Robust flow reconstruction from limited measurements via sparse representation

Jared L. Callaham[1,*] Kazuki Maeda,[2] and Steven L. Brunton[2]
[1]*Department of Applied Mathematics, University of Washington, Seattle, Washington 98195, USA*
[2]*Department of Mechanical Engineering, University of Washington, Seattle, Washington 98195, USA*

(Received 13 February 2019; published 30 October 2019)

In many applications it is important to estimate a fluid flow field from limited and possibly corrupt measurements. Current methods in flow estimation often use least squares regression to reconstruct the flow field, finding the minimum-energy solution that is consistent with the measured data. However, this approach may be prone to overfitting and sensitive to noise. To address these challenges we instead seek a sparse representation of the data in a library of examples. Sparse representation has been widely used for image recognition and reconstruction, and it is well-suited to structured data with limited, corrupt measurements. We explore sparse representation for flow reconstruction on a variety of fluid data sets with a wide range of complexity, including vortex shedding past a cylinder at low Reynolds number, a mixing layer, and two geophysical flows. In addition, we compare several measurement strategies and consider various types of noise and corruption over a range of intensities. We find that sparse representation has considerably improved the estimation accuracy and robustness to noise and corruption compared with least squares methods. We also introduce a sparse estimation procedure on local spatial patches for complex multiscale flows that preclude a global sparse representation. Based on these results, sparse representation is a promising framework for extracting useful information from complex flow fields with realistic measurements.

DOI: 10.1103/PhysRevFluids.4.103907

I. INTRODUCTION

Estimating the structure of a flow field from limited and noisy measurements is an important challenge in many engineering applications. For example, accurate estimation is central to active flow control [1–4], which has the potential to advance next-generation technology, ranging from fuel-efficient, low-drag automobiles [5] to high-efficiency turbines [6] and internal combustion engines [7]. The ability to reconstruct important flow features from restricted observations is also critical in applications as diverse as cardiac blood-flow modeling [8,9], ship wake identification [10], and climate science [11]. All of these applications rely on estimating the structure of complex fluid flows based on limited measurements. This work focuses on addressing this challenge by using techniques from machine learning and sparse representation [12], which have recently been applied to flow-field classification [13–16].

Modern experimental methods and the increasing scale and resolution of numerical simulations have led to an abundance of fluid-flow data. Although we are able to achieve unprecedented fidelity in measurement and simulation in laboratory settings, in applications we are typically limited to a few noisy sensors. The challenge in flow-field estimation is thus to synthesize the profusion of offline data and limited, unreliable online information. This synthesis relies on learning and

*jc244@uw.edu

representing the essential structure of the flow field by leveraging the physical behavior observed in past data. Fluid mechanics is not unique in having a wealth of data, however, and recent years have seen the rapid development of revolutionary machine learning techniques to leverage big data, particularly in image processing [17,18]. Since flow-field data are often discretized on a grid, many of these techniques can be applied to fluid mechanics with only minor modifications; for example, for flow-field classification [13] and estimation [19].

A common model-free approach to flow-field reconstruction is to represent the field as a linear combination of modes in a library, such as empirical eigenfunctions from proper orthogonal decomposition (POD) [20,21] or dynamic mode decomposition (DMD) [22–25]. Gappy POD was introduced by Everson and Sirovich [26] to repair corrupted or missing data and was adapted to flow-field reconstruction by Bui-Thanh *et al.* [27]. This method has been used to reconstruct unsteady flow fields around a cylinder [28] and an airfoil [29], arterial blood-flow data [8], and low-dimensional ocean velocity and temperature fields [30]. Podvin *et al.* [31] used a similar method to estimate POD coefficients for three-dimensional (3D) cavity flow from two-dimensional (2D) particle image velocimetry (PIV) data. Other leading methods for flow-field estimation include stochastic estimation and model-based observers, which are discussed in more detail in Sec. II. However, the majority of these approaches are based on least-squares regression, which may be prone to overfitting and sensitive to noise. Furthermore, although these methods minimize the kinetic-energy deviation between the predicted and actual measurements, this does not guarantee that the reconstruction will be globally optimal.

