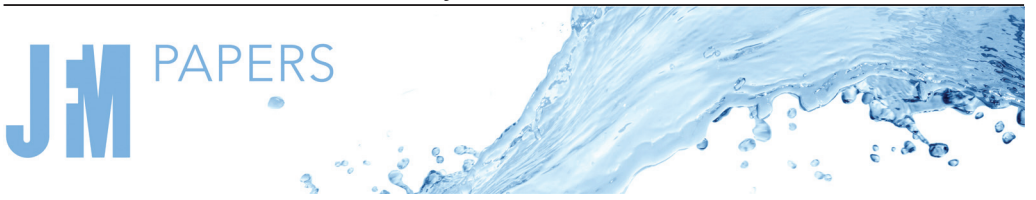
J. Fluid Mech. (2023), *vol*. 971, A3, doi:10.1017/jfm.2023.573



Multi-scale reconstruction of turbulent rotating flows with proper orthogonal decomposition and generative adversarial networks

Tianyi Li[1,2], Michele Buzzicotti[1], Luca Biferale[1,†], Fabio Bonaccorso[1], Shiyi Chen[2,3] and Minping Wan[2,3]

[1]Department of Physics and INFN, University of Rome 'Tor Vergata', Via della Ricerca Scientifica 1, 00133 Rome, Italy

[2]Guangdong Provincial Key Laboratory of Turbulence Research and Applications, Department of Mechanics and Aerospace Engineering, Southern University of Science and Technology, Shenzhen, Guangdong 518055, PR China

[3]Guangdong-Hong Kong-Macao Joint Laboratory for Data-Driven Fluid Mechanics and Engineering Applications, Southern University of Science and Technology, Shenzhen 518055, PR China

(Received 23 October 2022; revised 3 July 2023; accepted 10 July 2023)

Data reconstruction of rotating turbulent snapshots is investigated utilizing data-driven tools. This problem is crucial for numerous geophysical applications and fundamental aspects, given the concurrent effects of direct and inverse energy cascades. Additionally, benchmarking of various reconstruction techniques is essential to assess the trade-off between quantitative supremacy, implementation complexity and explicability. In this study, we use linear and nonlinear tools based on the proper orthogonal decomposition (POD) and generative adversarial network (GAN) for reconstructing rotating turbulence snapshots with spatial damages (inpainting). We focus on accurately reproducing both statistical properties and instantaneous velocity fields. Different gap sizes and gap geometries are investigated in order to assess the importance of coherency and multi-scale properties of the missing information. Surprisingly enough, concerning point-wise reconstruction, the nonlinear GAN does not outperform one of the linear POD techniques. On the other hand, the supremacy of the GAN approach is shown when the statistical multi-scale properties are compared. Similarly, extreme events in the gap region are better predicted when using GAN. The balance between point-wise error and statistical properties is controlled by the adversarial ratio, which determines the relative importance of the generator and the discriminator in the GAN training.

† Email address for correspondence: biferale@roma2.infn.it

Key words: machine learning, rotating turbulence

1. Introduction

The problem of reconstructing missing information, due to measurement constraints and lack of spatial/temporal resolution, is ubiquitous in almost all important applications of turbulence to laboratory experiments, geophysics, meteorology and oceanography (Le Dimet & Talagrand 1986; Bell *et al.* 2009; Torn & Hakim 2009; Krysta *et al.* 2011; Asch, Bocquet & Nodet 2016). For example, satellite imagery often suffers from missing data due to dead pixels and thick cloud cover (Shen *et al.* 2015; Zhang *et al.* 2018; Militino, Ugarte & Montesino 2019; Storer *et al.* 2022). In particle tracking velocimetry (PTV) experiments (Dabiri & Pecora 2020), spatial gaps naturally occur due to the use of a small number of seeded particles. Additionally, in particle image velocimetry (PIV) experiments, missing information can arise due to out-of-pair particles, object shadows or light reflection issues (Garcia 2011; Wang *et al.* 2016; Wen *et al.* 2019). Similarly, in many instances, the experimental probes are limited to assessing only a subset of the relevant fields, asking for a careful *a priori* engineering of the most relevant features to be tracked. Recently, many data-driven machine learning tools have been proposed to fulfil some of the previous tasks. Research using these black-box tools is at its infancy and we lack systematic quantitative benchmarks for paradigmatic high-quality and high-quantity multi-scale complex datasets, a mandatory step to making them useful for the fluid-dynamics community. In this paper, we perform a systematic quantitative comparison among three data-driven methods (no information on the underlying equations) to reconstruct highly complex two-dimensional (2-D) fields from a typical geophysical set-up, such as that of rotating turbulence. The first two methods are connected with a linear model reduction, the so-called proper orthogonal decomposition (POD) and the third is based on a fully nonlinear convolutional neural network (CNN) embedded in a framework of a generative adversarial network (GAN) (Goodfellow *et al.* 2014; Deng *et al.* 2019; Subramaniam *et al.* 2020; Buzzicotti *et al.* 2021; Guastoni *et al.* 2021; Kim *et al.* 2021; Buzzicotti & Bonaccorso 2022; Yousif *et al.* 2022). Proper orthogonal decomposition is widely used for pattern recognition (Sirovich & Kirby 1987; Fukunaga 2013), optimization (Singh *et al.* 2001) and data assimilation (Romain, Chatellier & David 2014; Suzuki 2014). To repair the missing data in a gappy field, Everson & Sirovich (1995) proposed gappy POD (GPOD), where coefficients are optimized according to the measured data outside the gap. By introducing some modifications to GPOD, Venturi & Karniadakis (2004) improved its robustness and made it reach the maximum possible resolution at a given level of spatio-temporal gappiness. Gunes, Sirisup & Karniadakis (2006) showed that GPOD reconstruction outperforms the Kriging interpolation (Oliver & Webster 1990; Stein 1999; Gunes & Rist 2008). However, GPOD is essentially a linear interpolation and thus is in trouble when dealing with complex multi-scale and non-Gaussian flows as the ones typical of fully developed turbulence (Alexakis & Biferale 2018) and/or large missing areas (Li *et al.* 2021). Extended POD (EPOD) was first used in Maurel, Borée & Lumley (2001) on the PIV data of a turbulent internal engine flow, where the POD analysis is conducted in a sub-domain spanning only the central rotating region but the preferred directions of the jet–vortex interaction can be clearly identified. Borée (2003) generalized the EPOD and reported that EPOD can be applied to study the correlation of any physical quantity in any domain with the projection of any measured quantity on its POD modes in the measurement domain. EPOD has many applications of flow sensing, where flow predictions are made based on remote probes

(Tinney, Ukeiley & Glauser 2008; Hosseini, Martinuzzi & Noack 2016; Discetti *et al.* 2019). For example, using EPOD as a reference of their CNN models, Guastoni *et al.* (2021) predicted the 2-D velocity-fluctuation fields at different wall-normal locations from the wall-shear-stress components and the wall pressure in a turbulent open-channel flow. EPOD also provides a linear relation between the input and output fields.

In recent years, CNNs have made a great success in computer vision tasks (Niu & Suen 2012; Russakovsky *et al.* 2015; He *et al.* 2016) because of their powerful ability of handling nonlinearities (Hornik 1991; Kreinovich 1991; Baral, Fuentes & Kreinovich 2018). In fluid mechanics, CNN has also been shown as a encouraging technique for data prediction/reconstruction (Fukami, Fukagata & Taira 2019; Güemes, Discetti & Ianiro 2019; Kim & Lee 2020; Li *et al.* 2023). Many researchers devote time to the super-resolution task, where CNNs are used to reconstruct high-resolution data from low-resolution data of laminar and turbulent flows (Liu *et al.* 2020; Subramaniam *et al.* 2020; Fukami, Fukagata & Taira 2021; Kim *et al.* 2021). In the scenario where a large gap exists, missing both large- and small-scale features, Buzzicotti *et al.* (2021) reconstructed for the first time a set of 2-D damaged snapshots of three-dimensional (3-D) rotating turbulence with GAN. Recent works show that CNN or GAN is also feasible to reconstruct the 3-D velocity fields with 2-D observations (Matsuo *et al.* 2021; Yousif *et al.* 2022). GAN consists of two CNNs, a generator and a discriminator. Previous preliminary research indicates that the introduction of discriminator significantly improves the high-order statistics of the prediction (Deng *et al.* 2019; Subramaniam *et al.* 2020; Buzzicotti *et al.* 2021; Güemes *et al.* 2021; Kim *et al.* 2021). Different from the previous work (Buzzicotti *et al.* 2021), here we attack the problem of data reconstruction with GAN by changing the ratio between the input measurements and the missing information. Furthermore, we present a first systematic assessment of the nonlinear vs linear reconstruction methods, by showing also results using two different POD-based methods. We discuss and present novel results concerning both point-based and statistical metrics. Moreover, the dependency of GAN properties on the adversarial ratio is also systematically studied. The adversarial ratio gauges the relative importance of the discriminator in comparison with the generator throughout the training process.

Two factors make the reconstruction difficult. First, turbulent flows have a large number of active degrees of freedom which grows with the turbulent intensity, typically parameterized by the Reynolds number. The second factor is the spatio-temporal gappiness, which depends on the area and geometry of the missing region. In the current work we conduct a first systematic comparative study between GPOD, EPOD and GAN on the reconstruction of turbulence in the presence of rotation, which is a paradigmatic system with both coherent vortices at large scales and strong non-Gaussian and intermittent fluctuations at small scales (Alexakis & Biferale 2018; Buzzicotti, Clark Di Leoni & Biferale 2018; Di Leoni *et al.* 2020).

Figure 1 displays some examples of the reconstruction task in this work. The aim is to fill the gap region with data close to the ground truth (figure 1*c*,*f*). A second long term goal would also be to systematically perform features ranking, which is understanding the quality of the supplied information on the basis of its performance in the reconstruction goal. The latter is connected to the sacred grail of turbulence: identifying the master degrees of freedom driving turbulent flow, connected also to control problems (Choi, Moin & Kim 1994; Lee *et al.* 1997; Gad-el Hak & Tsai 2006; Brunton & Noack 2015; Fahland *et al.* 2021). The study presented in this work is a first step towards a quantitative assessment of the tools that can be employed to ask and answer this kind of question. In order to focus on two paradigmatic realistic set-ups, we study two gap geometries, a central square gap (figure 1*a*,*d*) and random gappiness (figure 1*b*,*e*). The latter is related

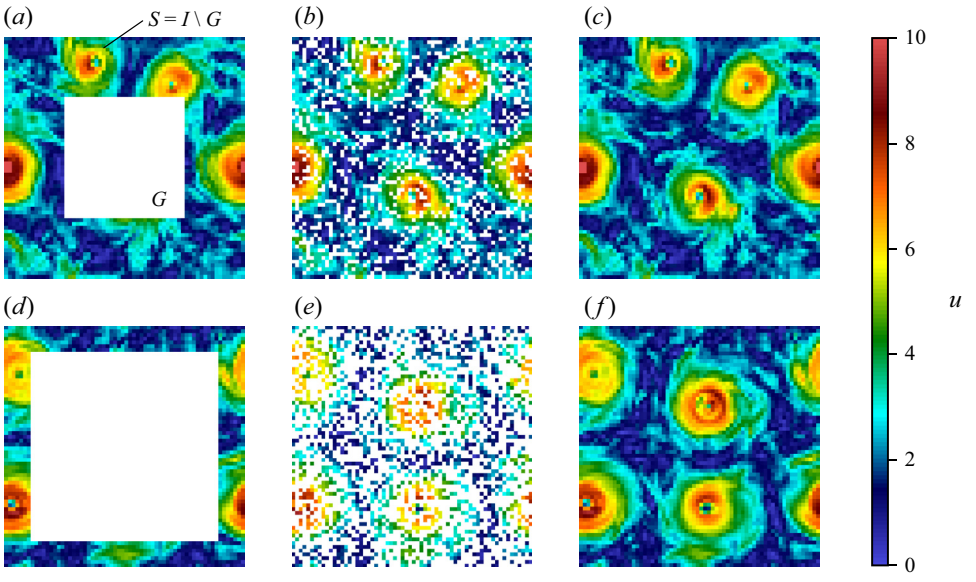


Figure 1. Examples of the reconstruction task on 2-D slices of 3-D turbulent rotating flows, where the colour code is proportional to the velocity module. Two gap geometries are considered: (*a*,*d*) a central square gap and (*b*,*e*) random gappiness. Gaps in each row have the same area and the corresponding ground truth is shown in (*c*,*f*). We denote G as the gap region and $S = I \setminus G$ as the known region, where I is the entire 2-D image.

to practical applications such as PTV and PIV. The gap area is also varied from a small to an extremely large proportion up to the limit where only one thin layer is supplied at the border, a seemingly impossible reconstruction task, for evaluation of the three methods in different situations. In a recent work, Clark Di Leoni *et al.* (2022) used physics-informed neural networks for reconstruction with sparse and noisy particle tracks obtained experimentally. As in practice the measurements are always noisy or filtered, we also investigate the robustness of the EPOD and GAN reconstruction methods.

The paper is organized as follows. Section 2.1 describes the dataset and the reconstruction problem set-up. The GPOD, EPOD and GAN-based reconstruction methods are introduced in §§ 2.2, 2.3 and 2.4, respectively. In § 3, the performances of POD- and GAN-based methods in turbulence reconstruction are systematically compared when there is one central square gap of different sizes. We address the dependency on the adversarial ratio for the GAN-based reconstruction in § 4 and show results for random gappiness from GPOD, EPOD and GAN in § 5. The robustness of EPOD and GAN to measurement noise and the computational cost of all methods are discussed in § 6. Finally, conclusions of the work are presented in § 7.

2. Methodology

2.1. *Dataset and reconstruction problem set-up*

For the evaluation of different reconstruction tools, we use a dataset from the TURB-Rot (Biferale *et al.* 2020) open database. The dataset used in this study is generated from a direct numerical simulation of the Navier–Stokes equations for homogeneous incompressible flow in the presence of rotation with periodic boundary conditions (Godeferd & Moisy 2015; Pouquet *et al.* 2018; Seshasayanan & Alexakis 2018; van Kan & Alexakis 2020; Yokoyama & Takaoka 2021). In a rotating frame of reference, both

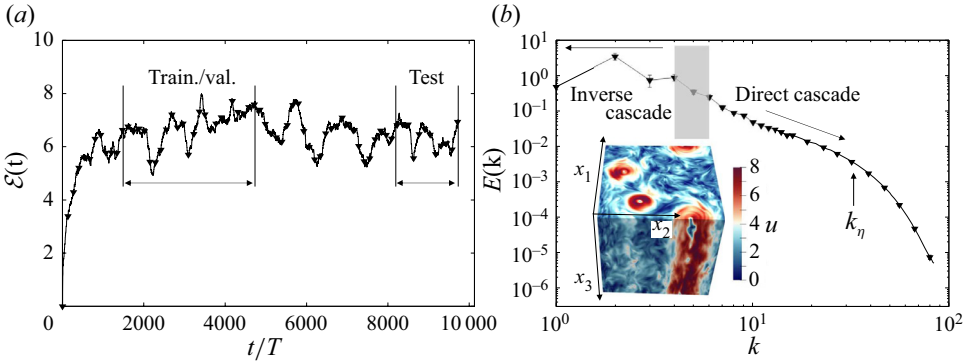


Figure 2. (*a*) Energy evolution of the turbulent flow, where we also show the sampling time periods for the training/validation and testing data. (*b*) The averaged energy spectrum. The grey area indicates the forcing wavenumbers, and k_η is the Kolmogorov dissipative wavenumber where $E(k)$ starts to decay exponentially. The inset shows an instantaneous visualization of the velocity module with the frame of reference for the simulation.

