504–507.

KELSHAW, D. & MAGRI, L. 2024 Proper latent decomposition, arXiv: 2412.00785.

LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. 1998 Gradient-based learning applied to document recognition. *Proc. IEEE* **86** (11), 2278–2324.

LI, Y., PERLMAN, E., WAN, M., YANG, Y., MENEVEAU, C., BURNS, R., CHEN, S., SZALAY, A. & EYINK, G. 2008 A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *J. Turbul.* **9**, N31.

LIU, Z., BECKERS, D. & ELDREDGE, J.D. 2024 Model-based reinforcement learning for control of strongly-disturbed unsteady aerodynamic flows. arXiv: 2408.14685.

LUCHTENBURG, D.M., GÜNTHER, B., NOACK, B.R., KING, R. & TADMOR, G. 2009 A generalized mean-field model of the natural and high-frequency actuated flow around a high-lift configuration. *J. Fluid Mech.* **623**, 283–316.

LUMLEY, J.L. 1967 The structure of inhomogeneous turbulent flows. In *Atmospheric Turbulence and Radio Wave Propagation* (ed. YAGLOM, A.M. & TATARSKI, V.I.), Nauka.

LUND, T.S., WU, X. & SQUIRES, K.D. 1998 Generation of turbulent inflow data for spatially-developing boundary layer simulations. *J. Comput. Phys.* **140** (2), 233–258.

MO, Y., TRAVERSO, T. & MAGRI, L. 2024 Decoder decomposition for the analysis of the latent space of nonlinear autoencoders with wind-tunnel experimental data. *Data Centric Engng* **5**, e38.

MOUSAVI, H. & ELDREDGE, J.D. 2025 Low-order flow reconstruction and uncertainty quantification in disturbed aerodynamics using sparse pressure measurements, arXiv: 2501.03406.

NOACK, B.R., AFANASIEV, K., MORZYNSKI, M., TADMOR, G. & THIELE, F. 2003 A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. *J. Fluid Mech.* **497**, 335–363.

RACCA, A., DOAN, N.A.K. & MAGRI, L. 2023 Predicting turbulent dynamics with the convolutional autoencoder echo state network. *J. Fluid Mech.* **975**, A2.

RUMELHART, D.E., HINTON, G.E. & WILLIAMS, R.J. 1986 Learning representations by back-propagation errors. *Nature* **322**, 533–536.

SMITH, L., FUKAMI, K., SEDKY, G., JONES, A. & TAIRA, K. 2024 A cyclic perspective on transient gust encounters through the lens of persistent homology. *J. Fluid Mech.* **980**, A18.

TOWNE, A. *et al.* 2023 A database for reduced-complexity modeling of fluid flows. *AIAA J.* **61** (7), 2867–2892.

TRAN, J., FUKAMI, K., INADA, K., UMEHARA, D., ONO, Y., OGAWA, K. & TAIRA, K. 2024 Aerodynamics-guided machine learning for design optimization of electric vehicles. *Commun. Engng* **3** (1), 174.

YOUSIF, M.Z., YU, L. & LIM, H. 2022 Physics-guided deep learning for generating turbulent inflow conditions. *J. Fluid Mech.* **936**, A21.