Inspired by the work of Wright *et al.* [12] on sparse representation for image recognition, we propose searching for a sparse representation in a library of example flow fields rather than a minimum-energy solution in the modal library, as shown schematically in Fig. 1. If the flow is statistically stationary and the library is sufficiently extensive, it may be possible to identify a sparse combination of the few most similar fields in the library that are consistent with the measurements. If such a sparse representation is available, this approach corresponds to searching in prior data for recurring coherent structures, which may be nonlinearly correlated with observations. Moreover, sparsity-promoting techniques are known to prevent overfitting and provide robustness to noisy and corrupt measurements, which are essential for flow-field estimation. Sparse representation has been previously applied to classify flow regimes in a library of POD modes based on limited sensors in the seminal work of Bright, Lin, and Kutz [13], and sparse flow classification has been extended in related work [14–16]. Our work builds on this framework, extending regime classification to full flow-field reconstruction and demonstrating sparse representation in a library of training examples instead of a modal basis. We show that, when the flows are more complex than previously studied, or the measurements are corrupted, this method significantly outperforms reconstruction with a POD library. Since highly complex flows may not have a sparse representation in terms of POD modes, this result reinforces the importance of sparsity for robustness to noise and accuracy of reconstruction. We extensively explore this method on several example systems of increasing complexity with various levels of measurement noise, including numerical and geophysical data sets and find that it produces more accurate and robust reconstructions than standard least-squares methods. Flow-field estimation continues to be an important and difficult challenge, but improvements in robustness and accuracy are consistent with previous work in sparse regression and highlight the value of sparsity in reconstruction methods.

The remainder of this work is organized as follows: Section II provides an overview of related work on flow-field estimation. In Sec. III, we describe the proposed method for flow-field reconstruction based on sparse representation. The four flow configurations used to test this method are described in detail in Sec. IV. Section V explores sparse reconstruction for flow-field reconstruction on these examples with various sampling strategies from corrupted measurements, demonstrating that sparse representation exhibits improved robustness and accuracy compared with least-squares estimation. These results include careful benchmark comparisons on four fluid flows, including two canonical flows and two geophysical flows: periodic vortex shedding past a cylinder at $\mathrm{Re} = 100$ (Sec. V A), a mixing layer at $\mathrm{Re} = 720$ (Sec. V B), global sea surface temperature fields

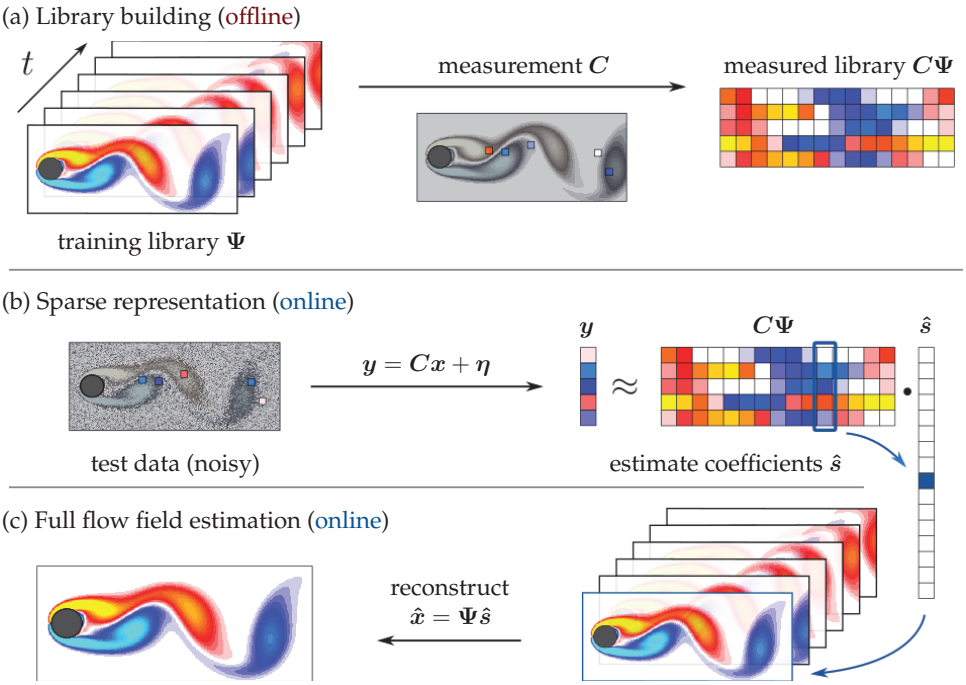

FIG. 1. Flow-field reconstruction process using sparse representation. (a) In offline library building the measurement operator $\boldsymbol{C}$ is applied to the training set $\boldsymbol{\Psi}$. (b) The sparse representation step solves the relaxed convex optimization problem (7) to estimate sparse coefficients $\hat{\boldsymbol{s}}$ which are consistent with the noisy measurements $\boldsymbol{y}$. (c) Finally, the full flow field is reconstructed as a linear combination of the training examples. The reconstructed field shown in panel (c) is the actual output of the sparse representation algorithm with the noisy test data and measurements shown in panel (b). Flow past a cylinder is discussed further in Sec. V A.