Coriolis and centripetal accelerations must be taken into account. However, the centrifugal force can be expressed as the gradient of the centrifugal potential and included in the pressure gradient term. In this way, the resulting equations will explicitly show only the presence of the Coriolis force, while the centripetal term is absorbed into a modified pressure (Cohen & Kundu 2004). The simulation is performed using a fully dealiased parallel pseudo-spectral code in a 3-D (x_1-x_2-x_3) periodic domain of size $[0, 2\pi)^3$ with 256 grid points in each direction, as shown in the inset of figure 2(*b*). Denoting $l_0 = 2\pi$ as the domain size, the Fourier spectral wavenumber is $\boldsymbol{k} = (k_1, k_2, k_3)$, where $k_1 = 2n_1\pi/l_0$ ($n_1 \in \mathbb{Z}$) and one can similarly obtain k_2 and k_3. The governing equations can be written as

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} + 2\boldsymbol{\Omega}\times\boldsymbol{u} = -\frac{1}{\rho}\nabla\tilde{p} + \nu\nabla^2\boldsymbol{u} + \boldsymbol{f}, \tag{2.1}$$

where $\boldsymbol{u}$ is the incompressible velocity, $\boldsymbol{\Omega} = \Omega\hat{\boldsymbol{x}}_3$ is the system rotation vector, $\tilde{p} = p + \frac{1}{2}\rho\|\boldsymbol{\Omega}\times\boldsymbol{x}\|^2$ is the pressure in an inertial frame modified by a centrifugal term, ν is the kinematic viscosity and $\boldsymbol{f}$ is an external forcing mechanism at scales around $k_f = 4$ via a second-order Ornstein–Uhlenbeck process (Sawford 1991; Buzzicotti *et al.* 2016). Figure 2(*a*) plots the energy evolution with time of the whole simulation. The energy spectrum $E(k) = \frac{1}{2}\sum_{k\le\|\boldsymbol{k}\|<k+1}\|\hat{\boldsymbol{u}}(\boldsymbol{k})\|^2$ averaged over time is shown in figure 2(*b*), where the grey area indicates the forcing wavenumbers. To enlarge the inertial range between k_f and the Kolmogorov dissipative wavenumber, $k_\eta = 32$, which is picked as the scale where $E(k)$ starts to decay exponentially, the viscous term $\nu\nabla^2\boldsymbol{u}$ in (2.1) is replaced with a hyperviscous term $\nu_h\nabla^4\boldsymbol{u}$ (Haugen & Brandenburg 2004; Frisch *et al.* 2008). We define an effective Reynolds number as $Re_{eff} = (k_0/k_\eta)^{-3/4} \approx 13.45$, with the smallest wavenumber $k_0 = 1$. A linear friction term $-\beta\boldsymbol{u}$ acting only on wavenumbers $\|\boldsymbol{k}\| \le 2$ is also used in right-hand side of (2.1) to prevent a large-scale condensation (Alexakis & Biferale 2018). As shown in figure 2(*a*), the flow reaches a stationary state with a Rossby number $Ro = \mathcal{E}^{1/2}/(\Omega/k_f) \approx 0.1$, where $\mathcal{E}$ is the kinetic energy. The integral length scale is $L = \mathcal{E}/\int kE(k)\,\mathrm{d}k \sim 0.15l_0$ and the integral time scale is $T = L/\mathcal{E}^{1/2} \approx 0.185$. Readers can refer to Biferale *et al.* (2020) for more details on the simulation.

Re_{eff}	Ro	L	$N_{x_1} \times N_{x_2}$	$\Delta t_s/T$	N_{train}	N_{valid}	T_{train}/T	N_{test}	T_{test}/T
13.45	0.1	$0.15l_0$	64×64	5.41	84 480	10 560	3243	20 480	865

Table 1. Description of the dataset used for the evaluation of reconstruction methods. Here, N_{x_1} and N_{x_2} indicate the resolution of the horizontal plane. The number of fields for training/validation/testing is denoted as $N_{train}/N_{valid}/N_{test}$. The sampling time periods for training/validation and testing are T_{train} and T_{test}, respectively.

The dataset is extracted from the above simulation as follows. First, we sampled 600 snapshots of the whole 3-D velocity field from time $t = 276$ up to $t = 876$ for training and validation, and we sampled 160 3-D snapshots from $t = 1516$ to $t = 1676$ for testing, as shown in figure 2(*a*). A sampling interval $\Delta t_s \approx 5.41T$ is used to decrease correlations in time between two successive snapshots.

To reduce the amount of data to be analysed, the resolution of sampled fields is downsized from 256^3 to 64^3 by a spectral low-pass filter

$$\boldsymbol{u}(\boldsymbol{x}) = \sum_{\|\boldsymbol{k}\| \le k_\eta} \hat{\boldsymbol{u}}(\boldsymbol{k}) \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}), \tag{2.2}$$

where the cutoff is the Kolmogorov dissipative wavenumber such as to only eliminate the fully dissipative degrees of freedom where the flow becomes smooth (Frisch 1995). Therefore, there is no loss of data complexity in this procedure and it also indicates that the measurement resolution corresponds to k_η.

For each downsized 3-D field, we selected 16 horizontal (x_1–x_2) planes at different x_3, each of which can be augmented to 11 (for training and validation) or 8 (for testing) different ones by randomly shifting it along both the x_1 and x_2 directions using periodic boundary conditions (Biferale *et al.* 2020). Therefore, a total of 176 or 128 planes can be obtained at each instant of time.

Finally, the 105 600 planes sampled at early times are randomly shuffled and used to constitute the train/validation split: 84 480 (80 %)/10 560 (10 %), which is used for the training process, while the other 20 480 planes sampled at later times are used for the testing process.

The parameters of the dataset are summarized in table 1. In the present study, we only reconstruct the velocity module, $u = \|\boldsymbol{u}\|$, which is always positive. Note that we restrict our study to 2-D horizontal slices in order to make contact with geophysical observation, although GPOD, EPOD and GAN are feasible to use with 3-D data.

We next describe the reconstruction problem set-up. Figure 3 presents an example of a gappy field, where I, G and S represent the whole region, the gap region and the known region, respectively. Given the damaged area A_G, we can define the gap size as $l = \sqrt{A_G}$. As shown in figure 1, two gap geometries are considered: (i) a square gap located at the centre and (ii) random gappiness which spreads over the whole region. Once the positions in G are determined, G is fixed for all planes over the training and the testing processes. Note that the GAN-based reconstruction can also handle the case where G is randomly changed for different planes (not shown). For a field $u(\boldsymbol{x})$ defined on I, we define the supplied measurements in S as $u_S(\boldsymbol{x}) = u(\boldsymbol{x})$ (with $\boldsymbol{x} \in S$), and the ground truth or the predicted field in G, as $u_G^{(t)}(\boldsymbol{x}) = u(\boldsymbol{x})$ or $u_G^{(p)}(\boldsymbol{x})$ (with $\boldsymbol{x} \in G$). The reconstruction models are 'learned' with the training data defined on the whole region I. Once the training process completed, one can evaluate the models by comparing the prediction and the ground truth in G over the test dataset.

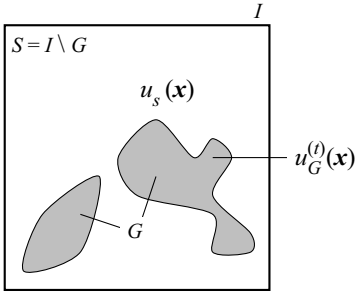


Figure 3. Schematic diagram of a gappy field.

2.2. *The GPOD reconstruction*

This section briefly presents the procedure of GPOD. The first step is to conduct POD analysis with the training data on the whole region I, namely solving the eigenvalue problem

$$\int_I R_I(\boldsymbol{x},\boldsymbol{y})\psi_n(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \lambda_n\psi_n(\boldsymbol{x}) \quad (\boldsymbol{x}\in I), \tag{2.3}$$

where

$$R_I(\boldsymbol{x},\boldsymbol{y}) = \langle u(\boldsymbol{x})u(\boldsymbol{y})\rangle \quad (\boldsymbol{x},\boldsymbol{y}\in I), \tag{2.4}$$

is the correlation matrix, given $\langle\cdot\rangle$ as the average over the training dataset. We denote λ_n as the eigenvalues and $\psi_n(\boldsymbol{x})$ as the POD eigenmodes, where $n = 1,\ldots,N_I$ and $N_I = N_{x_1}\times N_{x_2}$ is the number of points in I. For the homogeneous periodic flow considered in this study, it can be demonstrated that the POD modes correspond to Fourier modes, and their spectra are identical (Holmes *et al.* 2012). In all POD analyses of the present study, the mean of $u(\boldsymbol{x})$ is not removed. Any realization of the field can be decomposed as

$$u(\boldsymbol{x}) = \sum_{n=1}^{N_I} a_n\psi_n(\boldsymbol{x}) \quad (\boldsymbol{x}\in I), \tag{2.5}$$

with the POD coefficients

$$a_n = \int_I u(\boldsymbol{x})\psi_n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{2.6}$$

In the case when we have data only in S, the relation (2.6) cannot be used and one can adopt the dimension reduction by keeping only the first N' POD modes and minimizing the distance between the measurements and the linear POD decomposition (Everson & Sirovich 1995)

$$\tilde{E} = \int_S \left| u_S(\boldsymbol{x}) - \sum_{n=1}^{N'} a_n^{(p)}\psi_n(\boldsymbol{x}) \right|^2 \mathrm{d}\boldsymbol{x}, \tag{2.7}$$

to obtain the predicted coefficients $\{a_n^{(p)}\}_{n=1}^{N'}$. Then the GPOD prediction can be given as

$$u_G^{(p)}(\boldsymbol{x}) = \sum_{n=1}^{N'} a_n^{(p)}\psi_n(\boldsymbol{x}) \quad (\boldsymbol{x}\in G). \tag{2.8}$$

l	8	16	24	32	40	50	60	62
N' (s.g.)	72	45	21	21	13	13	13	13
N' (r.g.)	2334	2069	1726	1403	1039	551	98	56

Table 2. Summary of the optimal N' for the square gap (s.g.) and random gappiness (r.g.) with different sizes.

We optimize the value of N' during the training phase by requiring a minimum mean L_2 distance with the ground truth in the gap

$$\text{argmin}_{N'} \left\langle \int_G |u_G^{(t)}(\boldsymbol{x}) - u_G^{(p)}(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \right\rangle. \tag{2.9}$$

Table 2 summarizes the optimal N' used in this study. An analysis of reconstruction error for different N' is conducted in Appendix A. Let us notice that there also exists a different approach to select, frame by frame, a subset of POD modes to be used in the GPOD approach, based on Lasso, a regression analysis that performs mode selection with regularization (Tibshirani 1996). Results using this second approach do not show any significant improvement in a typical case of our study (see Appendix B).

2.3. *The EPOD reconstruction*

To use EPOD for flow reconstruction, we first compute the correlation matrix

$$R_S(\boldsymbol{x}, \boldsymbol{y}) = \langle u_S(\boldsymbol{x}) u_S(\boldsymbol{y}) \rangle \quad (\boldsymbol{x}, \boldsymbol{y} \in S), \tag{2.10}$$

and solve the eigenvalue problem

$$\int_S R_S(\boldsymbol{x}, \boldsymbol{y}) \phi_n(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \sigma_n \phi_n(\boldsymbol{x}) \quad (\boldsymbol{x} \in S), \tag{2.11}$$

to obtain the eigenvalues σ_n and the POD eigenmodes $\phi_n(\boldsymbol{x})$, where $n = 1, \ldots, N_S$ and N_S equals to the number of points in S. We remark that $\phi_n(\boldsymbol{x})$ are not Fourier modes, as the presence of the internal gap breaks the homogeneity. Any realization of the measured field in S can be decomposed as

$$u_S(\boldsymbol{x}) = \sum_{n=1}^{N_S} b_n \phi_n(\boldsymbol{x}) \quad (\boldsymbol{x} \in S), \tag{2.12}$$

where the nth POD coefficient is obtained from

$$b_n = \int_S u_S(\boldsymbol{x}) \phi_n(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{2.13}$$

Furthermore, with (2.12) and an important property (Borée 2003), $\langle b_n b_p \rangle = \sigma_n \delta_{np}$, one can derive the following identity:

$$\phi_n(\boldsymbol{x}) = \frac{\langle b_n u_S(\boldsymbol{x}) \rangle}{\sigma_n} \quad (\boldsymbol{x} \in S). \tag{2.14}$$

Here, we reiterate that $\langle \cdot \rangle$ denotes the average over the training dataset. Specifically, $\langle b_n u_S \rangle$ can be interpreted as $(1/N_{train}) \sum_{c=1}^{N_{train}} b_n^{(c)} u_S^{(c)}$, where the superscript c represents the index

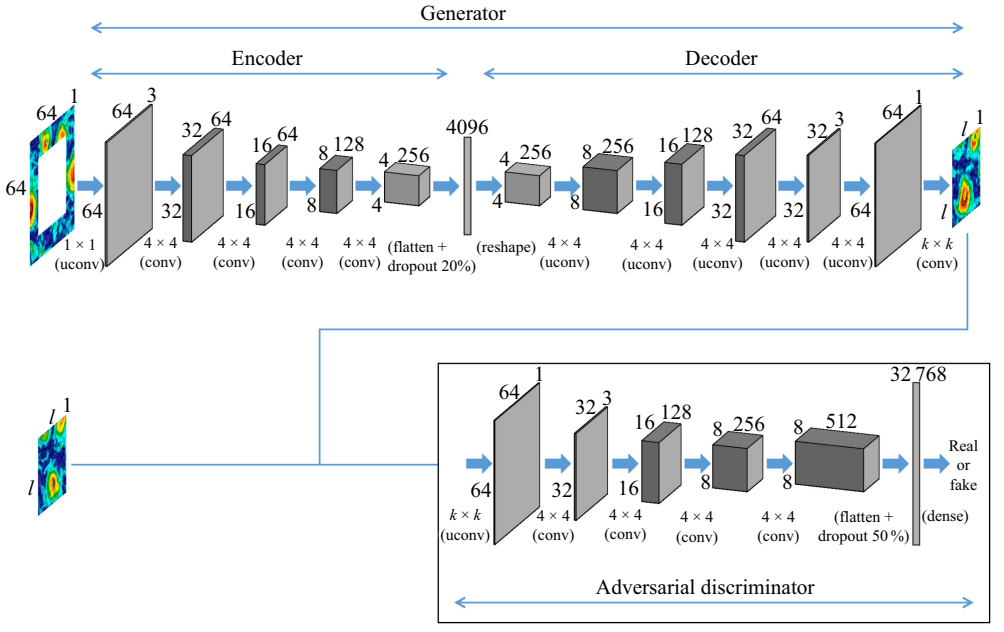


Figure 4. Architecture of generator and discriminator network for flow reconstruction with a square gap. The kernel size k and the corresponding stride s are determined based on the gap size l. Similar architecture holds for random gappiness as well.

of a particular snapshot. The extended POD mode is defined by replacing $u_S(\boldsymbol{x})$ with the field to be predicted $u_G^{(t)}(\boldsymbol{x})$ in (2.14)

$$\phi_n^E(\boldsymbol{x}) = \frac{\langle b_n u_G^{(t)}(\boldsymbol{x})\rangle}{\sigma_n} \quad (\boldsymbol{x} \in G). \tag{2.15}$$

Once we have obtained the set of EPOD modes (2.15) in the training process we can start the reconstruction of a test data with the measurement $u_S(\boldsymbol{x})$ from calculating the POD coefficients (2.13) and the prediction in G can be obtained as the correlated part with $u_S(\boldsymbol{x})$ (Borée 2003)

$$u_G^{(p)}(\boldsymbol{x}) = \sum_{n=1}^{N_S} b_n \phi_n^E(\boldsymbol{x}) \quad (\boldsymbol{x} \in G). \tag{2.16}$$

2.4. *The GAN-based reconstruction with context encoders*

In a previous work, Buzzicotti *et al.* (2021) used a context encoder (Pathak *et al.* 2016) embedded in GAN to generate missing data for the case where the total gap size is fixed, but with different spatial distributions. To generalize the previous approach to study gaps of different geometries and sizes, we extend previous GAN architecture by adding one layer at the start, two layers at the end of the generator and one layer at the start of the discriminator, as shown in figure 4. The generator is a functional $GEN(\cdot)$ first taking the damaged 'context', $u_S(\boldsymbol{x})$, to produce a latent feature representation with an encoder, and second with a decoder to predict the missing data, $u_G^{(p)}(\boldsymbol{x}) = GEN(u_S(\boldsymbol{x}))$. The latent feature represents the output vector of the encoder with 4096 neurons in figure 4, extracted

from the input with the convolutions and nonlinear activations. To constrain the predicted velocity module being positive, a rectified linear unit (ReLU) activation function is adopted at the last layer of generator. The discriminator acts as a 'referee' functional $D(\cdot)$, which takes either $u_G^{(t)}(\boldsymbol{x})$ or $u_G^{(p)}(\boldsymbol{x})$ and outputs the probability that the provided input ($u_G^{(t)}$ or $u_G^{(p)}$) belongs to the real turbulent ensemble. The generator is trained to minimize the following loss function:

$$\mathcal{L}_{GEN} = (1 - \lambda_{adv})\mathcal{L}_{MSE} + \lambda_{adv}\mathcal{L}_{adv}, \tag{2.17}$$

where the L_2 loss

$$\mathcal{L}_{MSE} = \left\langle \frac{1}{A_G} \int_G |u_G^{(p)}(\boldsymbol{x}) - u_G^{(t)}(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \right\rangle, \tag{2.18}$$

is the mean squared error (MSE) between the prediction and the ground truth. It is important to stress that, contrary to the GPOD case, the supervised L_2 loss is calculated only inside the gap region G. The hyper-parameter λ_{adv} is called the adversarial ratio and the adversarial loss is

$$\begin{aligned} \mathcal{L}_{adv} &= \langle \log(1 - D(u_G^{(p)})) \rangle = \int p(u_S) \log[1 - D(GEN(u_S))] \, \mathrm{d}u_S \\ &= \int p_p(u_G) \log(1 - D(u_G)) \, \mathrm{d}u_G, \end{aligned} \tag{2.19}$$

where $p(u_S)$ is the probability distribution of the field in S over the training dataset and $p_p(u_G)$ is the probability distribution of the predicted field in G given by the generator. At the same time, the discriminator is trained to maximize the cross-entropy based on its classification prediction for both real and predicted samples

$$\begin{aligned} \mathcal{L}_{DIS} &= \langle \log(D(u_G^{(t)})) \rangle + \langle \log(1 - D(u_G^{(p)})) \rangle \\ &= \int [p_t(u_G) \log(D(u_G)) + p_p(u_G) \log(1 - D(u_G))] \, \mathrm{d}u_G, \end{aligned} \tag{2.20}$$

where $p_t(u_G)$ is the probability distribution of the ground truth, $u_G^{(t)}(\boldsymbol{x})$. Goodfellow *et al.* (2014) further showed that the adversarial training between generator and discriminator with $\lambda_{adv} = 1$ in (2.17) minimizes the Jensen–Shannon (JS) divergence between the real and the predicted distributions, $\mathrm{JSD}(p_t \parallel p_p)$. Refer to (3.6) for the definition of the JS divergence. Therefore, the adversarial loss helps the generator to produce predictions that are statistically similar to real turbulent configurations. It is important to stress that the adversarial ratio λ_{adv}, which controls the weighted summation of $\mathcal{L}_{MSE}$ and $\mathcal{L}_{adv}$, is tuned to reach a balance between the MSE and turbulent statistics of the reconstruction (see § 4). More details about the GAN are discussed in Appendix C, including the architecture, hyper-parameters and the training schedule.

