Data-Centric Engineering (2026), 7: e44
doi:10.1017/dce.2026.10078

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Pairwise interactions in turbulence time series: essential measures for validating synthetic data

Saleh Rezaeiravesh[1], Daniele Massaro[2] and Philipp Schlatter[3]

[1]Department of Mechanical and Aerospace Engineering, The University of Manchester, UK
[2]Department of Mechanical Engineering, Massachusetts Institute of Technology, USA
[3]Institute of Fluid Mechanics (LSTM), Friedrich-Alexander University Erlangen-Nuremberg, Germany
Corresponding author: Saleh Rezaeiravesh; Email: saleh.rezaeiravesh@manchester.ac.uk

Received: 02 December 2025 UTC; **Revised:** 23 June 2026 UTC; **Accepted:** 16 July 2026 UTC

Keywords: information theory; pairwise interactions; turbulence time series; vector autoregression

Abstract

The synthesis of turbulence dynamics using machine learning, statistical, and dynamical system-based models has become an emerging research direction. Validation of synthetic turbulence time series is often based on marginal and second-order statistics, which may overlook whether the underlying interaction structure is preserved. In this work, we apply a set of linear and information-theoretic measures to evaluate interactions between pairs of turbulence time series obtained from a turbulent channel flow at various wall distances, as well as between their synthetic counterparts. To demonstrate the applicability of these measures, we assess synthetic time series generated from a high-order vector autoregressive (VAR) model. The results show that the VAR-generated time series preserve the linear interactions measured by auto- and cross-covariance functions, even for long-memory dynamics. However, when assessed by information-theoretic measures, such as time-delayed mutual information and transfer entropy, the agreement deteriorates. Despite under-predicting the magnitude of transfer entropy, the VAR-generated time series correctly capture the dominant direction of net information transfer in most of the pairs. Furthermore, when the time series embedding is optimally constructed, the interpretation of net transfer entropy is shown to be consistent with relevant physical interpretations of causal interactions reported in the turbulence literature. Future research should consider applying this framework to multivariate turbulence time series generated by other data-driven techniques to ensure more comprehensive validation.

Impact statement

When validating synthetic turbulence time series generated by data-driven models, assessments typically rely on statistical moments up to the second order. However, such measures alone are insufficient to determine whether temporal memory and the underlying pairwise dependencies between coupled time series are faithfully preserved. This work demonstrates the practical importance of a set of linear and information-theoretic interaction measures for evaluating pairwise relationships in turbulence time series. The measures are applied to velocity time series extracted from a turbulent channel flow simulation and to their synthetic counterparts generated by a high-order vector autoregressive model. By comparing these measures between simulation and synthetic data, and examining how their directional patterns are consistent with turbulence physics, the study highlights their value as complementary validation tools for data-driven turbulence time-series generation.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

1. Introduction

Turbulent flows are complex multiscale systems in which a wide range of spatio-temporal structures interact through various physical mechanisms (Jiménez, 2018; Jiménez, 2025; Massaro and Hussain, 2026). The state variables, such as velocity, pressure, and their derivative quantities, are interconnected in both space and time. These dependencies are encoded in high-dimensional joint probability distributions, whose exact forms are practically unobservable. Typically, we obtain samples drawn from these underlying distributions through experiments, scale-resolving simulations, and more recently, data-driven models. The degree to which the resulting data preserve the properties of the true physics depends on their level of contamination by error and uncertainty originating from various sources. In practice, we work with turbulence time series, that is, discrete samples of continuous flow quantities recorded at a finite set of timestamps. Several studies have provided strong evidence that the marginal distributions of turbulence time series are typically non-Gaussian, often skewed, heavy-tailed, or intermittent (Kovasznay et al., 1970; Castaing et al., 1990; Jiménez, 1998). This non-Gaussianity does not, however, pose a fundamental issue for most analyses in practice, as we are often interested in low-order turbulence statistics, such as mean velocity and Reynolds stresses. These quantities of interest have well-defined finite values, despite being affected by sampling uncertainty due to the finite size of time series (Oliver et al., 2014; Russo and Luchini, 2017; Rezaeiravesh et al., 2022; Xavier et al., 2024). In contrast, what is challenging is how to characterise intrinsic interactions among multivariate time series recorded at different spatial locations within a turbulent flow. Linear interactions are typically quantified using statistical measures such as correlation, autocorrelation, and cross-covariance functions, that is, second-order statistics. In practice, these can be estimated accurately for weakly statistically stationary time series (or after being converted to that), see, for example, Wei (2006). For quantifying both linear and non-linear interactions, and for performing detailed structure analysis of turbulence time series (stationarity is not needed), we can employ information-theoretic measures (Cover and Thomas, 2001; Kantz and Schreiber, 2003), such as Shannon entropy (Shannon, 1948), mutual information (Fraser and Swinney, 1986), and transfer entropy (Schreiber, 2000).

The central objective of this study is to establish a rigorous validation framework for synthetic turbulence time series based on pairwise interaction measures. While designed for synthetic data, this approach is applicable to time series generated by predictive models when the forecast horizon is sufficiently long. Rather than relying solely on marginal or second-order statistics, the framework evaluates whether synthetic data preserve both linear and information-theoretic dependencies present in the reference turbulence data. In particular, we employ velocity time series obtained from a computational fluid dynamics (CFD) simulation as the reference (ground truth) and for training a linear vector autoregressive (VAR) model. The VAR model serves as a baseline benchmark to demonstrate the framework, but it can be replaced by other data-driven models, examples of which are cited below. This objective is both timely and novel, given the recent expansion of data-driven models for forecasting and synthesising turbulence time series, and the pressing need for more thorough validation measures. Indeed, outside of turbulence research, comprehensive assessment of the quality of synthetic time series has recently gained attention; see, for example, Bahrpeyma et al. (2021), Stenger et al. (2024).

The generation of inflow conditions for scale-resolving simulations is one of the most important applications of synthetic turbulence time-series generation; see, for example, the review by Wu (2017). In recent years, several deep learning models have been utilised for this purpose, including an autoencoder-type convolutional neural network with a multilayer perceptron (Fukami et al., 2019), generative adversarial networks (Kim and Lee, 2020), a recurrent neural network (RNN) with LSTM (long short-term memory) (Yousif et al., 2022), a combination of transformer and generative models (Yousif et al., 2023), and diffusion models (Li et al., 2024; Liu et al., 2025). Similarly, some of the main approaches for predicting turbulence time series for canonical wall-bounded turbulence include RNN-LSTM (Nakamura et al., 2021; Matsubara et al., 2023), transformers (Drikakis et al., 2024), and echo state network (Racca et al., 2023). Alternatively, statistical (mainly autoregression) and spectral approaches have been used for synthesising turbulence time-series, see, for example, (Klein et al., 2003; Basu et al., 2004; Castro and

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

Paz, 2013; Patruno and Ricci, 2018; Gallego-Castillo et al., 2024; Xavier et al., 2024). These models typically impose spatio-temporal connectivities when constructing the model to preserve time-delayed dependence to an extent. In particular, Stanly et al. (2024) used a VAR model to synthesise the evolution of the temporal coefficients of the proper orthogonal decomposition in a boundary layer flow. The use of synthetic data at the inflow of another simulation resulted in achieving physical turbulence in a short adaptation length. In the present study, the same linear VAR model is used as a benchmark generator of synthetic turbulence time series from channel-flow data, allowing us to assess the extent to which linear and information-theoretic interaction structures are preserved.

Application of the information-theoretic measures (Shannon, 1948; Cover and Thomas, 2001; Kantz and Schreiber 2003) to turbulence has received considerable attention in recent years. Examples include the use of transfer entropy for identifying causality in turbulence structure in different parts of boundary layers (Lozano-Durán et al., 2019; Massaro et al., 2026; Zhang et al., 2025), flow control and subgrid-scale modelling (Lozano-Durán and Arranz, 2022), and driving computational algorithms (Massaro et al., 2023). To ensure the reliability of the information-theoretic analyses, the present study examines the sensitivity of the numerical estimates to estimator parameters and embedding choices, and employs surrogate-based significance tests following Pinto et al. (2024).

The remainder of this paper is organised as follows. Section 2 introduces the methodology, including the turbulence dataset, the VAR modelling approach, and the interaction measures. Section 3 presents the analysis of the linear and non-linear measures applied to the turbulent channel-flow data and the corresponding VAR-based synthetic time series.

2. Methodology

2.1. Turbulence time series data

The CFD-based time series data in this study are obtained from the wall-resolving simulation of a turbulent channel flow at a friction-based Reynolds number $\mathrm{Re}_\tau = 300$ (Figure 1a). This Reynolds number is defined as $\mathrm{Re}_\tau = u_\tau \delta / \nu$ where u_τ is the mean wall friction velocity, δ is the channel half-height, and ν is the kinematic viscosity. As detailed in Rezaeiravesh et al. (2021), the simulation has been performed by the open-source spectral-element solver, Nek5000 (Fischer et al., 2008). The computational mesh consists of fully structured elements with uniform spacing in the streamwise (x) and spanwise (z) directions and compression towards the walls. Each element contains $\left(N_p+1\right)^3$ Gauss–Lobatto–Legendre points (where $N_p = 7$ is the polynomial order), at which the discretised Navier–Stokes equations are solved. The first off-wall grid point is located at $y^+ = 0.55$, where $y^+ = (y/\delta)\,\mathrm{Re}_\tau$ denotes the wall-normal distance in wall units. Periodic conditions are applied in the streamwise and spanwise directions, where the average spacing between grid points is approximately 20 and 10 wall-units, respectively. After discarding the initial transient, a total of 104,377 time samples are collected for the velocity at 64 points in the wall-normal direction, y, across δ. The velocity values are averaged over the wall-parallel (xz) planes and sampled at constant time intervals $0.008\delta/U_b$, equivalent to $0.15\delta/u_\tau$, where U_b is the bulk velocity.

