OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 016026 (16pp)
https://doi.org/10.1088/1361-6501/ad9046

Turbulence statistics estimation across a step change in roughness via interpretable network-based modelling

Giovanni Iacobello[1,*], Marco Placidi[1], Shan–Shan Ding[1,2] and Matteo Carpentieri[1]

[1] School of Mechanical Engineering Sciences, University of Surrey, GU2 7XH, Guildford, United Kingdom
[2] Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, OX1 3PU Oxford, United Kingdom

E-mail: g.iacobello@surrey.ac.uk

Received 29 January 2024, revised 28 October 2024
Accepted for publication 8 November 2024
Published 20 November 2024

Abstract

This study proposes a data-driven methodology to complement existing time-series measurement tools for turbulent flows. Specifically, a cluster-based transition network model is employed for the estimation of velocity time traces and their corresponding statistics. The method is tested on a laboratory-modelled turbulent boundary layer over a step change in surface roughness, where velocity time series are recorded for training and validation purposes via Laser Doppler Anemometry. Results show that our approach can estimate velocity and momentum flux statistics within experimental uncertainty over a rough surface through an unsupervised approach, and across the step change in roughness through a semi-supervised variant. The friction velocity across the domain is also estimated with 10% relative error compared to the measured value. The proposed methodology is interpretable and robust against the main methodological parameters. A reliable data-driven framework is hence provided that can be integrated within existing laboratory setups to supplement or partially replace measurement systems, as well as to reduce wind tunnel running times.

Keywords: turbulence, boundary layer, transition networks, estimation

1. Introduction

Data-driven approaches to scientific discovery have witnessed a huge boost in applications and methodological variants in the last decade, with a significant impact on many disciplines. Fluid mechanics in particular, owing to its crucial role in a large variety of natural phenomena and industrial applications, has featured in several studies adopting data-driven approaches [1]. In recent years, a wealth of studies have tackled fluid mechanics problems from a data-driven perspective—e.g. through machine learning, Kalman filter-based, stochastic estimation-based, or Bayesian approaches, to name a few—both from a numerical and experimental perspective (the reader is referred to [2, 3] and references therein for an up to date overview).

Specifically, one of the potential benefits of data-driven approaches is the ability to support existing tools for data generation, namely, computational fluid dynamics (CFD) for numerical simulations and measurement tools for experiments [1–3]. The latter has recently been reviewed by Discetti and Liu [4], who provided a thorough overview of the state-of-the-art machine learning approaches to flow field measurements, highlighting achievements and perspectives of machine

* Author to whom any correspondence should be addressed.

learning applications with a particular emphasis on particle image velocimetry (PIV). In spite of the recent progress in data-driven fluid mechanics, however, some issues are yet to be fully resolved [2–4]; examples are, among others, improving the interpretability, accuracy, and robustness of data-driven methodologies, as well as reducing the amount of training data.

This study aims to contribute to the recent developments in the data-driven enhancement of flow measurements, by employing an interpretable and robust data-driven methodology for the estimation of time traces in turbulent flows and their corresponding statistics. Data from Laser Doppler Anemometers (LDA) are used for training and testing purposes, with the goal of assessing the proposed methodology's capability in estimating the third velocity component when the remaining two are known. This goal is motivated by the practical argument that two-component LDA systems—as well as other pointwise systems such as cross hot-wire anemometers—are cheaper and easier to set up in laboratory experiments compared to their three-component counterparts. This is particularly an issue in case of experiments on non-canonical flows, e.g. in the presence of bulky models, as in urban atmospheric flows [5]. Three-component systems require additional experimental complexity which, in the case of LDA systems (known to be non-obstructive measurement systems), implies multiple laser beams to be focused at the measurement volume, often limiting its applicability in many applications due to the inherent experimental setup limitations (e.g. due to blockage or the necessity to measure close to a surface). More generally, sensor intrusiveness is a widespread issue affecting several disciplines in measurement science beyond fluid mechanics, as the presence of sensors creates a disturbance (e.g. modifying the characteristics of the sensor output or introducing some extra noise) that needs to be (pre- or post-) processed.

Besides potential benefits associated with experimental constraints, the proposed data-driven approach can be used to significantly reduce wind tunnel running times, thus cutting down on measurement-related costs or supplementing two-dimensional measurements. As discussed in section 6, measurement time can be reduced by an order of magnitude with the approach presented herein by relying only on a limited amount of training data and still retaining a satisfactory level of accuracy (i.e. comparable with the experimental uncertainty).

To achieve the aforementioned goal, a transition network-based approach is employed to estimate velocity statistics of a turbulent boundary layer flow over a step-change in roughness (section 2). The term *estimation* here refers to the data-driven generation of velocity statistics in spatial locations of the flow that are *not* trained by (i.e. are unknown to) the network. Transition networks rely on a state-space representation of the dynamical system (here, turbulent boundary layer) and on transitional probabilities to switch from one state to another state [6], and they have recently drawn the attention of the fluid mechanics community alongside other complex network-based approaches [7, 8]. Transition networks have indeed recently been used as an alternative to more widespread machine learning architectures (specifically for flow reconstruction and estimation [9–15]), and successfully employed for flow reconstruction, estimation, as well as flow control [16–22].

Fernex *et al* [18] showed that the combination of state-space clustering and transition networks provides a robust data-driven model for reconstructing the dynamics of complex systems, including fluid flows. Cluster-based transition networks have successively been further enhanced and tested on various applications [21–26]. Like previous studies employing machine learning approaches to advance measurement science in various fields [27–29], the cluster-based network approach is considered to be generalisable to a large variety of applications, even beyond the realm of fluid mechanics. The main requirement of the method is the availability of a data-set of time series to construct the corresponding state space, while the specific features of the approach (e.g. accuracy or computational cost) will depend on the specific application. In this respect, the interpretability of the method is a significant feature that allows the method to be adapted in other research areas involving experimental measurements.

Specifically, unsteady aerodynamic load estimation was recently performed by Iacobello, Kaiser, & Rival [23] exploiting a weighted average-based (WAB) transition network with sparse sensors. The present builds upon these works, by adapting the WAB approach for turbulent velocity statistics (section 3). Compared to previous studies using transition networks, a new weight (ω_g, see section 3.4) is introduced here, which allows (i) to improve the statistics estimation accuracy of the unsupervised WAB approach, and (ii) to deliver a semi-supervised WAB approach for more complex scenarios.

A turbulent boundary layer flow developing over a step-change in roughness is used as a main test case, which is a representative scenario for complex flows over heterogeneous surfaces as occurring in many natural phenomena; these include, e.g. atmospheric flows over vegetation or urban canopies, as well as offshore winds encountering the coast line [30–33]. The demonstration of the capabilities of our methodology for the chosen flow field, therefore, can have tangible implications for the modelling of turbulent flows for the aforementioned applications. While this is a significant aspect of our chosen setup, it is worth noting that our methodology can be much more generally applied as demonstrated in other studies [18, 26].

Unsupervised and semi-supervised WAB approaches are employed to estimate the main statistics (i.e. mean value, standard deviation, higher-order moments) of vertical velocity w and momentum flux $u'w'$ (where u is the streamwise velocity and $'$ denotes zero-mean temporal fluctuations) across the boundary layer domain when only limited vertical profiles are known (section 2). The vertical velocity w is selected in this study as the target variable owing to its importance in understanding the development of the internal boundary layer, as well as due to its involvement in the momentum flux and,

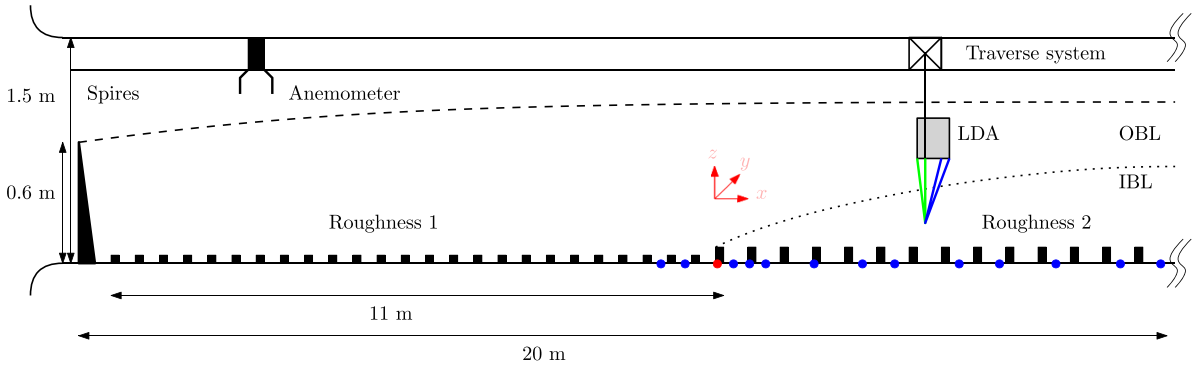

Figure 1. Schematic of the experimental setup with main dimensions. Dimensions are not to scale. OBL and IBL stand for Outer Boundary Layer and Internal Boundary Layer, respectively [30]. Blue dots indicate the measurements' positions, with the red dot indicating the origin of the right-hand coordinate system (shifted upward for ease of visualisation).

in turn, the evaluation of the friction velocity [30]. From a measurement science perspective, there are various challenges associated with the vertical velocity measurements in turbulent boundary layers (particularly at the interface of a step change in roughness), e.g. due to the orientation of the LDA probe and its laser beam pairs. Moreover, from a methodological point of view, w values tend to occupy a limited region in the state space, thus providing a representative example of dense state space (as discussed in section 5) making the accurate signal estimation a challenging task.

It is worth stressing that, for the proposed aim of reducing wind tunnel times, there is not a hierarchy in the importance of the velocity components, and the choice of the target variable should be dictated by problem-specific arguments (e.g. the flow setup) and/or by methodological (data-driven) considerations. For the flow considered here, the streamwise and spanwise velocity components turn out to be informative variables for our methodology, so they are used as training variables rather than as target variables.

Overall, the WAB approach is interpretable and robust against the main user-defined parameters (see section 4 and appendix A), making it a desirable candidate tool for data-driven enhancement of experimental measurements of fluid flows. While we do recognise that additional work is still necessary, as an ultimate goal, we envision exploiting cluster-based transition networks—integrated with machine learning architectures—for a robust, accurate, and (more importantly) generalisable data-driven representation of turbulent flows, which can eventually be used in everyday laboratory activities in support of existing measurement techniques.

2. Experimental facility and measurements

Experimental data were collected in the EnFlo wind tunnel within the Centre for Aerodynamics and Environmental Flow at the University of Surrey [34]. This is a unique open-return facility, with a test section of $20\,\mathrm{m} \times 3.5\,\mathrm{m} \times 1.5\,\mathrm{m}$ (streamwise $\times$ spanwise $\times$ wall-normal), which allows the modelling of atmospheric boundary layers with neutral and non-neutral thermal characteristics and has been extensively validated (e.g. see [5, 30, 35, 36] among others). Figure 1 shows a schematic of the wind tunnel's test section and the main features of the experimental setup.