3. Comparison between POD- and GAN-based reconstructions

To conduct a systematic comparison between POD- and GAN-based reconstructions, we start by studying the case with a central square gap of various sizes (see figure 1). All reconstruction methods are first evaluated with the predicted velocity module itself, which is dominated by the large-scale coherent structures. The predictions are further assessed from a multi-scale perspective, with the help of the gradient of the predicted velocity module, spectral properties and multi-scale flatness. Finally, the performance on predicting extreme events is studied for all methods.

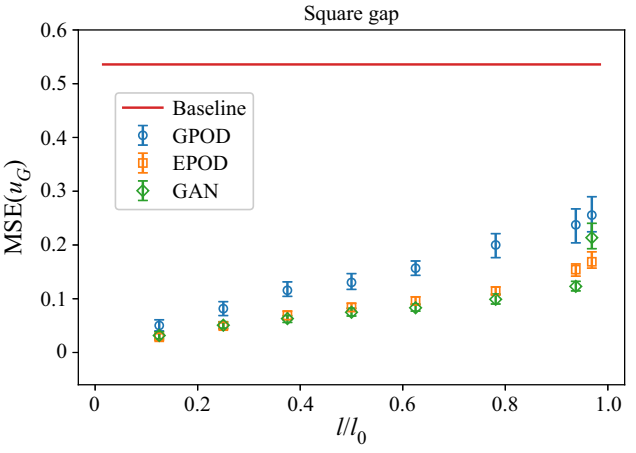


Figure 5. The MSE of the reconstructed velocity module from GPOD, EPOD and GAN in a square gap with different sizes. The abscissa is normalized by the domain size l_0. Red horizontal line represents the uncorrelated baseline.

3.1. *Large-scale information*

In this section, the predicted velocity module in the missing region is quantitatively evaluated. First we consider the reconstruction error and define the normalized MSE in the gap as

$$\mathrm{MSE}(u_G) = \frac{1}{E_{u_G}} \left\langle \frac{1}{A_G} \int_G |u_G^{(t)}(\boldsymbol{x}) - u_G^{(p)}(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \right\rangle, \tag{3.1}$$

where $\langle\cdot\rangle$ represents hereafter the average over the test data. The normalization factor is defined as

$$E_{u_G} = \sigma_G^{(p)} \sigma_G^{(t)}, \tag{3.2}$$

where

$$\sigma_G^{(p)} = \left\langle \frac{1}{A_G} \int_G |u_G^{(p)}|^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right\rangle^{1/2}, \tag{3.3}$$

and $\sigma_G^{(t)}$ is defined similarly. With the specific form of E_{u_G}, predictions with too small or too large energy will give a large MSE. To provide a baseline for MSE, a set of predictions can be made by randomly sampling the missing field from the true turbulent data. In other words, the baseline comes from uncorrelated predictions that are statistically consistent with the ground truth. The baseline value is around 0.54, see Appendix D. Figure 5 shows the MSE(u_G) from GPOD, EPOD and GAN reconstructions in a square gap with different sizes. The MSE is first calculated over data batches of size 128 (batch size used for GAN training), then the same calculation is repeated over 160 different batches, from which we calculate the MSE mean and its range of variation. The EPOD and GAN reconstructions provide similar MSEs except at the largest gap size, where GAN has a little bit larger MSE than EPOD. Besides, both EPOD and GAN have smaller MSEs than GPOD for all gap sizes. Figure 6 shows the probability density function (p.d.f.) of the spatially averaged L_2 error in the missing region for one flow configuration

$$\bar{\Delta}_{u_G} = \frac{1}{A_G} \int_G \Delta_{u_G}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{3.4}$$

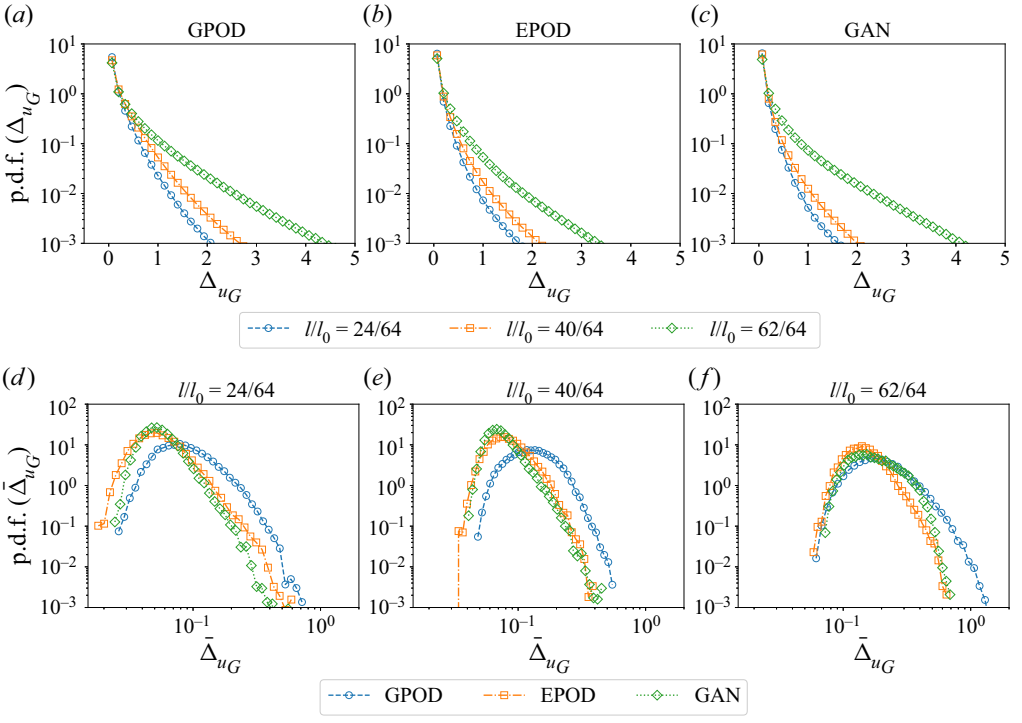


Figure 6. The p.d.f.s of the spatially averaged L_2 error over different flow configurations obtained from GPOD, EPOD and GAN for a square gap with sizes $l/l_0 = 24/64$, $40/64$ and $62/64$.

where

$$\Delta_{u_G}(\boldsymbol{x}) = \frac{1}{E_{u_G}} |u_G^{(p)}(\boldsymbol{x}) - u_G^{(t)}(\boldsymbol{x})|^2, \tag{3.5}$$

is the normalized point-wise L_2 error. The p.d.f.s are shown for three different gap sizes $l/l_0 = 24/64$, $40/64$ and $62/64$. Clearly, the p.d.f.s concentrating on regions of smaller $\bar{\Delta}_{u_G}$ correspond to the cases with smaller MSEs in figure 5. To further study the performance of the three tools, we plot the averaged point-wise L_2 error, $\langle \Delta_{u_G}(\boldsymbol{x}) \rangle$, for a square gap of size $l/l_0 = 40/64$ in figure 7. It shows that GPOD produces large $\langle \Delta_{u_G} \rangle$ all over the gap, while EPOD and GAN behave quite better, especially for the edge region. Moreover, GAN generates smaller $\langle \Delta_{u_G} \rangle$ than EPOD in the inner area (figure 7*b*,*c*). However, the L_2 error is naturally dominated by the more energetic structures (the ones found at large scales in our turbulent flows) and does not provide an informative evaluation of the predicted fields at multiple scales, which is also important for assessing the reconstruction tools for the turbulent data. Indeed, from figure 8 it is possible to see in a glimpse that the POD- and GAN-based reconstructions have completely different multi-scale statistics which are not captured by the MSE. Figure 8 shows predictions of an instantaneous velocity module field based on GPOD, EPOD and GAN methods compared with the ground truth solution. For all three gap sizes $l/l_0 = 24/64$, $40/64$ and $62/64$, GAN produces realistic reconstructions while GPOD and EPOD only generates blurry predictions. Besides, there are also obvious discontinuities between the supplied measurements and the GPOD predictions of the missing part. This is clearly due to the fact that the number of POD modes N' used for prediction in (2.8) is limited, as there are only N_S measured points available in (2.7) for each damaged data (thus $N' < N_S$).

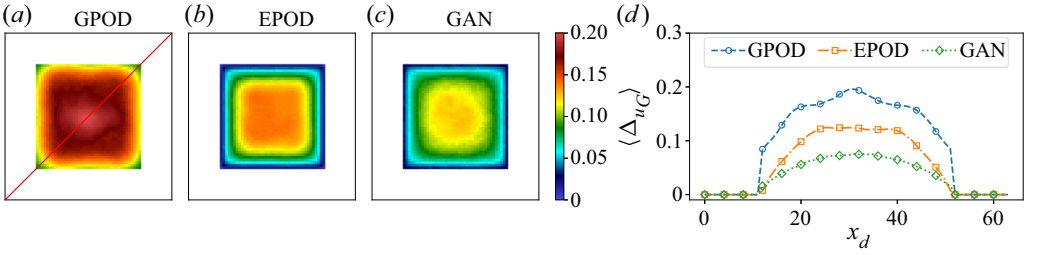


Figure 7. Averaged point-wise L_2 error obtained from (*a*) GPOD, (*b*) EPOD and (*c*) GAN for a square gap of size $l/l_0 = 40/64$. (*d*) Profiles of $\langle \Delta_{u_G} \rangle$ along the red diagonal line shown in (*a*), parameterized by x_d.

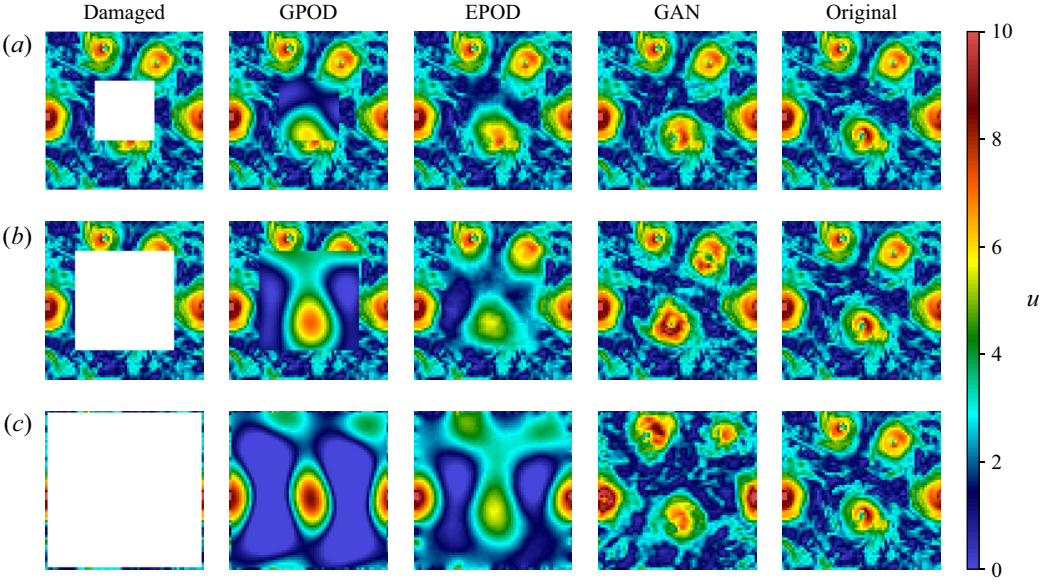


Figure 8. Reconstruction of an instantaneous field (velocity module) by the different tools for a square gap of sizes $l/l_0 = 24/64$ (*a*), $40/64$ (*b*) and $62/64$ (*c*). The damaged fields are shown in the first column, while the second to fourth columns show the reconstructed fields obtained from GPOD, EPOD and GAN. The ground truth is shown in the fifth column.

Moreover, minimizing the L_2 distance from ground truth in (2.9) results in solutions with almost the correct energy contents but without the complex multi-scale properties. Unlike GPOD using global basis defined on the whole region I, EPOD gives better results by considering the correlation between fields defined on two smaller regions, S and G. In this way the prediction (2.16) has the degrees of freedom equal to N_S, which are larger than those for GPOD. Therefore, EPOD can predict the large-scale coherent structures but is still limited in generating correct multi-scale properties. Specifically, when the gap size is extremely large, N_S is very small thus both GPOD and EPOD have small degrees of freedom to make realistic predictions.

To quantify the statistical similarity between the predictions and the ground truth, we can study the JS divergence, $\mathrm{JSD}(u_G) = \mathrm{JSD}(\mathrm{p.d.f.}(u_G^{(t)}) \parallel \mathrm{p.d.f.}(u_G^{(p)}))$, defined on the distribution of the velocity amplitude at one point, which is a marginal distribution of the whole p.d.f. of the real or predicted fields inside the gap, p_t or p_p. For distributions P and

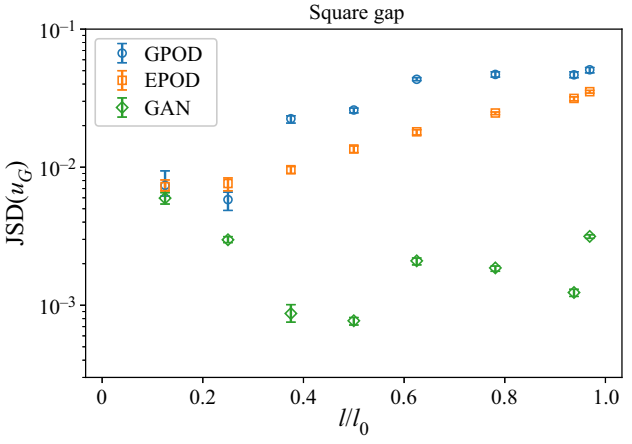


Figure 9. The JS divergence between p.d.f.s of the velocity module inside the missing region from the original data and the predictions obtained from GPOD, EPOD and GAN for a square gap with different sizes.