2.2. Vector autoregression (VAR)

Consider a set of K time series $\mathbf{X} = \left[X^{(1)}, X^{(2)}, \cdots, X^{(K)}\right]$ belonging to the same system with some underlying interactions and dependencies. A K-variate VAR model of order p can be used as a linear approximation to the dynamical systems described by the time series (Lütkepohl, 2005):

$$\mathbf{x}_t = \mathbf{A}_0 + \sum_{\ell=1}^{p} \mathbf{A}_\ell \mathbf{x}_{t-\ell} + \boldsymbol{\varepsilon}_t. \tag{1}$$

Here, $\mathbf{x}_t = \left[x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(K)}\right]^T$, $\mathbf{A}_\ell$ is a $K \times K$ coefficient matrix at lag ℓ, and the noise vector (innovation process) $\boldsymbol{\varepsilon}_t$ has a multivariate Gaussian distribution with covariance matrix $\mathbf{C}_\varepsilon$. Given a set of joint samples of length n for the time series, the coefficient and noise covariance matrices can be estimated using Least

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

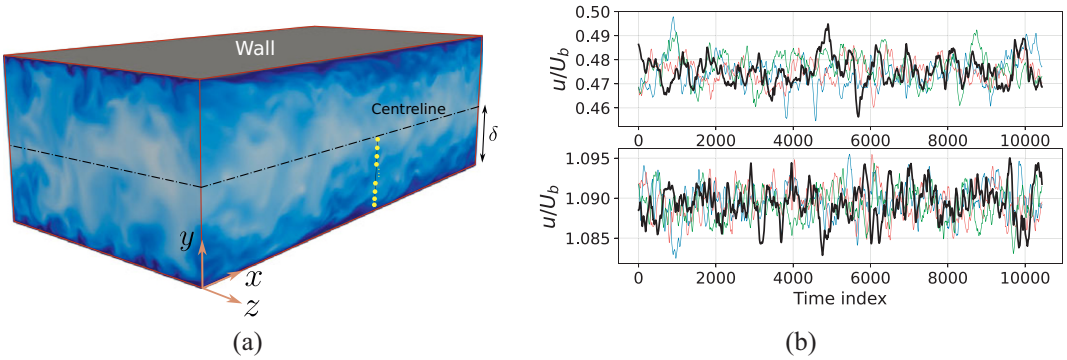


Figure 1. *(a) A snapshot of the streamwise velocity component, u, in the turbulent channel flow considered in this study. The yellow markers schematically indicate the sampling locations in the wall-normal direction, y. At each location, the streamwise velocity averaged over the xz-plane is sampled at a fixed rate to generate a simulation time series. (b) Example time series sampled at* $y^+ = 11$ *(top) and* $y^+ = 180$ *(bottom). The black lines denote the simulation data, and the coloured lines denote three realisations generated by the VAR model fitted to those data.*

Squares (LSE), Maximum Likelihood (MLE), or Bayesian methods. After construction, the VAR(p) model, as a stochastic process, can be used to synthesise samples for $\mathbf{X}$ of any length, with a temporal resolution identical to the training data.

2.3. Interaction measures

In general, there can be various measures of linear and non-linear interactions between each two components of a K-variate time series. The most standard measure of the statistical linear dependence between two time series X and Y is the cross-covariance function (CCF), which measures both concurrent and time-shifted covariances between the time series:

$$\gamma_{XY}(\pm\tau) = \mathrm{cov}\left(\mathrm{x_t}, \mathrm{y_{t\mp\tau}}\right) = \mathbb{E}\left[\left(\mathrm{x_t} - \mu_\mathrm{x}\right)\left(\mathrm{y_{t\mp\tau}} - \mu_\mathrm{y}\right)\right], \tag{2}$$

where $\tau > 0$ is the time shift (lead/lag), $\mathbb{E}[\cdot]$ denotes the expected value, and μ is the mean of a time series. The CCF is asymmetric, meaning that its value changes depending on whether X leading or lagging Y. The only symmetry is $\gamma_{XY}(\pm\tau) = \gamma_{YX}(\mp\tau)$. For $\tau = 0$, the standard cross-covariance between x_t and y_t is recovered, which is not necessarily the maximum value across all possible values of τ. In studies of turbulent flows, $\gamma_{XY}(0)$ is widely used since, when X and Y are velocity components, it represents a component of the Reynolds stress tensor. If both time series are identical, then the CCF (2) is reduced to an autocovariance function to measure how each sample in a time series covaries with its own past. Normalising the autocovariance function with the variance of the time series, the autocorrelation function (ACF) is derived:

$$\rho_X(\tau) = \mathrm{cov}(\mathrm{x_t}, \mathrm{x_{t-\tau}}) / \mathrm{cov}(\mathrm{x_t}, \mathrm{x_t}), \tag{3}$$

that is, symmetric. In practice, where there is only one realisation of a turbulent flow, the expectation in covariances is replaced by a sample-mean estimator. As a result, the CCF and ACF are intrinsically oscillatory at high time lags/leads τ.

Information theory (Cover and Thomas, 2001; Kantz and Schreiber, 2003) provides a set of measures for quantifying both linear and non-linear interactions between time series. A fundamental quantity is the Shannon entropy (Shannon, 1948), which provides a rigorous measure of the uncertainty or information

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

content of a random variable. For X, a continuous random variable (with discrete values), the differential entropy is defined as,

$$H(X) = -\int p(x)\log p(x)\,\mathrm{d}x, \tag{4}$$

where $p(X)$ is the probability density function (PDF) of X. Building on entropy, various measures of interactions between multiple random variables or time series can be obtained. Shannon (1948) proposed mutual information (MI) to measure the amount of shared information between two or more random variables. In connection with dynamical systems, the time-delayed MI was proposed by Fraser and Swinney (1986), which for two time series X and Y is defined as:

$$I_{XY}(\tau) = I(x_t, y_{t-\tau}) = \sum_{x_t, y_{t-\tau}} p(x_t, y_{t-\tau}) \log \frac{p(x_t, y_{t-\tau})}{p(x_t)p(y_{t-\tau})}, \tag{5}$$

where $\tau \geq 0$ is the time delay and $p(x_t, y_{t-\tau})$ is the joint PDF of x_t and the delayed y_t. This measures how much knowledge about the past of Y, that is, $y_{t-\tau}$, reduces uncertainty about the current state of X. Therefore, it is a measure of statistical dependence between X and Y considering both linear and non-linear effects in contrast to CCF (2). Clearly, $I_{XY}(\tau) \neq I_{YX}(\tau)$, meaning the time-delayed mutual information is, in general, asymmetric in time. However, because mutual information is symmetric under interchange of the two variables in the joint PDF in Equation (5)it does not, by itself, capture the directionality of information transfer. To obtain this, a conditional PDF was used by Schreiber (2000) who defined transfer entropy (TE) from X to Y as,

$$TE_{X\to Y}(m,\tau) = \sum_{x_t, y_t^m, x_t^m} p\left(y_t, y_t^m, x_t^m\right) \log \frac{p\left(y_t | y_t^m, x_t^m\right)}{p\left(y_t | y_t^m\right)}. \tag{6}$$

Given samples x_t and y_t from a single realisation, the time-delay embedding is used to reconstruct multi-dimensional state-space:

$$x_t^m = \left\{x_{t-(\ell-1)\tau}\right\}_{\ell=1}^{m}, \quad y_t^m = \left\{y_{t-(\ell-1)\tau}\right\}_{\ell=1}^{m}, \tag{7}$$

where $m \geq 1$ and $\tau \geq 1$ are the embedding dimension and embedding delay, respectively. The time-delay embedding used here follows the practical reconstruction procedure introduced by Packard et al. (1980), and is theoretically motivated by Takens' embedding theorem (Takens, 1981), which establishes conditions under which the attractor of a dynamical system can be reconstructed from delayed observations of a scalar time series. Note that in the original definition by Schreiber (2000), each time series can have a different m, but in this study they are assumed to be the same. The choice of m and τ can have a significant influence on the value of the estimated TE; see Tan et al. (2023), Wallot and Mønster (2018), and also the discussion in Section 3. According to Equation (6), $\mathrm{TE}_{X\to Y}$ quantifies how much additional information is learned from the past of X in addition to what is already known from the past of Y, for predicting a future state of Y. This interpretation is directly reflected in the conditional PDFs in Equation (6). Therefore, $\mathrm{TE}_{X\to Y}$ serves as a measure of statistical causality, indicating the extent to which X influences Y. However, the present bivariate TE is evaluated for the pair X and Y in isolation from the remaining time series. As a result, its causal interpretation is limited, since potential common drivers, redundant effects, and other multivariate dependencies are not taken into account. Nevertheless, it remains suitable for the purpose of the present study, that is, quantifying pairwise interactions between time series. In addition, compared with multivariate TE, the estimation of bivariate TE is more robust for a given dataset because of the lower dimensionality of the embedding space. Obviously, $\mathrm{TE}_{X\to Y}(m,\tau) \neq \mathrm{TE}_{Y\to X}(m,\tau)$ motivating the definition of the net TE as $\mathrm{TE}_{X\to Y} - \mathrm{TE}_{Y\to X}$ which indicates the dominant direction of information transfer in a time series pair.

Given a finite set of samples for two time series X and Y, computing MI and TE requires estimating various types of PDFs. The most basic approach is using binning to empirically estimate the PDFs, which,

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

however, can be very sensitive to various parameters, for example, the bin size. The estimates can be improved using the Kernel Density Estimation (KDE) method that returns smooth PDFs. This can, however, lead to filtering some details. In general, the KDE method is sensitive to bandwidth selection and may struggle with high-dimensional data; see, for example, Lee et al. (2012), Zambom and Dias (2013), and Appendix A. The most conventional method for estimating MI is based on k-nearest neighbour (KNN) distances, hereafter referred to as the KSG method after Kraskov et al. (2004). We use their estimator for the mutual information that was derived based on an expansion of different Shannon entropies (Cover and Thomas, 2001). It reads as,

$$I_{XY} \approx \psi(k) + \psi(n) - \langle \psi(n_X + 1) + \psi(n_Y + 1) \rangle, \tag{8}$$

where k is the number of nearest neighbours, $\psi(\cdot)$ is the digamma function, n is the length of the time series, $\langle \cdot \rangle$ denotes an average over n, and $n_X(t)$ specifies the number of $x_{t'}$ samples whose distance from x_t is less than a radius (computed in the method). The TE in Equation (6) can be written as a conditional time-delayed MI and expanded in terms of entropies (Frenzel and Pompe, 2007; Hlaváčková-Schindler et al., 2007):

$$TE_{X\to Y}(m,\tau) = I\left(y_t, x_t^m | y_t^m\right) = H\left(y_t, y_t^m\right) - H\left(y_t^m\right) - \left(H\left(y_t, x_t^m, y_t^m\right) - H\left(x_t^m, y_t^m\right)\right). \tag{9}$$

Based on this, Frenzel and Pompe (2007) extended the KSG estimator (Kraskov et al., 2004) to compute $\mathrm{TE}_{X\to Y}$ from,

$$TE_{X\to Y}(m,\tau) \approx \psi(k) + \langle \psi\left(n_{y_t^m} + 1\right) - \psi\left(n_{y_t, y_t^m} + 1\right) - \psi\left(n_{x_t^m y_t^m} + 1\right) \rangle, \tag{10}$$

where $n_{y_t^m}$, n_{y_t, y_t^m}, and $n_{x_t^m y_t^m}$ count the number of samples falling within a specific radius of the respective marginal space. Although the KSG estimators are non-parametric, their estimates can be affected by the choice of k. In general, smaller k values result in lower bias but more variance in the estimates (capturing local effects and letting the PDF vary over different regions), and vice versa (Stoļarovs et al., 2026). Gao et al. (2018) explained why KSG estimations are accurate for fixed values of k. Furthermore, there are methods to choose optimal k values to reduce the bias in the estimator, see, for example, Witter and Houghton (2024). In this study, we select $k = 10$ as a pragmatic intermediate value, rather than claiming it as an optimal choice across all time-series pairs. The sensitivity of the estimated information-theoretic measures to the choice of k is further detailed in Appendix A. Using the same value $k = 10$ for both the simulation and synthetic time-series pairs is consistent with the main objective of the study, namely validation, although its potential impact on the physical interpretation of the results should be acknowledged.