Artificially-thickened boundary layers are developed by employing a combination of Irwin spires [37] and roughness elements. A step change in surface roughness is introduced 11 m downstream (figure 1) of the tunnel inlet. The first roughness has $50\,\mathrm{mm} \times 16\,\mathrm{mm} \times 5\,\mathrm{mm}$ elements (width $\times$ height $\times$ depth) standing on the cross area of $50\,\mathrm{mm} \times 5\,\mathrm{mm}$ in 50% staggered pattern with a spacing of 510 mm and 360 mm in the spanwise and streamwise direction (both centre-to-centre), respectively. The second roughness has $80\,\mathrm{mm} \times 20\,\mathrm{mm} \times 2\,\mathrm{mm}$ elements, standing on the area of $80\,\mathrm{mm} \times 2\,\mathrm{mm}$ to form of 50% staggered pattern with spacing of 240 mm (centre-to-centre) both in the spanwise and streamwise directions. The origin of the coordinate system is set at the location of the step change, at the wall and along the centreline of the tunnel (as indicated by the red dot in figure 1).

The streamwise, spanwise, and vertical velocity components (u, v, w) are acquired by a Dantec Fibre Flow probe 3D Laser Doppler Anemometer (LDA) with a diameter of 27 mm and a focal length of 160 mm. The LDA measurement volume is estimated to be 1.57 mm long in the spanwise direction and with a diameter of 0.074 mm. Time series of the velocity components are recorded for three minutes and at a minimum rate of 100 Hz, and measurements are taken at 40 wall-normal locations starting at $z = 50\,\mathrm{mm}$ ($z/\delta \approx 0.08$, with δ the local boundary layer thickness) and ending at $z = 800\,\mathrm{mm}$ ($z/\delta \approx 1.3$).

The present LDA system has been extensively validated and employed for several scientific publications, where results have been cross-validated against previous works, and typical statistical error (uncertainty) in the mean velocity is $\pm 1\%$, while errors in turbulence quantities u', v', w' are $\pm 2\%$, $\pm 4\%$, and $\pm 7\%$, respectively [30, 38]. Throughout this work, the mean freestream wind speed was set to $U_\infty = 1.5\,\mathrm{m\,s^{-1}}$ as measured by a sonic anemometer mounted 5 m downstream of the tunnel inlet (figure 1). The corresponding Reynolds number based on the boundary layer thickness δ_0 of the approach

Table 1. Summary of test cases and the corresponding training $(x^{\mathrm{T}}, z^{\mathrm{T}})$ and validation $(x^{\mathrm{V}}, z^{\mathrm{V}})$ coordinates. The $A \ldots B$ notation refers to all the available values in-between A and B, while the $\forall$ symbol signifies all available coordinates (for z, they are 40 values). N^{T} and N^{V} indicate the total number of training and validating points, respectively.

Test case	Training coordinates (mm)	validation coordinates (mm)	N^{T}	N^{V}
1	$x^{\mathrm{T}} = \{720, 5880\}$, $\forall z^{\mathrm{T}}$	$x^{\mathrm{V}} = \{1320, \ldots, 4680\}$, $\forall z^{\mathrm{V}}$	80	240
2	$x^{\mathrm{T}} = \{720, 5880\}$, $z^{\mathrm{T}} = \{50, 125, 255, 365, 450, 560, 645, 745, 800\}$	$x^{\mathrm{V}} = \{1320, \ldots, 4680\}$, $\forall z^{\mathrm{V}}$	18	240
3	$x^{\mathrm{T}} = \{-760, 0, 5880\}$, $\forall z^{\mathrm{T}}$ and $x^{\mathrm{V}} = \{-380, 120, \ldots, 4680\}$, $z^{\mathrm{T}} = z_{\mathrm{opt}}$	$x^{\mathrm{T}} = \{-380, 120, \ldots, 4680\}$, $\forall z^{\mathrm{V}}$ except $z^{\mathrm{T}} = z_{\mathrm{opt}}$	131	429

flow is $Re_\delta = U_\infty \delta_0 / \nu \approx 4.5 \times 10^4$, where ν is the kinematic viscosity.

The setup is identical to that in reference [30, 38], which includes further details of the two surface conditions and the boundary layer development. The measurement locations used for this work are indicated by blue dots in figure 1; these comprise wall-normal velocity profiles measured on the tunnel centreline and spanning a streamwise range $-0.76\,\mathrm{m} \leqslant x \leqslant 5.880\,\mathrm{m}$ (i.e. from upstream to downstream of the discontinuity in surface roughness), as further discussed in table 1.

3. Cluster-based transition networks for instantaneous turbulence estimation

The weighted average-based (WAB) transition network described in reference [23] is employed in this work to estimate turbulence statistics. Time series of the three velocity components, (u, v, w), from the turbulent boundary layer over a roughness step change (see section 2) are used for both the training and as testing of the methodology.

The WAB method comprises four key stages as described below, where the first three stages refer to the *training* of the method (sections 3.1–3.3), while the last stage refers to its *validation* (i.e. turbulence estimation; section 3.4).

3.1. Data collection and preparation

Data are collected in the form of time series of the velocity signals, $u(t), v(t), w(t)$, at various spatial locations (as described in section 2). Specifically, time traces are grouped into vertical profiles corresponding to fixed x, y coordinates. However, since $y = const. = 0$ throughout this work, the y coordinate is neglected, the z coordinate is used as the independent variable, while the x coordinate is used as the configuration parameter. Various x locations, in fact, dictate the effect of the step-change in surface condition on the flow, as well as the boundary layer evolution in the streamwise direction (the Reynolds number dependence is, therefore, already implicitly embedded in the x dependence).

The dataset is split between time series used for training and time series used for validation. The superscripts $\bullet^{\mathrm{T}}$ and $\bullet^{\mathrm{V}}$ are hence used to refer to training and validation data, respectively. Three test cases are showcased to illustrate the estimation capabilities of the WAB approach as reported in table 1, which summarises the training and validation coordinates for each test case. The total number of training and validating points, N^{T} and N^{V}, are also reported in table 1. It is important to remark that the validation coordinates do not overlap with training coordinates for statistics *estimation* , namely, $(x^{\mathrm{T}}, z^{\mathrm{T}}) \neq (x^{\mathrm{V}}, z^{\mathrm{V}})$ (note that $\ldots$ notation in table 1 indicates a range of coordinates).

3.2. State-space definition and clustering

In this study, we build three-dimensional (3D) state spaces where axes correspond to the velocity components u, v, and w, as illustrated in figure 2(a). Accordingly, for all the N_t^{T} time steps used for training, a direct relation exists between a generic state in the state-space and a triplet of velocity values at a time t, i.e. $\{u(x^{\mathrm{T}}, z^{\mathrm{T}}; t), v(x^{\mathrm{T}}, z^{\mathrm{T}}; t), w(x^{\mathrm{T}}, z^{\mathrm{T}}; t)\}$. State clustering is then performed with a two-fold aim: (i) reducing the number of states, thus offering a reduced-order model of the system with a consequent computational cost benefit; (ii) capturing the representative states of the system (associated with the cluster centroids), thus providing a more robust representation of the time series. The latter goal is particularly relevant for experimental data, as they typically show a lower signal-to-noise ratio than numerically-simulated datasets (in which noise is typically artificially added), thereby state clustering tends to smooth such uncertainties across clusters [39].

Following previous works on cluster-based transition network modelling [18, 21, 23], the k-means++ algorithm is used for state clustering [40], leading to a set of N_C cluster centroids, $\boldsymbol{C}(h)$, $h = 1, \ldots, N_C$. These are, *de facto*, new points in the state space and, like raw states, a triplet of $\{u(t), v(t), w(t)\}$ values can be associated with each centroid $\boldsymbol{C}(h)$. N_C has to be set *a priori* for the k-means++ algorithm, with the constraint $N_C \leqslant N_t^{\mathrm{T}}$; see section 4 and appendix A for further details. For example, figure 2(b) shows a schematic of a 2D state-space, including a trajectory associated with a time series pair $\{u(t), v(t)\}$ and the corresponding discrete raw states (black dots), as well as cluster centroids $\boldsymbol{C}$ (red dots). It is worth mentioning that, in this study, clustering is performed on time series from individual training locations $\alpha = (x^{\mathrm{T}}, z^{\mathrm{T}})$ [18, 23], rather than aggregating all training data into the same state space [39]. Accordingly, the notation $\boldsymbol{C}_\alpha(h)$ is adopted to distinguish centroids belonging to specific measurement locations α.

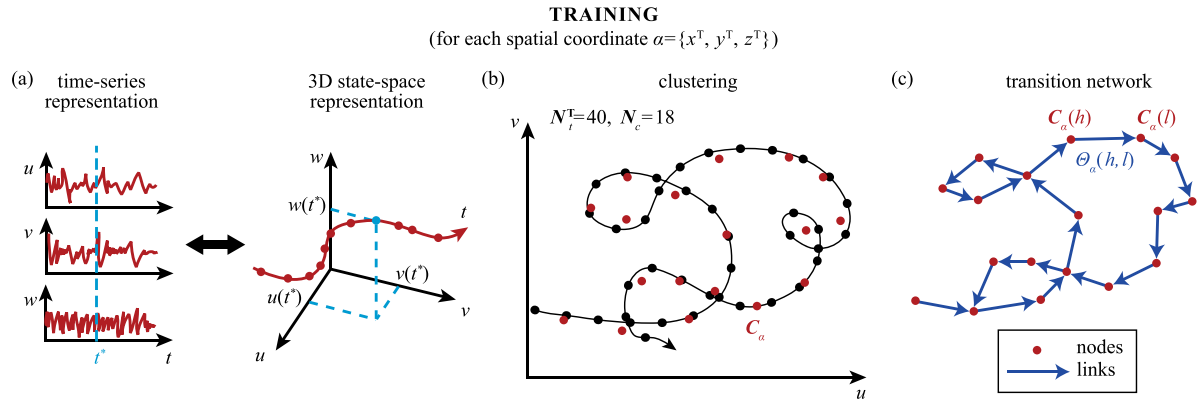

Figure 2. Schematic of the main training steps for the WAB approach. (a) Different data representations (left) as time series and (right) as equivalent state space. (b) An example of clustering in a 2D state space, with $N_t^{\mathrm{T}} = 40$ raw data points (black dots) and $N_C = 18$ clusters (red dots). The arrow indicates the direction of increasing time. (c) Transition network representation of the clustered data in (b), where arrows indicate the transition links (from a source node to a target node).