Q of a continuous random variable x, the JS divergence is a measure of their similarity

$$\mathrm{JSD}(P \parallel Q) = \tfrac{1}{2}\mathrm{KL}(P \parallel M) + \tfrac{1}{2}\mathrm{KL}(Q \parallel M), \tag{3.6}$$

where $M = \frac{1}{2}(P + Q)$ and

$$\mathrm{KL}(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \mathrm{d}x, \tag{3.7}$$

is the Kullback–Leibler divergence. A small JS divergence indicates that the two probability distributions are close and *vice versa*. We use the base 2 logarithm and thus $0 \leq \mathrm{JSD}(P \parallel Q) \leq 1$, with $\mathrm{JSD}(P \parallel Q) = 0$ if and only if $P = Q$. Similar to the MSE, the JS divergence is calculated using batches of data and 10 different batches are used to obtain its mean and range of variation. The batch size used to evaluate the JS divergence is now set at 2048, which is larger than that used for the MSE, in order to improve the estimation of the probability distributions. Figure 9 shows JSD(u_G) for the three reconstruction tools. We have found that GAN gives smaller JSD(u_G) than GPOD and EPOD by an order of magnitude over almost the full range of gap sizes, indicating that the p.d.f. of GAN prediction has a better correspondence to the ground truth. This is further shown in figure 10, where we present the p.d.f.s of the predicted velocity module for different gap sizes compared with that of the original data. Besides the imprecise p.d.f. shapes of GPOD and EPOD, we note that they are also predicting some negative values, which is unphysical for a velocity module. This problem is avoided in the GAN reconstruction, as a ReLU activation function has been used in the last layer of the generator.

3.2. *Multi-scale information*

This section reports a quantitative analysis of the multi-scale information reconstructed by the three methods. We first study the gradient of the predicted velocity module in the missing region, $\partial u_G/\partial x_1$. Figure 11 plots MSE($\partial u_G/\partial x_1$), which is similarly defined as (3.1), and we can see that all methods produce MSE($\partial u_G/\partial x_1$) with values much larger than those of MSE(u_G). Moreover, GAN shows similar errors to GPOD at the largest gap size and to EPOD at small gap sizes. However, the MSE itself is not enough for a

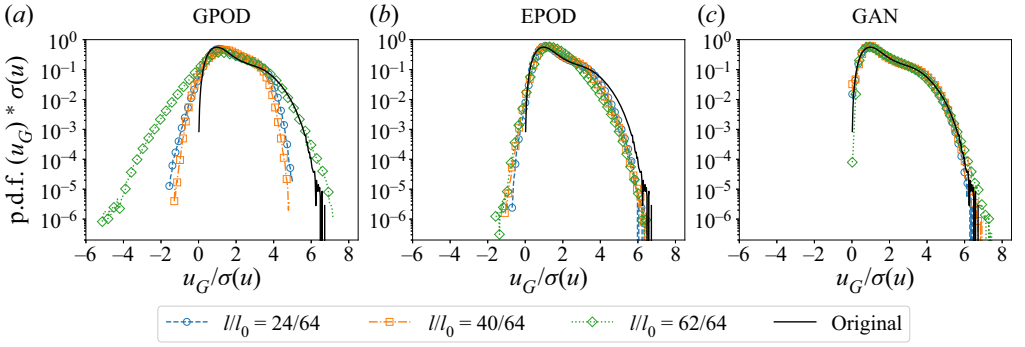


Figure 10. The p.d.f.s of the velocity module in the missing region obtained from GPOD, EPOD and GAN for a square gap with different sizes. The p.d.f. of the original data over the whole region is plotted for reference and $\sigma(u)$ is the standard deviation of the original data.

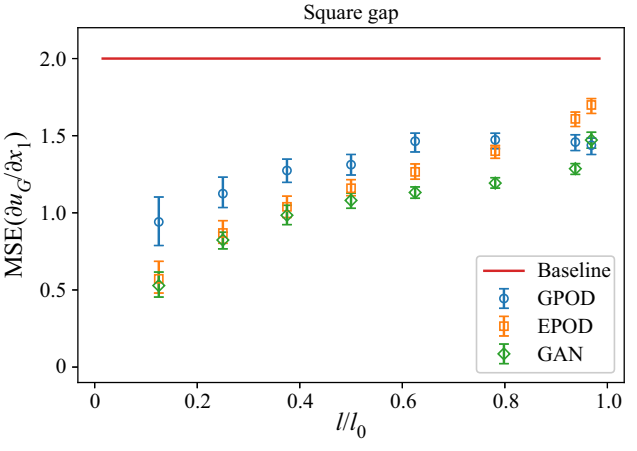


Figure 11. The MSE of the gradient of the reconstructed velocity module from GPOD, EPOD and GAN in a square gap with different sizes. Red horizontal line represents the uncorrelated baseline.

comprehensive evaluation of the reconstruction. This can be easily understood again by looking at the gradient of different reconstructions shown in figure 12. It is obvious that GAN predictions are much more 'realistic' than those of GPOD and EPOD, although their values of MSE$(\partial u_G/\partial x_1)$ are close. Indeed, if both fields are highly fluctuating, even a small spatial shift between the reconstruction and the true solution would result in a significantly larger MSE. This is exactly the case of GAN predictions, where we can see that they have obvious correlations with the ground truth but the MSE is large because of its sensitivity to small spatial shifting. On the other hand, the GPOD or EPOD solutions are inaccurate, having too small spatial fluctuations even with a similar MSE when compared with the GAN. As done above for the velocity amplitude, we further quantify the quality of the reconstruction by looking at the JS divergence between the two p.d.f.s in figure 13. For other metrics to assess the quality of the predictions see, e.g. Wang *et al.* (2004); Wang & Simoncelli (2005) and Li *et al.* (2023). Figure 13 confirms that GAN is able to well predict the p.d.f. of $\partial u_G/\partial x_1$ while GPOD and EPOD do not have this ability. Moreover, GPOD produces comparable JSD$(\partial u_G/\partial x_1)$ to EPOD. The above conclusions are further supported in figure 14, which shows p.d.f.s of $\partial u_G/\partial x_1$ from the predictions and the ground

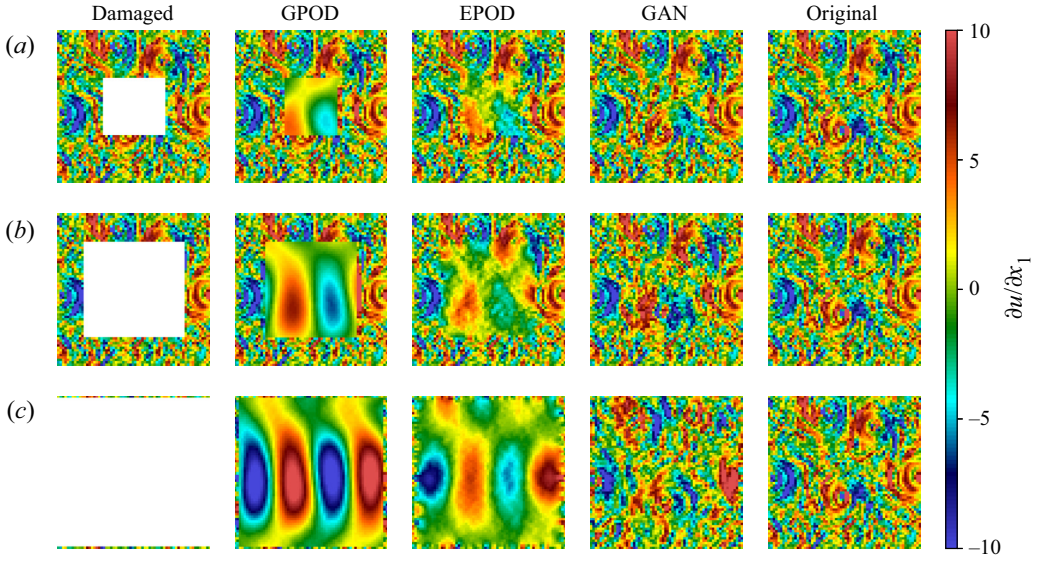


Figure 12. The gradient of the velocity module fields shown in figure 8. The first column shows the damaged fields with a square gap of sizes $l/l_0 = 24/64$ (a), $40/64$ (b) and $62/64$ (c). Note that, for the maximum gap size, $l/l_0 = 62/64$, we have only one velocity layer on the vertical borders where we do not supply any information on the gradient. The second to fifth columns plot the gradient fields obtained from GPOD, EPOD, GAN and the ground truth.

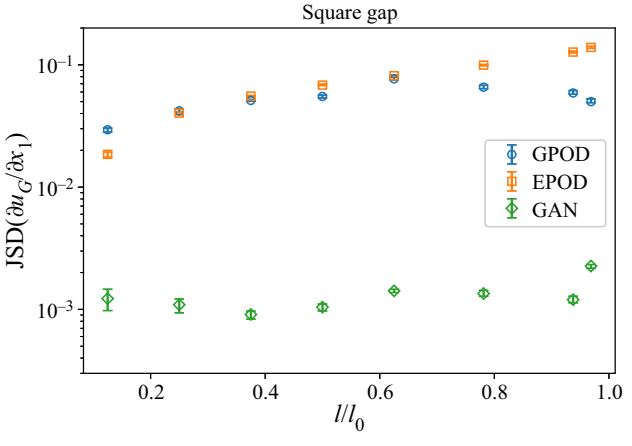


Figure 13. The JS divergence between p.d.f.s of the gradient of the reconstructed velocity module inside the missing region from the original data and the predictions obtained from GPOD, EPOD and GAN for a square gap with different sizes.

truth. We next compare the scale-by-scale energy budget of the original and reconstructed solutions in figure 15, with the help of the energy spectrum defined over the whole region

$$E(k) = \sum_{k \le \|\boldsymbol{k}\| < k+1} \frac{1}{2} \langle \hat{u}(\boldsymbol{k}) \hat{u}^*(\boldsymbol{k}) \rangle, \tag{3.8}$$

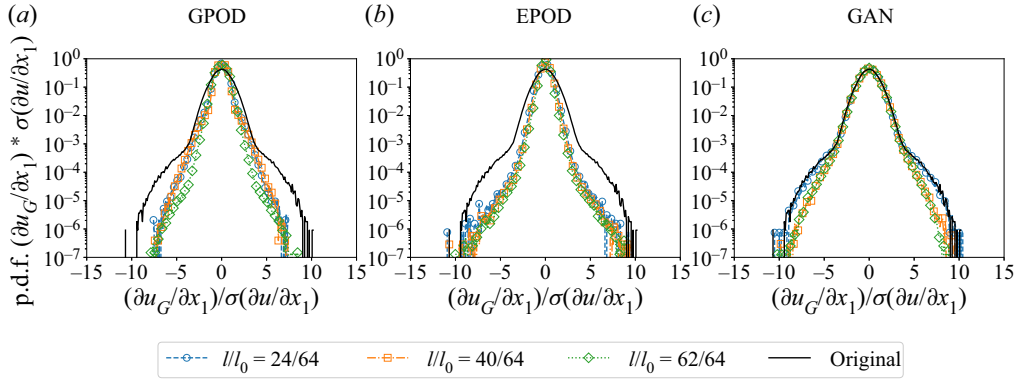


Figure 14. The p.d.f.s of the gradient of the reconstructed velocity module in the missing region obtained from GPOD, EPOD and GAN for a square gap with different sizes. The p.d.f. of the original data over the whole region is plotted for reference and $\sigma(\partial u/\partial x_1)$ is the standard deviation of the original data.

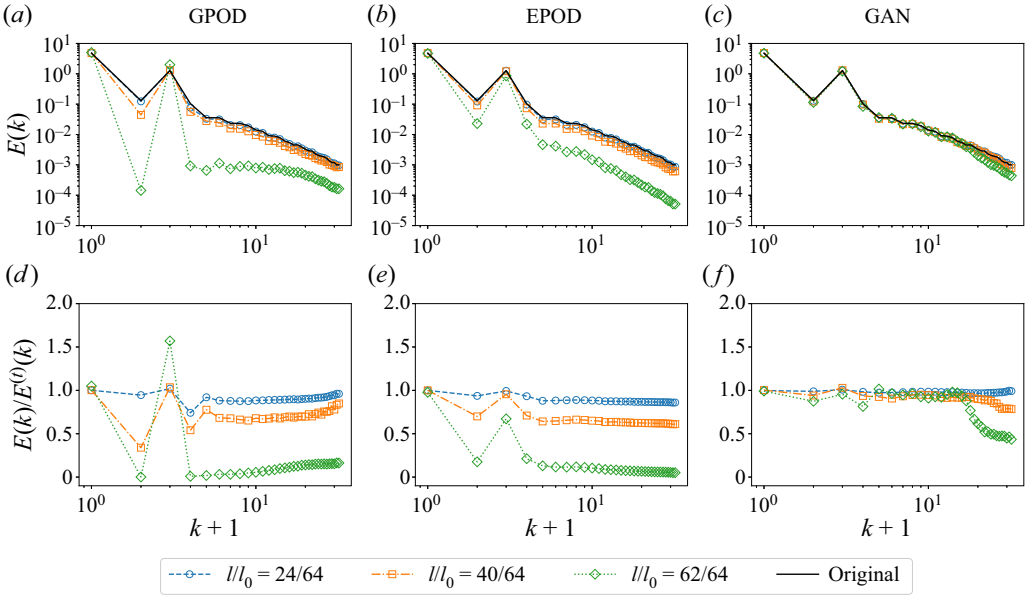


Figure 15. Energy spectra of the original velocity module and the reconstructions obtained from GPOD, EPOD and GAN for a square gap of different sizes (*a–c*). The corresponding $E(k)/E^{(t)}(k)$ is shown in (*d–f*), where $E(k)$ and $E^{(t)}(k)$ are the spectra of the reconstructions and the ground truth, respectively.

where $\boldsymbol{k} = (k_1, k_2)$ is the wavenumber, $\hat{u}(\boldsymbol{k})$ is the Fourier transform of velocity module and $\hat{u}^*(\boldsymbol{k})$ is its complex conjugate. To highlight the reconstruction performance as a function of the wavenumber, we also show the ratio between the reconstructed and the original spectra, $E(k)/E^{(t)}(k)$ for the three different gap sizes in the second row of figure 15. Figure 16 plots the flatness of the reconstructed fields,

$$F(r) = \langle (\delta_r u)^4 \rangle / \langle (\delta_r u)^2 \rangle, \tag{3.9}$$

where $\delta_r u = u(\boldsymbol{x} + \boldsymbol{r}) - u(\boldsymbol{x})$ and $\boldsymbol{r} = (r, 0)$. The angle bracket in (3.9) represents the average over the test dataset and over $\boldsymbol{x}$, for which $\boldsymbol{x}$ or $\boldsymbol{x} + \boldsymbol{r}$ fall in the gap. The flatness of the ground truth calculated over the whole region is also shown for reference. We remark

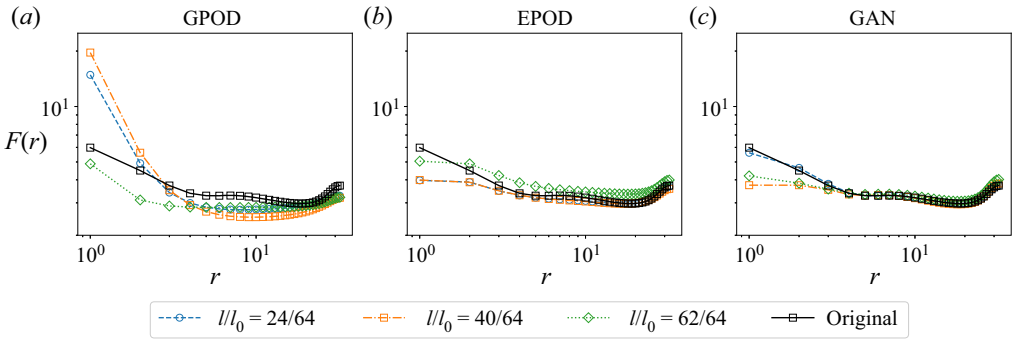


Figure 16. The flatness of the original field and the reconstructions obtained from GPOD, EPOD and GAN for a square gap of different sizes.

that the flatness is used to characterize the intermittency in the turbulence community (see Frisch 1995). It is determined by the two-point p.d.f.s, p.d.f.$(\delta_r u)$, connected to the distribution of the whole real or generated fields inside the gap, p_t or p_p. Figures 15 and 16 show that GAN performs well to reproduce the multi-scale statistical properties, except at small scales for large gap sizes. However, GPOD and EPOD can only predict a good energy spectrum for the small gap size $l/l_0 = 24/64$ but fail at all scales for both the energy spectrum and flatness at gap sizes $l/l_0 = 40/64$ and $62/60$.