3. Results and discussion

A set of 22 non-equispaced points (a subset of the original 64 points) in the wall-normal direction of the channel flow described in Section 2.1 is selected, at which time series of plane-averaged streamwise velocity are considered, see Figure 1a. The corresponding wall-normal locations span from $y^+ = 0.55$ (adjacent to the wall) to $y^+ = 298$ (the channel centre), with multiple sampling points located within the viscous sublayer, buffer layer, logarithmic layer, and wake region of the turbulent boundary layer (Pope, 2000). The time series contain 10,437 samples, obtained by down-sampling the original CFD data by a factor of 10, resulting in a temporal resolution of $0.08\delta/U_b$. The resulting 22 time series create 231 distinct interacting pairs, for which the interaction measures in Section 2.3 are evaluated using an in-house Python code. To reduce overall computational cost, a message-passing interface (MPI) parallelisation is implemented, where the computations associated with each pair of time series are assigned to a separate rank.

A VAR model of order $p = 60$ is trained using the above 22-variate time series, employing the least-squares estimator available in statsmodels (Seabold and Perktold, 2010). Example time series generated by the VAR model at two sampling locations are shown in Figure 1b. It is noteworthy that applying the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) yields suitable VAR orders of 20 and 15, respectively. However, as the VAR model is

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

used here for synthetic sample generation to reproduce the interaction structure of the time series, we adopt $p = 60$ as a pragmatic choice to retain sufficiently long temporal memory. Importantly, repeating the analyses in the following sections with $p = 30$ does not lead to any discernible change in the results. The only noticeable difference is that $p = 60$ yields a slightly better agreement of the mean cross-covariance function with the simulation data at higher lags (see Figure 3).

Using the trained VAR model, 50 independent realisations of synthetic time series, each of length 10,437, are generated. The strategy in the subsequent analyses is to compare the empirical mean of the pairwise interaction measures obtained from the 50 synthetic realisations with the corresponding measures computed from the simulation data (ground truth) used for training of the VAR model. The ensemble of synthetic realisations also allows for estimating 95% confidence intervals for the interaction measures between the VAR-generated time series.

3.1. Linear interactions

Figure 2a shows the matrix of mean correlation coefficients for the time series pairs obtained from the VAR model (the values between identical time series are excluded from all matrix plots). As expected, the time series show a high positive correlation with the neighbouring wall distances. For $y^+ \lesssim 25$ (almost up to the end of the buffer layer), a high concurrent (i.e., no time-delay) linear dependency between time series is observed, as is known in the classical boundary layer theory, see, for example, Pope (2000). It is interesting to spot two regions of small/zero correlation encompassing pairs with highly negative correlations. In both regions, the time series up to $y^+ \lesssim 27$ are involved, which show small/no linear correlation with the dynamics near the channel centre as well as with $50 \lesssim y^+ \lesssim 80$. As shown in Figure 2b, the same sets of interacting time series correspond to most of the high-uncertainty correlations of synthetic data. In contrast, for highly correlated pairs, the uncertainty in the correlation coefficient of the synthetic time series is low. As expected from a linear model, the VAR model perfectly preserves the correlation coefficient as compared to the ground truth, as inferred from the scatter plot in Figure 2c.

Next, we examine additional second-order linear measures that account for time-delayed dependencies, specifically the ACF and CCF. In Figure 3, these measures are evaluated for two pairs of time series with highly negative and positive correlation coefficients. The ACF is a measure of linear dependency of a

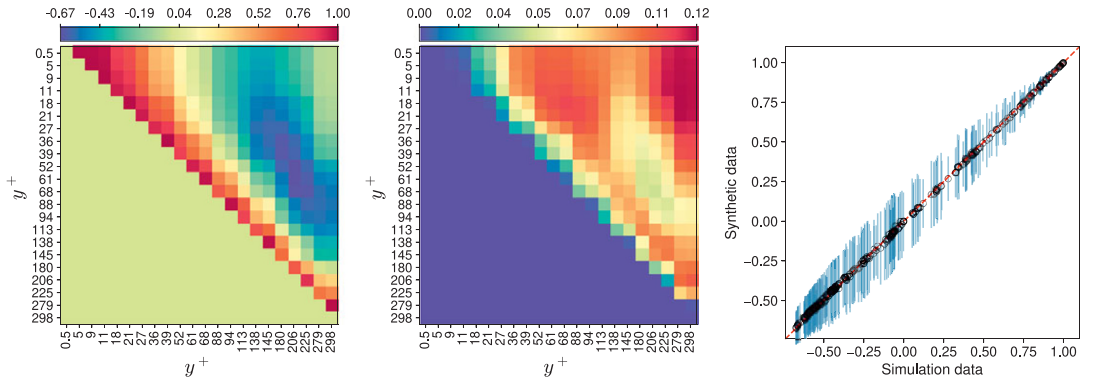


Figure 2. *(a, b) Matrices of the mean and standard deviation of the correlation coefficient obtained from the VAR synthetic data, and (c) scatter plot of the correlation coefficient between the 231 time series pairs within the synthetic and simulation data. In (a, b), the y^+ values of the 22 sampling points are shown on the axes. In (c), the bars present 95% confidence intervals around the mean values of the synthetic data, due to the stochastic nature of the VAR model. In each figure in this study, the colourbar limits are set to the minimum and maximum values of the corresponding dataset.*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

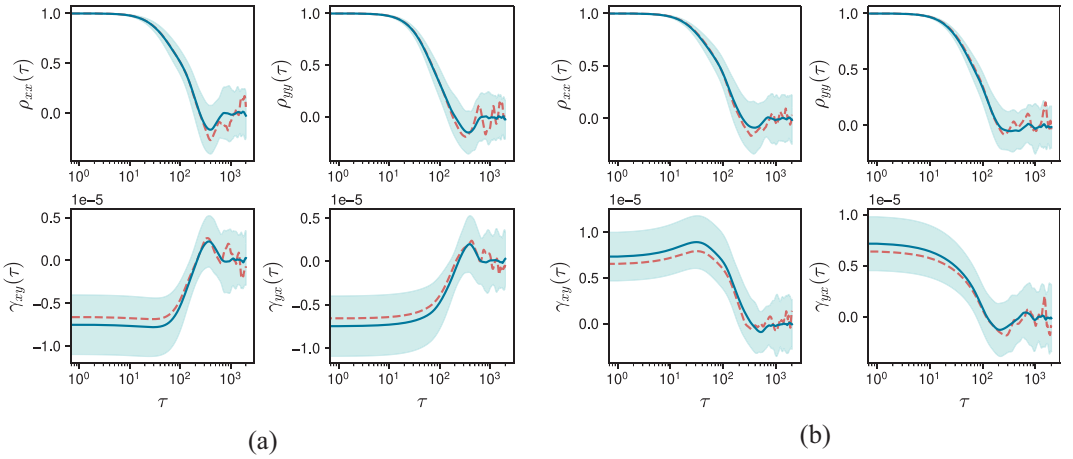


Figure 3. *Linear time-delayed interactions between time series located at (a) $y^+ = 36, 180$ and (b) $y^+ = 61, 88$. In each subplot, the top and bottom rows show the ACF and CCF versus time lag τ, respectively. Dashed lines: simulation data, solid lines: mean of the synthetic data, and shaded area:* 95% *confidence interval for the synthetic data.*

time series with its own past, and indicates the time scale over which linear memory lasts. Similarly, the CCF measures how the past (or future) of a time series linearly covaries with the future (or past) of another time series. For the time series analysed here, the ACFs start from 1 and decay to zero at high lags, although they may show non-monotonic decay at higher lags. The latter can be an indication of a dominant time-periodic structure in the time series' past. The time it takes for the ACF of a time series to reach zero can be taken as an integral time scale required to make samples uncorrelated to their past. As visible in Figure 3, such time can be significantly long, of the order of a few hundred time lags. Such long-lasting memories are, however, known in the turbulence community and, in practice, lead to slow convergence of turbulence statistics; see, for example, Oliver et al. (2014), Rezaeiravesh et al. (2025), Xavier et al. (2024).

Depending on the time series pair, the plots of CCFs can exhibit more diverse patterns. For any pair, the zero-lag value satisfies $\gamma_{XY}(0) = \gamma_{YX}(0)$, which represents the cross-covariance between the two concurrent time series. At higher lags, however, asymmetries in the CCF may arise. For instance, in Figure 3b, a "bump" is observed at higher τ values for $\gamma_{XY}(\tau)$, suggesting the existence of event(s) in the past of Y that highly linearly covaries with the future of X (but not the other way around, in this specific case). This aligns with earlier observations by Massaro et al. (2023). Across all considered time series pairs, the VAR model reliably reproduces the ACF and CCF, showing good agreement with simulation data. Note that due to the empirical average over the VAR realisations, the intrinsic oscillations in the ACF and CCF at high lags of the synthetic data exhibit lower amplitude than those observed in the simulation data. Furthermore, the empirical uncertainty in the ACF increases with τ, with the highest values at non-physical oscillations around zero ACF. In contrast, for the CCF, the uncertainty in the VAR values is high across all lags, in addition to a small deviation between the mean CCF of VAR and the sample-estimated CCF of the simulation data. It is noteworthy that the Fourier transform of the ACF and CCF leads to power spectral and cross-spectral densities, respectively (see, e.g., Park, 2018), that are instrumental in turbulence studies. Therefore, the excellent agreement observed here between the linear time-delayed covariances of the VAR-synthetic and simulation time series is also expected to hold in the frequency domain.

