Data reconstruction of turbulent flows with Gappy POD, Extended POD and Generative Adversarial Networks

Tianyi Li¹,², Michele Buzzicotti¹, Luca Biferale¹ †, Fabio Bonaccorso¹, Shiyi Chen²,³ and Minping Wan²,³

¹Department of Physics and INFN, University of Rome “Tor Vergata”, Via della Ricerca Scientifica 1, 00133, Rome, Italy
²Guangdong 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
³Guangdong-Hong Kong-Macao Joint Laboratory for Data-Driven Fluid Mechanics and Engineering Applications, Southern University of Science and Technology, Shenzhen 518055, PR China

(Received xx; revised xx; accepted xx)

Three methods are used to reconstruct two-dimensional instantaneous velocity fields in a turbulent flow under rotation. The first two methods both use the linear proper orthogonal decomposition (POD), which are Gappy POD (GPOD) and Extended POD (EPOD), while the third one reconstructs the flow using a fully non-linear Convolutional Neural Network embedded in a Generative Adversarial Network (GAN). First, we show that there is always an optimal number of modes regarding a specific gap for the GPOD with dimension reduction. Moreover, adopting a Lasso regularizer for GPOD provides comparable reconstruction results. In order to systematically compare the applicability of the three tools, we consider a square gap at changing the size. Results show that compared with POD-based methods, GAN reconstruction not only has a smaller $L^2$ error, but also better turbulent statistics of both the velocity module and the velocity module gradient. This can be attributed to the ability of nonlinearity expression of the network and the presence of adversarial loss during the GAN training. We also investigate effects of the adversarial ratio, which controls the compromising between the $L^2$ error and the statistical properties. Finally, we assess the reconstruction on random gappiness. All methods perform well for small- and medium-size gaps, while GAN works better when the gappiness is large.

Key words: Authors should not enter keywords on the manuscript, as these must be chosen by the author during the online submission process and will then be added during the typesetting process (see Keyword PDF for the full list). Other classifications will be added at the same time.

MSC Codes (Optional) Please enter your MSC Codes here

1. Introduction

The problem of reconstructing missing information, due to measurements constraints and lack of spatial/temporal resolution, is ubiquitous in almost all important applications of turbulence to laboratory experiments, geophysics, meteorology and oceanography (Asch et al. 2016; Le Dimet...
& Talagrand 1986; Torn & Hakim 2009; Bell et al. 2009; Krysta et al. 2011). 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 et al. 2019). In particle image velocimetry (PIV) experiments, instances of missing information come from the issue of out-of-pair particles, object shadows or the light reflection (Garcia 2011; Wang et al. 2016; Wen et al. 2019). Similarly, in many instances, the experimental probes are limited to assess only a subset of the relevant fields, asking for a careful apriori 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 make them useful for the fluid-dynamics community. In this paper, we perform a first systematic quantitative comparison among three data-driven methods (no information on the underlying equations) to reconstruct highly complex two-dimensional (2D) fields from a typical geophysical set-up, as the one of rotating turbulence. The first two methods are connected with a linear biased model reduction, the so-called Proper Orthogonal Decomposition (POD) and the third is based on a fully non-linear unbiased Convolutional Neural Network (CNN) embedding in a framework of Generative Adversarial Network (GAN) (Deng et al. 2019; Subramaniam et al. 2020; Buzzicotti et al. 2021; Kim et al. 2021; Guastoni et al. 2021; Yousif et al. 2022). POD is widely used for pattern recognition (Sirovich & Kirby 1987; Fukunaga 2013), optimization (Singh et al. 2001) and data assimilation (Romain et al. 2014; Suzuki 2014). To repair the missing data in a gappy field, Everson & Sirovich (1995) proposed 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 et al. (2006) showed that GPOD reconstruction outperforms the Kriging interpolation (Oliver & Webster 1990; Myers 2002; 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).

EPOD was first used in Maurel et al. (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 et al. 2008; Hosseini et al. 2016; Discetti et al. 2019). For example, using EPOD as a reference of their CNN models, Guastoni et al. (2021) predicted the 2D 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 et al. 2018). In fluid mechanics, CNN has also been shown as a promising technique for data reconstruction. Many researches devote 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 et al. 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 2D damaged snapshots of three-dimensional (3D) rotating turbulence with a GAN consisting of two CNNs, a generator and a discriminator. Previous preliminary researches indicate that the introduction of discriminator, the relative importance of which is called the adversarial ratio, significantly
improves the high-order statistics of the prediction (Deng et al. 2019; Subramaniam et al. 2020; Buzzicotti et al. 2021; Kim et al. 2021; Güemes et al. 2021). Here we extend the work (Buzzicotti et al. 2021) with a more refined statistical investigation of the GAN, including its dependency on the hyper-parameters, on the gap shape and more importantly, with a direct comparison against the POD-based reconstruction methods.

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 et al. 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) and (f)). A second long term goal would also be to systematically perform features ranking: 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 freedoms driving turbulent flow, connected also to control problems (Choi et al. 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 questions. In order to concentrate on two paradigmatic realistic set-ups, we study two gap geometries which are one central square gap (figure 1(a,d)) and random gappiness (figure 1(b,e)). 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 reconstructing task, for evaluation of the three methods on different situations.

The paper is organized as follows. Section 2.1 describes the database used to evaluate the reconstruction methods in this study. The GPOD, EPOD and GAN-based reconstruction methods are discussed in §2.2, §2.3 and §2.4, respectively. In §3, the performances of POD- and GAN-
based methods on 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. Finally, conclusions of the work are presented in §6.

2. Methodology

2.1. Dataset

For the evaluation of different reconstruction tools, we use incompressible forced homogeneous rotating turbulence with the following governing equations

\[ \nabla \cdot \mathbf{u} = 0, \]
\[ \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} + 2\mathbf{\Omega} \times \mathbf{u} = -\frac{1}{\rho} \nabla \bar{p} + \nu \nabla^2 \mathbf{u} + \mathbf{f}, \]

where \( \mathbf{u} \) is the velocity, \( \mathbf{\Omega} = \Omega \hat{\mathbf{x}}_3 \) is the system rotation vector, \( \bar{p} = p + \frac{1}{2}\rho ||\mathbf{\Omega} \times \mathbf{x}||^2 \) is the pressure in an inertial frame modified by a centrifugal term, \( \nu \) is the kinematic viscosity and \( \mathbf{f} \) is an external forcing mechanism.

The dataset is obtained from TURB-Rot (Biferale et al. 2020), an open database generated from direct numerical simulations (DNS) in a 3D (\( x_1 - x_2 - x_3 \)) periodic domain of size \( [0, 2\pi]^3 \) with 256 grid points in each direction. A Gaussian forcing mechanism is adopted in spectral space around \( k_f = 4 \) with a fixed energy input rate. The viscous term \( \nu \nabla^2 \mathbf{u} \) in equation (2.2) is replaced with a hyperviscous term \( \nu_h \nabla^4 \mathbf{u} \) to enlarge the inertial range. Moreover, a large scale friction term \( -\beta \mathbf{u} \) is used in r.h.s. to reach a stationary state with a Rossby number \( Ro = E_{tot}^{1/2} / k_f \mathbf{\Omega} \approx 0.1 \), where \( E_{tot} \) is the kinetic energy.