(Sec. V C), and data from a Gulf of Mexico ocean model (Sec. V D). On the mixing layer and Gulf of Mexico ocean data, we demonstrate that, when a globally sparse representation is not available, the flow field can be more accurately estimated with a superposition of local reconstructions and show that this improvement is facilitated by enhanced sparsity of the local representations. In Sec. VI we summarize the main results and discuss limitations of the method, and we conclude with Sec. VII. To promote reproducible research, all code is available online.[1]

II. PRIOR WORK IN FLOW-FIELD RECONSTRUCTION

Because of its far-reaching applications, flow-field estimation is a rich field with seminal works spanning the past half century. Here, we provide a brief overview of some of the most relevant related works, which are organized into three broad groups: stochastic estimation, model-based observers, and library-based reconstruction. Each of these methodologies approaches the problem with a slightly different motivation, but many modern studies are driven by the overarching goals of estimation and control.

Stochastic estimation (SE) was introduced by Adrian [32] to study coherent structures in turbulence. SE is a statistical technique that estimates a quantity of interest in the flow as a conditional average given the measurements. Expanding the conditional average in a Taylor series

[1] github.com/jcallaham/robust-flow-reconstruction

and minimizing the mean-square estimation error yields a functional dependence between the observation and flow-field variables determined by unconditional statistics, such as the two-point correlation tensor. SE has been extended to the estimation of POD coefficients [33], spectral coefficients [34,35], and inclusion of time-delayed measurements [36,37]. The stochastic estimation method has been used to study isotropic turbulence [32,38,39], turbulent boundary layers [40,41], axisymmetric jets [33,35], the backwards-facing step [42–44], open cavities [45,46], and for feedback in closed-loop control of flow separation over an airfoil [47].

In another approach to flow-field estimation, an observer dynamical system is used to evolve the estimate of the system state according to a reduced-order model while measurements provide feedback used to improve the estimate. The model may be linear; for example, based on dynamic mode decomposition (DMD) [22–25], with an estimate maintained by Kalman filtering [16,48,49]. The model could also be nonlinear, based on a Galerkin projection of the Navier-Stokes equations onto a set of POD modes [50–52], or the result of model identification [53,54]. Recent work has investigated the use of data assimilation techniques (e.g., particle filters or ensemble Kalman filters) to estimate the mean flow [55–57] or the full flow field [58–62]. In any case, the accuracy of observer-based methods depends on the quality of the reduced-order model, so there is inevitably a trade-off between low-latency and high-accuracy models. There are a number of excellent reviews of modal decomposition and model reduction in fluids [63,64].

A third category of model-free flow-field estimation takes advantage of large offline data sets via library-based reconstruction. Often the flow field will be discretized and reshaped into a high-dimensional vector, which is then approximated by a linear combination of modes in a library [64]. The modes may be generic (e.g., a Fourier or wavelet basis) or tailored to the particular flow (such as POD or DMD modes), which each have advantages for sensor-based flow reconstruction [65]. Advanced data-driven algorithms such as K-SVD [66,67] and GOBAL [68] may also be used. Gappy POD [26] is a popular library-based method, where the library consists of POD modes and coefficients are estimated by least-squares regression based on limited or masked data [27,29,46]. Podvin *et al.* [31] used gappy POD to estimate POD coefficients for a 3D flow past a cavity from 2D PIV data, choosing the number of measurements to equal the number of modes, resulting in the inversion of a square matrix. In a related approach, Yu and Hesthaven [19] use deep learning to estimate POD coefficients, while Fukami *et al.* and Erichson *et al.* have both studied direct estimation of flow fields with neural networks [69,70].

Linear dimensionality reduction techniques, such as POD and DMD, have well-known limitations, including the inability to efficiently capture symmetries in the data, such as traveling waves or rotating structures [71,72]. An alternative approach involves identifying a *nonlinear* coordinate transformation onto a low-dimensional manifold, for example using an autoencoder neural network with nonlinear activation functions [73]. With advances in deep learning, these nonlinear embedding approaches are becoming more mainstream [74–78]. More generally, deep learning is a powerful emerging technique to represent multiscale flow structure and model turbulence closure [79–82], although it generally requires tremendous amounts of training data and may be prone to overfitting unless care is taken to constrain the models with known physics. In a sense, deep learning may be considered a sophisticated nonlinear interpolation scheme that leverages a large library of historical examples [83]. However, linear dimensionality reduction is still widely used, in part because of its simple formulation in terms of linear algebra, which enables fast computations and methodological extensions [72].