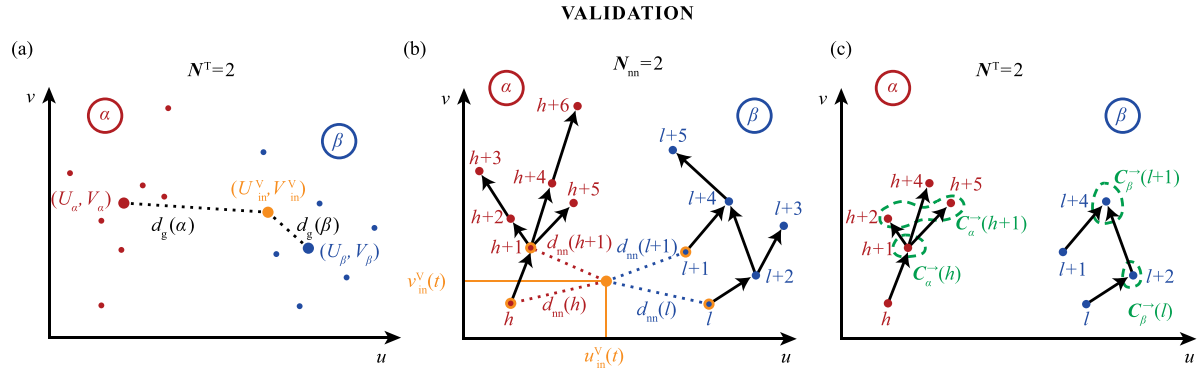

Figure 3. Schematic of the main validation steps for the WAB approach. (a) Example of 2D state space with $N^{\mathrm{T}} = 2$ training configurations, α and β. For each configuration, instantaneous states (small dots) and mean states (large dots) are shown. The mean state of the input data, $(U_{\mathrm{in}}^{\mathrm{V}}, V_{\mathrm{in}}^{\mathrm{V}})$, is also illustrated as a large orange dot, together with the Euclidean distances d_{g} (dotted black lines) to (U_α, V_α) and (U_β, V_β). (b) The same state space as in (a), illustrating the location of the instantaneous input data $(u_{\mathrm{in}}^{\mathrm{V}}(t), v_{\mathrm{in}}^{\mathrm{V}}(t))$ (orange dot), the transition networks (black arrows) for both configurations α and β, as well as the Euclidean distances d_{nn} (dotted coloured lines). In this example, the number of nearest neighbours is $N_{\mathrm{nn}} = 2$, corresponding to nodes h and $h+1$ for configuration α and nodes l and $l+1$ for configuration β (these nodes are highlighted with orange circles). (c) Maximum-probability target nodes, $\boldsymbol{C}_i^{\rightarrow}$, are identified for the same state space in (b) through green dashed contours. Black arrows in (b) and (c) indicate the transition network links (from source node to target node).

3.3. Transition network construction

Once system states are captured via the state-space clustering, the dynamical features of the system need to be embedded in the method. For this purpose, a transition network model is employed [18, 23], where the network nodes represent the cluster centroids and a link between two nodes, h and l, is set to be proportional to the probability, $\boldsymbol{\Theta}_\alpha(h,l)$, of switching from h to l [23]. Figure 2(c) shows the transition network corresponding to the exemplifying trajectory of figure 2(b). Specifically, transition probabilities are computed by counting how many times raw states (figure 2(b)) transit from a cluster $\boldsymbol{C}_\alpha(h)$ to a cluster $\boldsymbol{C}_\alpha(l)$.

Two main deliverables are hence obtained by training the WAB method using the velocity signals at each $\alpha = (x^{\mathrm{T}}, z^{\mathrm{T}})$ location of the physical domain: a list of centroids $\boldsymbol{C}_\alpha(h)$ associated with (u, v, w) coordinates in the state space, and a transition network in the form of a probability matrix, $\boldsymbol{\Theta}_\alpha(h,l)$, for each pair of centroids $\boldsymbol{C}_\alpha(h)$ and $\boldsymbol{C}_\alpha(l)$.

3.4. WAB estimation

The last stage of the WAB procedure consists of the estimation step. Here, we seek to estimate the statistics of a velocity component, $w_{\mathrm{est}}(t)$, at validation coordinates, $(x^{\mathrm{V}}, z^{\mathrm{V}}) \neq (x^{\mathrm{T}}, z^{\mathrm{T}})$. The time series of the remaining two components, $u_{\mathrm{in}}^{\mathrm{V}}(t)$ and $v_{\mathrm{in}}^{\mathrm{V}}(t)$, are known during the estimation stage and are provided as input data. A reference vertical velocity, $w_{\mathrm{ref}}(t)$, was also measured but only used for performance comparison, thus not being exploited for the estimation procedure.

The weighted average in equation (1) provides the value of $w_{\mathrm{est}}(t)$ at any time $t > 0$:

$$w_{\mathrm{est}}(t) = \frac{\sum_i \sum_j \omega_{\mathrm{g}}(i)\, \omega_{\mathrm{d}}(i,j)\, \overline{W}(i,j)}{\sum_i \sum_j \omega_{\mathrm{g}}(i)\, \omega_{\mathrm{d}}(i,j)}, \tag{1}$$

where $i = 1, \ldots, N^{\mathrm{T}}$ and $j = 1, \ldots, N_{\mathrm{nn}}$, with $N_{\mathrm{nn}} \leqslant N_C$ the number of nearest-neighbour centroids used for the weighted-average calculation. To better understand how w_{est} is calculated, the key steps involved in the estimation process are

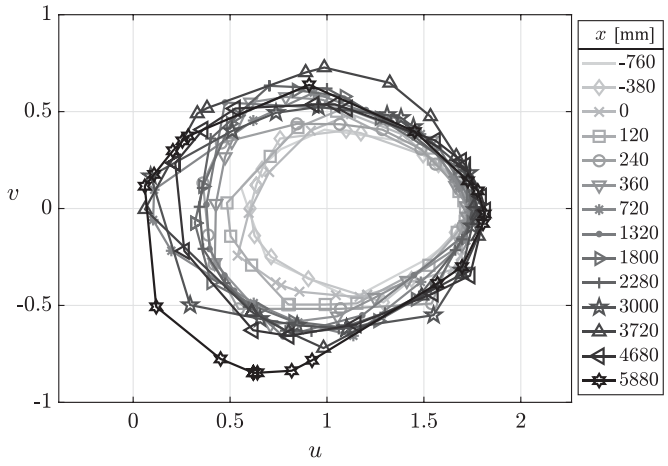

Figure 4. Two-dimensional projection of the 3D state space for the current experimental data, with closed lines showing the convex hulls (i.e. boundaries) of all instantaneous u, v velocity values. Each closed line corresponds to the convex hull for all velocity values collected at a given x coordinate (vertical profiles), as reported in the legend.

illustrated in figure 3, while the three quantities appearing in equation (1)—i.e. ω_g, ω_d, and $\overline{\mathcal{W}}$—are described in the following:

- A global weight $\omega_g(i)$ is introduced in the present study to explicitly consider the proximity between the mean velocity values of the input data, (U^V_{in}, V^V_{in}), and of all the trained time series, (U^T_α, V^T_α). As shown in figure 4, in fact, velocity values for all configurations significantly overlap in the state space, making the mean value estimation particularly challenging. Specifically, the weight is evaluated as

$$\omega_g(i) = \left(\frac{d_{opt}(i)}{\sum_i d_{opt}(i)} \frac{d_g(i)}{\sum_i d_g(i)} \right)^{-\epsilon_g}, \qquad (2)$$

  where $\epsilon_g \geqslant 0$ is a user-defined exponent whose value will be discussed in section 4, while $d_g(i)$ is the Euclidean distance between the *mean* states for the input data (U^V_{in}, V^V_{in}) and the *mean* states of the ith trained configuration (U^T_i, V^T_i). For example, in figure 3(a), $d_g(i)$ is illustrated for two configurations $i = \{\alpha, \beta\}$, as dotted lines.
  An additional distance function, d_{opt}, is also used in equation (2) for semi-supervised estimation. Results are firstly shown for $d_{opt} = const.$ (see sections 5.1 and 5.2), thus not affecting the weighted average. A more complex case where $d_{opt} \neq const.$ will support the estimation performance is described in section 5.3, where more details on d_{opt} are provided.
- A second weight is defined as

$$\omega_d(i,j) = \left(\frac{1}{d_{nn}(i,j)} \right)^{\epsilon_{nn}}, \qquad (3)$$

  where d_{nn} is the Euclidean distance between the *instantaneous* state of the input data, (u^V_{in}, v^V_{in}), and the N_{nn} nodes (i.e. nearest-neighbour centroids) for each trained configuration i. For example, $N_{nn} = 2$ in figure 3(b), such that nodes $j = \{h, h+1\}$ and $j = \{l, l+1\}$ are the two nearest-neighbours nodes from configurations $i = \alpha$ and $i = \beta$, respectively, to the input state $(u^V_{in}(t), v^V_{in}(t)$; orange dot). $\omega_d(i,j)$ is the main weight of the WAB approach, as also discussed in reference [23], with the exponent $\epsilon_{nn} \geqslant 1$ used to enhance distances in the state space [18].
- The last quantity in equation (1) is $\overline{\mathcal{W}}(i,j)$. Its definition requires the introduction of the concept of *maximum-probability targets*, $\boldsymbol{C}^{\rightarrow}_i(j)$, of a generic node j, that is the subset of nodes with the maximum transition probability from node j. This concept is illustrated in figure 3(c) as green dashed contour lines for nodes $h+1$ and l (belonging to configurations $i = \alpha$ and $i = \beta$, respectively), that is, $\boldsymbol{C}^{\rightarrow}_\alpha(h+1) = \{h+2, h+5\}$ and $\boldsymbol{C}^{\rightarrow}_\alpha(l) = \{l+2\}$. In fact, the WAB method selects target nodes based on the maximum probability rather than following random probability transitions [23].
  The w values associated with $\boldsymbol{C}^{\rightarrow}_i(j)$ nodes are stored into a matrix $\mathcal{W}(i,j,k)$, with k equal to the number of nodes in $\boldsymbol{C}^{\rightarrow}_i(j)$; e.g. $k = 2$ for $\boldsymbol{C}^{\rightarrow}_\alpha(h+1)$ in figure 3(c). $\mathcal{W}(i,j,k)$ is averaged over all k values, leading to the matrix $\overline{\mathcal{W}}(i,j)$ that, hence, represents the most probable w values to transit into, by starting from the nearest neighbours of the input state (u^V_{in}, v^V_{in}) [18, 23]; figures 3(b) and (c).

Overall, the weighted average in equation (1) combines information from the mean velocities through ω_g, from the instantaneous velocities through ω_d, and from the temporal turbulence dynamics via the transition matrix exploited in determining $\overline{\mathcal{W}}$. Further details on the parameters involved in the proposed methodology, as well as their interpretation, are provided in section 4.

4. Parameters selection and interpretation

A key feature of the proposed WAB approach is its interpretability. In fact, the clustered state-space representation of the dynamical system (here, turbulent flow) leads to an intuitive and straightforward interpretation of the quantities and parameters involved in the training and validation steps. This striking feature, therefore, provides transition network-based methods with a clear advantage against other data-driven frameworks. In this section, we highlight the main parameters involved in our methodology and their interpretation, allowing one to tune them based on the application of interest.

Overall, four parameters need to be specified, and they are reported in table 2 together with their interpretation. Thanks to their high degree of interpretability, these parameters can, therefore, be easily tuned depending on the specific application. For example, high N_{nn} values could be helpful when the data in the state space are sparse, thus requiring additional nearest-neighbour nodes in the weighted average.

Table 2. List of main methodological parameters (first column), the value used in this work (second column), as well as their interpretation (third column).