Although GPOD, EPOD and GAN are feasible to 3D data, in order to make contact with geophysical observation we restrict here to 2D horizontal slices in the \( x_1 - x_2 \) plane. Moreover, velocity is downsized from \( 256 \times 256 \) to \( 64 \times 64 \) by a spectral low-pass filter

\[ \tilde{\mathbf{u}}(x) = \sum_{||k||<k_\eta} \hat{\mathbf{u}}(k) e^{ik \cdot x}, \]

where \( k_\eta = 32 \) (see later for comments about limitations on the input data-size). In the present study we only reconstruct the velocity module, \( \mathbf{u} = ||\mathbf{u}|| \), and the dataset is divided as Train/Validation/Test split: 84480 (73.1%)/10560 (9.1%)/20480 (17.7%).

2.2. GPOD reconstruction

In this section we briefly present the procedure of GPOD. First, we use a training dataset, \( \{\mathbf{u}^c(x)\}_{c=1}^{N_{\text{train}}} \), to obtain the POD modes of the ensemble, \( \{\psi_n(x)\}_{n=1}^{N} \). POD modes are eigenvectors of the correlation matrix

\[ R(x,y) = \frac{1}{N_{\text{train}}} \sum_{c=1}^{N_{\text{train}}} \mathbf{u}^c(x)\mathbf{u}^c(y) \quad (x,y \in I). \]

Therefore, the number of POD modes \( N \) is limited by the size of the image. During the reconstruction stage, one can make a prediction inside the gap by writing the linear POD expansion:

\[ \mathbf{u}_{\text{pred}}(x) = \sum_{n=1}^{N} a_n \psi_n(x) \quad (x \in G), \]

where the POD coefficients, \( \{a_n\}_{n=1}^{N} \), are obtained from the least squares estimates with the available information in the area outside the gap, \( S \). Note that the number of measurements in \( S \)
is always smaller than $N$, which equals to the whole image size. Therefore, there does not exist a unique least squares estimate, resulting in infinite variance and the model cannot be used at all.

Two approaches are used to handle this problem. The first is dimension reduction (DR) as used in Everson & Sirovich (1995), where we only keep the first $N'$ terms in (2.5). The POD coefficients are computed by minimizing the following error

$$
\tilde{E} = \int_S \left[ u_{\text{true}}(x) - \sum_{n=1}^{N'} a_n \psi_n(x) \right]^2 \, dx.
$$

Further details are discussed in Appendix A, where the error analysis of the reconstruction is also conducted. The main result of this analysis is that although a larger $N'$ leads to a smaller truncation error in the POD expansion, there exists a trade-off against the number of POD coefficients that need to be computed from the available information supplied by the data in $S$. Hence, the optimal $N'$ has to be determined empirically from the training dataset for each specific gap.

The second approach is conducting the linear regression with regularization. Instead of only keeping the leading POD modes, all POD modes are used for optimization but the objective function is modified as

$$
\tilde{E}_{L_1} = \int_S \left[ u_{\text{true}}(x) - \sum_{n=1}^{N} a_n \psi_n(x) \right]^2 \, dx + \alpha \sum_{n=1}^{N} |a_n|.
$$

Here we add a Lasso (Tibshirani 1996) regularizer to make the $L_1$ norm of POD coefficients small. Lasso 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 $\alpha$ controls regularization strength and we estimate $\alpha$ by fivefold cross-validation (Efron & Tibshirani 1994) in this study. Results of GPOD with Lasso for a square gap are shown in Appendix B, where it also indicates that the GPOD with DR and with Lasso regularizer are comparable. Therefore, only the former is used for comparison with EPOD and GAN in the following of this paper.

### 2.3. EPOD reconstruction

To use EPOD for flow reconstruction, we first conduct POD analysis of the training dataset only on the known region by solving the eigenvalue problem of the correlation matrix

$$
R(x, y) = \frac{1}{N_{\text{train}}} \sum_{c=1}^{N_{\text{max}}} u^c(x) u^c(y) \quad (x, y \in S),
$$

where the corresponding eigenvalues and POD modes are $\{\lambda_n\}_{n=1}^{\tilde{N}}$ and $\{\phi_n(x)\}_{n=1}^{\tilde{N}}$, respectively. In this study, the number of POD modes $\tilde{N}$ is set as its maximum possible value, which equals to the number of points in $S$. Therefore, any instance of the training dataset can be expanded as

$$
u^c(x) = \sum_{n=1}^{\tilde{N}} b^c_n \phi_n(x) \quad (x \in S),
$$

where $\{b^c_n\}_{n=1}^{\tilde{N}}$ represents the corresponding POD coefficients. In the next step, the EPOD modes for the velocity module in the gap region are obtained as

$$
\phi_n^c(x) = \frac{1}{N_{\text{train}}} \sum_{c=1}^{N_{\text{max}}} b^c_n u^c(x) / \lambda_n \quad (x \in G).
$$
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.

Borée (2003) demonstrates that the only part of the velocity module inside the gap correlated with the one outside is predicted as

$$u_{\text{pred}}(x) = \sum_{n=1}^{N} b_n \phi_n^c(x) \quad (x \in G),$$

where we used the POD coefficients, $\{b_n\}_{n=1}^{N}$, obtained from the supplied data in $S$.

2.4. GAN-based reconstruction with Context Encoders

In a previous work, Buzzicotti et al. (2021) used a context encoder embedding in GAN (Goodfellow et al. 2014) to generate missing data in the same turbulent database but exploring only the case where the total gap size is fixed. In order to extend and generalize the previous approach to study gaps of different sizes, here we have started from a similar GAN architecture and added one layer at the start and two layers at the end of the generator and one layer at the start of the discriminator. The generator has an encoder-decoder architecture (Pathak et al. 2016). It first takes the damaged data and produces a latent feature representation by the encoder part. Then the decoder learns to generalize the latent features and generates candidates for the missing data. Figure 2 shows the architectures of the generator and discriminator. Each convolution (up-convolution) layer is followed by a Leaky Rectified Linear Unit (ReLU) activation function.