3.3. *Extreme events*

In this section, we focus on the ability of the different methods to reconstruct extreme events inside the gap for each frame. In figure 17 we present the scatter plots of the largest values of velocity module or its gradient measured in the gap region from the original data and the predicted fields generated by GPOD, EPOD or GAN. On top of each panel we report the scatter plot correlation index, defined as

$$c = \langle 1 - |\sin\theta| \rangle, \tag{3.10}$$

where $|\sin\theta| = \|\boldsymbol{U} \times \boldsymbol{e}\|/\|\boldsymbol{U}\|$ with θ as the angle between the unit vector $\boldsymbol{e} = (1/\sqrt{2}, 1/\sqrt{2})$ and $\boldsymbol{U} = (\max(u_G^{(p)}), \max(u_G^{(t)}))$. The vector $\boldsymbol{U}$ for $\partial u_G/\partial x_1$ can be similarly defined. It is obvious that $c \in [0, 1]$ and $c = 1$ corresponds to a perfect prediction in terms of the extreme events. Figure 17 shows that, for both extreme values of the velocity module and its gradient, GAN is the least biased while the other two methods tend to underestimate them.

4. Dependency of GAN-based reconstruction on the adversarial ratio

As shown by the previous results, GAN is certainly superior regarding metrics evaluated in this study. This supremacy is given by the fact that, with the nonlinear CNN structure of the generator, GAN optimizes the point-wise L_2 loss and minimizes the JS divergence between the probability distributions of the real and generated fields with the help of the adversarial discriminator (see § 2.4). To study the effects of the balancing between the above two objectives on reconstruction quality, we have performed a systematic scanning of the GAN performances for changing adversarial ratio λ_{adv}, the hyper-parameter controlling the relative importance of L_2 loss and adversarial loss of the generator, as shown in (2.17). We consider a central square gap of size $l/l_0 = 40/64$ and train the GAN with different

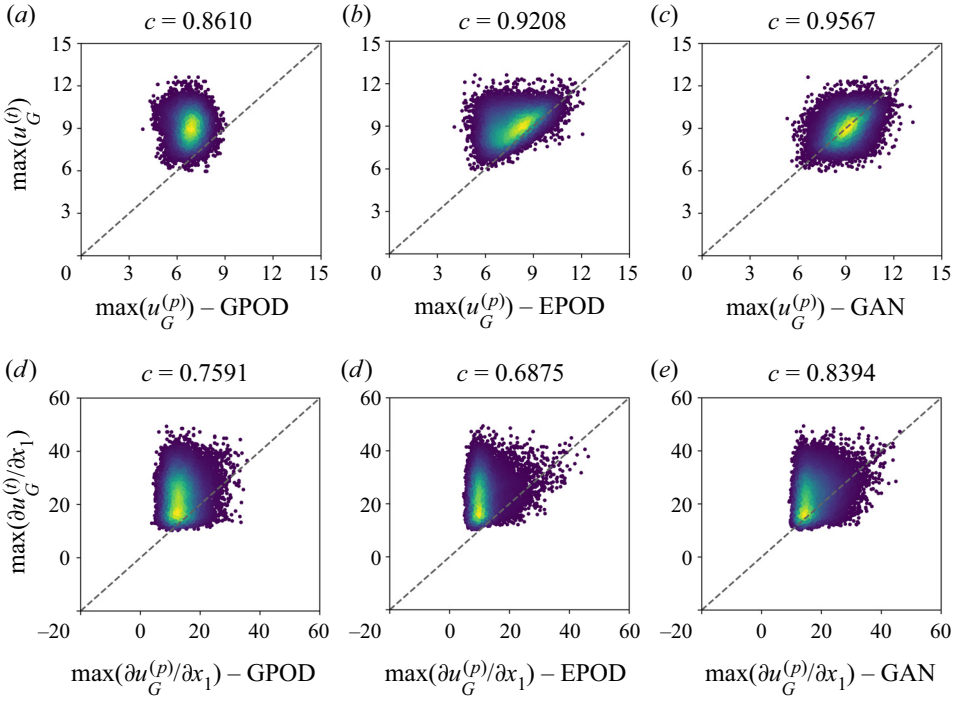


Figure 17. Scatter plots of the maximum values of velocity module (*a–c*) and its gradient (*d–f*) in the missing region obtained from the original data and the one produced by GPOD, EPOD or GAN for a square gap of size $l/l_0 = 40/64$. Colours are proportional to the density of points in the scatter plot. The correlation indices are shown on top of each panel.

adversarial ratios, where $\lambda_{adv} = 10^{-4}$, 10^{-3}, 10^{-2} and 10^{-1}. Table 3 shows the values of MSE(u_G) and JSD(u_G) obtained at different adversarial ratios.

It is obvious that the adversarial ratio controls the balance between the point-wise reconstruction error and the predicted turbulent statistics. As the adversarial ratio increases, the MSE increases while the JS divergence decreases. p.d.f.s of the predicted velocity module from GANs with different adversarial ratios are compared with that of the original data in figure 18, which shows that the predicted p.d.f. gets closer to the original one with a larger adversarial ratio. The above results clearly show that there exists an optimal adversarial ratio to satisfy the multi-objective requirements of having a small L_2 distance and a realistic p.d.f.. In the limit of vanishing λ_{adv}, the GAN outperforms GPOD and EPOD in terms of MSE, but falls behind them concerning JS divergence (table 3).

5. Dependency on gap geometry: random gappiness

Things change when looking at a completely different topology of the damages. Here, we study case (ii) in § 2.1, where position points are removed randomly in the original domain I, without any spatial correlations. Because the random gappiness is easier for interpolation than a square gap of the same size, all reconstruction methods show good and comparable results in terms of the MSE, the JS divergence and p.d.f.s for velocity module (figures 19 and 20). For almost all damaged densities, POD- and GAN-based methods give small values of MSE(u_G) and JSD(u_G). However, when the total damaged region area is extremely large, GPOD and EPOD are not able to reconstruct the field at large

λ_{adv}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	GPOD	EPOD
MSE$(u_G)(\times 10^2)$	$6.2^{+0.8}_{-0.7}$	$7.8^{+0.6}_{-0.7}$	$8.3^{+0.7}_{-0.6}$	$9.0^{+0.7}_{-0.8}$	15^{+1}_{-1}	$9.2^{+0.7}_{-0.7}$
JSD$(u_G)(\times 10^3)$	65^{+1}_{-1}	$2.6^{+0.2}_{-0.1}$	$2.1^{+0.1}_{-0.1}$	$2.06^{+0.05}_{-0.11}$	43^{+1}_{-1}	18^{+1}_{-1}

Table 3. The MSE and the JS divergence between p.d.f.s for the original and generated velocity module inside the missing region, obtained from GAN with different adversarial ratios for a square gap of size $l/l_0 = 40/64$. The results for GPOD and EPOD are provided as well for comparison. The MSE and JS divergence are computed over different test batches, specifically of sizes 128 and 2048, respectively. From these computations, we obtain both the mean values and the error bounds.

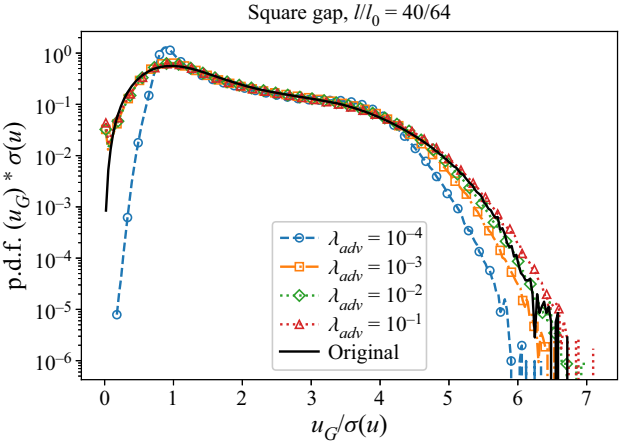


Figure 18. The p.d.f.s of the reconstructed velocity module inside the gap region, which is obtained from GAN with different adversarial ratios, for a square gap of size $l/l_0 = 40/64$.

wavenumbers while GAN still works well because of the adversarial training, as shown by the energy spectra in figure 21. Figure 22 shows the reconstruction of the velocity module and the corresponding gradient fields for random gappiness with two extremely large sizes. It is obvious that GPOD and EPOD only predict the large-scale structure while GAN generates reconstructions with multi-scale information. To make a comparison between the random gappiness case and the super-resolution task, we can compare the effect of random gappiness of sizes $l/l_0 = 60/64$ and $62/64$, similar to a downsampling of the original field by approximately a factor of 3 and 4 for each spatial direction, respectively.

6. Dependency on measurement noise and computational costs

So far, we have investigated the reconstruction of turbulent data without noise and therein the measurement resolution equals to the Kolmogorov scale, as shown in (2.2). However, field measurements are usually noisy. The noise can come from the errors encountered experimentally and/or a lack of resolution, such as for the filtered data in PIV. In this section, we evaluate the robustness of EPOD and GAN methods by considering a scenario where the magnitude of the velocity module remains unchanged, while its phase is randomly perturbed for wavenumbers above a threshold value, k_n. We estimate the noise level in the physical space, represented as NL(k_n), as the MSE of the noisy data with respect to the original fields. Contrary to (3.1), which is averaged over the gap region G, the

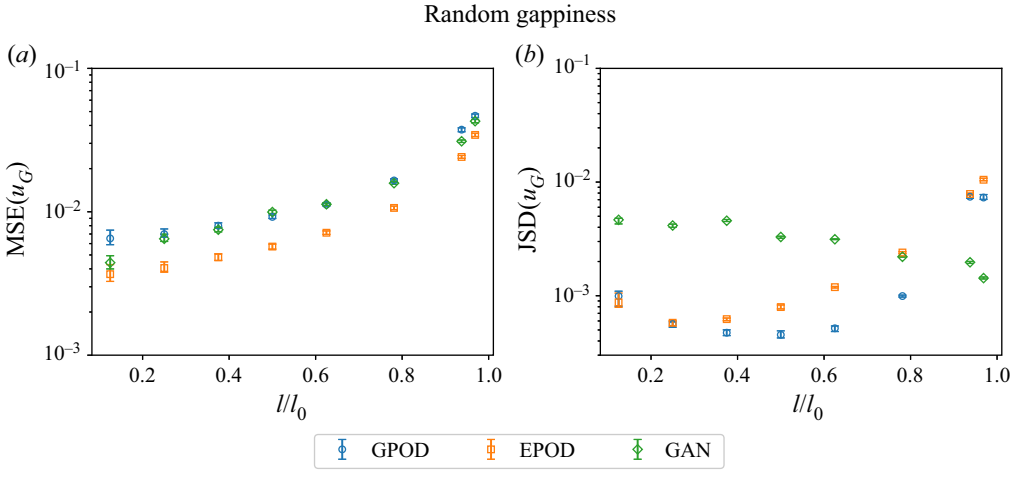


Figure 19. The MSE (a) and the JS divergence (b) between p.d.f.s for the original and generated velocity module inside the missing region, obtained from GPOD, EPOD and GAN for random gappiness with different sizes.

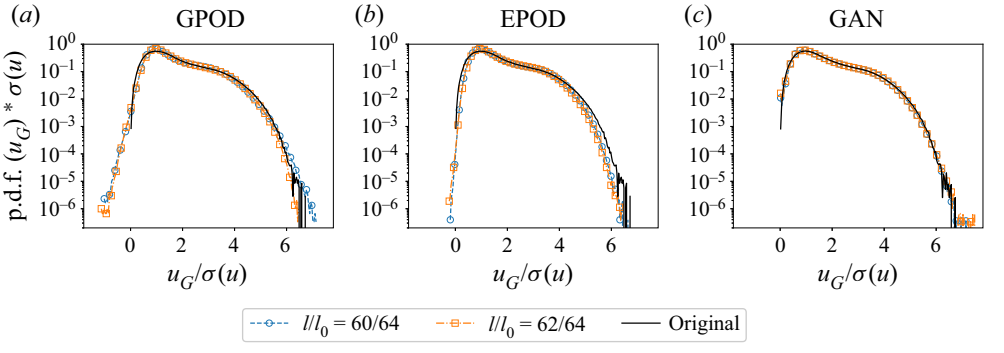


Figure 20. The p.d.f.s of the velocity module in the missing region obtained from GPOD, EPOD and GAN for random gappiness with different sizes. The p.d.f. of the original data over the whole region is plotted for reference and $\sigma(u)$ is the standard deviation of the original data.

noise level in this case is averaged across the entire domain I. Given the noise properties, we have $\mathrm{NL}(k_n) \sim \sum_{\|\boldsymbol{k}\| \geq k_n} E(k)$ (see figure 2b). The noisy measurements in the known region S are then fed into the reconstruction models, which have already been trained with the noiseless data. It is important to remark that the predictions are evaluated with the ground truth with no noise. Figures 23 and 24 show the MSE(u_G) and JSD(u_G) obtained from EPOD and GAN with input measurements of different noise levels for a square gap of sizes $l/l_0 = 24/64$ and $40/64$, respectively. The results obtained with the noiseless input are also shown with black symbols at the right end of each panel. Both MSE and JS divergence of the velocity module are more sensitive to the most energetic large-scale properties, while they change slightly when the noise is applied at $k_n \geq 3$, indicating the robustness of both approaches for these cases. Both EPOD and GAN predict drastically worse results when the noise is applied at $k_n \leq 2$.

Concerning the computational cost of the three methods here studied, we remark that during the training process of GPOD one first conducts a singular value decomposition (SVD) to solve (2.3), with a computational cost $\sim O(N_{train}^2 N_I)$. This operation is

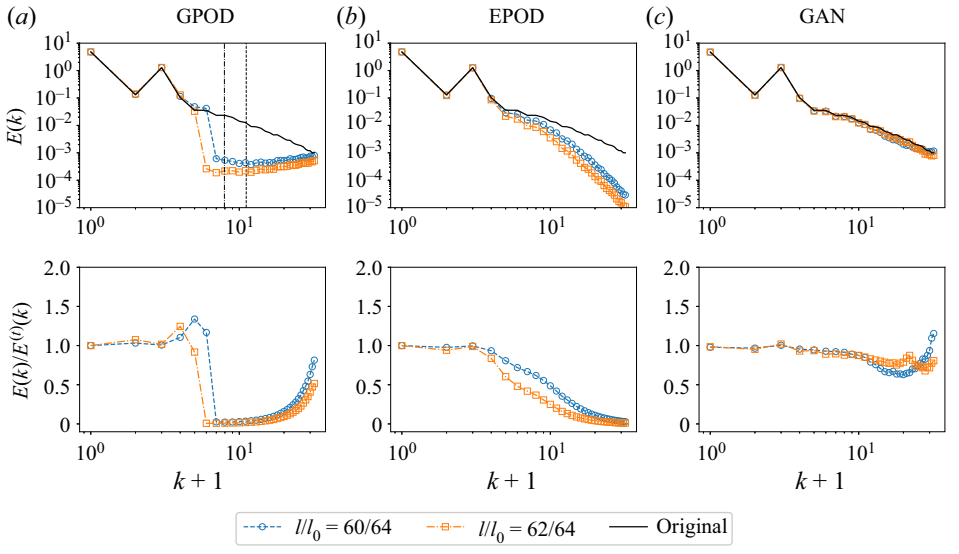


Figure 21. Energy spectra of the original velocity module and the reconstructions obtained from GPOD, EPOD and GAN for random gappiness with different sizes (*a–c*). The corresponding $E(k)/E^{(t)}(k)$ is shown in (*d–f*), where $E(k)$ and $E^{(t)}(k)$ are the spectra of the reconstructions and the ground truth, respectively. The vertical dashed and dash-dot lines respectively indicate the wavenumbers corresponding to the data resolutions $l/l_0 = 60/64$ and $62/64$. These wavenumbers are calculated as k_η/d, where d is the corresponding downsampling rate.

followed by an optimization of N' by scanning all possible values ($1 \leq N' \leq N_S$). This second step adds an extra computational cost of $O(N_S^2 N_I N_{train})$ for the required linear algebra operations as discussed in Appendix A, see (2.7), (A14), (A15) and (A16). The GPOD testing process is conducted according to (A14) and (A15) with a computational cost of $O(N' N_I N_{test})$. The training process of EPOD is computationally cheaper than that of GPOD. The cost can be estimated as $O(N_{train}^2 N_S)$ considering a SVD to solve (2.11) and the linear algebra operations in (2.13) and (2.15). The testing process of EPOD consists of carrying out (2.13) and (2.16), with a computational cost of $O(N_{test} N_S N_I)$. The GAN is the most computationally expensive method. It has approximately 5×10^6 ($\gg N_{train}$) trainable parameters, which are involved in the forward and the backward propagation for all the training data in one epoch. Moreover, hundreds of epoch are required for the convergence of GAN. However, benefiting from the GPU hardware, GAN training requires only 4 hours on an A100 Nvidia GPU. Once trained, all methods are highly efficient in performing reconstruction. It is important to emphasize that any improvement over existing methods is valuable, regardless of the computational cost involved. Even when computational resources are not a constraint, GPOD and EPOD cannot further improve the accuracy of the reconstruction. This limitation is attributed to the linear estimation of the flow state inherent in these methods. Nevertheless, there is still potential for further improvement of the GAN results, as numerous hyper-parameters remain to be fine tuned. These hyper-parameters include aspects such as the depth of the networks, the dimension of the latent feature, etc.