3.2. Information-theoretic measures

The time-delayed mutual information (5) can measure the amount of shared information between a pair of time series. While $I_{XY}(\tau)$ is comparable to the CCF $\gamma_{XY}(\tau)$, it can capture both linear and non-linear

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

dependencies between X and Y (Shannon, 1948). The time-delayed MI has been previously evaluated for a few turbulent flows, see Arranz and Lozano-Durán (2024), Cerbus and Goldburg (2013), Granero-Belinchon (2018). For the present data, the time-delayed MI was found to be, in general, a decaying function of τ (see, e.g., Figure 5a), where the decay rate could be regarded as the rate at which information (dependencies) between two time series is lost at different τ. In particular, we are mainly interested in the maximum time-delayed MI, which is heuristically computed by searching over a set of candidate values of τ. The motivation for this choice is that the lag at which the strongest time-delayed dependence occurs varies across interacting pairs, while the maximum of the time-delayed MI serves as a compact summary measure. Although this choice may be more sensitive to noise than MI evaluated at a prescribed lag, it avoids introducing an arbitrary common lag across all pairs and is applied consistently to both the simulation and VAR data. In Figure 4a, the mean value of this measure is shown for different pairs of time series synthesised by the VAR model. The vertical and horizontal axes show the y^+ value of the source and target time series, respectively. The plot suggests a high level of symmetry in maximum TE, evident from the values on both sides of the main diagonal. In addition to high values of maximum MI between spatially close locations, a set of high values is observed below $y^+ \approx 21$. For the time series very close to the wall (at $y^+ = 0.5$), the maximum MI remains small when the y^+ of the pair increases beyond ≈ 27. This is in contrast to the behaviour of the correlation coefficient in Figure 2a.

According to the scatter plot in Figure 4b, the mean of the maximum MI obtained from the synthetic data agrees relatively well with those from the simulation data. Remarkably, this indicates that the PDFs in Equation (5) estimated from the VAR predictions closely approximate those of the simulation data, hence resulting in comparable values of the summation. The accuracy of the maximum MI predicted by the VAR data is further assessed in Figure 4c, via evaluating the following normalised error:

$$R = \left(\overline{\max_\tau \mathrm{I}_{\mathrm{XY}}(\tau)_{\mathrm{VAR}}} - \max_\tau \mathrm{I}_{\mathrm{XY}}(\tau)_{\mathrm{Simulation}} \right) / \max_\tau \mathrm{I}_{\mathrm{XY}}(\tau)_{\mathrm{Simulation}}, \tag{11}$$

where the overbar stands for the empirical averaging over the repetitions of the VAR predictions. For most of the pairs, the error is positive, indicating over-prediction of time-delayed MI by the VAR synthesised data, which could be translated to spurious shared information. Additionally, the error plot is, to a large extent, asymmetric and shows that the highest errors more frequently occur when a time series at $y^+ \geq 88$ interacts with a time series in the same region or very close to the wall.

Next, we consider bivariate TE as a measure of non-linear statistical causality between a pair of time series. A key to accurate estimation of TE for a dynamical system is constructing a valid time-delay embedding; see Equation (7). We consider multiple candidate values of the embedding dimension m to

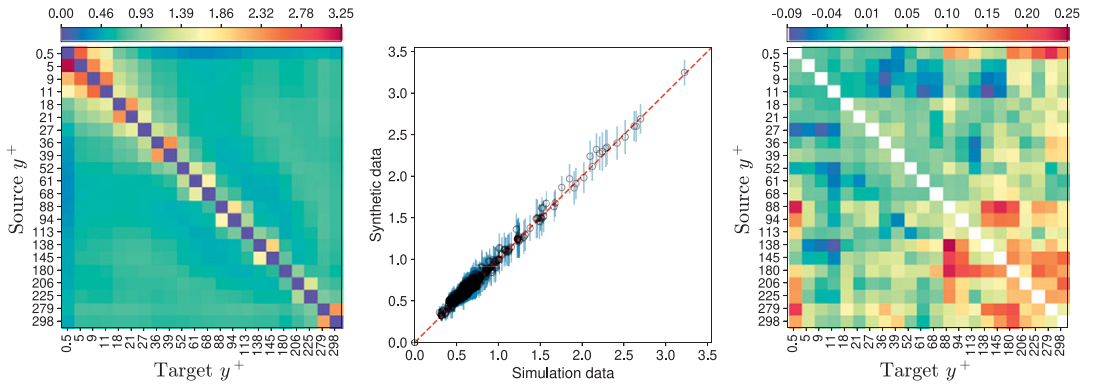


Figure 4. *(a) Matrix of the mean maximum MI estimated from synthetic data, (b) scatter plot of the maximum MI obtained from the synthetic and simulation data, and (c) matrix of the normalised error (equation 11).*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

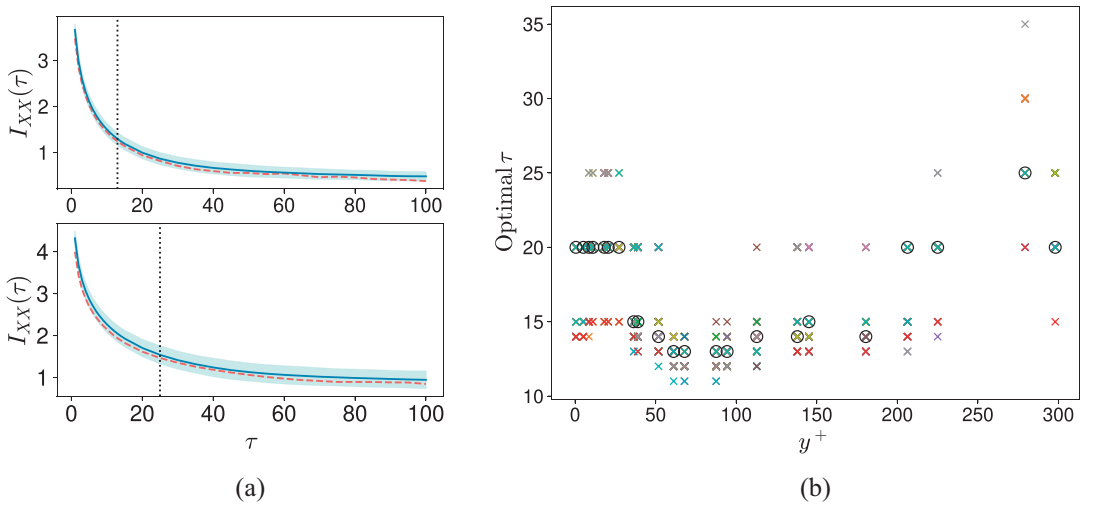


Figure 5. *(a) Auto-mutual information function (AMIF) for a pair of time series associated with (top) $y^+ = 88$ and (bottom) $y^+ = 279$. The dashed lines show the estimation from the simulation data, and the solid lines represent the mean of the AMIF from the synthetic data (along with the 95% confidence interval). The dotted vertical lines specify the optimal value of embedding delay τ. (b) The optimal τ for time series at the y^+ values considered in the analyses: ∘ corresponds to the simulation data and × are associated with the samples generated from VAR model. These optimal values of τ have been used for estimation of TE in Figures 6–9.*

compute TE across several wall-normal locations. For high-order dynamical systems like turbulent flows, selecting the time delay τ is non-trivial and critically affects the results, see Appendix B. To optimise τ, several methods have been suggested in the literature for time series in various applications, see Wallot and Mønster (2018), Tan et al. (2023), and references therein. For chaotic dynamical systems, Fraser and Swinney (1986) showed choosing τ based on the local minimum of $I_{XX}(\tau)$, the time-delayed MI of a time series with its own past, aka, auto-mutual information function (AMIF), can be a better option compared to the local minimum or first zero of the ACF of X. Further, relying on ACF could be misleading, noting its linearity and intrinsic oscillations shown in Figure 3, which require a significant number of samples to obtain a smooth ACF (Massaro et al., 2023; Xavier et al., 2024; Rezaeiravesh et al., 2025). For the present time series, the $I_{XX}(\tau)$ is observed to be monotonically decaying with τ reaching a saturation-like status (plateau) at high lags, see Figure 5a. Note that the integral time scale from an ACF, Figure 3, is larger than the corresponding time scale computed from AMIF (i.e., the ACF decays more slowly than the AMIF with τ). This can be a strong indicator of the linear statistical dependence in the time series considered here being longer-term than the non-linear ones. For monotonically decaying AMIFs, the optimal embedding delay τ for estimating TE has been suggested to be chosen as the lowest lag above which $I_{XX}(\tau)/I_{XX}(0)$ drops below $1/e$, see Wallot and Mønster (2018) and Chapter 9 in Kantz and Schreiber (2003). Following this, the optimal τ for computing TE between different time series pairs from both simulation and VAR data are obtained, see Figure 5b. These optimal embedding delays, τ_{opt}, are clearly larger than the default value of 1, and also vary with the distance from the wall. Considering the simulation data, for y^+ up to ≈ 39, τ_{opt} is 20, then it drops to smaller values up to $y^+ \approx 200$, and above this, it has a larger value of 25 before dropping back to 20 at the channel centre. Note that the value of τ_{opt} at a y^+ can be different among realisations of the VAR model.

Figure 6 shows the variation of TE for four time series pairs as a function of the embedding dimension m. In general, the maximum of TE for all pairs was observed to occur at $m = 1$ to 3 (not shown here). Interestingly, the highest deviation between the simulation and synthetic data, as well as the highest

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

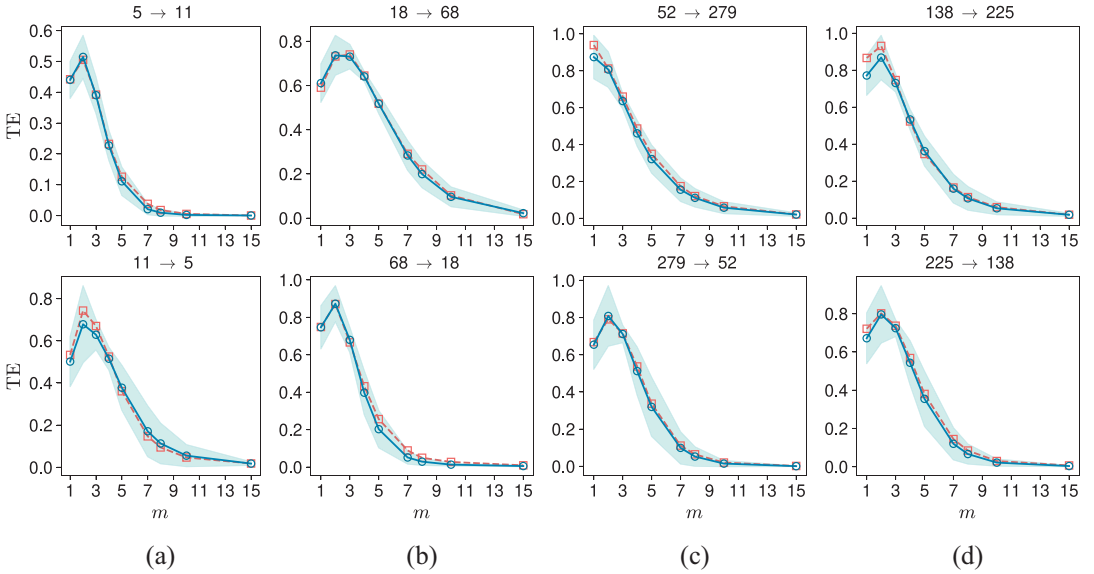


Figure 6. *Variation of the TE with embedding dimension m, estimated using the simulation data (dashed lines) and synthetic data (mean value: solid line) for time series pairs with y^+ equal to (a)* $(5,11)$*, (b)* $(18,68)$*, (c)* $(52,279)$*, and (d)* $(138,225)$*. The maximum values of TE are used in Equation (2) and later shown in Figure 9.*

uncertainty in the values obtained from synthetic data, are found to be around the maximum TE. This measure (in both directions) computed between all 231 time series pairs synthesised by the VAR model is shown in Figure 7a. The maximum TE for most pairs is considerably high (roughly more than 0.5). Importantly, the asymmetry with respect to the main diagonal suggests a dominant direction of pairwise interaction, see below. For interacting time series in the viscous sublayer, $y^+ \leq 5$, the maximum TE is lower than that obtained between the rest of the pairs. Moreover, the maximum TE from and to the closest point to the wall is generally low. To assess whether the estimated TEs are statistically significant, we