The generator is trained to minimize the following loss function:

$$L_{\text{GEN}} = (1 - \lambda_{\text{adv}}) L_{\text{MSE}} + \lambda_{\text{adv}} L_{\text{adv}},$$

which is a weighted summation of the mean squared loss ($L_2$ loss), $L_{\text{MSE}}$, and the adversarial loss, $L_{\text{adv}}$. The weight $\lambda_{\text{adv}}$ is called the adversarial ratio. The $L_2$ loss handles the continuity between generated data and the context, which is in favour of blurry predictions without high-
frequency information. It is important to stress that on the contrary of the GPOD case, here the supervised $L_2$ loss is calculated only inside the gap region $G$. At the same time, the adversarial loss helps the generator to produce predictions that are statistically similar to real turbulent configurations. Specifically, it minimizes the divergence between the generated and the true probability distributions (Goodfellow et al. 2014; Nowozin et al. 2016). Denoting the predicted and original data respectively as $u_{\text{pred}}$ and $u_{\text{true}}$, the mean squared error is defined as

$$\mathcal{L}_{\text{MSE}}(u_{\text{pred}}, u_{\text{true}}) = \frac{1}{m(G)} \int_G [u_{\text{pred}}(x) - u_{\text{true}}(x)]^2 \, dx,$$

(2.13)

where $m(G)$ is the number of points in $G$. The adversarial loss is provided by the discriminator output $D$, and it is defined as,

$$\mathcal{L}_{\text{adv}} = \log \left( 1 - D \left( u_{\text{pred}} \right) \right).$$

(2.14)

At the same time, the discriminator is trained such as to maximize the logistic loss based on the classification prediction of both real and generated samples,

$$\mathcal{L}_{\text{DIS}} = \log \left( D \left( u_{\text{true}} \right) \right) + \log \left( 1 - D \left( u_{\text{pred}} \right) \right).$$

(2.15)

More details about the architecture and training of the GAN are discussed in Appendix C.

3. Comparison between POD- and GAN-based reconstructions

To conduct a systematic comparison between POD- and GAN-based reconstructions, we start by assessing their performances on a square gap of various sizes located at the center of the $64 \times 64$ field. Model evaluation is carried out first for large-scale information and subsequently from a multi-scale perspective, based on mean squared errors (MSE) and turbulence statistics. Finally, we analyze the performance of predicting extreme events.

3.1. Large-scale information

In this section, the prediction of velocity module is quantitatively evaluated over the full test dataset. First we consider the reconstruction error and define the normalized MSE in the gap as

$$\text{MSE}_{\text{gap}}(u) = \frac{1}{E_u} \left( \frac{1}{m(G)} \int_G [u_{\text{pred}}(x) - u_{\text{true}}(x)]^2 \, dx \right).$$

(3.1)

where the notation

$$\langle \cdot \rangle = \frac{1}{N_{\text{test}}} \sum_{c=1}^{N_{\text{test}}} \langle \cdot \rangle$$

represents the average over the $N_{\text{test}} = 20480$ test planes and we define the normalization factor as

$$E_u = u_{\text{true}}^{\text{rms}} \cdot u_{\text{pred}}^{\text{rms}}.$$  

(3.2)

The $\text{rms}$ refers to root-mean-squared (r.m.s.) quantities in the gap, for example,

$$u_{\text{pred}}^{\text{rms}} = \left( \frac{1}{m(G)} \int_G u_{\text{pred}}^2(x) \, dx \right)^{1/2}.$$  

(3.3)

Note that we use $u_{\text{pred}}^{\text{rms}}$ in the denominator to make sure that a prediction with too small or too large energy gives a large MSE. To provide a baseline MSE, we make a new set of reconstructions by replacing the missing data with velocity fields sampled randomly from the test dataset. In other words the baseline gives the error obtainable when reconstructing with a statistically consistent field but not corresponding to the specific realization. The baseline gives a value around 0.5358,
Figure 3. The MSE of the reconstructed velocity module from GPOD, EPOD and GAN for a square gap with different sizes \( l \). The abscissa is normalized by the image size \( l_0 = 64 \).

Figure 4. PDF of the point-wise \( L_2 \) error obtained from GPOD, EPOD and GAN for a square gap with different sizes.

see Appendix D. Figure 3 shows the MSE of velocity module from GPOD, EPOD and GAN reconstructions for the square gap with different sizes, which are normalized by the image size \( l_0 = 64 \). The error bar represents the spread between the minimum and maximum \( L_2 \) error calculated on subsets of batch size equal to 128. 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 4 shows the PDF of point-wise \( L_2 \) error,

\[
\Delta u(x_1, x_2) = \frac{1}{E_u} \left[ u_{\text{pred}}(x_1, x_2) - u_{\text{true}}(x_1, x_2) \right]^2, \tag{3.4}
\]

for three different gap sizes \( l/l_0 = 24/64, 40/64 \) and \( 62/64 \). The PDFs are consistent with the results presented in figure 3. We can observe that the PDFs’ tails of the point-to-point error fluctuations are slightly reduced for the cases with smaller MSEs. To further study the capacity of the three tools, we plot the averaged point-wise \( L_2 \) error, \( \langle \Delta u(x_1, x_2) \rangle \), for a square gap of size \( l/l_0 = 40/64 \) in figure 5. Gappy produces large \( \langle \Delta u \rangle \) all over the gap, while EPOD and GAN behave better near the gap edge. Moreover, GAN generates smaller \( \langle \Delta u \rangle \) than EPOD in the inner area. 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 information on the multi-scale physics.
Figure 5. Averaged point-wise $L_2$ error obtained from (a) GPOD, (b) EPOD and (c) GAN for a square gap of size $l/l_0 = 40/64$, where the optimal $N' = 12$ is used in GPOD reconstruction. (d) Profiles of $\langle \Delta u \rangle$ along the red diagonal line shown in (a).

Figure 6. Reconstruction visualization of velocity module for GPOD, EPOD and GAN for a square gap of sizes $l/l_0 = 24/64$ (1st row), $40/64$ (2nd row) and $62/64$ (3rd row), where the optimal $N' = 20$, 12 and 12 are used in GPOD reconstruction, respectively. The last column plots generated and ground truth profiles along the vertical line shown in the first column.

of the reconstructed fields. Indeed, from figure 6 it is possible to see in a glimpse that the POD- and GAN-based reconstructions have a completely different multi-scale statistics which is not captured by the mean squared error. Figure 6 shows predictions of an instantaneous velocity 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 generate blurry predictions. Besides, there are also obvious discontinuities between context and the gap for the GPOD prediction. This is clearly due to the fact that minimizing the $L_2$ distance from ground truth can result in multi-minimal quasi-optimal solution where different ‘physics’ coexist, with almost the same ‘energetic’ contents but strongly different multi-scale properties. GPOD is unable to capture the complexity of this physical background and blatantly fail to deliver realistic flow conditions when the gap is big. EPOD only keeps the correlated part with the information outside the gap. Therefore, it can predict the large-scale coherent structures but is incapable of generating correct multi-scale properties. Specifically, when the gap size is extremely large, the number of EPOD modes is very small (equals to the number of points outside the gap), thus EPOD has a limited degree of freedoms to make realistic predictions. To quantify the above statements, we proceed with a statistical analysis of the reconstructed velocity for all methods, using the Jensen–Shannon (JS) divergence. The JS divergence is a measure of
the similarity between two probability distributions, which is defined as
\[
\text{JSD}(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M),
\]
(3.5)