Parameter	Value	Interpretation
N^{T}	Table 1	N^{T} is the number of trained configurations, i.e. the size of the training dataset. Although this parameter is fixed *a priori* for unsupervised estimation, a subset of the N^{T} configurations can be selected during the estimation stage if some features of the estimated variable are known *a priori* (semi-supervised approach).
N_C/N_t^{T}	0.75	N_C is the number of cluster centroids, and its definition usually follows from a sensitivity analysis (see A). Generally, larger N_C values provide a better state-space representation, hence a better estimation performance. However, if high noise levels are present in the training data, a very high N_C value can negatively affect the estimation performance, as cluster centroids cannot reliably capture the main system states while heavily relying on noisy values (one of the benefits of clustering is indeed to smoothen out noise effects; section 3.2).
N_{nn}	2	N_{nn} is the number of nearest-neighbour nodes (for each of the N^{T} configurations) to be considered for the instantaneous weighted average, with $N_{\mathrm{nn}} \geqslant 1$ (see section 3.4). N_{nn} is typically of the order of unity to guarantee accuracy [23]. Larger N_{nn} values would include farther nodes in the state space with respect to the input data, $(u_{\mathrm{in}}^{\mathrm{V}}, v_{\mathrm{in}}^{\mathrm{V}})$, typically leading to less accurate estimations.
$\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{nn}} \geqslant 1$	$\frac{1}{10} N^{\mathrm{T}}$	ϵ_{g} and ϵ_{nn} exponents are used to get more accurate results [18], by enhancing the distances associated with weights ω_{g} and ω_{d} in equation (1) (see definitions in section 3.4). Since both weights depend on the number of trained configurations N^{T}, the effect of ϵ_{g} and ϵ_{nn} has to depend on N^{T} as well (see the Value column) to guarantee methodological consistency across different estimation scenarios (see table 1). Moreover, here, we set $\epsilon_{\mathrm{g}} = \epsilon_{\mathrm{nn}}$ to ensure consistency in the order of magnitude between the two weights, and upper-bound them as $(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{nn}}) \leqslant 15$ to mitigate numerical issues associated with large exponents.

Besides these four methodological parameters, two additional quantities have to be set: the temporal length of the time traces used for training, N_t^{T}, and the corresponding signal length used for validation, N_t^{V}. They are not WAB parameters in the sense that they mainly depend on the measurement capabilities, but they can be changed as part of the data preprocessing (i.e. section 3.1). In this work, $N_t^{\mathrm{T}}/N_t = 0.3$ (i.e. 30% of the available measured signals), and $N_t^{\mathrm{V}}/N_t = 1$. In this way, we showcase the WAB capabilities in estimating long signals by using shorter trained signals.

Lastly, it is worth mentioning that, while the simplest transition networks only account for probabilities over a single time step (commonly referred to as a Markov chain [6]), long-term memory effects can be included by considering higher-order conditional probabilities [18]. In this study, for simplicity, we do not consider any long-term memory effects as this would require the definition of additional parameters, and thus increase the complexity and computational cost of the methodology.

5. Results

This section reports the results of the application of the proposed network-based methodology on the three test cases summarised in table 1. In the first test case, we estimate w statistics downstream of the roughness change (i.e. $x > 0$). This represents our reference, as we illustrate the method capabilities on a simpler, yet challenging, setup of rough-wall turbulence. The second test case is a variation of the first test case in which a sparser training dataset is utilised (see N^{T} in table 1). Both these cases are described in section 5.1.

Additional results on the estimated statistics of $u'w'$, including an estimate of the friction velocity u_*, for the first test case are also reported in section 5.2, thus showing the capabilities to infer statistics of derivative variables. The third and last test case is the most challenging scenario, as it aims at estimating the w and $u'w'$ statistics across the step-change in roughness (hence, $x \lessgtr 0$), as discussed in section 5.3. For all cases, training is performed on the first 60 seconds of the

measured signals, while validation is performed by using the full three-minute signals.

To assess the WAB estimation capabilities, four errors are computed for the main statistical quantities:

- The normalised error on the mean value of w:

$$\mathcal{E}_W = \frac{|W_{\text{est}}\left(x^{\text{V}}, z^{\text{V}}\right) - W_{\text{ref}}\left(x^{\text{V}}, z^{\text{V}}\right)|}{\Delta W_{\text{ref}}\left(x^{\text{V}}\right)}, \tag{4}$$

  where

$$\Delta W_{\text{ref}}\left(x^{\text{V}}\right) = \max_{z^{\text{V}}}\left[W_{\text{ref}}\left(x^{\text{V}}, z^{\text{V}}\right)\right] - \min_{z^{\text{V}}}\left[W_{\text{ref}}\left(x^{\text{V}}, z^{\text{V}}\right)\right] \tag{5}$$

  is the maximum variation in the mean value at a given x^{V} coordinate.
- The normalised error on the standard deviation of w:

$$\mathcal{E}_{\sigma_w} = \frac{|\sigma_{w,\text{est}}\left(x^{\text{V}}, z^{\text{V}}\right) - \sigma_{w,\text{ref}}\left(x^{\text{V}}, z^{\text{V}}\right)|}{\Delta\sigma_{w,\text{ref}}\left(x^{\text{V}}\right)}, \tag{6}$$

  where $\Delta\sigma_{w,\text{ref}}$ is equivalent to the expression in equation (5) for the standard deviation σ_w.
- The relative error on the higher-order moments of w':

$$\mathcal{E}_{S_{w'}} = \frac{1}{N_{\text{bins}}}\sum_{i=1}^{N_{\text{bins}}}\left|\frac{\log\left[S_{w',\text{est}}\left(i\right)\right] - \log\left[S_{w',\text{ref}}\left(i\right)\right]}{\log\left[S_{w',\text{ref}}\left(i\right)\right]}\right|, \tag{7}$$

  where $S_{w'}(i; x^{\text{V}}, z^{\text{V}})$ is the reliability function (also referred to as the survival function), namely, the complementary cumulative distribution function of w', while $N_{\text{bins}} = 101$ is the number of bins used to calculate $S_{w'}$.
  The reliability function is chosen here in place of the probability density function (PDF) to estimate the error in the higher-order moments, because S shows smoother distribution tails (due to its cumulative nature) while PDF tails tend to be very noisy. Specifically, the logarithm of S is taken to emphasise the role of the tails in the error (hence of higher-order moments), and the velocity fluctuations (w') are used to discriminate against the effect of the error on the mean value (W), as this is already quantified by $\mathcal{E}_W$.
- The normalised root-mean-square (RMS) deviation of w:

$$\mathcal{E}_{\text{RMS}_w} = \frac{\sqrt{\frac{1}{N_t^{\text{T}}}\sum_t\left(w_{\text{est}}\left(t\right) - w_{\text{ref}}\left(t\right)\right)^2}}{\max_t\left[w_{\text{ref}}\right] - \min_t\left[w_{\text{ref}}\right]}, \tag{8}$$

  which is used to get an average difference between the measured velocity values (w_{ref}) and estimated ones (w_{est}). The normalisation by the maximum variation is taken, here, instead of the mean value, W_{est}, because this occurs very close to zero, thus leading to diverging errors (a known issue in relative error calculation).

All the errors introduced above are a function of the validation coordinates, x^{V} and z^{V}, and are extended to the momentum flux estimation by replacing w with $u'w'$ in equations (4)–(8). Moreover, the errors require a reference value for comparison purposes, which are obtained here from the measured w_{ref} values with their uncertainty. The high accuracy of the LDA system provides us with confidence that w_{ref} values can be considered as *true* values. Such uncertainties can propagate into the estimated velocity values via equation (1), but its exact quantification is non-trivial due to the nonlinearities of the WAB approach. A systematic analysis will hence be performed in future endeavours. In this work, a preliminary and complementary analysis is reported on the method sensitivity to artificially-added noise for the first test case in table 1: as shown in appendix C, the turbulence statistics errors do not significantly increase even in the extreme case of added noise.

5.1. Vertical velocity statistics over rough-wall turbulence

Figures 5(a) and (b) show the vertical profiles of the mean wall-normal velocity and its standard deviation for the first test case in table 1. Black dots in figure 5 correspond to the mean and standard deviation values of the training data, red dots indicate the reference values W_{ref} and $\sigma_{w,\text{ref}}$ used in equations (4)–(6) at validation coordinates $(x^{\text{V}}, z^{\text{V}})$, while blue dots are the estimated values at the same spatial coordinates. Although figure 5 only provides a qualitative picture of the estimation performance, the mean and standard deviation along z are well captured across the full x range considered. Particularly, while standard deviation values do not significantly change with x (figure 5(b)), the vertical profiles of W (figure 5(a)) show a significant variation from $x = 720$ mm to $x = 5880$ mm. This implies that the WAB methodology has to discriminate—in an unsupervised way—what the most informative training data are via weights ω_g and ω_d.

On this note, the overshoots for W_{ref} observed in figure 5(a) (e.g. at $x = 1320$ mm) are likely due to the fact that the method is giving more emphasis (or, in other words, is giving more 'weight' in the weighted average) to the training data at $x = 720$ mm around the overshoot. The reason why such overshoots occur, therefore, relates to the interpolatory and unsupervised nature of the method, which is unaware of the spatial coordinates associated with each velocity triplet (u, v, w) in the state space, thus automatically seeking for the most reliable cluster in the state space to give more weight to.

A quantitative analysis is reported in figure 6, which shows aggregated statistics over the vertical direction z for the four errors in equations (4)–(8), as a function of the validation coordinate x^{V}. Specifically, the average behaviour (solid lines) is shown, as well as the 25th percentile (dashed lines) and the 75th percentile (dash-dotted lines). As expected, the estimation accuracy is not constant with x^{V} due to the variability in the velocity profiles along the streamwise direction (i.e. the flow evolves along x). Figure 6(a) shows that the error on W, $\mathcal{E}_W$, is about 12% on average with significant bounds (1st and 3rd quartiles) in-between 4% and 30%. The error in the standard deviation, $\mathcal{E}_{\sigma_w}$, is shown in figure 6(b): the WAB approach well captures the fluctuating behaviour of the w time series

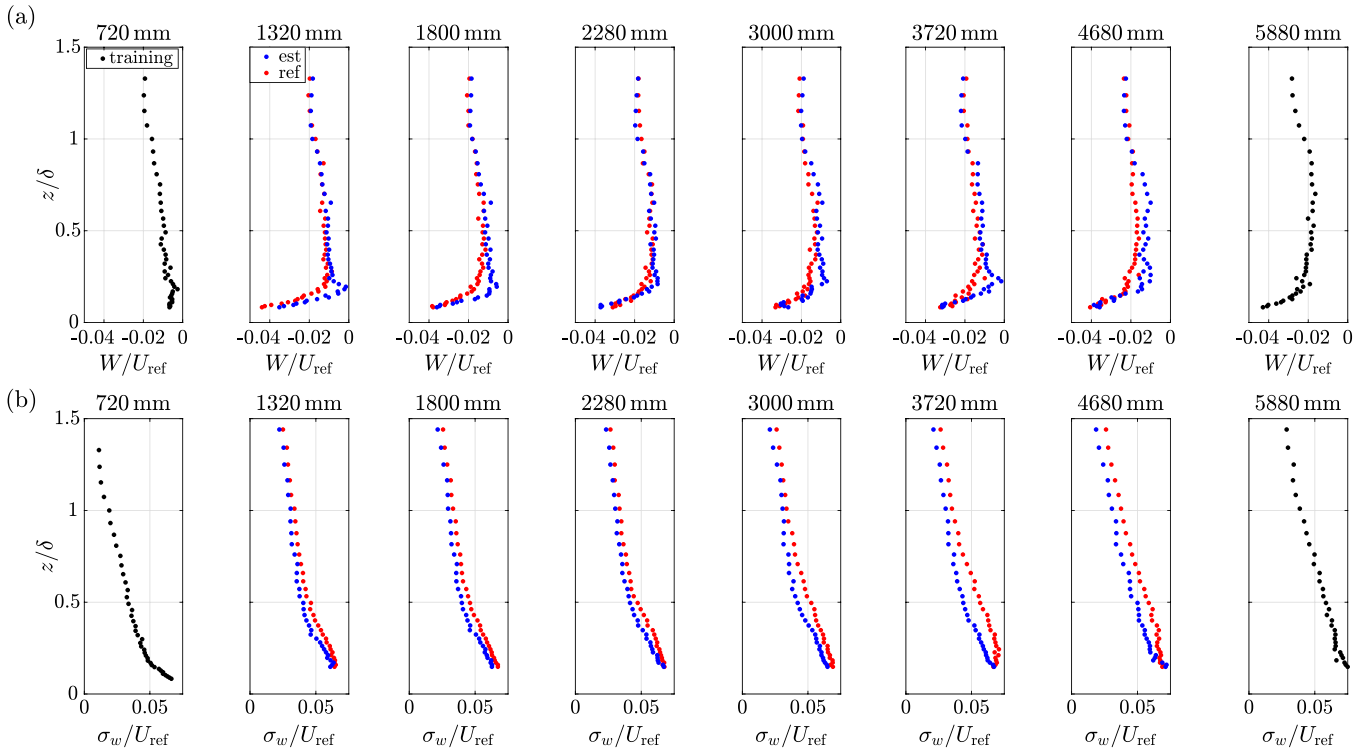

Figure 5. Vertical profiles of (a) mean vertical velocity W/U_{ref}, and (b) standard deviation of vertical velocity $\sigma_w/U_{\mathrm{ref}}$ for test case 1. Training measurement points are shown as black dots, estimated points are shown as blue dots, and the corresponding measured points are shown as red dots (reference values). Plots are for increasing x (in mm) from left to right.