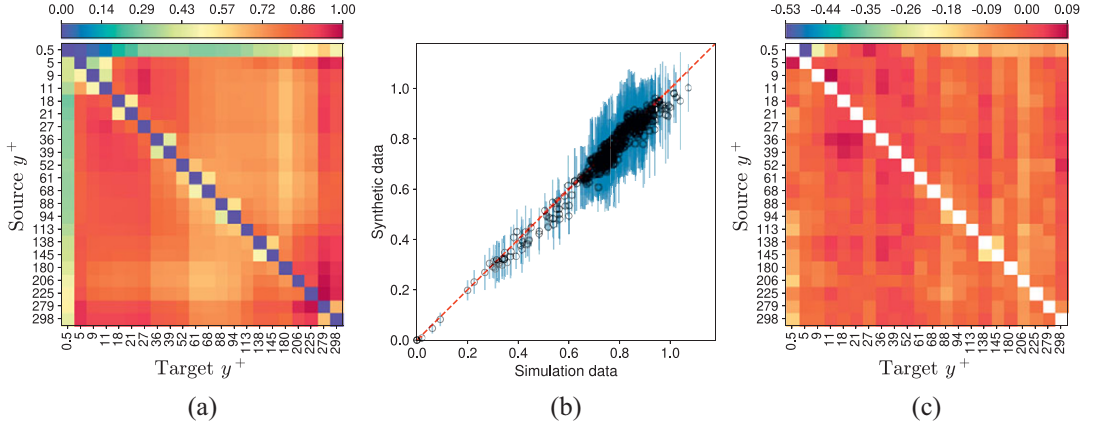


Figure 7. *(a) Mean of the maximum TE estimated from the VAR synthetic data, (b) scatter plot of the maximum TE estimated from the VAR and simulation data, and (c) normalised error in the maximum TE estimated from the VAR synthetic data.*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

performed surrogate-based hypothesis testing. In particular, we considered the row-shuffling surrogate proposed by Pinto et al. (2024), in which, for $\mathrm{TE}_{X\to Y}$ the rows of the source-history embedding, that is, the matrix associated with x_t^m in Equation (7), are randomly permuted while the target-history and target-future samples are kept fixed. This destroys the row-wise alignment between the source-history vectors and the target-side samples used in the TE estimator. The corresponding null hypothesis is that there is no directed dependence from the source past to the target future beyond the target's own past, that is, that the observed TE can be explained by chance association under this destroyed alignment. For each interacting pair of simulation time series, 200 such surrogates were generated, and the TE in both directions were estimated using the same settings as for the original time series. For all interacting pairs considered in this study, and for $k = 5, 10$ and 30 in the KNN estimator, none of the surrogate TE values exceeded the TE estimated from the original data, yielding the minimum Monte Carlo p-value of $1/201 \approx 0.00498$. This indicates a clear rejection of the row-shuffling null hypothesis for all pairs.

Compared to the correlation coefficient and time-delayed mutual information (Figures 2 and 4), the mean maximum TE from the VAR-generated time series shows a larger deviation from the corresponding simulation estimates, as illustrated in Figure 7b. Such deviations are expected because the VAR model is linear and assumes multivariate Gaussian noise, while physical turbulence time series are, in general, non-linear and non-Gaussian (Castaing et al., 1990; Jiménez, 1998). Remarkably, the overall performance of the VAR model in predicting the TE with this level of accuracy is above expectations. The matrix of the normalised error in maximum TE, computed similarly to Equation (11), is illustrated in Figure 7c. For most pairs, the maximum TE obtained from the VAR model is smaller than that of the simulation data, indicating that a part of the information transfer that should have been non-linear has not been captured by the VAR model. To further examine whether and where the interaction between pairs of simulation time series contains a non-linear component, an additional surrogate-based hypothesis test was performed. We employed the Bivariate Iterative Amplitude Adjusted Fourier Transform (Bi-IAAFT) surrogate proposed by Pinto et al. (2024), extending the IAAFT surrogate of Schreiber and Schmitz (1996). The associated null hypothesis is that the time series X and Y arise from processes that are exclusively linearly correlated. A Bi-IAAFT surrogate preserves the histograms of X and Y, as well as their power spectra and cross-spectrum. For each interacting pair, 200 Bi-IAAFT surrogates were generated, and the TE was estimated using the same settings as for the original data. The resulting rejection/non-rejection matrix at the 5% significance level for the default choice $k = 10$ is shown in Figure 8, while the corresponding results for $k = 5$ and $k = 30$ are provided in Appendix A. For most interacting pairs, the null hypothesis is rejected,

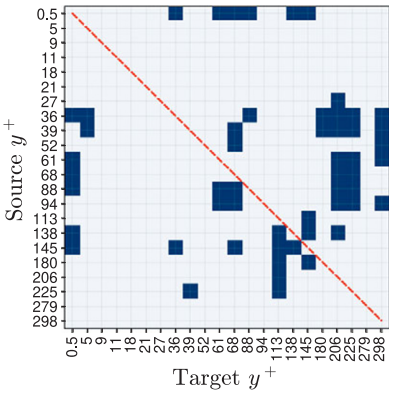


Figure 8. *Matrix of rejection (light colour)/non-rejection (dark colour) outcomes for the Bi-IAAFT surrogate test (Pinto et al., 2024) applied to TE between all interacting pairs of the simulation time series, using* $k = 10$*. Entries marked as rejected correspond to p-value less than* 0.05 *under the null hypothesis that the paired time series are exclusively linearly correlated. The dashed line shows the main diagonal of the matrix.*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

indicating that the corresponding TE values cannot be fully explained by preserved linear second-order structure alone and therefore suggesting a non-linear contribution to the interaction. Furthermore, the asymmetry of the matrix with respect to the main diagonal indicates that the evidence of a non-linear contribution to the pairwise interaction can itself be direction-dependent.

Despite the discrepancy between the TE values of the synthetic and simulation time series, we aim to investigate how much the synthetic data can preserve the correct dominant direction of information transfer between each time series pair. The net transfer entropy is defined in analogy with the flux in Massaro et al. (2026). For this, we use normalised maximum TE which, for each interacting pair from the simulation data, is defined as,

$$\mathrm{Net}\max_m \mathrm{TE}_{X\to Y}(\mathrm{pair};m) = \frac{\max_m \mathrm{TE}_{X\to Y}(\mathrm{pair};m) - \max_m \mathrm{TE}_{Y\to X}(\mathrm{pair};m)}{\max_{\mathrm{pair}}\left(\max_m \mathrm{TE}_{X\to Y}(\mathrm{pair};m) - \max_m \mathrm{TE}_{Y\to X}(\mathrm{pair};m)\right)}. \tag{12}$$

A similar definition is used for the synthetic data, using the $\max_m$TE values averaged over realisations. Here, $\max_{\mathrm{pair}}$ denotes the maximum of the numerator in Equation (12) over all interacting pairs considered. Hence, the normalisation is performed globally over the full set of interacting pairs, but separately for the simulation and VAR-synthetic datasets. The resulting measure is plotted in Figure 9. There is a strong agreement between the two plots regarding the positions of non-zero values, indicating that the VAR model can effectively preserve the direction of the net information transfer for most of the time series pairs. A cautious conclusion could be that linear mechanisms dominate the underlying information transfer in these interactions. In contrast to the direction, the intensity of information flows is not necessarily preserved (under-predicted by the synthetic data). Further, compared to the ground truth, the VAR model is unable to capture the variation in the magnitude of the net maximum TE, and returns more uniform values in the plot.

The insights from Figure 9 are consistent with established findings in the turbulence literature regarding the interactions and directional relationships between different regions of a boundary layer. This suggests that statistical causality, quantified here by bivariate transfer entropy measuring directed pairwise interactions between such regions, can show consistency with relevant physical interpretations. For clarity, we examine zones *A*–*D*, whose boundaries are determined by the respective ranges of the source y^+: *A*: $y^+ \lesssim 21$, *B*: $27 \lesssim y^+ \lesssim 39$, *C*: $52 \lesssim y^+ \lesssim 180$, and *D*: $206 \lesssim y^+ \lesssim 298$. The values of the net maximum TE in zone *A* imply that the information transfer from the wall is strong only up to the location of maximum turbulent kinetic energy production ($y^+ \lesssim 12$); see also Massaro et al. (2026). This agrees with the intrusive causality studies by Flores and Jiménez (2006), Mizuno and Jiménez (2013), where the turbulence away from the wall could sustain after removing the near-wall region from simulations. Zone *B*

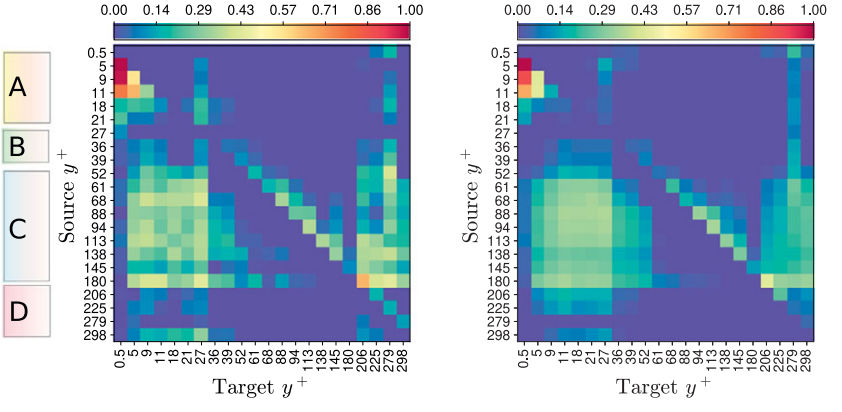


Figure 9. *Normalised net maximum TE defined in Equation (12) evaluated using the (a) simulation and (b) synthetic data (only the mean values are shown).*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

corresponds to the buffer layer, and does not show any significant systematic directed interaction. Zone *C* can be considered as the logarithmic layer acting as an information source. As this layer is the overlap between the inner and outer regions of the boundary layer, its interactions are expected to be significant with different regions in the boundary layer. This is confirmed here, as Zone *C* exhibits directed pairwise interaction towards the near-wall region (up to $y^+ \approx 39$), the wake region ($206 \lesssim y^+ \lesssim 298$), and also higher target y^+ within $52 \lesssim y^+ \lesssim 180$ (visible in the non-zero upper diagonal elements in the plot). The net information transfer from the logarithmic layer towards Zones *A* and *B* is consistent with the baseline of hybrid and wall-modelled large eddy simulations (Larsson et al., 2016), where the information from the core flow is used to model/reconstruct unresolved effects near the wall. Interestingly, the high net TE from Zone *D* to the near-wall region ($y^+ \lesssim 27$) could provide evidence of top-down interactions that likely reflect the near-wall footprints (amplitude modulation) of the large structures; see Marusic et al. (2010), Mathis et al. (2009). It is emphasised again that a more in-depth physical interpretation would require the computation of multivariate TE.