where \( M = \frac{1}{2} (P + Q) \) and
\[
D(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \, dx
\]
(3.6)
is the Kullback–Leibler (KL) divergence. A small JS divergence indicates that the two probability distributions are close and vice versa. We use the base 2 logarithm and thus \( 0 \leq \text{JSD}(P \parallel Q) \leq 1 \). Figure 7 shows the JS divergence between PDFs of the velocity module for the ground truth and the two predictions measured inside the gap, \( \text{JSD}(u) = \text{JSD}((PDF(u_{\text{true}})) \parallel (PDF(u_{\text{pred}}))) \). We divide the test data into 10 batches of size 2048, calculate the mean JS divergence over these batches and we indicate with error bars its range of fluctuation. From this analysis we have found that GAN gives smaller JS divergence than GPOD and EPOD by an order of magnitude over almost the full range of gap sizes, indicating that the PDF of GAN prediction has a better correspondence to the ground truth. This is further shown in figure 8 where we present the PDFs of the reconstructed velocity fields for different gap sizes compared with the original data. Besides the wrong PDF shapes of GPOD and EPOD, we note that they are also predicting some negative values, nonphysical for a velocity module. The same problem is avoided by the GAN architecture, using a ReLU activation function 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 consider the analysis of the velocity gradients, \( \partial u / \partial x_1 \). Figure 9 plots MSE of the reconstructed gradients, in this case we can see that all methods produce errors much larger than in the velocity reconstruction. Moreover, GAN shows similar errors with GPOD at the largest gap size and with EPOD at small gap sizes. However, also in this case the MSE itself is not enough to describe the full picture. This can be easily understood by looking again at the visualization of the gradients reconstruction shown in figure 10. That GPOD and EPOD generate clearly much worse predictions than the GAN. The reason is that introducing more realistic
turbulent fluctuations the GAN is very sensitive to any small shifting between the predictions and the true solutions which can lead to large MSE. On the other hand even the GPOD or EPOD solution is inaccurate, having too small oscillations and it ends up with a similar MSE when compared with the GAN results. To evaluate the statistical quality of the three reconstructions we use again the JS divergence between gradients PDFs from original and generated data only inside the gap regions, as reported in figure 11. These results confirm that GAN is able to well predict the PDF of $\partial u / \partial x_1$ while GPOD and EPOD do not have this ability. Moreover, GPOD produces smaller JSD($\partial u / \partial x_1$) than EPOD. Figure 12 support further such conclusion. To compare the scale-by-scale energy budget of the reconstructed and the original solutions, in figure 13 we have measured the energy spectrum for the different fields, namely

$$E(k) = \sum_{k \leq |k| < k+1} \langle \hat{u}(k)\hat{u}^*(k) \rangle,$$  

where $k = (k_1, k_2)$ is the wave number, $\hat{u}(k)$ is the Fourier transform of velocity module and $\hat{u}^*(k)$ is its complex conjugate. To highlight the reconstruction performance as a function of the wave number, we also show the ratio between the reconstructed and the original spectra, $E(k)/E_{\text{true}}(k)$ for the three different gap sizes on the second row of the same figure 13.
Figure 10. Reconstruction visualization of velocity module gradient for GPOD, EPOD and GAN for a square gap of sizes \( l/l_0 = 24/64 \) (1st row), 40/64 (2nd row) and 62/64 (3rd row), where the optimal \( N' = 20, 12 \) and 12 are used in GPOD reconstruction, respectively. 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.

Figure 11. The JS divergence between PDFs of the velocity module gradient inside the reconstructed region calculated from the original data and the reconstruction from GPOD, EPOD and GAN for a square gap with different sizes.

compare higher-order statistical observables, in figure 14 we show the flatness of the different fields,

\[
F(r) = \frac{\langle (\delta_r u)^4 \rangle}{\langle (\delta_r u)^2 \rangle},
\]

where \( \delta_r u = u(x + r) - u(x) \) and \( r = (r, 0) \). The angle brackets in the definition of the flatness represents the average over the test dataset and over \( x \), for which \( x \) or \( x + r \) fall in the gap. Figures 13 and 14 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 fails at all scales for both the energy spectrum and flatness at gap sizes \( l/l_0 = 40/64 \) and 62/64.
Figure 12. PDF of the velocity module gradient in the reconstructed region obtained from GPOD, EPOD and GAN for a square gap with different sizes, where $\sigma(\partial u / \partial x_1)$ is the standard deviation of the original data.

Figure 13. Energy spectrum of the reconstructed velocity module over the whole region obtained from GPOD, EPOD and GAN for a square gap of different sizes (the 1st row). The corresponding $E(k)/E_{true}(k)$ is shown on the 2nd row, where $E(k)$ and $E_{true}(k)$ are the reconstructed and original spectra, respectively.

Figure 14. The flatness of the reconstructed velocity module obtained from GPOD, EPOD and GAN for a square gap of different sizes.
To evaluate the ability of the different techniques in reproducing correctly the extreme events missing in the damaged turbulent configurations, in figure 15 we present the scatter plots of the largest velocity and gradient values 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

\[ r = \langle 1 - |\sin \theta| \rangle, \]

where \( \theta \) is the angle between \( \mathbf{U} = (\max(x_{\text{pred}}), \max(x_{\text{true}})) \) and the unit vector \( \mathbf{e} = (1/\sqrt{2}, 1/\sqrt{2}) \), and

\[ |\sin \theta| = \frac{||\mathbf{U} \times \mathbf{e}||}{||\mathbf{U}||} = \frac{|\max(x_{\text{pred}}) - \max(x_{\text{true}})|}{\sqrt{2 \left[ \max(x_{\text{pred}})^2 + \max(x_{\text{true}})^2 \right]}}. \]

where \( x \) indicates generically either the velocity or the gradient values. It is obvious that \( r \in [0, 1] \) and \( r = 1 \) corresponds to a perfect prediction. In figure 15 it shows that GAN is the best and EPOD performs better than GPOD in terms of properly reproducing the correct maximum velocity module inside the gap. Moreover, GAN also outperforms the other two methods for the maximum velocity gradient.

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

As shown by the previous results, GAN is certainly superior whenever we ask questions concerning multi-scale reconstruction properties. It is natural to ask whether this supremacy comes from an ad-hoc balancing between the two adversarial machines or it is already embedded.
in the non-linear CNN structure of the generator only. In order to answer this question we have performed a systematic scanning of the GAN performances at changing the adversarial ratio \( \lambda_{adv} \), the hyper-parameter controlling the relative importance of \( L_2 \) and adversarial loss of the generator, as shown in equation (2.12). To study the effects of \( \lambda_{adv} \) on GAN reconstruction of turbulent data, we consider a central square gap of size \( l/l_0 = 40/64 \) and train the GAN with different adversarial ratios, where \( \lambda_{adv} = 10^{-4}, 10^{-3}, 10^{-2} \) and \( 10^{-1} \). Table 1 shows the MSE and JS divergence of the PDF for the velocity module as functions of the adversarial ratio. 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. PDFs of the GAN predicted velocity module with different adversarial ratios are compared with that of the original data in figure 16, which shows that the predicted PDF 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 PDF. In the limit of vanishing \( \lambda_{adv} \), results of GAN shows that the nonlinear CNN is still superior to GPOD but is inferior to EPOD in the absence of a critic discriminator. Therefore, there is still room for the improvement of the generator architecture and we can consider combining GAN with EPOD in the future work.