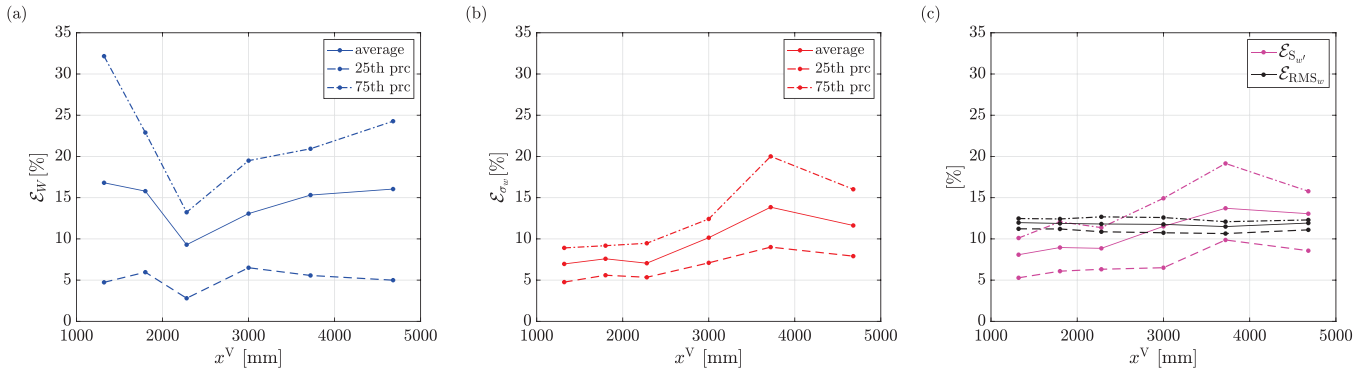

Figure 6. Estimation statistics of w signals as a function of the streamwise locations x^{V} for test case 1. (a) Average normalised error for the mean, $\mathcal{E}_W$. (b) Average normalised error for the standard deviation, $\mathcal{E}_{\sigma_w}$. (c) Normalised root-mean-square deviation, $\mathcal{E}_{\mathrm{RMS}_w}$ (black), and relative error on the higher-order moments of w', $\mathcal{E}_{S_{w'}}$ (magenta). The 25th and 75th percentiles for each quantity are also reported as dashed and dot-dashed lines, respectively, alongside average values (solid lines).

across the full domain, as errors are bounded to 15% with narrow 1st–3rd quartile bands. Figures 6(a) and (b), therefore, confirm the good estimation capabilities of the WAB approach in a challenging scenario of experimental noisy data and complex turbulence dynamics. Errors up to 20% of the reference value are indeed considerable acceptable as they are within a margin of 10% on top of the typical experimental uncertainty, especially considering the unsupervised (i.e. fully data-driven) nature of the data-driven approach.

The two remaining errors, $\mathcal{E}_{S_{w'}}$ and $\mathcal{E}_{\mathrm{RMS}_w}$, are reported figure 6(c). In both cases, errors are nearly constant with x^{V} and around 12%. These results point out that higher-order moments (as captured by the reliability function S) are well estimated (in line with the good estimation of the standard deviation in figure 6(b)), and instantaneous values $w_{\mathrm{est}}(t)$ do not significantly depart (on average) from the measured values $w_{\mathrm{ref}}(t)$. Note that errors in figures 6(a)–(c) are much lower than the maximum possible errors (not shown), which are at least one order of magnitude larger than reported estimation errors.

Results for the second test case are discussed next, as shown in figure 7. This case exploits the same training and validation streamwise coordinates as test case 1, but fewer vertical coordinates are used for training (see table 1). By doing so,

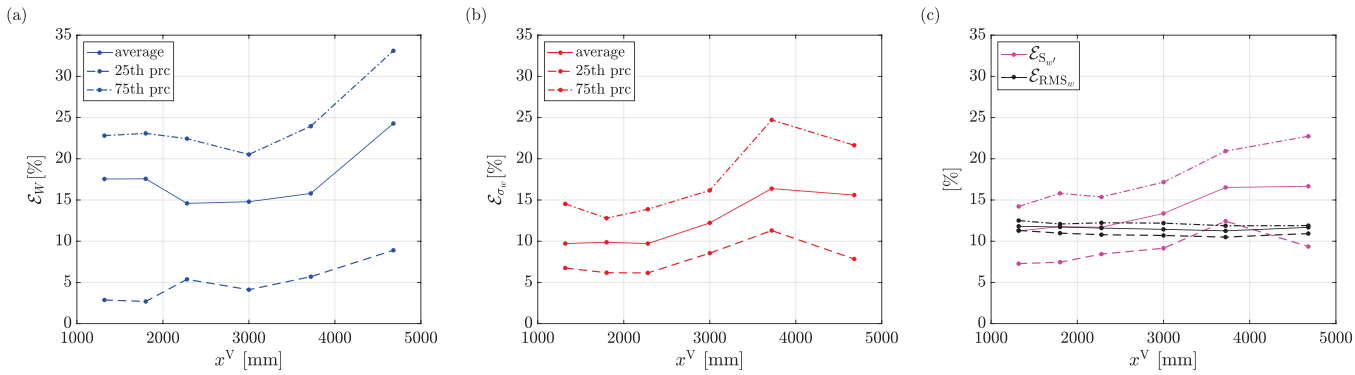

Figure 7. Estimation statistics of w signals as a function of x^{V} for test case 2 (sparse training). Panels content is equivalent to figure 6.

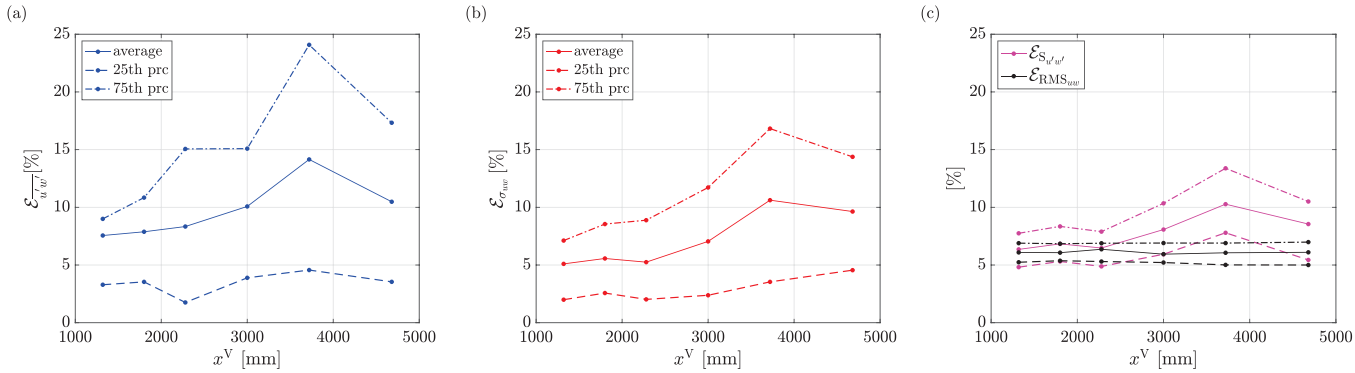

Figure 8. Estimation statistics of momentum flux signals $u'w'$ as a function of x^{V} for test case 1. Panels content is equivalent to figure 6.

we show the method's capability in estimating the full velocity profiles (i.e. all 40 vertical measurement points z^{V}) by using a sparser training dataset (that is, $N^{\mathrm{T}} = 18$ instead of $N^{\mathrm{T}} = 80$), thus reducing the amount of information used to perform the signal estimation. Figure 7 shows that, although the number of training data is significantly reduced (by a factor $80/18 \approx 4.44$) the error levels only slightly worsen when compared with results for test case 1 (figure 6). This outcome indicates that the proposed approach is robust against the size of the training dataset, as we estimated velocity statistics for $N^{\mathrm{V}} = 240$ points at various x-z coordinates by using velocity from only $N^{\mathrm{T}} = 18$ points. This offers a tantalising prospect to supplement experimental data with transition network-based methodologies while retaining acceptable uncertainty on first and second-order velocity quantities.

5.2. *Momentum-flux statistics over rough-wall turbulence*

The ability to estimate w statistics well has implications for turbulence analysis, since w is involved in the evaluation of useful quantities such as the momentum flux, $u'w'$, or the internal boundary layer thickness (via the velocity variance profile) [30]. Understanding the behaviour of $u'w'$ is important, not only to assess how momentum flux changes but also because the friction velocity, u_*, can be directly inferred from $\overline{u'w'}(x,z)$ (where the overbar indicates time averaging) as $u_*(x) = \sqrt{-\overline{u'w'}_0}$ [30, 41]. Here $\overline{u'w'}_0$ is the value of momentum flux at the roughness height, which is obtained via linear interpolation of the $\overline{u'w'}(z)$ values in the range $50\,\mathrm{mm} \leqslant z \leqslant 88\,\mathrm{mm}$, namely, where $\overline{u'w'}(z)$ is approximately constant with z (see also [30] for an analogous procedure on the same setup).

Figure 8 shows the four errors (equations (4)–(8)) for the estimated momentum flux

$$u'w'(t) = \left(u^{\mathrm{V}}_{\mathrm{in}}(t) - U^{\mathrm{V}}_{\mathrm{in}}\right)\left(w_{\mathrm{est}}(t) - W_{\mathrm{est}}\right), \tag{9}$$

which is given by the product of the measured (input) streamwise velocity fluctuations and the estimated vertical velocity fluctuations. Estimated mean values of the momentum flux (figure 8(a)) are consistent with the corresponding mean value estimates for the single variable w (figure 6(a)). Even better estimates for the standard deviation (figure 8(b)) and the higher-order moments (figure 8(c)) are obtained compared to those of the single variable w shown in figure 6(b) and figure 6(c), respectively.