7. Conclusions

In this work, two linear POD-based approaches, GPOD and EPOD, are compared against GAN, consisting of two adversarial nonlinear CNNs, to reconstruct 2-D damaged

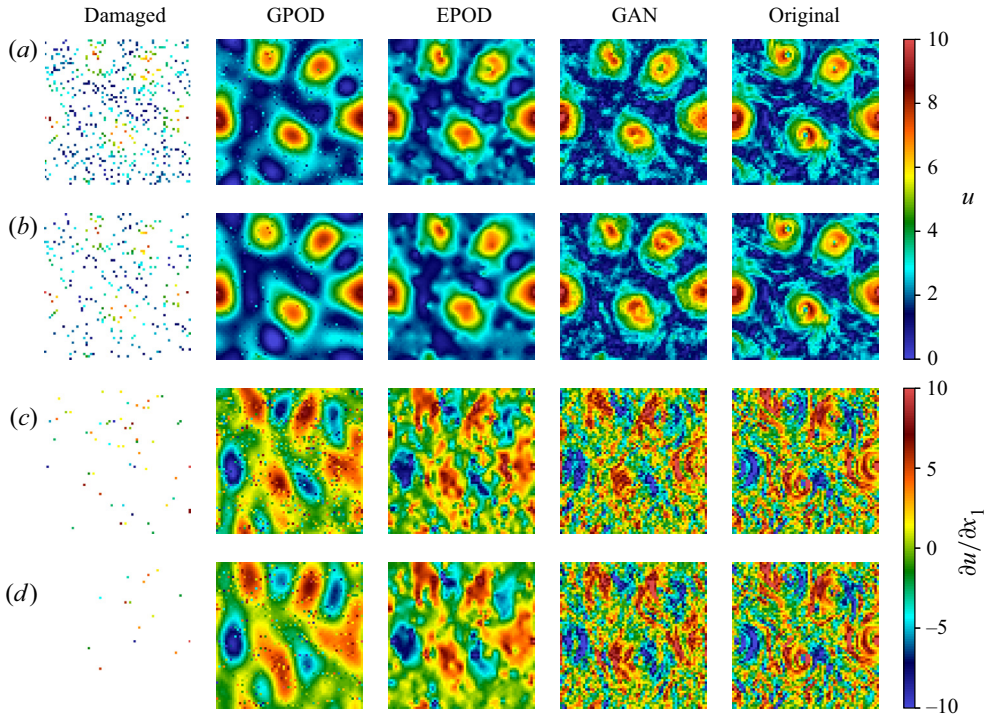


Figure 22. Reconstruction of an instantaneous field (velocity module) by the different tools for random gappiness of sizes $l/l_0 = 60/64$ (*a*) and $l/l_0 = 62/64$ (*b*). The corresponding gradient fields are shown in (*c*,*d*). The damaged fields are shown in the first column, while the second to fourth columns show the fields obtained from GPOD, EPOD and GAN. The ground truth is shown in the fifth column.

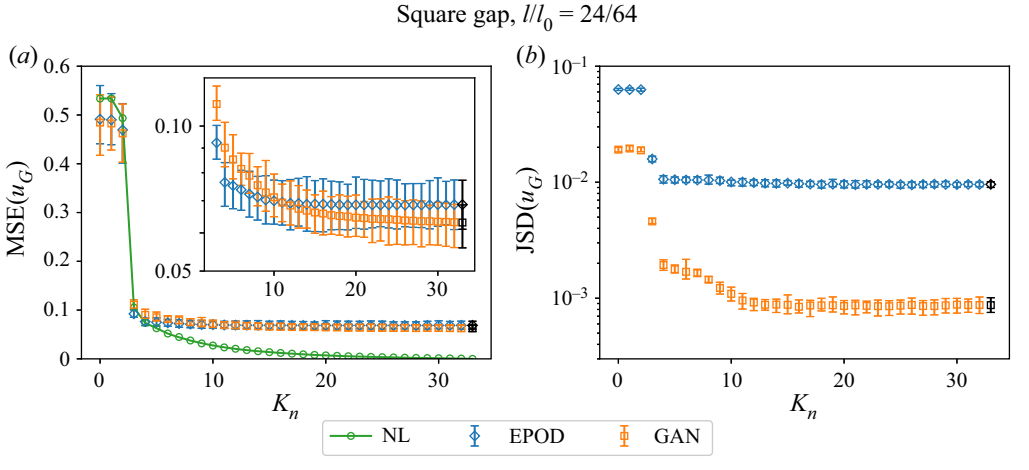


Figure 23. The MSE (*a*) and the JS divergence (*b*) between p.d.f.s for the original and generated velocity module inside the missing region, obtained from EPOD and GAN with input measurements of different noise levels for a square gap of size $l/l_0 = 24/64$. The results obtained with the noiseless input are plotted with black symbols at the right end of each panel. Results obtained with noiseless input are represented by black symbols, positioned on the right-hand side of each panel. The estimate of the noise level introduced in the physical space is given by the green curve (NL). The inset box presents the MSE on a log–lin scale.

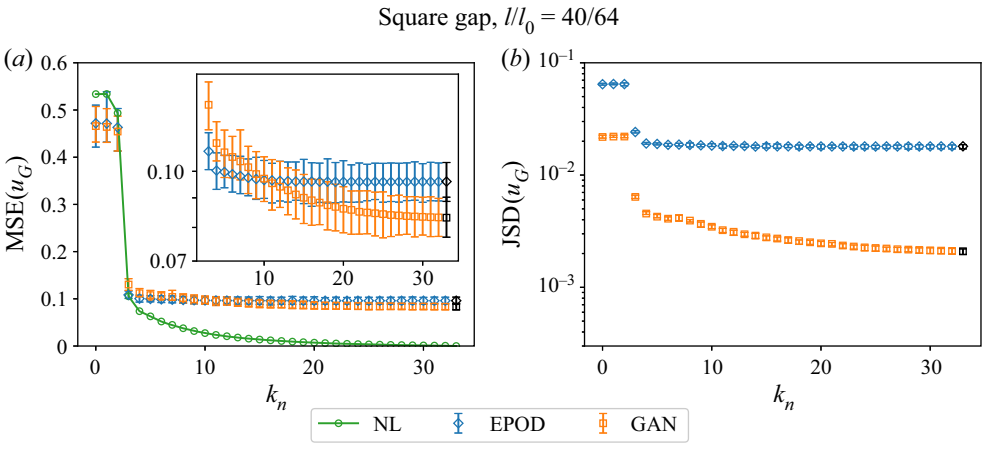


Figure 24. The same as figure 23 but for a square gap of size $l/l_0 = 40/64$.

fields taken from a database of 3-D rotating turbulent flows. Performances have been quantitatively judged on the basis of (i) L_2 distance between each the ground truth and the reconstructed field, (ii) statistical validations based on JS divergence between the one-point p.d.f.s, (iii) spectral properties and multi-scale flatness and (iv) extreme events for a single frame. For one central square gap the GAN approach is proven to be superior to GPOD and EPOD, when both MSE and JS divergence are simultaneously considered, in particular for large gap sizes where the missing of multi-scale information makes the task extremely difficult. Moreover, GAN predictions are also better in terms of the energy spectra and flatness, as well as for the predicted extreme events. In the presence of random damages, the three approaches give similar results except for the case of extreme gappiness, where GAN is leading again.

GPOD always generates 'discontinuous' predictions with respect to the supplied measurements. This is because GPOD only minimizes the L_2 distance and the optimal number of POD modes used is usually much smaller than the number of measured points. On the other hand, EPOD considers the correlation between the fields inside and outside the gap and its predictions have a number of degrees of freedom equal to the number of measured points. Compared with GPOD, EPOD is less computationally demanding and generates better predictions. When the gap is extremely large, neither GPOD nor EPOD gives satisfying predictions as they have too few degrees of freedom.

With the help of adversarial training, GAN can optimize a multi-objective problem, minimizing simultaneously the L_2 distance frame by frame and the JS divergence between the real and generated distributions of the whole fields in the missing region. Furthermore, we show that for GAN reconstructions, large adversarial ratios undermine the MSE but improve the generated statistical properties and *vice versa*.

In terms of the potential for practical applications of the three tools analysed in this study, we have demonstrated that both EPOD and GAN exhibit robust properties when faced with noisy multi-scale measurements. It is also worth noting that, in many applications, gaps can also arise in the Fourier space. This typically occurs when we encounter measurement noise or modelling limitations at high wavenumbers. In such situations, we face a super-resolution problem where we need to reconstruct the missing small-scale information.

Our work is a first step toward the setting up of benchmarks and grand challenges for realistic turbulent problems with interest in geophysical and laboratory applications,

where the lack of measurements obstructs our capability to fully control the system. Many questions remain open, connected to the performance of different GAN architectures, and the difficulty of having *a priori* estimates of the deepness and complexity of the GAN architecture as a function of the complexity of the physics, in particular concerning the quantity and the geometry (two- or three-dimensional) of the missing information. Furthermore, little is known about the performance of the data-driven models as a function of the Reynolds or Rossby numbers, and the possibility of supplying physics information to help to further improve the network's performances.

Funding. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 882340); the MUR-FARE project R2045J8XAW; the EuroHPC2023 programme (grant no. EHPC-REG-2021R0049); the NSFC (grant nos 12225204, 91752201 and 11988102); Shenzhen Science & Technology Program (grant no. KQTD20180411143441009); and the Department of Science and Technology of Guangdong Province (grant nos 2019B21203001, 2020B1212030001).

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

Author ORCIDs.
Tianyi Li https://orcid.org/0000-0002-7979-4426;
Luca Biferale https://orcid.org/0000-0001-8767-9092;
Minping Wan https://orcid.org/0000-0001-5891-9579.

Appendix A. Error analysis of GPOD reconstruction

Consider the whole region I, the gap region G and the known region S as sets of positions.

Given $m(\cdot)$ as a function returning the number of elements in a set, we can define

$$\left.\begin{aligned} I &= \{x_1, x_2, \ldots, x_{m(I)}\}, \\ G &= \{x_k \mid x_k \text{ in the gap}\} = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{m(G)}}\}, \\ S &= I \setminus G = \{x_{j_1}, x_{j_2}, \ldots, x_{j_{m(S)}}\}, \end{aligned}\right\} \tag{A1}$$

and

$$\left.\begin{aligned} \boldsymbol{x} &= \begin{bmatrix} x_1 & x_2 & \cdots & x_{m(I)} \end{bmatrix}^{\mathrm{T}}, \\ \bar{\boldsymbol{x}} &= \begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_{m(G)}} \end{bmatrix}^{\mathrm{T}}, \\ \tilde{\boldsymbol{x}} &= \begin{bmatrix} x_{j_1} & x_{j_2} & \cdots & x_{j_{m(S)}} \end{bmatrix}^{\mathrm{T}}. \end{aligned}\right\} \tag{A2}$$

With N' as the number of POD modes kept for dimension reduction, the POD decomposition

$$u(\boldsymbol{x}) = \sum_{n=1}^{N_I} a_n \psi_n(\boldsymbol{x}) = \sum_{n=1}^{N'} a_n \psi_n(\boldsymbol{x}) + \sum_{n=N'+1}^{N_I} a_n \psi_n(\boldsymbol{x}), \tag{A3}$$

can be written in the vector form

$$\boldsymbol{u} = \boldsymbol{X}\boldsymbol{a} = \boldsymbol{X}'\boldsymbol{a}' + \boldsymbol{r}', \tag{A4}$$

where the definitions of $\boldsymbol{u}$, $\boldsymbol{X}$, $\boldsymbol{X}'$, $\boldsymbol{a}$, $\boldsymbol{a}'$ and $\boldsymbol{r}'$ are shown below

$$\boldsymbol{u} = u(\boldsymbol{x}) = \begin{bmatrix} u(x_1) & u(x_2) & \cdots & u(x_{m(I)}) \end{bmatrix}^{\mathrm{T}}, \tag{A5}$$

$$\boldsymbol{X} = \begin{bmatrix} \psi_1(\boldsymbol{x}) & \psi_2(\boldsymbol{x}) & \cdots & \psi_{N_I}(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_{N_I}(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_{N_I}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{m(I)}) & \psi_2(x_{m(I)}) & \cdots & \psi_{N_I}(x_{m(I)}) \end{bmatrix}, \tag{A6}$$

$$\boldsymbol{X}' = \begin{bmatrix} \psi_1(\boldsymbol{x}) & \psi_2(\boldsymbol{x}) & \cdots & \psi_{N'}(\boldsymbol{x}) \end{bmatrix}, \tag{A7}$$

$$\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{N_I} \end{bmatrix}^{\mathrm{T}}, \tag{A8}$$

$$\boldsymbol{a}' = \begin{bmatrix} a_1 & a_2 & \cdots & a_{N'} \end{bmatrix}^{\mathrm{T}}, \tag{A9}$$

$$\boldsymbol{r}' = \begin{bmatrix} \psi_{N'+1}(\boldsymbol{x}) & \psi_{N'+2}(\boldsymbol{x}) & \cdots & \psi_{N_I}(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} a_{N'+1} & a_{N'+2} & \cdots & a_{N_I} \end{bmatrix}^{\mathrm{T}}. \tag{A10}$$

Here, $(\cdot)'$ is connected to truncating the POD space with N_I modes to the leading N' modes. Before moving on to GPOD reconstruction, we denote

$$\bar{\boldsymbol{u}} = u(\bar{\boldsymbol{x}}), \quad \tilde{\boldsymbol{u}} = u(\tilde{\boldsymbol{x}}), \tag{A11a,b}$$

and

$$\left.\begin{aligned} \bar{\boldsymbol{X}} &= \begin{bmatrix} \psi_1(\bar{\boldsymbol{x}}) & \psi_2(\bar{\boldsymbol{x}}) & \cdots & \psi_{N_I}(\bar{\boldsymbol{x}}) \end{bmatrix} \\ &= \begin{bmatrix} \psi_1(x_{i_1}) & \psi_2(x_{i_1}) & \cdots & \psi_{N_I}(x_{i_1}) \\ \psi_1(x_{i_2}) & \psi_2(x_{i_2}) & \cdots & \psi_{N_I}(x_{i_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{i_{m(G)}}) & \psi_2(x_{i_{m(G)}}) & \cdots & \psi_{N_I}(x_{i_{m(G)}}) \end{bmatrix}, \\ \tilde{\boldsymbol{X}} &= \begin{bmatrix} \psi_1(\tilde{\boldsymbol{x}}) & \psi_2(\tilde{\boldsymbol{x}}) & \cdots & \psi_{N_I}(\tilde{\boldsymbol{x}}) \end{bmatrix} \\ &= \begin{bmatrix} \psi_1(x_{j_1}) & \psi_2(x_{j_1}) & \cdots & \psi_{N_I}(x_{j_1}) \\ \psi_1(x_{j_2}) & \psi_2(x_{j_2}) & \cdots & \psi_{N_I}(x_{j_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{j_{m(S)}}) & \psi_2(x_{j_{m(S)}}) & \cdots & \psi_{N_I}(x_{j_{m(S)}}) \end{bmatrix}. \end{aligned}\right\} \tag{A12}$$

Besides, $\bar{\boldsymbol{X}}'$, $\tilde{\boldsymbol{X}}'$, $\bar{\boldsymbol{r}}'$ and $\tilde{\boldsymbol{r}}'$ can be similarly defined.