5. Dependency on gap geometry: random gappiness

Things change again when looking at a completely different topology of the damages. Here we study the case when pixels are removed randomly in the original domain \( I \), without any spatial
Figure 17. The MSE (left) and the JS divergence (right) between PDFs of the original and generated velocity modules inside the reconstructed region, obtained from GPOD, EPOD and GAN for random-pixel gap with different sizes.

Figure 18. PDF of the velocity module in the reconstructed region obtained from GPOD, EPOD and GAN for random-pixel gap with different sizes. Note that results of GPOD and EPOD are not physically correct because there exist negative tails for the positive velocity module.

correlations. We define the gap size of random gappiness as the square root of the damaged area. Because the random-pixel gap is easier for interpolating than a square gap of the same size, all reconstruction methods show good and comparable results in terms of MSE and PDFs of velocity module (figures 17 and 18). For almost all damaged densities, POD- and GAN-based methods give small MSE and JS divergence. However, when the total damaged region area is extremely large, GPOD and EPOD are not able to reconstruct the field at large wave numbers while GAN still works well because of the adversarial training. For extremely large random gappiness, figure 19 shows the reconstructed energy spectrum and figure 20 shows reconstructions of velocity module and velocity module gradient. It is obvious that GPOD and EPOD only predict the large-scale structure while GAN generates reconstructions with multi-scale information.

6. Conclusions

In this work, two linear POD-based approaches, GPOD and EPOD, are compared against a fully non-linear and unbiased convolutional network, based on a generative adversarial architecture, to reconstruct 2D damaged images taken from a database of rotating turbulent flows. Performances have been quantitatively judged on the the basis of (i) large-scale features based on the $L_2$ distance
between the ground truth and the reconstructed field and (ii) multi-scale properties given by the distance between the original and the reconstructed probability distribution functions of the local velocity module and of its spatial gradient. For one central square gap the GAN approach is proved to be superior, with small MSE and JS divergence, in particular for the large gap sizes where multi-scale features play a major role. GAN reconstructs the velocity field with better multi-scale properties in terms of the PDF of the velocity module gradient, energy spectrum and
flatness, as well as the prediction of the extreme events inside the missing area. In the presence of random damages, the three approaches give competent results except for the case of extreme gappiness where GAN is leading again. Furthermore, two GPOD approaches are compared, which use different approaches to determine the POD coefficients. The first one is keeping only the leading POD modes which is known as dimension reduction. An error analysis demonstrated that the optimal number of POD modes is controlled by the balance between different error components. As a result, for each specific gap, the optimal number of POD modes should be determined from a set of training data prior to the prediction step. The second approach uses a Lasso regularizer to select a subset of POD modes during the linear regression. Results show that both approaches provide a similar MSE and JS divergence. Dependency of GAN-based model on the adversarial ratio is also quantitatively studied. The adversarial ratio controls the compromise between having an ‘optimal’ point-wise reconstruction error and a realistic turbulent statistics also concerning extreme and rare events. Large adversarial ratios undermine the MSE while are good for generated statistical properties and vice versa. Our work is a first step toward the set-up of benchmarks and grand challenges for realistic ‘hard’ turbulent problem with interest for geophysical and laboratory applications, where the lack of measurements obstructs the capability to fully control the system. Many questions remain open, connected to the difficulty of having apriori estimates of the deepness and complexity of the GAN architecture as a function of the complexity of the physics, in particular concerning the amplitude of the missing data (here limited to medium-size gaps) and to the geometry (2D or 3D). As a result, little is known about the performance of the data-driven models as a function of the Reynolds or Rossby numbers, and/or on the possibility to supply physics information to help to further improve the network’s performances. Furthermore, problems such as remote sensing of velocity via pressure signals at the wall can be, in principle, attacked with the same tools.

Appendix A. GPOD (DR) reconstruction and error analysis

Figure 21 shows a schematic diagram of a gappy field. Denote $I$, $G$ and $S$ as point sets composed of points in the whole field, points inside and outside the gap region, respectively. Given $m(\cdot)$ as a function returning the number of points in a set, we can define

\[
I = \{x_1, x_2, \ldots, x_{m(I)}\},
G = \{x_k \mid x_k \text{ in the gap} = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{m(G)}}\},
S = I \setminus G = \{x_{j_1}, x_{j_2}, \ldots, x_{j_{m(S)}}\},
\]

(A1)
\[ x = \begin{bmatrix} x_1 & x_2 & \cdots & x_{m(I)} \end{bmatrix}^T, \]
\[ \bar{x} = \begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_{m(G)}} \end{bmatrix}^T, \]
\[ \tilde{x} = \begin{bmatrix} x_{j_1} & x_{j_2} & \cdots & x_{j_{m(S)}} \end{bmatrix}^T. \]  

(2)

Expand a full field \( u(x) \) on POD modes \( \{\psi(x)\}_{n=1}^{n=N} \),

\[
 u(x) = \sum_{n=1}^{N} a_n \psi_n(x) = \sum_{n=1}^{N} a_n \psi_n(x) + \sum_{n=N'+1}^{N} a_n \psi_n(x), \]  

which can be written as

\[
 u = Xa = X'a' + r', \]  

where \( N \) is the total number of POD modes and \( N' \) is the number of POD modes adopted for approximation. The definitions of \( X, X', a, a' \) and \( r' \) are shown below,

\[
 X = \begin{bmatrix} \psi_1(x) & \psi_2(x) & \cdots & \psi_N(x) \\ \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_N(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{m(I)}) & \psi_2(x_{m(I)}) & \cdots & \psi_N(x_{m(I)}) \end{bmatrix}, \]  

(5)

\[
 X' = \begin{bmatrix} \psi_1(x) & \psi_2(x) & \cdots & \psi_{N'}(x) \\ \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_{N'}(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_{N'}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_{m(I)}) & \psi_2(x_{m(I)}) & \cdots & \psi_{N'}(x_{m(I)}) \end{bmatrix}, \]  

(6)

\[
 a = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}^T, \]  

(7)

\[
 a' = \begin{bmatrix} a_1 & a_2 & \cdots & a_{N'} \end{bmatrix}^T, \]  

(8)

\[
 r' = r'(x) = \begin{bmatrix} \psi_{N'+1}(x) & \psi_{N'+2}(x) & \cdots & \psi_N(x) \end{bmatrix} \begin{bmatrix} a_{N'+1} & a_{N'+2} & \cdots & a_N \end{bmatrix}^T. \]  

(9)

Before we move on to GPOD reconstruction, we introduce two operators, \((\cdot)\) and \(\tilde{(\cdot)}\). For \( u = u(x) \),

\[
 \tilde{u} = u(\bar{x}), \quad \tilde{u} = u(\tilde{x}), \]  