Based on these outcomes, the friction velocity is also estimated. Figure 9 shows the relative error

$$\mathcal{E}_{u_*} = \frac{\left(u_{*,\mathrm{est}} - u_{*,\mathrm{ref}}\right)}{u_{*,\mathrm{ref}}}, \tag{10}$$

as a function of x^{V}, where $u_{*,\mathrm{ref}}$ is the reference value of u_* obtained from experimentally-measured u and w signals. The estimated friction velocity, $u_{*,\mathrm{est}}$, is underestimated (as

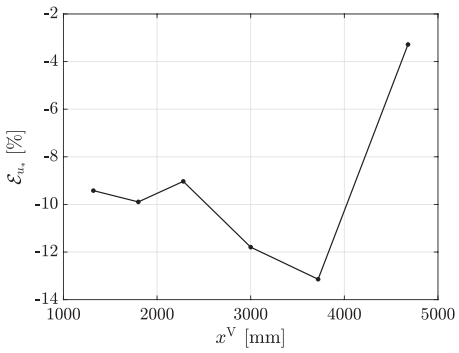

Figure 9. Relative error, $\mathcal{E}_{u_*}$, for the estimated friction velocity u_* at various streamwise locations, x^{V}, for test case 1.

the relative error is always negative) but is at most 15% the reference values. Since the friction velocity does not significantly change with x for this test case, an average friction velocity across all the x^{V} can be computed, resulting in a relative error on the friction velocity of about -9.4%; this is comparable to the uncertainty in the estimation of the friction velocity via the original methodology, as discussed in [41].

5.3. Velocity and momentum-flux statistics over step-change roughness

Estimation results for velocity profiles taken at various x locations across the step change (test case 3 in table 1) are discussed in this section. This is the most challenging test case because a more complex flow dynamics is considered in both the training and the validation, owing to significant flow evolution across the step change in surface condition. Since velocity values tend to accumulate within the same state-space region as shown in figure 4, the WAB methodology needs to discern the correct flow state in a 'forest' of incorrect states. A dense state space can, in fact, promote the so-called *ambiguity issue* [23], occurring when multiple target states (here w_{est}) can be associated with the same input data (here $u^{\mathrm{V}}_{\mathrm{in}}$ and $v^{\mathrm{V}}_{\mathrm{in}}$). More specifically, as explained in sections 3.4 and 5.1, the estimation of the mean value of the w velocity component, W, is a particularly challenging task in this work. The ω_{g} weight—see equation (2)—is indeed introduced to account for this challenge and reduce the mean value estimation error. Although ω_{g} comprises two terms, d_{g} and d_{opt}, only d_{g} has been used so far, while d_{opt} was set to an arbitrary constant, thus not affecting the estimation process. Due to the high ambiguity in estimating the mean values for this third test case, a semi-supervised approach is adopted here, which considers a non-constant d_{opt} in equation (2).

The core idea behind the semi-supervised approach is to quantify the degree of similarity between the mean velocity profiles in the training dataset, i.e. $\{U,V,W\}(x^{\mathrm{T}},z^{\mathrm{T}})$, and the velocity profiles at the validation coordinate, i.e. $\{U,V,W\}(x^{\mathrm{V}},z_{\mathrm{opt}})$, where z_{opt} indicates the subset of vertical coordinates used for such a comparison. For simplicity, only a single z_{opt} coordinate is used in this work, which results to be $z_{\mathrm{opt}} = 50\,\mathrm{mm}$ (i.e. $z_{\mathrm{opt}}/\delta \approx 0.083$; see appendix B), i.e. the closest measurement point to the wall. This implies that $z_{\mathrm{opt}} = 50\,\mathrm{mm}$ is the most informative coordinate, namely, it is the coordinate that better identifies the different features of the mean velocity profiles $\{U,V,W\}$ at various x locations. The user's knowledge of the flow evolution along x is exploited here but the approach estimates w signals without any data labelling, thus justifying the *semi-supervised* terminology. Further details on the non-constant d_{opt} values, and how the z_{opt} coordinate is automatically picked from the training data, are reported in appendix B.

Figures 10(a) and (b) show the velocity profiles of the mean and the standard deviation of w for the third test case, in analogy with figure 5 for the first test case. The standard deviation is very-well captured for all validation coordinates $(x^{\mathrm{V}},z^{\mathrm{V}})$, figure 10(b), and most of the full vertical profiles of W are well captured except for the region in close proximity to the wall at x^{V} equal to 120 mm and 240 mm (figure 10(a)). This is the region immediately after the step change in surface conditions, which experiences significant variations both in the shape of the vertical profile of W as well as its values (suddenly switching from negative values at $x^{\mathrm{V}} = 120\,\mathrm{mm}$ to positive values at $x^{\mathrm{V}} = 240\,\mathrm{mm}$, and again negative values at $x^{\mathrm{V}} = 360\,\mathrm{mm}$). Additional z_{opt} coordinates could help mitigate this issue, at the cost of a more supervised approach.

A more quantitative analysis is provided in figure 11, which shows the three errors on the mean, standard deviation, and higher-order moments of w_{est} for the third test case. The three errors $\mathcal{E}_W$ (figure 11(a)), $\mathcal{E}_{\sigma_w}$ (figure 11(b)), and $\mathcal{E}_{\mathrm{S}_w}$, (figure 11(c)), are shown for the semi-supervised approach (filled dots) and compared with those obtained with the unsupervised approach (i.e. for $d_{\mathrm{g}} = const.$ and $z_{\mathrm{opt}} = \{\varnothing\}$; open circles). While estimation performance is slightly worse, as expected, for all statistics compared to the reference test case 1 (section 5.1), the use of a semi-supervised approach clearly leads to an improvement in the estimation of the statistics even by using only 1 (out of 40) additional coordinates, z_{opt}.

This is particularly the case of the mean value (figure 11(a)), whose estimation improves over the second rough wall ($x > 0$). The reason for such an improvement can be associated with the fact that the velocity profiles at streamwise training coordinates $x^{\mathrm{T}} = \{-760, 0\}$ are much different than those over the second rough wall (e.g. at $x^{\mathrm{T}} = 5880$). The WAB approach, therefore, relies on $2/3$ of training profiles whose behaviour is different than those over the downstream roughness. The use of a semi-supervised approach, therefore, supports the WAB method in re-weighting the contribution to the estimation from each training profile. Summarising, a semi-supervised approach can easily—and effectively—improve the estimation capabilities by adding only a few additional training points

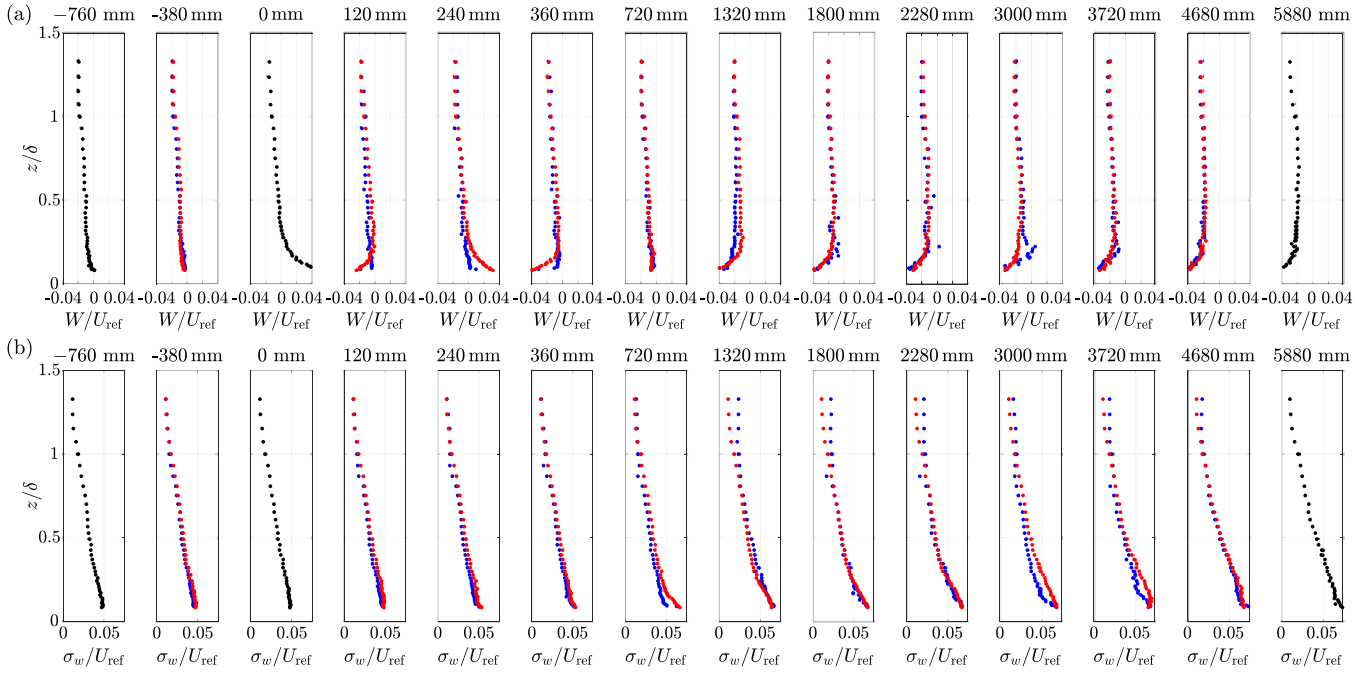

Figure 10. Vertical profiles of (a) mean vertical velocity W/U_{ref}, and (b) standard deviation of vertical velocity $\sigma_w/U_{\mathrm{ref}}$ for test case 3. Training measurement points are shown as black dots, estimated points are shown as blue dots, and the corresponding measured points are shown as red dots (reference values). Plots are for increasing x (in mm) from left to right.

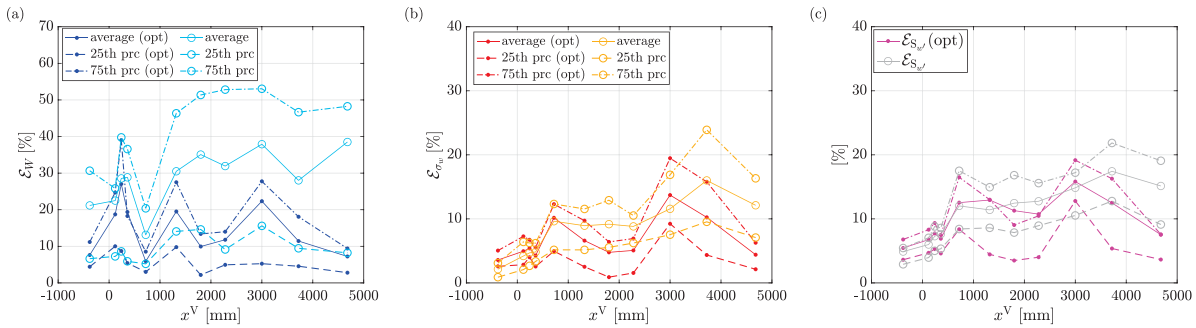

Figure 11. Estimation statistics of w signals as a function of the streamwise locations x^{V} for test case 3 (filled dots), compared with the corresponding results with a constant d_{g} (open circles). (a) Average normalised error for the mean, $\mathcal{E}_W$. (b) Average normalised error for the standard deviation, $\mathcal{E}_{\sigma_w}$. (c) Relative error on the higher-order moments of w', $\mathcal{E}_{S_{w'}}$ (magenta). The 25th and 75th percentiles for each quantity are also reported as dashed and dot-dashed lines, respectively, alongside average values (solid lines).

provided that those are taken at informative coordinates (i.e. z_{opt} coordinates).