The estimation and reconstruction algorithms described above are generally based on ℓ_2 optimization, which suffers from the same limitations as standard least-squares parameter estimation; sparsity-promoting methods have emerged as a principled way to address these shortcomings by regularizing the regression [72,84–86]. Sparse representation in a library takes advantage of known structure in the data and is robust to measurement corruption [12]. If the coefficient vector of the modal representation is sparse in the sense that it has relatively few nonzero entries, the coefficients can be recovered from surprisingly few measurements with efficient tools, such as matching pursuit [87–90] or by ℓ_1 minimization of the coefficient vector [91,92], under certain assumptions. If the

library consists of generic modes, such as a discrete cosine transform (DCT) basis, then recovery based on ℓ_1 minimization is known as compressed sensing (CS) [93–95]. Compressed sensing has been used in fluid mechanics to reconstruct a signal in a linear-duct acoustic problem [96], identify dominant frequencies in low-dimensional projections of sub-Nyquist rate PIV data [97], and to find a compact representation for wall-bounded turbulence [98]. Although these results are promising, general flows are often *not sparse enough* to take advantage of compressed sensing, requiring prohibitively many measurements and expensive computations that do not scale well.

The sparsifying library does not need to be universal, however. Sparse representation in a data-driven POD basis was used to classify the Reynolds number for flow past a cylinder [13,15]. Bai *et al.* [99] similarly demonstrated CS in a POD library to reconstruct PIV data. Wright *et al.* [12] proposed a straightforward alternative to modal libraries, such as DCT and POD, in the sparse representation for classification (SRC) algorithm for facial recognition. In SRC, an image of an individual is downsampled and approximated with a sparse representation in terms of a library consisting of *the training data itself*, which contains some example images of the same person. The coefficients corresponding to the test individual will naturally be of greater magnitude, indicating the identity of the subject. SRC is robust against noise, corruption, or occlusion of the image, and has been applied for early diagnosis of Alzheimer's disease [100], segmentation of MRI images [101], automatic detection and classification of brain tumors [102], music genre categorization [103], and dolphin whistle classification [104]. Although SRC was introduced for classification, the success of sparse representation in a library of the training data has far reaching applications, including for flow-field reconstruction, as will be explored here. A similar idea has been used in the field of analog weather forecasting, where identifying the most similar past conditions (i.e., finding a sparse representation of current conditions in terms of past conditions) provides information that enables future predictions [105–107].

III. SPARSE REPRESENTATION OF A FLOW FIELD IN A LIBRARY

In this work, we investigate the utility of sparse representation for flow-field reconstruction in a library of historical flow-field data, exploring robustness and scaling with flow complexity. This section provides the methodological foundations for the results that follow. We describe the general library-based signal recovery framework in Sec. III A, including reconstruction from sparse representation. In Sec. III B we introduce our method for sparse-representation-based flow-field reconstruction.

A. Library-based signal recovery

Here we provide the general problem statement and notation for sensor-based reconstruction of a high-dimensional state in a library. Given a discretized state vector $\boldsymbol{x} \in \mathbb{R}^n$, for example representing the fluid velocity or vorticity field at a set of grid points, and linear measurements $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x}$, with $\boldsymbol{y} \in \mathbb{R}^p$ and $p \ll n$, we seek an estimate $\hat{\boldsymbol{x}}$ of the full signal. We assume that the state $\boldsymbol{x}$ can be accurately expressed as a linear combination of library elements $\{\boldsymbol{\psi}_j\}$, $j = 1, 2, \ldots, r$ with $\boldsymbol{\psi}_j \in \mathbb{R}^n$, so that

$$\boldsymbol{x} \approx \boldsymbol{\Psi}\boldsymbol{s}, \tag{1}$$

for some coefficient vector $\boldsymbol{s} \in \mathbb{R}^r$, where columns in the library $\boldsymbol{\Psi} \in \mathbb{R}^{n\times r}$ are the vectors $\boldsymbol{\psi}_j$. The reconstruction problem reduces to estimating the coefficients $\hat{\boldsymbol{s}}$ that satisfy

$$\boldsymbol{y} \approx \boldsymbol{C}\boldsymbol{\Psi}\hat{\boldsymbol{s}}. \tag{2}$$