(10)
and for $X$,

$$
\tilde{X} = \begin{bmatrix}
\psi_1(\tilde{x}) & \psi_2(\tilde{x}) & \cdots & \psi_N(\tilde{x}) \\
\psi_1(x_{i_1}) & \psi_2(x_{i_1}) & \cdots & \psi_N(x_{i_1}) \\
\psi_1(x_{i_2}) & \psi_2(x_{i_2}) & \cdots & \psi_N(x_{i_2}) \\
\vdots & \vdots & \ddots & \vdots \\
\psi_1(x_{i_{m(G)})} & \psi_2(x_{i_{m(G)})} & \cdots & \psi_N(x_{i_{m(G)})} \\
\end{bmatrix}
$$

(A11)

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

Everson & Sirovich (1995) minimize the error outside the gap,

$$
\tilde{E} = \|\tilde{u} - \tilde{X}'a'\| = \int_S \left| u(x) - \sum_{n=1}^{N'} a_n \psi_n(x) \right|^2 \, dx,
$$

(A12)

and achieve a best fit of $a'$ (Penrose 1956; Planitz 1979),

$$
a'_b = \arg \min_{a'} \tilde{E} = \tilde{X}'u + \left( I' - \tilde{X}'\tilde{X}' \right) w',
$$

(A13)

where $I' \in \mathbb{R}^{N' \times N'}$ is an identity matrix, $\tilde{X}'$ is the Moore-Penrose inverse of $\tilde{X}'$, and $w' \in \mathbb{R}^{N' \times 1}$ is an arbitrary vector. The reconstructed field in the gap is obtained from

$$
\tilde{u}_b = \tilde{X}'a'_b,
$$

(A14)

and the reconstruction error is

$$
\tilde{E} = \|\tilde{u} - \tilde{u}_b\| = \|\tilde{X}' \left( \left( I' - \tilde{X}'\tilde{X}' \right) \left( a' - w' \right) - \tilde{X}'\tilde{r}' \right)\| = \|\tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3\|,
$$

(A15)

where

$$
\tilde{e}_1 = \tilde{X}' \left( I' - \tilde{X}'\tilde{X}' \right) (a' - w'), \quad \tilde{e}_2 = -\tilde{X}'\tilde{X}'\tilde{r}', \quad \tilde{e}_3 = \tilde{r}'.
$$

(A16)

Equations (A15) and (A16) show that the quality of reconstruction depends on three terms, $\tilde{e}_1$, $\tilde{e}_2$ and $\tilde{e}_3$. We calculate their contributions to the reconstruction error

$$
\tilde{C}_1 = \|\tilde{e}_1\|, \quad \tilde{C}_2 = \|\tilde{e}_2\|, \quad \tilde{C}_3 = \|\tilde{e}_3\|.
$$

(A17)

For a square gap with $l/l_0 = 8/64$ and $40/64$, figure 22 shows $\langle \tilde{C}_1 \rangle$, $\langle \tilde{C}_2 \rangle$, $\langle \tilde{C}_3 \rangle$ and $\langle \tilde{E} \rangle$ as functions of $N'$, where $\langle \cdot \rangle = \frac{1}{N_{\text{test}}} \sum_{c=1}^{N_{\text{test}}} (\cdot)$ represents the average over $N_{\text{test}} = 20480$ test data.

Values are normalized by $m(G)E_u$, where $E_u$ is the normalization factor given in equation (3.2). We use $w' = 0$ to calculate $\langle \tilde{C}_1 \rangle$ and $\langle \tilde{E} \rangle$ without loss of generality.

Figure 22 shows that $\langle \tilde{C}_1 \rangle$ is always zero when $N'$ is smaller than a threshold, $N'_c$, because in equation (A16) $\tilde{X}'$ is invertible and thus $I' - \tilde{X}'\tilde{X}' = 0$. The arbitrariness of $w'$ only takes effect when $N'$ is larger than the threshold, in which case $\tilde{X}'$ is not invertible and $\langle \tilde{C}_1 \rangle$ is not zero. When $N'$ increases from zero, $\langle \tilde{C}_3 \rangle$ always decreases as it represents the truncation error of POD expansion, while $\langle \tilde{C}_2 \rangle$ increases at $N' < N'_c$ and decreases at $N' > N'_c$. From the trade-off between different error components, we can obtain an optimal $N' = N'_{\text{opt}}$ with the smallest
Appendix B. Comparison between DR and Lasso for GPOD reconstruction

The two GPOD methods introduced in §2.2 are compared for a square gap of size $l/l_0 = 40/64$. Figure 23 (left) shows the PDF of the estimated value of $\alpha$ over the test data for Lasso regression. Table 2 shows that the GPOD reconstruction with DR and Lasso give similar MSE and JS divergence of the PDF for the velocity module. Figure 23 (right) also shows that the reconstructed PDFs of the velocity module are comparable. The difference between DR and Lasso is illustrated by the reconstruction sample in figure 24. It shows that DR can better capture the large-scale structure while Lasso provides more multi-scale structures. Figure 24 also shows
Square gap, $l/l_0 = 40/64$

![PDF graph](image)

**Figure 23.** PDF of the estimated value of $\alpha$ over the test data for Lasso regression (left) and PDFs of the velocity module in the reconstructed region obtained from GPOD reconstruction with DR and Lasso (right) for a square gap of size $l/l_0 = 40/64$.

### Table 2

|           | $\text{MSE}_{\text{gap}}(u)$ | $\text{JSD}(u)$               |
|-----------|-------------------------------|--------------------------------|
| DR        | $0.1683(-0.0153, +0.0123)$    | $0.04769(-0.00067, +0.00080)$ |
| Lasso     | $0.1708(-0.0165, +0.0168)$    | $0.04854(-0.00110, +0.00063)$ |

**Table 2.** The MSE and the JS divergence between PDFs of the original and generated velocity modules inside the reconstructed region, obtained from GPOD reconstruction with DR and Lasso for a square gap of size $l/l_0 = 40/64$.

![Reconstruction visualization](image)

**Figure 24.** Reconstruction visualization of velocity module (the 1st row) and velocity module gradient (the 2nd row) for the two GPOD approaches for a square gap of size $l/l_0 = 40/64$, where the optimal $N' = 12$ is used for DR. The last column plots the spectra of reconstructed POD coefficients obtained from DR and Lasso for the flow configuration shown on the left.

The corresponding spectra of POD coefficients obtained from DR and Lasso, where DR gives a nonzero spectrum up to $N' = 12$ and Lasso selects both large- and small-scale modes.

### Appendix C

This appendix contains details of the architecture (figure 2) and training of the GAN. For square gap with different sizes $l = 8, 16, 24, 32, 40, 50, 60$ and $62$, we use different kernel sizes of the last layer of generator and the first layer of discriminator, $k = 8, 4, 18, 2, 25, 15, 5$ and
The adversarial loss as a function of epoch (left) and PDFs of the original data and the predicted data at different epochs (right). Results are obtained from the training process of the GAN for a square gap of size $l/l_0 = 40/64$.