4. Conclusions

In this study, we evaluate both linear and information-theoretic pairwise interactions between streamwise-velocity time series in a turbulent channel flow. The analysis considered CFD simulation data together with an ensemble of 50 synthetic datasets generated by a VAR model trained on the simulation data. The evaluation of the interactions between an extensive set of 231 time series pairs provides a detailed picture of the variation in different statistical measures across the entire turbulent channel flow boundary layer. The VAR synthetic data can completely preserve both concurrent and time-delayed linear relationships measured by autocorrelation and cross-covariance functions. Such agreement is expected, since the VAR formulation explicitly incorporates linear dependencies over multiple time lags. Moreover, these observations confirm that the selected model order ($p = 60$) is appropriate. It is noteworthy that the VAR model with order 30 was also examined, which did not show a significant impact on the results.

The VAR synthetic data retained the time-delayed mutual information with a reasonable fidelity, demonstrating partial preservation of non-linear dependencies, despite slight spurious over-predictions. Transfer entropy, however, showed systematic under-predictions for most pairs relative to the simulation data. Despite this, the normalised net maximum transfer entropy of the VAR synthetic data was found to match relatively well with those obtained from the simulation time series. This measure can be utilised for identifying the direction of the dominant information flow within each time series pair. Therefore, an important conclusion is that a VAR model, with sufficiently high order, can properly predict the direction of information transfer despite inaccurately predicting its intensity. This suggests that linear interactions with longer time scales are dominant in the information transfer. Overall, the results indicate that a VAR model can effectively function as a model for turbulence time series synthesis, although its limitations should be acknowledged.

Transfer entropy is commonly used as a statistical measure of causality. In this study, we observed that bivariate TE, despite its limitations, can yield patterns that are consistent with relevant physical causality reported in the turbulence literature, particularly across different regions of boundary layers (Jiménez, 2013; Lozano-Durán et al., 2019; Martínez-Sánchez et al., 2024; Massaro et al., 2026; Zhang et al., 2025). Specifically, this study showed the importance of employing optimal delays for accurate estimation of TE. For optimisation, we used the auto-mutual information function, which showed a shorter integral time scale than the autocorrelation function. According to Appendix B, using the default embedding delay can significantly affect the computed values of transfer entropy, particularly for VAR synthetic data. In addition, the choice of estimator for the non-linear interaction measures was found to be important. The results presented in the main body of the paper were obtained using the KSG estimators (Kraskov et al., 2004; Frenzel and Pompe, 2007) with $k = 10$. As shown in Appendix A, varying the value of k has some influence on the results, see also Stoļarovs et al., 2026, although this effect is considerably smaller than the impact of switching to a KDE-based estimator.

Given the non-intrusive nature and significance of the analyses performed in this study, we recommend applying the same methodology to synthetic or forecasted turbulence time series generated by other

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

models. This is crucial for assessing the degree to which such data preserve the underlying physical interactions and statistical dependencies found in simulations or experiments.

The present study can be extended in several directions. To better capture the rare and extreme turbulent events, moving-average terms can be incorporated into the VAR model. The velocity time series used here were spatially averaged over the homogeneous direction of the channel flow, which might influence the performance of the VAR model. Repeating the analyses particularly at higher Reynolds numbers where scale separation is more pronounced should also be considered. Finally, this study focused exclusively on pairwise transfer entropy. To enhance physical interpretability and identify the actual drivers within the turbulent causal network, future works will incorporate multivariate transfer entropy.

Data availability statement. The source code and turbulent channel flow time series data used to compute the interactions in this study are available at: https://github.com/simFlowDataLab/pairwiseValidation_turbulenceTS (Rezaeiravesh et al., 2026).

Acknowledgements. This work used the ARCHER2 UK National Supercomputing Service (https://www.archer2.ac.uk). The authors would like to thank the UK EPSRC for providing computational time for the computations on ARCHER2 via the UK Turbulence Consortium (EP/D44073/1 and EP/G06958/1).

Author contribution. Conceptualisation: S.R., D.M., P.S.; Methodology: S.R., D.M.; Software: S.R.; Data curation: S.R.; Visualisation: S.R.; Formal analysis: S.R., D.M.; Writing original draft: S.R; Writing—review & editing: D.M., P.S. All authors approved the final submitted draft.

Funding statement. D.M. acknowledges the financial support provided by the Knut and Alice Wallenberg Foundation.

Competing interests. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Arranz G and Lozano-Durán A (2024) Informative and non-informative decomposition of turbulent flow fields. *Journal of Fluid Mechanics 1000*. https://doi.org/10.1017/jfm.2024.1007.

Bahrpeyma F, **Roantree M**, **Cappellari P**, **Scriney M and McCarren A** (2021) A methodology for validating diversity in synthetic time series generation. *MethodsX 8*, 101459. https://doi.org/10.1016/j.mex.2021.101459.

Basu S, **Foufoula-Georgiou E and Porté-Agel F** (2004) Synthetic turbulence, fractal interpolation, and large-eddy simulation. *Physical Review E 70*(2). https://doi.org/10.1103/PhysRevE.70.026310.

Castaing B, **Gagne Y and Hopfinger E** (1990) Velocity probability density functions of high Reynolds number turbulence. *Physica D: Nonlinear Phenomena 46*(2), 177–200. https://doi.org/10.1016/0167-2789(90)90035-N.

Castro HG and Paz RR (2013) A time and space correlated turbulence synthesis method for large eddy simulations. *Journal of Computational Physics 235*, 742–763. https://doi.org/10.1016/j.jcp.2012.10.035.

Cerbus RT and Goldburg WI (2013) Information content of turbulence. *Physical Review E 88*(5). https://doi.org/10.1103/PhysRevE.88.053012.

Cover TM and Thomas JA (2001) *Elements of Information Theory*. Wiley. https://doi.org/10.1002/0471200611.

Drikakis D, **Kokkinakis IW**, **Fung D and Spottswood SM** (2024) Self-supervised transformers for turbulent flow time series. *Physics of Fluids 36*(6). https://doi.org/10.1063/5.0214468.

Fischer PF, **Lottes JW and Kerkemeier SG** (2008) *Nek5000 Web page*. http://nek5000.mcs.anl.gov.

Flores O and Jiménez J (2006) Effect of wall-boundary disturbances on turbulent channel flows. *Journal of Fluid Mechanics 566*, 357–376. https://doi.org/10.1017/S0022112006001534.

Fraser AM and Swinney HL (1986) Independent coordinates for strange attractors from mutual information. *Physical Review A 33*(2), 1134–1140. https://doi.org/10.1103/PhysRevA.33.1134.

Frenzel S and Pompe B (2007) Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters 99*(20). https://doi.org/10.1103/physrevlett.99.204101.

Fukami K, **Nabae Y**, **Kawai K and Fukagata K** (2019) Synthetic turbulent inflow generator using machine learning. *Physical Review Fluids 4*(6). https://doi.org/10.1103/PhysRevFluids.4.064603.

Gallego-Castillo C, **Elagamy M**, **Cuerva-Tejero A**, **Lopez-Garcia O and Avila S** (2024) Synthesis of realistic non-homogeneous non-Gaussian turbulent wind fields. *Journal of Physics: Conference Series 2767*(5), 052019. https://doi.org/10.1088/1742-6596/2767/5/052019.

Gao W, **Oh S and Viswanath P** (2018) Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory 64*(8), 5629–5661. https://doi.org/10.1109/TIT.2018.2807481.

Granero-Belinchon C (2018) *Multiscale Information Transfer in Turbulence.* PhD dissertation, Université de Lyon.

Hlaváčková-Schindler K, **Paluš M**, **Vejmelka M and Bhattacharya J** (2007) Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports 441*(1), 1–46.

Jiménez J (1998) Turbulent velocity fluctuations need not be Gaussian. *Journal of Fluid Mechanics 376*, 139–147. https://doi.org/10.1017/S0022112098002432.

Jiménez J (2013) How linear is wall-bounded turbulence? *Physics of Fluids 25*(11). https://doi.org/10.1063/1.4819081.

Jiménez J (2018) Coherent structures in wall-bounded turbulence. *Journal of Fluid Mechanics 842*, P1.

Jiménez J (2025) Chaos, coherence, and turbulence. *Physical Review Fluids 10*(10). https://doi.org/10.1103/lkvl-dlcn.

Kantz H and Schreiber T (2003) *Nonlinear Time Series Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755798.

Kim J and Lee C (2020) Deep unsupervised learning of turbulence for inflow generation at various Reynolds numbers. *Journal of Computational Physics 406*, 109216. https://doi.org/10.1016/j.jcp.2019.109216.

Klein M, **Sadiki A and Janicka J** (2003) A digital filter based generation of inflow data for spatially developing direct numerical or large eddy simulations. *Journal of Computational Physics 186*(2), 652–665. https://doi.org/10.1016/S0021-9991(03)00090-1.

Kovasznay L, **Kibens V and Blackwelder R** (1970) Large-scale motion in the intermittent region of a turbulent boundary layer. *Journal of Fluid Mechanics 41*, 283–325.

Kraskov A, **Stögbauer H and Grassberger P** (2004) Estimating mutual information. *Physical Review E 69*(6). https://doi.org/10.1103/PhysRevE.69.066138.