To conduct GPOD reconstruction, we minimize the error in the measurement region S given by (2.7)

$$\tilde{E} = \int_S |u(\boldsymbol{x}) - \sum_{n=1}^{N'} a_n^{(p)} \psi_n(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} = \|\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{X}}' \boldsymbol{a}'^{(p)}\|^2, \tag{A13}$$

and the best fit of coefficients is given as (Penrose 1956; Planitz 1979)

$$\boldsymbol{a}'^{(p)} = \tilde{\boldsymbol{X}}'_{+} \tilde{\boldsymbol{u}} + (\boldsymbol{I}' - \tilde{\boldsymbol{X}}'_{+} \tilde{\boldsymbol{X}}') \boldsymbol{w}', \tag{A14}$$

where $\boldsymbol{I}' \in \mathbb{R}^{N' \times N'}$ is an identity matrix, $\tilde{\boldsymbol{X}}'_{+}$ is the pseudo-inverse of $\tilde{\boldsymbol{X}}'$ satisfying the Moore–Penrose conditions and $\boldsymbol{w}' \in \mathbb{R}^{N' \times 1}$ is an arbitrary vector. Then the reconstructed

field in G is obtained from

$$\bar{\boldsymbol{u}}^{(p)} = \bar{\boldsymbol{X}}'\boldsymbol{a}'^{(p)}, \tag{A15}$$

and the reconstruction error is

$$\bar{E} = \|\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}^{(p)}\|^2 = \|\bar{\boldsymbol{X}}'[(\boldsymbol{I}' - \tilde{\boldsymbol{X}}'_+\tilde{\boldsymbol{X}}')(\boldsymbol{a}' - \boldsymbol{w}') - \tilde{\boldsymbol{X}}'_+\tilde{\boldsymbol{r}}'] + \bar{\boldsymbol{r}}'\|^2 = \|\bar{\boldsymbol{e}}_1 + \bar{\boldsymbol{e}}_2 + \bar{\boldsymbol{e}}_3\|^2, \tag{A16}$$

where

$$\bar{\boldsymbol{e}}_1 = \bar{\boldsymbol{X}}'(\boldsymbol{I}' - \tilde{\boldsymbol{X}}'_+\tilde{\boldsymbol{X}}')(\boldsymbol{a}' - \boldsymbol{w}'), \quad \bar{\boldsymbol{e}}_2 = -\bar{\boldsymbol{X}}'\tilde{\boldsymbol{X}}'_+\tilde{\boldsymbol{r}}', \quad \bar{\boldsymbol{e}}_3 = \bar{\boldsymbol{r}}'. \tag{A17a–c}$$

Equations (A16) and (A17a–c) show that the reconstruction error depends on three terms, the contributions of which can be calculated as

$$\bar{C}_1 = \|\bar{\boldsymbol{e}}_1\|^2, \quad \bar{C}_2 = \|\bar{\boldsymbol{e}}_2\|^2, \quad \bar{C}_3 = \|\bar{\boldsymbol{e}}_3\|^2. \tag{A18a–c}$$

For a square gap with sizes $l/l_0 = 8/64$ and $40/64$, figure 25 shows $\langle\bar{C}_1\rangle$, $\langle\bar{C}_2\rangle$, $\langle\bar{C}_3\rangle$ and $\langle\bar{E}\rangle$ as functions of N', where the angle brackets represent the average over training data. Quantities are normalized by $A_G E_{u_G}$. It shows that $\langle\bar{C}_1\rangle$ is always zero when N' is smaller than a threshold, N'_c, because, in this case, $\tilde{\boldsymbol{X}}'$ is invertible (with a condition number less than 10^{10}) and thus $\boldsymbol{I}' - \tilde{\boldsymbol{X}}'_+\tilde{\boldsymbol{X}}' = \boldsymbol{0}$ in (A17a–c). The arbitrariness of $\boldsymbol{w}'$ only takes effect when N' is larger than N'_c, in which case $\tilde{\boldsymbol{X}}'$ is not invertible and $\langle\bar{C}_1\rangle$ is not zero. In figure 25 we use the grey area to indicate this range of N' and plot $\langle\bar{C}_1\rangle$ and $\langle\tilde{E}\rangle$ with $\boldsymbol{w}' = \boldsymbol{0}$. When N' increases from zero, $\langle\bar{C}_3\rangle$ always decreases as it represents the truncation error of POD expansion, while $\langle\bar{C}_2\rangle$ increases at $N' < N'_c$ and decreases at $N' > N'_c$. Because of the trade-off between different error components, there exists an optimal N' with the smallest reconstruction error $\langle\bar{E}\rangle$, which will be used in the testing process.

Appendix B. The GPOD reconstruction with Lasso regularization

Different from using dimension reduction (DR) to keep only the leading POD modes, GPOD can use the complete POD decomposition for reconstruction

$$u_G^{(p)}(\boldsymbol{x}) = \sum_{n=1}^{N_I} a_n^{(p)}\psi_n(\boldsymbol{x}) \quad (\boldsymbol{x} \in G), \tag{B1}$$

and minimize the distance between the measurements and the POD decomposition with the help of Lasso regularization (Tibshirani 1996)

$$\tilde{E}_{L_1} = \int_S \left| u_S(\boldsymbol{x}) - \sum_{n=1}^{N_I} a_n^{(p)}\psi_n(\boldsymbol{x}) \right|^2 \mathrm{d}\boldsymbol{x} + \alpha \sum_{n=1}^{N_I} |a_n^{(p)}|. \tag{B2}$$

Lasso penalizes the L_1 norm of the coefficients and tends to produce some coefficients that are exactly zero, which is similar to finding a best subset of POD modes that does not necessarily consist of the leading ones. The hyper-parameter α controls regularization strength and we estimate α by fivefold cross-validation (Efron & Tibshirani 1994) with the data in S during the reconstruction process.

With this approach, we conducted a reconstruction experiment for a square gap of size $l/l_0 = 40/64$ for illustration and it is not our intention to perform a systematic

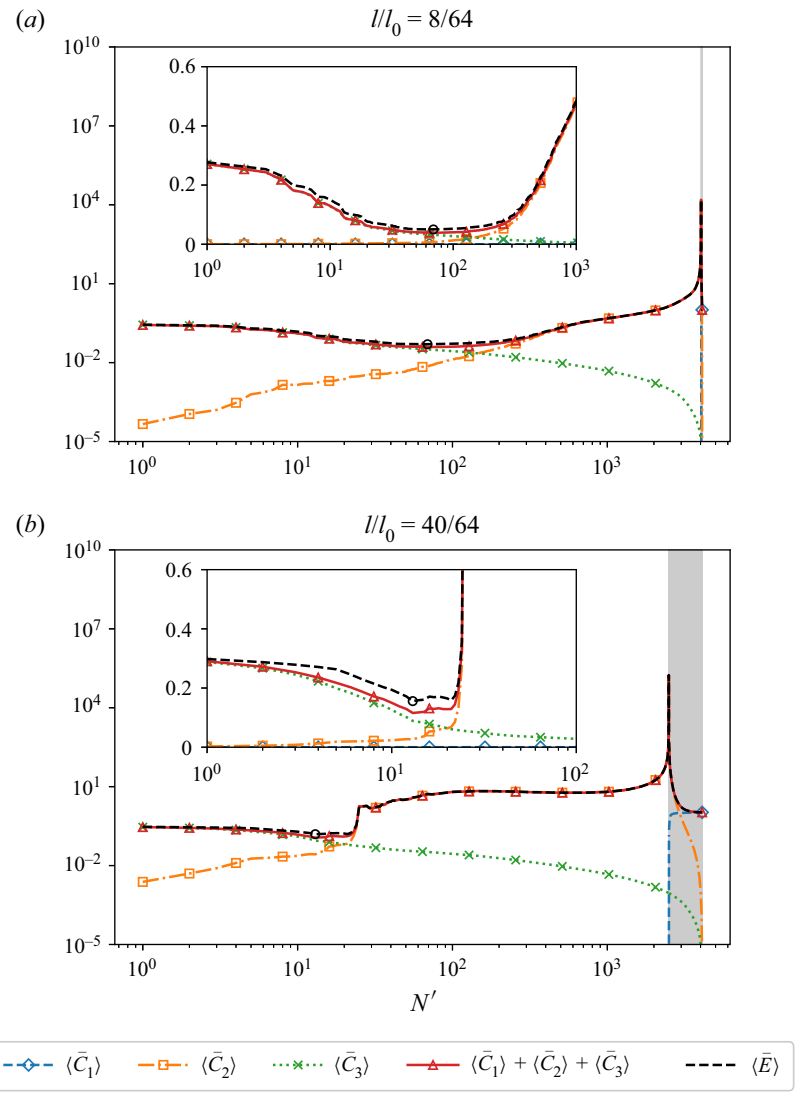


Figure 25. The GPOD reconstruction error with its different contributions as functions of the number of POD modes, for a square gap with different gap sizes. The corresponding plots in the lin–log scale are shown in the insets. The black circles indicate the optimal N' with the smallest reconstruction error. The range of N' where the arbitrariness of $\boldsymbol{w}'$ takes effect is indicated by the grey area.

investigation of changing the geometry and area of the gap. Figure 26(*a*) shows the p.d.f. of the estimated value of α over the test data for Lasso regression.Table 4 shows that the GPOD reconstructions with DR and Lasso give similar values of MSE(u_G) and JSD(u_G). Figure 26(*b*) also shows that the p.d.f.s of their predicted velocity module are comparable. The difference between DR and Lasso can be illustrated by the spectra of the predicted POD coefficients of an instantaneous field with a square gap of size $l/l_0 = 40/64$, as shown in figure 27(*a*). The DR gives a non-zero spectrum up to $N' = 12$, while Lasso selects both large- and small-scale modes with a wide range of indices. This can be further shown with the reconstruction in figure 27(*b*), where DR only predicts 'smooth' structures given by the leading POD modes and Lasso generates predictions with multiple scales.

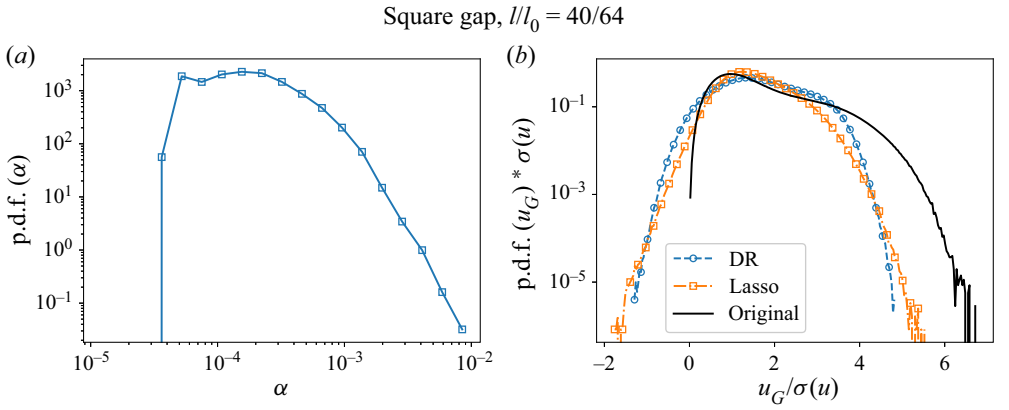


Figure 26. The p.d.f. of the estimated value of α over the test data for Lasso regression (a) and p.d.f.s of the velocity module from the ground truth and that from the missing region obtained from GPOD with DR and Lasso (b) for a square gap of size $l/l_0 = 40/64$.

	MSE(u_G)	JSD(u_G)
DR	$0.17^{+0.01}_{-0.02}$	$0.048^{+0.001}_{-0.001}$
Lasso	$0.20^{+0.02}_{-0.02}$	$0.049^{+0.001}_{-0.001}$

Table 4. The MSE and the JS divergence between p.d.f.s for the original and generated velocity module inside the missing region, obtained from GPOD with DR and Lasso for a square gap of size $l/l_0 = 40/64$. The mean value and the error bound are calculated over test batches of size 128 for MSE and 2048 for JS divergence.

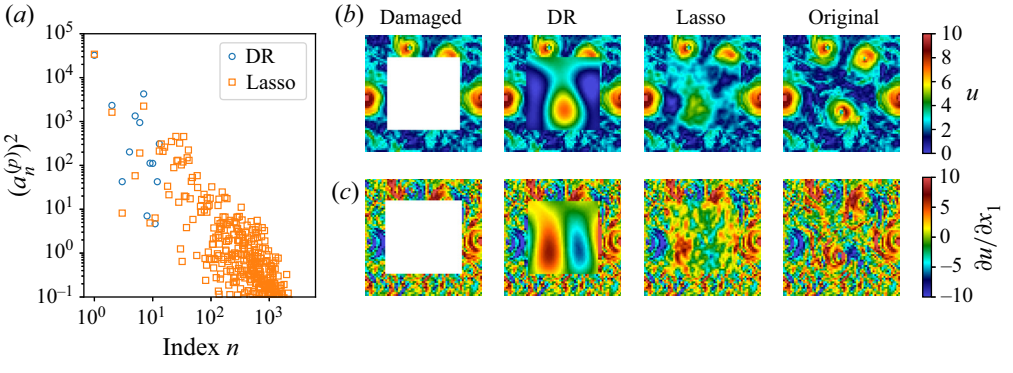


Figure 27. The spectra of the predicted POD coefficients obtained from the GPOD with DR and Lasso for an instantaneous field with a square gap of size $l/l_0 = 40/64$ (a). The corresponding damaged, reconstructed and original velocity module fields with their gradient fields are shown on the right.

Appendix C

This appendix contains details of the GAN used in this study, of which the architecture is shown in figure 4. For a square gap with different sizes $l = 8$, 16, 24, 32, 40, 50, 60 and 62, we use different kernel sizes for the last layer of the generator and the first layer of the discriminator, $k = 8$, 4, 18, 2, 25, 15, 5 and 3. This can be obtained from the relation

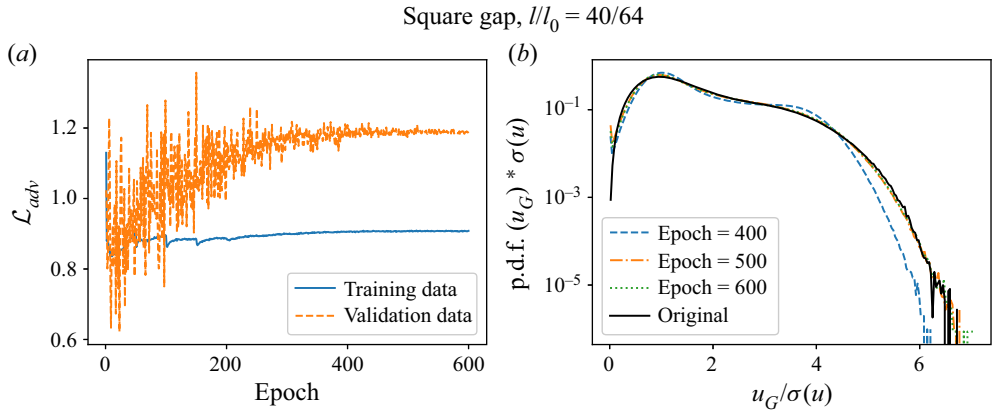


Figure 28. The adversarial loss as a function of epoch (*a*) and p.d.f.s of the predictions in the missing region at different epochs and the one of the ground truth over the whole region for the validation data (*b*). Results are obtained from the training process of the GAN for a square gap of size $l/l_0 = 40/64$.

for the corresponding unpadded convolution (up-convolution) layer, $l = (64 - k)/s + 1$, as both k and the stride s are integers. For random gappiness, we use $l = 64$ and $k = 1$ but the L_2 loss is only computed in the gap. Moreover, the whole output of generator is used as the input of discriminator. To generate positive output (velocity module), ReLU is adopted as the activation function for the last layer of generator. The negative slope of the leaky ReLU activation function is empirically chosen as 0.2 for other convolution (up-convolution) layers. As illustrated in § 4, we can pick the adversarial ratio to obtain a good compromise between MSE and the reconstructed turbulent statistics, which gives $\lambda_{adv} = 10^{-2}$ for a central square gap and $\lambda_{adv} = 10^{-3}$ for random gappiness.

We train the generator and discriminator together with Adam optimizer (Kingma & Ba 2014), where the learning rate of generator is twice that of discriminator. To improve the stability of training, a staircase-decay schedule is adopted to the learning rate. It decays with a rate of 0.5 every 50 epochs for 11 times, corresponding to the maximum epoch equal to 600. We choose a batch size of 128 and the initial learning rate of generator as 10^{-3}. Figure 28 shows the training process of the GAN for a $l/l_0 = 40/64$ central square gap. As training proceeds, the adversarial loss saturates at fixed values (figure 28*a*), while the predicted p.d.f. gets closer to the ground truth for the validation data (figure 28*b*). This indicates the training convergence.