For random gappiness, we use $l = 64$ and $k = 1$ but the $L_2$ loss is only computed in the gap. To generate positive output, ReLU is adopted as the activation function for the last layer of generator. A hyper-parameter $\alpha = 0.2$ is chosen for the leaky ReLU activation function of other convolution (up-convolution) layers. We use the adversarial ratio $\lambda_{adv} = 10^{-2}$ for a central square gap and $\lambda_{adv} = 10^{-3}$ for random gappiness as they give good compromise between MSE and reconstructed turbulent statistics.

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 25 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 25(a)), while the predicted PDF gets closer to the ground truth (figure 25(b)). This indicates the training convergence.

Appendix D.

To simplify the notation, we denote $x$ as the ground truth, $u_{true}$ (or $\partial u_{true}/\partial x_1$), and $y$ as the generated data, $u_{pred}$ (or $\partial u_{pred}/\partial x_1$). The average over gap region and the test data is indicated by

$$\langle \cdot \rangle = \frac{1}{N_{\text{test}}} \sum_{c=1}^{N_{\text{test}}} \frac{1}{m(G)} \int_G \langle \cdot \rangle \, dx.$$  \hfill (D1)

Equation (3.1) can be rewritten as below,

$$\text{MSE}_{\text{gap}} = \frac{\langle (x - y)^2 \rangle}{\sqrt{\langle x^2 \rangle} \sqrt{\langle y^2 \rangle}} = \frac{\langle x^2 \rangle - 2\langle xy \rangle + \langle y^2 \rangle}{\sqrt{\langle x^2 \rangle} \sqrt{\langle y^2 \rangle}}.$$  \hfill (D2)

Since the generated data is randomly sampled from the original one, they are completely uncorrelated thus $\langle xy \rangle = \langle x \rangle \langle y \rangle$. Besides, $x$ and $y$ have the same statistics, indicating that $\langle x \rangle = \langle y \rangle$ and $\langle x^2 \rangle = \langle y^2 \rangle$. Therefore, one can obtain that

$$\text{MSE}_{\text{gap}} = \frac{2(\langle x^2 \rangle - \langle x \rangle^2)}{\langle x^2 \rangle} = 2 \left( 1 - \frac{\langle x \rangle^2}{\langle x^2 \rangle} \right).$$  \hfill (D3)
Therefore, we have the estimate $\text{MSE}_{\text{gap}}(u) \approx 0.5358$ from the mean value and the mean energy of the velocity module. For the velocity module gradient, we have $\langle \chi \rangle = 0$ from the periodicity and thus $\text{MSE}_{\text{gap}}(\partial u/\partial x_1) \approx 2$.

REFERENCES

ALEXAKIS, ALEXANDROS & BIFERALE, LUCA 2018 Cascades and transitions in turbulent flows. Physics Reports 767, 1–101.

ASCH, MARK, BOCQUET, MARC & NODET, MAELLE 2016 Data assimilation: methods, algorithms, and applications. SIAM.

BARAL, CHITTA, FUENTES, OLC & KREINOVIICH, VLADIK 2018 Why deep neural networks: a possible theoretical explanation. In Constraint programming and decision making: theory and applications, pp. 1–5. Springer.

BELL, MICHAEL J, LEFEVBRE, MICHEL, LE TRAON, PIERRE-YVES, SMITH, NEVILLE & MILMER-BECKER, KIRSTEN 2009 Godae: the global ocean data assimilation experiment. Oceanography 22 (3), 14–21.

BIFERALE, LUCA, Bonaccorso, fabio, Buzzicotti, Michele & di LeonI, P CLARK 2020 Turb-rot: a large database of 3d and 2d snapshots from turbulent rotating flows. arXiv preprint arXiv:2006.07469.

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

Brunton, Steven L & NOACK, Bernd R 2015 Closed-loop turbulence control: Progress and challenges. Applied Mechanics Reviews 67 (5).

Buzzicotti, Michele, Bonaccorso, Fabio, Di LeonI, P CLARK & BIFERALE, LUCA 2021 Reconstruction of turbulent data with deep generative models for semantic inpainting from turb-rot database. Physical Review Fluids 6 (5), 050503.

Buzzicotti, Michele, CLARK Di LeonI, PATRICIO & BIFERALE, LUCA 2018 On the inverse energy transfer in rotating turbulence. The European Physical Journal E 41 (11), 1–8.

Choi, Haecheon, MoIn, Parviz & Kim, John 1994 Active turbulence control for drag reduction in wall-bounded flows. Journal of Fluid Mechanics 262, 75–110.

Deng, Zhiwen, He, Chuangxin, Liu, Yingzheng & Kim, Kyung Chun 2019 Super-resolution reconstruction of turbulent velocity fields using a generative adversarial network-based artificial intelligence framework. Physics of Fluids 31 (12), 125111.

Di LeonI, P CLARK, ALEXAKIS, ALEXANDROS, BIFERALE, L & Buzzicotti, M 2020 Phase transitions and flux-loop metastable states in rotating turbulence. Physical Review Fluids 5 (10), 104603.

Discetti, Stefano, Bellani, Gabriele, Örlü, Ramis, Serpieri, Jacopo, Vila, Carlos SANMIGUEL, RAIOLA, Marco, ZENG, XIAOBO, Mascotelli, Lucia, Talamelli, Alessandro & IANIRO, Andrea 2019 Characterization of very-large-scale motions in high-re pipe flows. Experimental Thermal and Fluid Science 104, 1–8.

Efron, Bradley & Tibshirani, Robert J 1994 An introduction to the bootstrap. CRC press.

Everson, Richard & Sirovich, Lawrence 1995 Karhunen–loeve procedure for gappy data. JOSA A 12 (8), 1657–1664.

Fahlund, Georg, Stroh, Alexander, Frohnapfel, Bettina, Atzori, Marco, Vinuesa, Ricardo, Schlatter, Philipp & GATTI, Davide 2021 Investigation of blowing and suction for turbulent flow control on airfoils. AIAA Journal 59 (11), 4422–4436.

Fukami, KAI, Fukagata, Kaji & Taira, Kunihiko 2021 Machine-learning-based spatio-temporal super resolution reconstruction of turbulent flows. Journal of Fluid Mechanics 909.

Fukunaga, Keinosuke 2013 Introduction to statistical pattern recognition. Elsevier.

Garcia, Damien 2011 A fast all-in-one method for automated post-processing of piv data. Experiments in fluids 50 (5), 1247–1259.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron & Bengio, Yoshua 2014 Generative adversarial nets. Advances in neural information processing systems 27.

Guastoni, Luca, güemes, alejandro, IANIRO, Andrea, Discetti, Stefano, Schlatter, Philipp, Azizpour, Hossein & Vinuesa, Ricardo 2021 Convolutional-network models to predict wall-bounded turbulence from wall quantities. Journal of Fluid Mechanics 928. Güemes, Alejandro, Discetti, Stefano, IANIRO, Andrea, Sirmacek, Beril, Azizpour, Hossein &
Vinuesa, Ricardo 2021 From coarse wall measurements to turbulent velocity fields through deep learning. *Physics of Fluids* 33 (7), 075121.