Larsson J, **Kawai S**, **Bodart J and Bermejo-Moreno I** (2016) Large eddy simulation with modeled wall-stress: Recent progress and future directions. *Mechanical Engineering Reviews 3*(1). https://doi.org/10.1299/mer.15-00418.

Lee J, **Nemati S**, **Silva I**, **Edwards BA**, **Butler JP and Malhotra A** (2012) Transfer entropy estimation and directional coupling change detection in biomedical time series. *Biomedical Engineering Online 11*(1), 19. https://doi.org/10.1186/1475-925x-11-19.

Li T, **Biferale L**, **Bonaccorso F**, **Scarpolini MA and Buzzicotti M** (2024) Synthetic Lagrangian turbulence by generative diffusion models. *Nature Machine Intelligence 6*(4), 393–403. https://doi.org/10.1038/s42256-024-00810-0.

Liu XY, **Parikh MH**, **Fan X**, **Du P**, **Wang Q**, **Chen YF and Wang JX** (2025) CoNFiLD-inlet: Synthetic turbulence inflow using generative latent diffusion models with neural fields. *Physical Review Fluids 10*(5). https://doi.org/10.1103/PhysRevFluids.10.054901.

Lozano-Durán A and Arranz G (2022) Information-theoretic formulation of dynamical systems: Causality, modeling, and control. *Physical Review Research 4*(2). https://doi.org/10.1103/PhysRevResearch.4.023195.

Lozano-Durán A, **Bae HJ and Encinar MP** (2019) Causality of energy-containing eddies in wall turbulence. *Journal of Fluid Mechanics 882*. https://doi.org/10.1017/jfm.2019.801.

Lütkepohl H (2005) *New Introduction to Multiple Time Series Analysis*. Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-540-27752-1.

Martínez-Sánchez A, **Arranz G and Lozano-Durán A** (2024) Decomposing causality into its synergistic, unique, and redundant components. *Nature Communications 15*, 9296.

Marusic I, **Mathis R and Hutchins N** (2010) Predictive model for wall-bounded turbulent flow. *Science 329*(5988), 193–196. https://doi.org/10.1126/science.1188765.

Massaro D and Hussain F (2026) Interpreting coherence in wall turbulence. *Nature Reviews Physics 8*, 276–290. https://doi.org/10.1038/s42254-026-00929-6.

Massaro D, **Rezaeiravesh S and Schlatter P** (2023) On the potential of transfer entropy in turbulent dynamical systems. *Scientific Reports 13*(1). https://doi.org/10.1038/s41598-023-49747-1.

Massaro D, **Rezaeiravesh S and Schlatter P** (2026) Causally coherent structures in turbulent dynamical systems. *Physical Review E 113*(3), L032201. https://doi.org/10.1103/16fy-qvdp.

Mathis R, **Monty JP**, **Hutchins N and Marusic I** (2009) Comparison of large-scale amplitude modulation in turbulent boundary layers, pipes, and channel flows. *Physics of Fluids 21*(11). https://doi.org/10.1063/1.3267726.

Matsubara K, **Mitsuishi A**, **Iwamoto K and Murata A** (2023) Prediction of pulsating turbulent pipe flow by deep learning with generalization capability. *International Journal of Heat and Fluid Flow 104*, 109214. https://doi.org/10.1016/j.ijheatfluidflow.2023.109214.

Mizuno Y and Jiménez J (2013) Wall turbulence without walls. *Journal of Fluid Mechanics 723*, 429–455. https://doi.org/10.1017/jfm.2013.137.

Nakamura T, **Fukami K**, **Hasegawa K**, **Nabae Y and Fukagata K** (2021) Convolutional neural network and long short-term memory based reduced order surrogate for minimal turbulent channel flow. *Physics of Fluids 33*(2). https://doi.org/10.1063/5.0039845.

Oliver T, **Malaya N**, **Ulerich R and Moser R** (2014) Estimating uncertainties in statistics computed from direct numerical simulation. *Physics of Fluids 26*. https://doi.org/10.1063/1.4866813.

Packard NH, **Crutchfield JP**, **Farmer JD and Shaw RS** (1980) Geometry from a time series. *Physical Review Letters 45*(9), 712–716. https://doi.org/10.1103/PhysRevLett.45.712.

Park KI (2018) *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer International Publishing. https://doi.org/10.1007/978-3-319-68075-0.

Patruno L and Ricci M (2018) A systematic approach to the generation of synthetic turbulence using spectral methods. *Computer Methods in Applied Mechanics and Engineering 340*, 881–904. https://doi.org/10.1016/j.cma.2018.06.028.

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

Pinto H, **Lazic I**, **Antonacci Y**, **Pernice R**, **Gu D**, **Barà C**, **Faes L and Rocha AP** (2024) Testing dynamic correlations and nonlinearity in bivariate time series through information measures and surrogate data analysis. *Frontiers in Network Physiology 4*. https://doi.org/10.3389/fnetp.2024.1385421.

Pope SB (2000) *Turbulent Flows, 10th Printing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840531.

Racca A, **Doan NAK and Magri L** (2023) Predicting turbulent dynamics with the convolutional autoencoder echo state network. *Journal of Fluid Mechanics 975*. https://doi.org/10.1017/jfm.2023.716.

Rezaeiravesh S, **Gscheidle C**, **Peplinski A**, **Garcke J and Schlatter P** (2025) In-situ estimation of time-averaging uncertainties in turbulent flow simulations. *Computer Methods in Applied Mechanics and Engineering 433*, 117511. https://doi.org/10.1016/j.cma.2024.117511.

Rezaeiravesh S, **Massaro D and Schlatter P** (2026) *Source Code and Data for Computing Pairwise Interactions in Simulated and VAR-Synthetic Turbulence Time Series.* https://doi.org/10.5281/zenodo.21481711. Available at https://github.com/simFlowDataLab/pairwiseValidation_turbulenceTS.

Rezaeiravesh S, **Vinuesa R and Schlatter P** (2021) On numerical uncertainties in scale-resolving simulations of canonical wall turbulence. *Computers & Fluids 227*, 105024. https://doi.org/10.1016/j.compfluid.2021.105024.

Rezaeiravesh S, **Xavier D**, **Vinuesa R**, **Yao J**, **Hussain F and Schlatter P** (2022) Estimating uncertainty of low- and high-order turbulence statistics in wall turbulence. In *12th International Symposium on Turbulence and Shear Flow Phenomena*, Japan.

Russo S and Luchini P (2017) A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages. *Journal of Computational Physics 347*, 328–340. https://doi.org/10.1016/j.jcp.2017.07.005.

Schreiber T (2000) Measuring information transfer. *Physical Review Letters 85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461.

Schreiber T and Schmitz A (1996) Improved surrogate data for nonlinearity tests. *Physical Review Letters 77*(4), 635–638. https://doi.org/10.1103/PhysRevLett.77.635.

Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464. http://www.jstor.org/stable/2958889.

Seabold S and Perktold J (2010) statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference.*

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal 27*(3), 379–423.

Stanly R, **Du S**, **Xavier D**, **Perez A**, **Mukha T**, **Markidis S**, **Rezaeiravesh S and Schlatter P** (2024) Generating synthetic turbulence with vector autoregression of proper orthogonal decomposition time coefficients. *Journal of Fluid Mechanics 1000*. https://doi.org/10.1017/jfm.2024.1034.

Stenger M, **Bauer A**, **Prantl T**, **Leppich R**, **Hudson N**, **Chard K**, **Foster I and Kounev S** (2024) Thinking in categories: A survey on assessing the quality for time series synthesis. *Journal of Data and Information Quality 16*(2), 1–32. https://doi.org/10.1145/3666006.

Stoļarovs K, **Toosi S**, **Yang L**, **Massaro D**, **Garcia E**, **Yao J**, **Mihajlović M**, **Hussain F**, **Rezaeiravesh S and Schlatter P** (2026) Quantifying interaction among POD modes in turbulent flows using transfer entropy. *Journal of Physics: Conference Series 3230*(1), 012028. https://doi.org/10.1088/1742-6596/3230/1/012028.

Takens F (1981) Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*. Berlin Heidelberg: Springer, pp. 366–381. https://doi.org/10.1007/BFb0091924.

Tan E, **Algar S**, **Corrêa D**, **Small M**, **Stemler T and Walker D** (2023) Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology. *Chaos: An Interdisciplinary Journal of Nonlinear Science 33*(3). https://doi.org/10.1063/5.0137223.

Wallot S and Mønster D (2018) Calculation of average mutual information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in Matlab. *Frontiers in Psychology 9*. https://doi.org/10.3389/fpsyg.2018.01679.

Wei W (2006) *Time Series Analysis: Univariate and Multivariate Methods*. Time Series Analysis: Univariate and Multivariate Methods. Pearson Addison Wesley. Available at https://books.google.co.uk/books?id=aY0QAQAAIAAJ.

Witter J and Houghton C (2024) *Nearest-Neighbours Estimators for Conditional Mutual Information.* arXiv: 2403. 00556 [cs.IT]. Available at https://arxiv.org/abs/2403.00556.

Wu X (2017) Inflow turbulence generation methods. *Annual Review of Fluid Mechanics 49*(1), 23–49. https://doi.org/10.1146/annurev-fluid-010816-060322.

Xavier D, **Rezaeiravesh S and Schlatter P** (2024) Autoregressive models for quantification of time-averaging uncertainties in turbulent flows. *Physics of Fluids 36*(10). https://doi.org/10.1063/5.0211541.

Yousif MZ, **Yu L and Lim H** (2022) Physics-guided deep learning for generating turbulent inflow conditions. *Journal of Fluid Mechanics 936*. https://doi.org/10.1017/jfm.2022.61.

Yousif MZ, **Zhang M**, **Yu L**, **Vinuesa R and Lim H** (2023) A transformer-based synthetic-inflow generator for spatially developing turbulent boundary layers. *Journal of Fluid Mechanics 957*. https://doi.org/10.1017/jfm.2022.1088.

Zambom AZ and Dias R (2013) A review of kernel density estimation with applications to econometrics. *International Econometric Review (IER) 5*(1), 20–42.

Zhang J, **Zhu Z**, **Wang L and Hu R** (2025) *Causal Analysis of Inner and Outer Motions in Near-Wall Turbulence.* Available at https://arxiv.org/abs/2404.08907.