We conclude this section by showing, in figure 12, the relative error $\mathcal{E}_{u_*}$ on the friction velocity $u_{*,\mathrm{est}}$, obtained by the estimation of the momentum flux $u'w'$ as in equation (10). Figure 12 shows that the friction velocity is well captured across the full streamwise domain for the semi-supervised approach (filled dots), as the error along the x coordinates is, on average, below the experimental uncertainty bound of 10 %. In contrast, as expected from figure 11, the relative error in the fully unsupervised approach (open circles in figure 12) is larger, especially downstream of the step change in roughness.

6. Discussion and conclusions

In this work, a weighted-average transition network-based algorithm is employed and adapted for the estimation of turbulence statistics in an experimentally modelled turbulent boundary layer over a step change in surface roughness. The present data-driven method—referred to as WAB—is an unsupervised approach since all training data are not labelled and the algorithm automatically discerns which are the most relevant training data. This approach has been employed in previous work for the estimation of unsteady loads [23], and it is now generalised for turbulence statistics estimation.

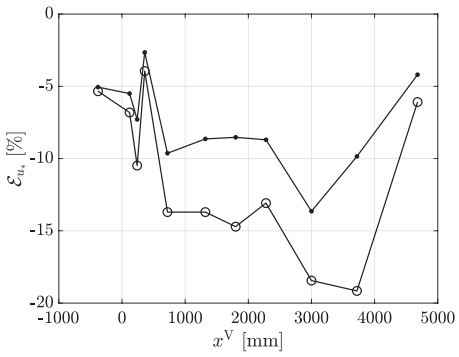

Figure 12. Relative error, $\mathcal{E}_{u_*}$, for the estimated friction velocity u_* at various streamwise locations, x^{V}, for test case 3 (filled dots). Open circles correspond to results on the same test case but with a constant d_{g} (unsupervised WAB).

The WAB approach is first tested for the estimation of vertical velocity statistics over a rough-wall turbulent boundary layer (test case 1 in table 1). Results show that, despite the experimental uncertainties in both the training and validation datasets, the WAB approach can capture with good accuracy the mean, standard deviation, and higher-order moments of the vertical velocity, w, as well as of the momentum flux, $u'w'$. Specifically, the friction velocity is estimated with good accuracy, with an average relative error below 10%. Furthermore, good results are obtained even when very sparse training data are used, as the ratio between training and validation data points is 7.5% (see test case 2 in table 1). Estimation errors in the range $10\% - -20\%$ give us confidence that further refinements to the methodology can bring the errors down within experimental uncertainty, which is the desired performance for a data-driven methodology operating on experimental data.

A more challenging scenario is also considered, in which the flow across the step change in roughness is estimated (test case 3 in table 1). Due to the complexity of this test case, a semi-supervised approach is employed here by exploiting the newly-introduced weight, ω_{g}. This semi-supervised approach consists of adding a limited number of training points at estimation coordinates, thus providing support for the WAB approach in discerning the most informative training data. Results for this test case indicate that the proposed semi-supervised approach can well estimate turbulence statistics (both for w and $u'w'$) with only one additional training coordinate (i.e. z_{opt}).

Overall, the proposed WAB approach is easily interpretable—as motivated in section 4—and robust against the main methodological parameters involved (e.g. see appendix A). These features allowed us to confidently estimate velocity statistics over three minutes by using only one minute of training data, i.e. $N_t^{\mathrm{V}}/N_t^{\mathrm{T}} \approx 3$. Moreover, the ratio between the signal measurement time and its estimation time is $N_t^{\mathrm{V}}/N_{\mathrm{comp}} \approx 12$, where $N_{\mathrm{comp}} \approx 15\,\mathrm{s}$ is the computational time for the estimation of statistics in a single validation coordinate $(x^{\mathrm{V}}, z^{\mathrm{V}})$ for $N_t^{\mathrm{V}} \approx 180\,\mathrm{s}$ (computational time is assessed on a M2 Pro, 16 GB RAM, using MATLAB® 2023b, 12 parallel cores). In other words, the WAB approach can help save one order of magnitude in experimental measurement time when one velocity component cannot be measured (either due to instrumentation or setup limitations) while retaining three key features of data-driven modelling: interpretability, robustness against parameters, and low level of supervision.

In conclusion, although transition networks have only recently been employed for signal reconstruction and estimation in fluid mechanics [18, 21, 23, 39], and some issues still need to be fully addressed (e.g. when sensors are very sparse, or signal-to-noise ratio is poor), we believe transition networks represent a valid alternative to other data-driven approaches. Future work will aim at addressing the aforementioned issues by testing and improving the methodology on a variety of other turbulence test cases. Uncertainty propagation from reference values to estimated values will also be addressed in follow-up studies. Moreover, future efforts will integrate transition networks with machine learning architectures (e.g. graph-neural networks), to exploit the mutual benefits of both approaches, thus boosting the methodology capabilities in addressing measurement science issues, while supporting existing time-series measurement tools.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.15126/surreydata.900993.

Acknowledgment

The authors acknowledge the support received by EPSRC under Agreement No. EP/V010921/1 ‘Fluid dynamics of Urban Tall-building clUsters for Resilient built Environments (FUTURE)’ and NERC under the Agreement No. NE/W002825/1 Across-Scale processeS in URban Environments (ASSURE).

Appendix A. Parametric analysis on the number of clusters

This appendix reports the effect of the number of clusters, N_C, of the k-means algorithm (see section 3.2) on the estimation results. N_C is an important—yet arbitrary—parameter which needs to be defined *a priori*. For this parametric analysis, we use the reference test case 1 as illustrated in figure A1 where estimation errors are reported for three increasing values of number of clusters: $N_C = \{0.3, 0.75, 0.9\}$. The error on the mean value (figure A1(a)) slightly decreases with N_C, as the WAB approach benefits from additional states (hence, additional information) in the estimation of the mean value. In contrast, the error on the standard deviation (figure A1(b)) tends to slightly increase with N_C: this behaviour is expected

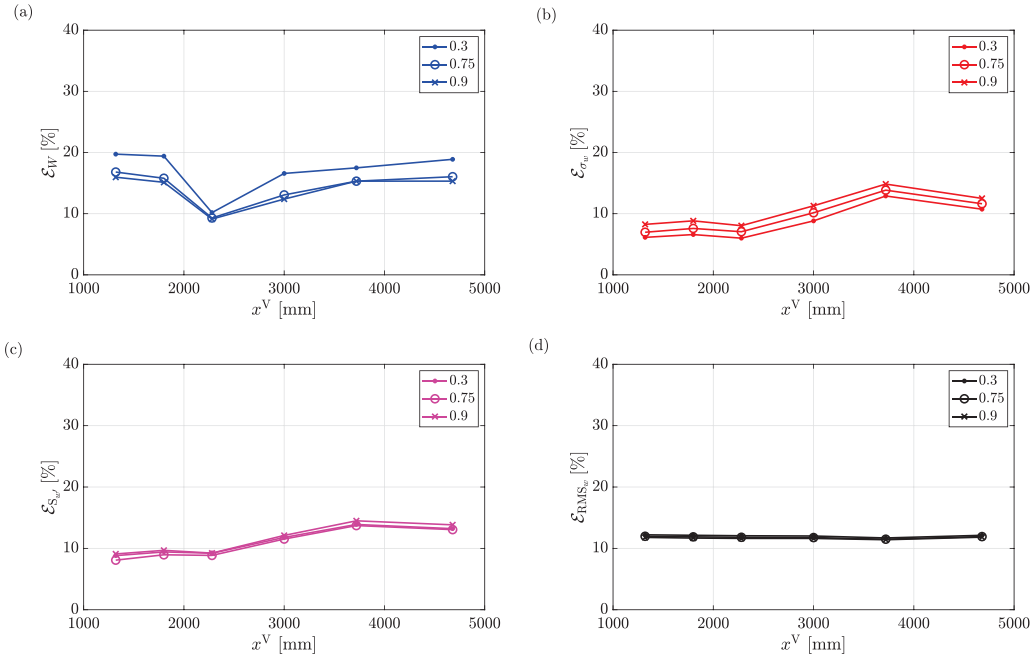

Figure A1. Estimation statistics of w signals as a function of the streamwise locations x^{V} and various number of clusters, $N_C = \{0.3, 0.75, 0.9\}$, for test case 1. (a) Average normalised error for the mean, $\mathcal{E}_W$. (b) Average normalised error for the standard deviation, $\mathcal{E}_{\sigma_w}$. (c) Relative error on the higher-order moments of w', $\mathcal{E}_{\mathrm{S}_{w'}}$. (d) Normalised root-mean-square deviation, $\mathcal{E}_{\mathrm{RMS}_w}$.

because the increase in N_C implies that clustering does not appropriately capture the noise, thus impacting the estimation of velocity fluctuations. To balance the opposite effect of N_C on the estimation of the mean value and the standard deviation, $N_C = 0.75$ is selected as a reference value throughout this work. Nevertheless, mild variations in aggregate errors in figures A1(a) and (b)—corroborated by the negligible changes in the error on the higher-order moments (figure A1(c)) and the root-mean-square error (figure A1(d)) with N_C—support the robustness of the WAB approach against the number of clusters.

Appendix B. Methodological details on the semi-supervised approach

Herewith, we provide details for the semi-supervised approach employed to estimate velocity statistics across the step change (section 5.3). First, the most informative coordinates, z_{opt}, have to be found. These are N_{opt} coordinates that extend the training dataset by including velocity signals at the coordinate $(x^{\mathrm{V}}, z_{\mathrm{opt}})$ (see table 1, test case 3). The additional velocity time series measured at $(x^{\mathrm{V}}, z_{\mathrm{opt}})$ are then used to assess the degree of similarity between training and validation mean velocity profiles. For simplicity, and to reduce to the minimum the additional amount of information required, we set $N_{\mathrm{opt}} = 1$. The degree of similarity is explicitly computed as:

$$\mathcal{S}_{\mathrm{opt}}\left(z^{\mathrm{T}}\right) = \frac{1}{3}\left(\sigma^2_U + \sigma^2_V + \sigma^2_W\right), \tag{B.1}$$

where $\sigma^2_{U,V,W}(z^{\mathrm{T}})$ are the variance of the mean velocities U, V, W calculated across the training coordinates x^{T}. Larger $\mathcal{S}_{\mathrm{opt}}(z^{\mathrm{T}})$ values indicate that the mean value of three velocity components at z^{T} are significantly variable, therefore time series at z^{T} provide a high amount of information (i.e. are more discriminant). z_{opt} is, therefore, automatically obtained as the vertical coordinate (or, more generally, the first N_{opt} coordinates if $N_{\mathrm{opt}} > 1$) with the highest $\mathcal{S}_{\mathrm{opt}}$ value.

Once z_{opt} is obtained, it is possible to explicitly calculate the similarity distance d_{opt} appearing in equation (2) as follows:

$$d_{\mathrm{opt}}\left(x^{\mathrm{T}}\right) = \frac{1}{3}\sum_{j=1}^{3}\left|\frac{\mathcal{U}_j\left(x^{\mathrm{T}}, z_{\mathrm{opt}}\right)}{\mathcal{U}_j\left(x^{\mathrm{V}}, z_{\mathrm{opt}}\right)} - 1\right|, \tag{B.2}$$

where $\mathcal{U}_j = U, V, W$ labels the three mean velocity components. Equation (B.2) compares the mean velocity values at the training coordinates $(x^{\mathrm{T}}, z_{\mathrm{opt}})$ with those at the newly-added training coordinate $(x^{\mathrm{V}}, z_{\mathrm{opt}})$. If mean velocities are of comparable values at z_{opt}, then the ratio in equation (B.2) tends to 1 and $d_{\mathrm{opt}}(x^{\mathrm{T}})$ decreases. In turn, a smaller d_{opt} acts as a shorter Euclidean distance in the state space (e.g. like d_{g} in equation (2)), thus enhancing the weight ω_{g}. In other words, smaller $d_{\mathrm{opt}}(x^{\mathrm{T}})$ suggests to the algorithm that mean velocity profiles at a generic validation streamwise coordinate x^{V} are more likely to resemble (or have a higher degree of similarity with) mean velocity profiles at the specific training streamwise coordinate x^{T}.