Gunes, Hasan & Rist, Ulrich 2008 On the use of kriging for enhanced data reconstruction in a separated transitional flat-plate boundary layer. *Physics of Fluids* 20 (10), 104109.

Gunes, Hasan, Sirisup, Sirod & Karniadakis, George Em 2006 Gappy data: To krig or not to krig? *Journal of Computational Physics* 212 (1), 358–382.

Gad-el-Hak, Mohamed & Tsai, Her Mann 2006 Transition and turbulence control, , vol. 8. World Scientific.

He, Kai-Ming, Zhang, Xiangyu, Ren, Shaoqing & Sun, Jian 2016 Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hornik, Kurt 1991 Approximation capabilities of multilayer feedforward networks. *Neural networks* 4 (2), 251–257.

Hosseini, Zahra, Martinuzzi, Robert J & Noack, Bernd R 2016 Modal energy flow analysis of a highly modulated wake behind a wall-mounted pyramid. *Journal of Fluid Mechanics* 798, 717–750.

Kim, Hyojin, Kim, Junhyuk, Won, Sungjin & Lee, Changhoon 2021 Unsupervised deep learning for super-resolution reconstruction of turbulence. *Journal of Fluid Mechanics* 910.

Kingma, Diederik P & Ba, Jimmy 2014 Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kreinovich, Vladik Ya 1991 Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem. *Neural networks* 4 (3), 381–383.

Krysta, Monika, Blayo, Eric, Cosme, Emmanuel & Verron, Jacques 2011 A consistent hybrid variational-smoothing data assimilation method: Application to a simple shallow-water model of the turbulent midlatitude ocean. *Monthly weather review* 139 (11), 3333–3347.

Le Dimet, Françoise-Xavier & Talagrand, Olivier 1986 Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography* 38 (2), 97–110.

Lee, Changhoon, Kim, John, Babcock, David & Goodman, Rodney 1997 Application of neural networks to turbulence control for drag reduction. *Physics of Fluids* 9 (6), 1740–1747.

Li, Tianyi, Buzzicotti, Michele, Biferale, Luca, Wan, Minping & Chen, Shiyi 2021 Reconstruction of turbulent data with gappy pod method. *Chinese Journal of Theoretical and Applied Mechanics* 53 (10), 2703–2711.

Liu, Bo, Tang, Jiupeng, Huang, Haibo & Lu, Xi-Yun 2020 Deep learning methods for super-resolution reconstruction of turbulent flows. *Physics of Fluids* 32 (2), 025105.

Maurel, S, Borée, J & Lumley, JL 2001 Extended proper orthogonal decomposition: Application to jet/vortex interaction. *Flow, Turbulence and Combustion* 67 (2), 125–136.

Militino, Ana F, Ugarte, MD & Montesino, M 2019 Filling missing data and smoothing altered data in satellite imagery with a spatial functional procedure. *Stochastic Environmental Research and Risk Assessment* 33 (10), 1737–1750.

Myers, DE 2002 Interpolation of spatial data: some theory for kriging.

Ni, Xiaoxiao & Suen, Ching Y 2012 A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition* 45 (4), 1318–1325.

Nowozin, Sebastian, Cseke, Botond & Tomioka, Ryota 2016 f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems* 29.

Oliver, Margaret A & Webster, Richard 1990 Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System* 4 (3), 313–332.

Pathak, Deepak, Krahenbuhl, Philippe, Donahue, Jeff, Darrell, Trevor & Efros, Alexei A 2016 Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544.

Penrose, Roger 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. *The Mathematical Gazette* 63 (425), 181–185.

Romain, Leroux, Chatellier, Ludovic & David, Laurent 2014 Bayesian inference applied to spatio-temporal reconstruction of flows around a naca0012 airfoil. *Experiments in fluids* 55 (4), 1–19.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang,
ZHENG, KARPATHY, ANDREI, KHOSLA, ADITYA, BERNSTEIN, MICHAEL & OTHERS 2015 Imagenet large scale visual recognition challenge. International journal of computer vision 115 (3), 211–252.

SHEN, HUANFENG, LI, XINGHUA, CHENG, QING, ZENG, CHAO, YANG, GANG, LI, HUIFANG & ZHANG, LIANGPEI 2015 Missing information reconstruction of remote sensing data: A technical review. IEEE Geoscience and Remote Sensing Magazine 3 (3), 61–85.

SINGH, SAIHENDRA N, MYATT, JAMES H, ADDINGTON, GREGORY A, BANDA, SIVA & HALL, JAMES K 2001 Optimal feedback control of vortex shedding using proper orthogonal decomposition models. J. Fluids Eng. 123 (3), 612–618.

SIROVICH, LAWRENCE & KIRBY, MICHAEL 1987 Low-dimensional procedure for the characterization of human faces. Josa a 4 (3), 519–524.

SUBRAMANIAM, AKSHAY, WONG, MAN LONG, BORKER, RAUNAK D, NIMMAGADDA, SRAYYA & LELE, SANJIVA K 2020 Turbulence enrichment using physics-informed generative adversarial networks. arXiv preprint arXiv:2003.01907 .

SUZUKI, TAKAO 2014 Pod-based reduced-order hybrid simulation using the data-driven transfer function with time-resolved ptv feedback. Experiments in fluids 55 (8), 1–17.

TIBSHIRANI, ROBERT 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 (1), 267–288.

TINNEY, CE, UKEILEY, LS & GLAUSER, MARK N 2008 Low-dimensional characteristics of a transonic jet. part 2. estimate and far-field prediction. Journal of Fluid Mechanics 615, 53–92.

TORN, RYAN D & HAKIM, GREGORY J 2009 Ensemble data assimilation applied to rainex observations of hurricane katrina (2005). Monthly weather review 137 (9), 2817–2829.

VENTURI, DANIELE & KARNAIADKIS, GEORGE Em 2004 Gappy data and reconstruction procedures for flow past a cylinder. Journal of Fluid Mechanics 519, 315–336.

WANG, CHENGYUE, GAO, Qi, WANG, HONGPING, WEI, RUNJIE, LI, TIAN & WANG, JINJUN 2016 Divergence-free smoothing for volumetric piv data. Experiments in Fluids 57 (1), 15.

WEN, XIN, LI, ZIYAN, PENG, Di, ZHOU, WENWU & LIU, YINGZHENG 2019 Missing data recovery using data fusion of incomplete complementary data sets: A particle image velocimetry application. Physics of Fluids 31 (2), 025105.

YOUSEF, MUSTAFA Z, YU, LINQI, HOYAS, SERGIO, VINUESA, RICARDO & LIM, HEECHANG 2022 A deep-learning approach for reconstructing 3d turbulent flows from 2d observation data. arXiv preprint arXiv:2208.05754 .

ZHANG, QUAN, YUAN, QIANGQIANG, ZENG, CHAO, LI, XINGHUA & WEI, YANCONG 2018 Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing 56 (8), 4274–4288.