Appendix D

To simplify the notation, we denote x and y, respectively, as the predicted and original fields inside the missing region. Here, we use the angle brackets as the average over the gap and over the test data

$$\langle \cdot \rangle = \frac{1}{N_{test}} \sum_{c=1}^{N_{test}} \left(\frac{1}{A_G} \int_G (\cdot)^c \, \mathrm{d}\boldsymbol{x} \right), \tag{D1}$$

where c is the index of a flow frame in the test data. As shown in § 3.1, the baseline MSE comes from uncorrelated predictions that are statistically consistent with the ground truth. Therefore, we have $\langle xy \rangle = \langle x \rangle \langle y \rangle$ because of the uncorrelation and the statistical consistency gives that $\langle x \rangle = \langle y \rangle$ and $\langle x^2 \rangle = \langle y^2 \rangle$. With the simplified notation, (3.1) can be

rewritten as

$$\mathrm{MSE} = \frac{\langle (x-y)^2 \rangle}{\sqrt{\langle x^2 \rangle}\sqrt{\langle y^2 \rangle}} = \frac{\langle x^2 \rangle - 2\langle xy \rangle + \langle y^2 \rangle}{\sqrt{\langle x^2 \rangle}\sqrt{\langle y^2 \rangle}}. \tag{D2}$$

Then, using the relations above, we can obtain the baseline MSE

$$\mathrm{MSE}_b = \frac{2(\langle x^2 \rangle - \langle x \rangle^2)}{\langle x^2 \rangle} = 2\left(1 - \frac{\langle x \rangle^2}{\langle x^2 \rangle}\right). \tag{D3}$$

For the velocity module, with its mean value $\langle u_G \rangle$ and the mean energy $\langle u_G^2 \rangle$ in the gap, we have the estimate $\mathrm{MSE}_b(u_G) \approx 0.5358$. For the gradient, as $\langle \partial u_G / \partial x_1 \rangle = 0$ resulted from the periodicity, one can obtain $\mathrm{MSE}_b(\partial u_G / \partial x_1) \approx 2$.

References

Alexakis, A. & Biferale, L. 2018 Cascades and transitions in turbulent flows. *Phys. Rep.* **767**, 1–101.

Asch, M., Bocquet, M. & Nodet, M. 2016 *Data Assimilation: Methods, Algorithms, and Applications*. SIAM.

Baral, C., Fuentes, O. & Kreinovich, V. 2018 Why deep neural networks: a possible theoretical explanation. In *Constraint Programming and Decision Making: Theory and Applications* (ed. M. Ceberio & V. Kreinovich), Studies in Systems, Decision and Control, vol. 100, pp. 1–5. Springer.

Bell, M.J., Lefebvre, M., Le Traon, P.-Y., Smith, N. & Wilmer-Becker, K. 2009 Godae: the global ocean data assimilation experiment. *Oceanography* **22** (3), 14–21.

Biferale, L., Bonaccorso, F., Buzzicotti, M. & di Leoni, P.C. 2020 Turb-rot. A large database of 3d and 2d snapshots from turbulent rotating flows. arXiv:2006.07469.

Borée, J. 2003 Extended proper orthogonal decomposition: a tool to analyse correlated events in turbulent flows. *Exp. Fluids* **35** (2), 188–192.

Brunton, S.L. & Noack, B.R. 2015 Closed-loop turbulence control: progress and challenges. *Appl. Mech. Rev.* **67** (5), 050801.

Buzzicotti, M., Bhatnagar, A., Biferale, L., Lanotte, A.S. & Ray, S.S. 2016 Lagrangian statistics for Navier–Stokes turbulence under Fourier-mode reduction: fractal and homogeneous decimations. *New J. Phys.* **18** (11), 113047.

Buzzicotti, M. & Bonaccorso, F. 2022 Inferring turbulent environments via machine learning. *Eur. Phys. J.* E **45** (12), 102.

Buzzicotti, M., Bonaccorso, F., Di Leoni, P.C. & Biferale, L. 2021 Reconstruction of turbulent data with deep generative models for semantic inpainting from turb-rot database. *Phys. Rev. Fluids* **6** (5), 050503.

Buzzicotti, M., Clark Di Leoni, P. & Biferale, L. 2018 On the inverse energy transfer in rotating turbulence. *Eur. Phys. J.* E **41** (11), 1–8.

Choi, H., Moin, P. & Kim, J. 1994 Active turbulence control for drag reduction in wall-bounded flows. *J. Fluid Mech.* **262**, 75–110.

Clark Di Leoni, P., Agarwal, K., Zaki, T., Meneveau, C. & Katz, J. 2022 Reconstructing velocity and pressure from sparse noisy particle tracks using physics-informed neural networks. arXiv:2210.04849.

Cohen, I.M. & Kundu, P.K. 2004 *Fluid Mechanics*. Elsevier.

Dabiri, D. & Pecora, C. 2020 *Particle Tracking Velocimetry*, vol. 785. IOP Publishing.

Deng, Z., He, C., Liu, Y. & Kim, K.C. 2019 Super-resolution reconstruction of turbulent velocity fields using a generative adversarial network-based artificial intelligence framework. *Phys. Fluids* **31** (12), 125111.

Di Leoni, P.C., Alexakis, A., Biferale, L. & Buzzicotti, M. 2020 Phase transitions and flux-loop metastable states in rotating turbulence. *Phys. Rev. Fluids* **5** (10), 104603.

Discetti, S., Bellani, G., Örlü, R., Serpieri, J., Vila, C.S., Raiola, M., Zheng, X., Mascotelli, L., Talamelli, A. & Ianiro, A. 2019 Characterization of very-large-scale motions in high-*Re* pipe flows. *Exp. Therm. Fluid Sci.* **104**, 1–8.

Efron, B. & Tibshirani, R.J. 1994 *An Introduction to the Bootstrap*. CRC Press.

Everson, R. & Sirovich, L. 1995 Karhunen–Loeve procedure for gappy data. *J. Opt. Soc. Am.* A **12** (8), 1657–1664.

Fahland, G., Stroh, A., Frohnapfel, B., Atzori, M., Vinuesa, R., Schlatter, P. & Gatti, D. 2021 Investigation of blowing and suction for turbulent flow control on airfoils. *AIAA J.* **59** (11), 4422–4436.

FRISCH, U. 1995 *Turbulence: The Legacy of A.N. Kolmogorov*. Cambridge University Press.
FRISCH, U., KURIEN, S., PANDIT, R., PAULS, W., RAY, S.S., WIRTH, A. & ZHU, J.-Z. 2008 Hyperviscosity, galerkin truncation, and bottlenecks in turbulence. *Phys. Rev. Lett.* **101** (14), 144501.
FUKAMI, K., FUKAGATA, K. & TAIRA, K. 2019 Super-resolution reconstruction of turbulent flows with machine learning. *J. Fluid Mech.* **870**, 106–120.
FUKAMI, K., FUKAGATA, K. & TAIRA, K. 2021 Machine-learning-based spatio-temporal super resolution reconstruction of turbulent flows. *J. Fluid Mech.* **909**, A9.
FUKUNAGA, K. 2013 *Introduction to Statistical Pattern Recognition*. Elsevier.
GAD-EL HAK, M. & TSAI, H.M. 2006 *Transition and Turbulence Control*, vol. 8. World Scientific.
GARCIA, D. 2011 A fast all-in-one method for automated post-processing of PIV data. *Exp. Fluids* **50** (5), 1247–1259.
GODEFERD, F.S. & MOISY, F. 2015 Structure and dynamics of rotating turbulence: a review of recent experimental and numerical results. *Appl. Mech. Rev.* **67** (3), 030802.
GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. & BENGIO, Y. 2014 Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**.
GUASTONI, L., GÜEMES, A., IANIRO, A., DISCETTI, S., SCHLATTER, P., AZIZPOUR, H. & VINUESA, R. 2021 Convolutional-network models to predict wall-bounded turbulence from wall quantities. *J. Fluid Mech.* **928**, A27.
GÜEMES, A., DISCETTI, S. & IANIRO, A. 2019 Sensing the turbulent large-scale motions with their wall signature. *Phys. Fluids* **31** (12), 125112.
GÜEMES, A., DISCETTI, S., IANIRO, A., SIRMACEK, B., AZIZPOUR, H. & VINUESA, R. 2021 From coarse wall measurements to turbulent velocity fields through deep learning. *Phys. Fluids* **33** (7), 075121.
GUNES, H. & RIST, U. 2008 On the use of kriging for enhanced data reconstruction in a separated transitional flat-plate boundary layer. *Phys. Fluids* **20** (10), 104109.
GUNES, H., SIRISUP, S. & KARNIADAKIS, G.E. 2006 Gappy data: to krig or not to krig? *J. Comput. Phys.* **212** (1), 358–382.
HAUGEN, N.E.L. & BRANDENBURG, A. 2004 Inertial range scaling in numerical turbulence with hyperviscosity. *Phys. Rev.* E **70** (2), 026405.
HE, K., ZHANG, X., REN, S. & SUN, J. 2016 Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
HOLMES, P., LUMLEY, J.L., BERKOOZ, G. & ROWLEY, C.W. 2012 *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press.
HORNIK, K. 1991 Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4** (2), 251–257.
HOSSEINI, Z., MARTINUZZI, R.J. & NOACK, B.R. 2016 Modal energy flow analysis of a highly modulated wake behind a wall-mounted pyramid. *J. Fluid Mech.* **798**, 717–750.
VAN KAN, A. & ALEXAKIS, A. 2020 Critical transition in fast-rotating turbulence within highly elongated domains. *J. Fluid Mech.* **899**, A33.
KIM, H., KIM, J., WON, S. & LEE, C. 2021 Unsupervised deep learning for super-resolution reconstruction of turbulence. *J. Fluid Mech.* **910**, A29.
KIM, J. & LEE, C. 2020 Prediction of turbulent heat transfer using convolutional neural networks. *J. Fluid Mech.* **882**, A18.
KINGMA, D.P. & BA, J. 2014 Adam: a method for stochastic optimization. arXiv:1412.6980.
KREINOVICH, V.Y. 1991 Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem. *Neural Netw.* **4** (3), 381–383.
KRYSTA, M., BLAYO, E., COSME, E. & VERRON, J. 2011 A consistent hybrid variational-smoothing data assimilation method: application to a simple shallow-water model of the turbulent midlatitude ocean. *Mon. Weath. Rev.* **139** (11), 3333–3347.
LE DIMET, F.-X. & TALAGRAND, O. 1986 Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus* A **38** (2), 97–110.
LEE, C., KIM, J., BABCOCK, D. & GOODMAN, R. 1997 Application of neural networks to turbulence control for drag reduction. *Phys. Fluids* **9** (6), 1740–1747.
LI, T., BUZZICOTTI, M., BIFERALE, L. & BONACCORSO, F. 2023 Generative adversarial networks to infer velocity components in rotating turbulent flows. *Eur. Phys. J.* E **46**, 31.
LI, T., BUZZICOTTI, M., BIFERALE, L., WAN, M. & CHEN, S. 2021 Reconstruction of turbulent data with gappy pod method. *Chin. J. Theor. Appl. Mech.* **53** (10), 2703–2711.
LIU, B., TANG, J., HUANG, H. & LU, X.-Y. 2020 Deep learning methods for super-resolution reconstruction of turbulent flows. *Phys. Fluids* **32** (2), 025105.

MATSUO, M., NAKAMURA, T., MORIMOTO, M., FUKAMI, K. & FUKAGATA, K. 2021 Supervised convolutional network for three-dimensional fluid data reconstruction from sectional flow fields with adaptive super-resolution assistance. arXiv:2103.09020.

MAUREL, S., BORÉE, J. & LUMLEY, J.L. 2001 Extended proper orthogonal decomposition: application to jet/vortex interaction. *Flow Turbul. Combust.* **67** (2), 125–136.

MILITINO, A.F., UGARTE, M.D. & MONTESINO, M. 2019 Filling missing data and smoothing altered data in satellite imagery with a spatial functional procedure. *Stoch. Environ. Res. Risk Assess.* **33** (10), 1737–1750.

NIU, X.-X. & SUEN, C.Y. 2012 A novel hybrid CNN–SVM classifier for recognizing handwritten digits. *Pattern Recognit.* **45** (4), 1318–1325.

OLIVER, M.A. & WEBSTER, R. 1990 Kriging: a method of interpolation for geographical information systems. *Intl J. Geogr. Inform. Syst.* **4** (3), 313–332.

PATHAK, D., KRAHENBUHL, P., DONAHUE, J., DARRELL, T. & EFROS, A.A. 2016 Context encoders: feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544. IEEE.

PENROSE, R. 1956 On best approximate solutions of linear matrix equations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 52, pp. 17–19. Cambridge University Press.

PLANITZ, M. 1979 3. Inconsistent systems of linear equations. *Math. Gaz.* **63** (425), 181–185.

POUQUET, A., ROSENBERG, D., MARINO, R. & HERBERT, C. 2018 Scaling laws for mixing and dissipation in unforced rotating stratified turbulence. *J. Fluid Mech.* **844**, 519–545.

ROMAIN, L., CHATELLIER, L. & DAVID, L. 2014 Bayesian inference applied to spatio-temporal reconstruction of flows around a NACA0012 airfoil. *Exp. Fluids* **55** (4), 1–19.

RUSSAKOVSKY, O., *et al.* 2015 Imagenet large scale visual recognition challenge. *Intl J. Comput. Vis.* **115** (3), 211–252.

SAWFORD, B.L. 1991 Reynolds number effects in lagrangian stochastic models of turbulent dispersion. *Phys. Fluids* A **3** (6), 1577–1586.

SESHASAYANAN, K. & ALEXAKIS, A. 2018 Condensates in rotating turbulent flows. *J. Fluid Mech.* **841**, 434–462.

SHEN, H., LI, X., CHENG, Q., ZENG, C., YANG, G., LI, H. & ZHANG, L. 2015 Missing information reconstruction of remote sensing data: a technical review. *IEEE Geosci. Remote Sens. Mag.* **3** (3), 61–85.

SINGH, S.N., MYATT, J.H., ADDINGTON, G.A., BANDA, S. & HALL, J.K. 2001 Optimal feedback control of vortex shedding using proper orthogonal decomposition models. *J. Fluids Engng* **123** (3), 612–618.

SIROVICH, L. & KIRBY, M. 1987 Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am.* A **4** (3), 519–524.

STEIN, M.L. 1999 *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

STORER, B.A., BUZZICOTTI, M., KHATRI, H., GRIFFIES, S.M. & ALUIE, H. 2022 Global energy spectrum of the general oceanic circulation. *Nat. Commun.* **13** (1), 5314.

SUBRAMANIAM, A., WONG, M.L., BORKER, R.D., NIMMAGADDA, S. & LELE, S.K. 2020 Turbulence enrichment using physics-informed generative adversarial networks. arXiv:2003.01907.

SUZUKI, T. 2014 Pod-based reduced-order hybrid simulation using the data-driven transfer function with time-resolved PTV feedback. *Exp. Fluids* **55** (8), 1–17.

TIBSHIRANI, R. 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* B **58** (1), 267–288.

TINNEY, C.E., UKEILEY, L.S. & GLAUSER, M.N. 2008 Low-dimensional characteristics of a transonic jet. Part 2. Estimate and far-field prediction. *J. Fluid Mech.* **615**, 53–92.

TORN, R.D. & HAKIM, G.J. 2009 Ensemble data assimilation applied to rainex observations of Hurricane Katrina (2005). *Mon. Weath. Revi.* **137** (9), 2817–2829.

VENTURI, D. & KARNIADAKIS, G.E. 2004 Gappy data and reconstruction procedures for flow past a cylinder. *J. Fluid Mech.* **519**, 315–336.

WANG, C., GAO, Q., WANG, H., WEI, R., LI, T. & WANG, J. 2016 Divergence-free smoothing for volumetric PIV data. *Exp. Fluids* **57** (1), 15.

WANG, Z., BOVIK, A.C., SHEIKH, H.R. & SIMONCELLI, E.P. 2004 Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13** (4), 600–612.

WANG, Z. & SIMONCELLI, E.P. 2005 Translation insensitive image similarity in complex wavelet domain. In *Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, vol. 2, pp. II–573. IEEE.

WEN, X., LI, Z., PENG, D., ZHOU, W. & LIU, Y. 2019 Missing data recovery using data fusion of incomplete complementary data sets: a particle image velocimetry application. *Phys. Fluids* **31** (2), 025105.

YOKOYAMA, N. & TAKAOKA, M. 2021 Energy-flux vector in anisotropic turbulence: application to rotating turbulence. *J. Fluid Mech.* **908**, A17.

YOUSIF, M.Z., YU, L., HOYAS, S., VINUESA, R. & LIM, H. 2022 A deep-learning approach for reconstructing 3D turbulent flows from 2D observation data. arXiv:2208.05754.

ZHANG, Q., YUAN, Q., ZENG, C., LI, X. & WEI, Y. 2018 Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **56** (8), 4274–4288.