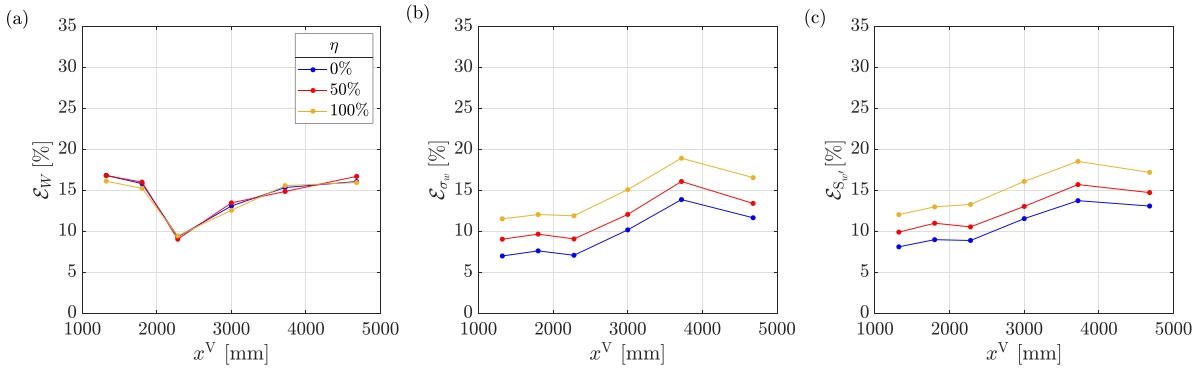

Figure C1. Average estimations errors as in figure 6 (blue lines), and corresponding results for added Gaussian noise with 50% ($\eta = 0.5$; red lines) and 100% ($\eta = 1$; orange lines) standard deviation with respect to original data.

Appendix C. Sensitivity analysis for increasing noise

Despite its interpretability, the proposed method exploits non-linear relations thereby providing *a priori* estimates of uncertainties is a very hard task, which would require a dedicated study. However, similar to the case of traditional measurement tools, we can assess the response of the data-driven method to various uncertainties (noise) in the data. A representative example is reported in this appendix for the first test case of table 1, where results are reported for increasing noise intensity. Noise is added to the velocity values u, v, w as follows: $u^* = u + \eta\sigma_u G$, $v^* = v + \eta\sigma_v G$, $w^* = w + \eta\sigma_w G$, where $*$ superscript indicates the modified velocity components, G is a random standard Gaussian noise (zero mean and unitary variance), η is a proportionality factor, and $\sigma_(u,v,w)$ are the original standard deviation values of u,v, and w. The proposed methodology is hence applied to the u^*, v^*, w^* signals for varying values of η. Figure C1 shows that the average error in the estimation of the mean remains almost unaffected, while the error for the standard deviation (panel b) and higher order moments (panel c) increases with the noise intensity, η, as expected. However, errors do not significantly increase even in the extreme case of added noise with the same standard deviation of the original data (100% case in figure C1).

ORCID iDs

Giovanni Iacobello https://orcid.org/0000-0002-0954-8545
Marco Placidi https://orcid.org/0000-0001-5105-8980
Shan–Shan Ding https://orcid.org/0000-0001-9673-1294
Matteo Carpentieri https://orcid.org/0000-0002-8968-9339

References

[1] Brunton S L, Noack B R and Koumoutsakos P 2020 Machine learning for fluid mechanics *Annu. Rev. Fluid Mech.* **52** 477–508
[2] Vinuesa R and Brunton S L 2022 Enhancing computational fluid dynamics with machine learning *Nat. Comput. Sci.* **2** 358–66
[3] Vinuesa R, Brunton S L and McKeon B J 2023 The transformative potential of machine learning for experiments in fluid mechanics *Nat. Rev. Phys.* **5** 536–45
[4] Discetti S and Liu Y 2022 Machine learning for flow field measurements: a perspective *Meas. Sci. Technol.* **34** 021001
[5] Castro I *et al* 2017 Measurements and computations of flow in an urban street system *Bound.-Layer Meteorol.* **162** 207–30
[6] Campanharo A S, Sirer M I, Malmgren R D, Ramos F M, Amaral L A N and Perc M 2011 Duality between time series and networks *PLoS One* **6** e23378
[7] Iacobello G, Ridolfi L and Scarsoglio S 2021 A review on turbulent and vortical flow analyses via complex networks *Physica* A **563** 125476
[8] Taira K and Nair A G 2022 Network-based analysis of fluid flows: progress and outlook *Prog. Aerosp. Sci.* **131** 100823
[9] Rabault J, Kolaas J and Jensen A 2017 Performing particle image velocimetry using artificial neural networks: a proof-of-concept *Meas. Sci. Technol.* **28** 125301
[10] Chen J, Raiola M and Discetti S 2022 Pressure from data-driven estimation of velocity fields using snapshot PIV and fast probes *Exp. Therm. Fluid Sci.* **136** 110647
[11] Chen D, Kaiser F, Hu J, Rival D E, Fukami K and Taira K 2023 Sparse pressure-based machine learning approach for aerodynamic loads estimation during gust encounters *AIAA J.* **1** 1–16
[12] Erichson N B, Mathelin L, Yao Z, Brunton S L, Mahoney M W and Kutz J N 2020 Shallow neural networks for fluid flow reconstruction with limited sensors *Proc. R. Soc.* A **476** 20200097
[13] Maulik R, Fukami K, Ramachandra N, Fukagata K and Taira K 2020 Probabilistic neural networks for fluid flow surrogate modeling and data recovery *Phys. Rev. Fluids* **5** 104401
[14] Hou W, Darakananda D and Eldredge J D 2019 Machine-learning-based detection of aerodynamic disturbances using surface pressure measurements *AIAA J.* **57** 5079–93
[15] Fukami K, Fukagata K and Taira K 2023 Super-resolution analysis via machine learning: a survey for fluid flows *Theor. Comput. Fluid Dyn.* **37** 1–24
[16] Kaiser E, Noack B R, Cordier L, Spohn A, Segond M, Abel M, Daviller G, Östh J, Krajnović S and Niven R K 2014 Cluster-based reduced-order modelling of a mixing layer *J. Fluid Mech.* **754** 365–414
[17] Nair A G, Yeh C-A, Kaiser E, Noack B R, Brunton S L and Taira K 2019 Cluster-based feedback control of turbulent post-stall separated flows *J. Fluid Mech.* **875** 345–75

[18] Fernex D, Noack B R and Semaan R 2021 Cluster-based network modeling—from snapshots to complex dynamical systems *Sci. Adv.* **7** eabf5006
[19] Foroozan F, Guerrero V, Ianiro A and Discetti S 2021 Unsupervised modelling of a transitional boundary layer *J. Fluid Mech.* **929** A3
[20] Li H, Fernex D, Semaan R, Tan J, Morzyński M and Noack B R 2021 Cluster-based network model *J. Fluid Mech.* **906** A21
[21] Hou C, Deng N and Noack B R 2022 Trajectory-optimized cluster-based network model for the sphere wake *Phys. Fluids* **34** 085110
[22] Deng N, Noack B R, Morzyński M and Pastur L R 2022 Cluster-based hierarchical network model of the fluidic pinball – cartographing transient and post-transient, multi-frequency, multi-attractor behaviour *J. Fluid Mech.* **934** A24
[23] Iacobello G, Kaiser F and Rival D E 2022 Load estimation in unsteady flows from sparse pressure measurements: application of transition networks to experimental data *Phys. Fluids* **34** 025105
[24] Hou C, Deng N and Noack B R 2024 Dynamics-augmented cluster-based network model *J. Fluid Mech.* **988** A48
[25] Colanera A, Deng N, Chiatto M, de Luca L and Noack B R 2024 arXiv:2407.01109
[26] Noack B, Deng N, Pastur L, Maceda G C and Hou C 2024 Cluster globally, model locally: clusterwise modeling of nonlinear dynamics *Research Square Preprint* (https://doi.org/10.21203/rs.3.rs-4583139/v1)
[27] Vallejo M, De La Espriella C, Gómez-Santamaría J, Ramírez-Barrera A F and Delgado-Trejos E 2019 Soft metrology based on machine learning: a review *Meas. Sci. Technol.* **31** 032001
[28] Urbas U, Vlah D and Vukašinović N 2021 Machine learning method for predicting the influence of scanning parameters on random measurement error *Meas. Sci. Technol.* **32** 065201
[29] Farhat M H, Chiementin X, Chaari F, Bolaers F and Haddar M 2021 Digital twin-driven machine learning: ball bearings fault severity classification *Meas. Sci. Technol.* **32** 044006
[30] Ding S-S, Placidi M, Carpentieri M and Robins A 2023 Neutrally- and stably-stratified boundary layers adjustments to a step change in surface roughness *Exp. Fluids* **64** 86
[31] Antonia R and Luxton R 1971 The response of a turbulent boundary layer to a step change in surface roughness part 1. Smooth to rough *J. Fluid Mech.* **48** 721–61
[32] Antonia R and Luxton R 1972 The response of a turbulent boundary layer to a step change in surface roughness. Part 2. Rough-to-smooth *J. Fluid Mech.* **53** 737–57
[33] Cheng H and Castro I P 2002 Near-wall flow development after a step change in surface roughness *Bound.-Layer Meteorol.* **105** 411–32
[34] Iacobello G, Placidi M, Ding S and Carpentieri M 2024 *Dataset for the article "Turbulence statistics estimation across a step change in roughness via interpretable network-based modellin"* University of Surrey Open Research Repository (https://doi.org/10.15126/surrey data.900993)
[35] Marucci D, Carpentieri M and Hayden P 2018 On the simulation of thick non-neutral boundary layers for urban studies in a wind tunnel *Int. J. Heat Fluid Flow* **72** 37–51
[36] Marucci D and Carpentieri M 2020 Stable and convective boundary-layer flows in an urban array *J. Wind Eng. Ind. Aerodyn.* **200** 104140
[37] Irwin H 1981 The design of spires for wind simulation *J. Wind Eng. Ind. Aerodyn.* **7** 361–6
[38] Ding S, Carpentieri M, Robins A and Placidi M 2024 Statistical properties of neutrally and stably stratified boundary layers in response to an abrupt change in surface roughness *J. Fluid Mech.* **986** A4
[39] Kaiser F, Iacobello G and Rival D E 2022 Aerodynamic state estimation from sparse sensor data by pairing Bayesian statistics with transition networks *Aiaa Scitech 2022 Forum* p 1669
[40] Arthur D and Vassilvitskii S 2006 k-means++: The advantages of careful seeding (Tech. Rep.) *Stanford Infolab* **8090** 778
[41] Connelly J S, Schultz M P and Flack K A 2005 Velocity-defect scaling for turbulent boundary layers with a range of relative roughness *Exp. Fluids* **40** 188–95