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

Appendices

Appendix A: Impact of the estimator on the information-theoretic measures

The mutual information and transfer entropy in this study are estimated using the KSG estimator from Equations (8) and (10), respectively. These estimators can be sensitive to the value of k. The results obtained from default $k = 10$ in Figures 4, 7, and 9, can be compared to the case of using $k = 5$ and $k = 30$ in Figure A1 (the first and second rows, respectively). The general pattern in the maximum MI and TE is not affected, and particularly the main features in the net maximum TE are largely preserved. However, the upper bound for the maximum TE is significantly reduced with increasing k. In such cases, the estimator will consider larger neighbourhoods around each sample over which the density is assumed to be constant. Therefore, for high values of k, there should

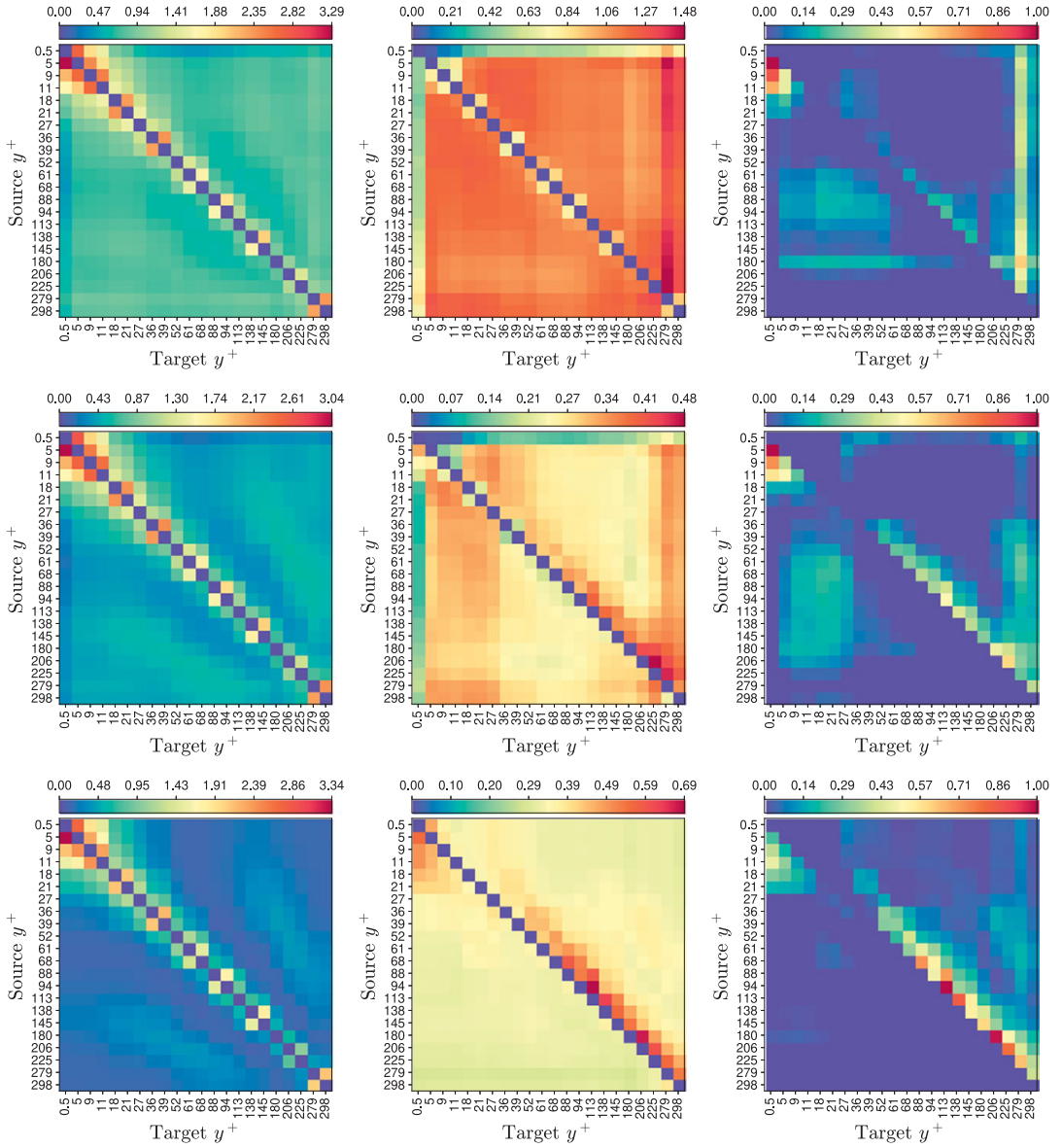


Figure A1. *Mean of the (left column) maximum time-delayed MI, (middle column) maximum TE, and (right column) normalised net maximum TE estimated from VAR synthetic data using (top row) the KSG estimator with $k = 5$, (middle row) the KSG estimator with $k = 30$, and (bottom row) the KDE estimator.*

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

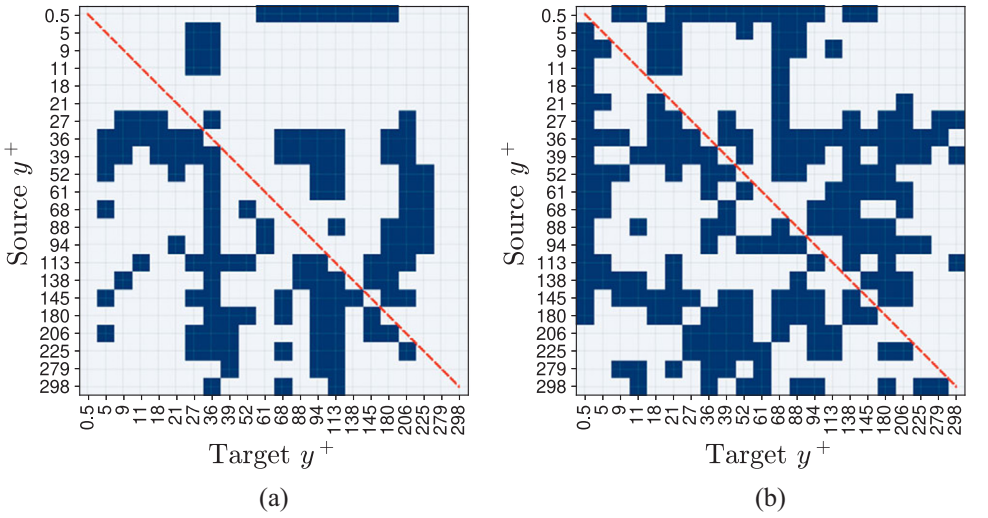


Figure A2. *Matrix of rejection (light colour)/non-rejection (dark colour) outcomes for the Bi-IAAFT surrogate test (Pinto et al., 2024) applied to TE between all interacting pairs of the simulation time series, using (a) $k = 5$ and (b) $k = 30$. Entries marked as rejected correspond to p-value less than 0.05 under the null hypothesis that the paired time series are exclusively linearly correlated. The dashed line shows the main diagonal of the matrix.*

be a smoothing effect resulting in larger bias and lower variance. Although the selection of k is not, in general, trivial and could be optimised (Witter and Houghton, 2024), in this study we choose $k = 10$. In contrast, the use of the KDE estimator, the bottom row in Figure A1, can lead to over-smoothing and large bias in the estimations. In particular, the maximum TE and its net value are different from those by the KSG estimator even with $k = 30$.

The choice of k also affects surrogate-based hypothesis testing. For the source-row-shuffling surrogate (Pinto et al., 2024), the associated null hypothesis is rejected for all three values $k = 5, 10$, and 30, indicating that the estimated TE values remain statistically significant under this null. For the Bi-IAAFT surrogate (Pinto et al., 2024), however, the effect of k is more pronounced. Comparing Figure 8 with Figure A2 shows that increasing k reduces the number of pairs for which the Bi-IAAFT null hypothesis is rejected. This suggests that larger values of k make departures from the linear-surrogate null harder to detect, which is consistent with the stronger smoothing effect induced by larger neighbourhoods in the KSG estimator.

Additional sensitivity analyses with representative interacting pairs in Section 3, using different sample lengths n and several values of k, did not reveal a single uniformly optimal choice of k. Nevertheless, they showed that the main qualitative conclusions, particularly the sign of the net maximum TE for the representative pairs considered, are robust for $10 \leq k \leq 40$ and $n = 10{,}437$. This further supports the use of $k = 10$ as a pragmatic intermediate choice. Furthermore, because the same value of k is used consistently for both the simulation and the VAR-synthetic data, any residual estimator bias is of secondary importance for the comparative aim of the present study.

Appendix B: Impact of the embedding delay value on TE estimates

In Section 3, for computing the TE, the embedding delay τ in Equation (7) was optimised for all the interacting time series. This is a crucial step, noting that the default value of $\tau = 1$ in Figure B1 leads to largely under-predicted TEs compared to Figure 7. Moreover, the pattern of the net maximum TE is largely affected, and for many pairs, the maximum TE values of the synthetic data are found to be significantly underpredicted compared to those of the simulation data. These clearly show that $\tau = 1$ is too small to provide any new information about the past of a source time series. This redundancy is, however, intuitive, reminding the slow decay in both ACF and AMIF in Figures 3 and 5a, respectively.

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press

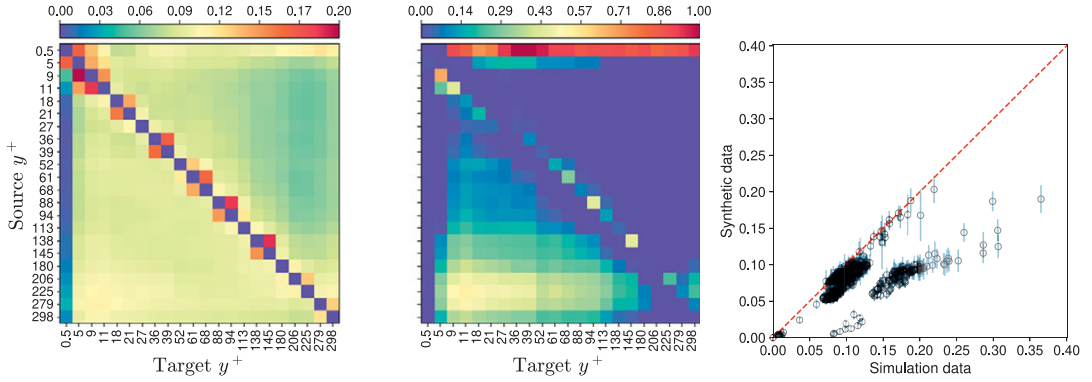


Figure B1. *Mean of (a) maximum TE and (b) net maximum TE obtained from the VAR synthetic data. (c) Scatter plot of the maximum TE obtained from the VAR synthetic data and simulation data. All TE are estimated using embedding delay* $\tau = 1$ *in Equation (7). The KSG estimator with* $k = 10$ *is used.*

Cite this article: Rezaeiravesh S, Massaro D and Schlatter P (2026). Pairwise interactions in turbulence time series: essential measures for validating synthetic data. *Data-Centric Engineering*, 7, e44. doi:10.1017/dce.2026.10078

https://doi.org/10.1017/dce.2026.10078 Published online by Cambridge University Press