In other words, we seek an estimate that produces measurements $\hat{\boldsymbol{y}} = \boldsymbol{C}\boldsymbol{\Psi}\hat{\boldsymbol{s}}$ consistent with actual observations $\boldsymbol{y}$. As described earlier, the library $\boldsymbol{\Psi}$ may comprise a modal basis, such as Fourier, wavelets, POD, or DMD modes, or it may be chosen to contain examples of flow fields from training data. We also explore reconstruction for different classes of measurement matrix $\boldsymbol{C}$, although it is also possible to tailor this matrix for a given library $\boldsymbol{\Psi}$ for improved reconstruction [65].

In practice, solving for $\hat{\boldsymbol{s}}$ in Eq. (2) must be formulated as an optimization problem, since $\boldsymbol{C\Psi}$ is not typically a square matrix. For instance, in the overdetermined case where $p > r$ and there are more measurements than library elements, we may choose to solve for the least-squares solution

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{y} - \boldsymbol{C\Psi s}\|_2. \tag{3}$$

It is generally useful to modify the least-squares regression by adding a regularization term to prevent overfitting and promote robustness to noise and outliers in the data:

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{y} - \boldsymbol{C\Psi s}\|_2 + \lambda \|\boldsymbol{s}\|_q. \tag{4}$$

A choice of $q = 2$, corresponding to Tikhonov or ridge regression, penalizes high-variance solutions. For instance, Buffoni [52] employed this regularization to estimate POD coefficients in a nonlinear observer. A choice of $q = 1$ (LASSO regression) promotes a sparse representation [108]. The regularization parameter λ can be tuned to adjust the strength of this term. Other choices of q are possible, but $q = 1$ and $q = 2$ are the most common since they can be solved with convex optimization [109], which scales well to large problems.

For high-dimensional and multiscale data, it is often the case that there are fewer available measurements than modes in the library, leading to an underdetermined problem with $p < r$. In this case the appropriate optimization problem is

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_q \text{ subject to } \boldsymbol{y} = \boldsymbol{C\Psi s}. \tag{5}$$

Again, $q = 2$ leads to the minimum-energy solution consistent with measured data, while $q = 1$ leads to a sparse representation in the library $\boldsymbol{\Psi}$.

If the coefficient vector $\boldsymbol{s}$ is known to be sparse, and it is assumed to have exactly K nonzero elements (i.e., the vector $\boldsymbol{s}$ is *K-sparse*), this problem can be formulated as

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{y} - \boldsymbol{C\Psi s}\|_2 \text{ subject to } \|\boldsymbol{s}\|_0 = K, \tag{6}$$

where $\|\boldsymbol{s}\|_0$ is the number of nonzero entries of $\boldsymbol{s}$. Although an estimate of the sparsity K may not generally be available, there are many efficient algorithms such as OMP [87–89] or CoSaMP [90] that can solve this problem more efficiently than Eq. (5).

The representation problem in Eq. (5) can be applied to noisy sensor measurements $\boldsymbol{y}$. In this case, the measurements are $\mathbf{y} = \mathbf{Cx} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a noise vector. The optimization problem then relaxes the equality constraint:

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_q \text{ subject to } \|\boldsymbol{y} - \boldsymbol{C\Psi s}\|_2 < \epsilon, \tag{7}$$

where ϵ is an error tolerance. If the measurements have independent and identically distributed Gaussian noise [i.e., $\boldsymbol{\eta} \in \mathbb{R}^p$ with $\eta_i \sim \mathcal{N}(0, \sigma)$], ϵ may be chosen as a multiple of the total noise $\sigma\sqrt{p}$. When noise is introduced artificially, we nondimensionalize the noise level σ by the rms fluctuations of the field variable in the training set.

The relaxation in Eq. (7) assumes that σ is relatively small compared with typical fluctuations in the measurement vector $\boldsymbol{y}$. Wright *et al.* [12] describe a method for sparse representation, with $q = 1$, to handle sparse corruption with large amplitude, where some unknown fraction ρ of random entries in $\boldsymbol{y}$ suffer from uniformly distributed corruption over the full range of observed values. That is, the measurement $\boldsymbol{y}$ is now $\boldsymbol{y} = \boldsymbol{Cx} + \boldsymbol{e}$, where $\boldsymbol{e}$ has ρp nonzero entries. The optimization problem is extended to also identify $\boldsymbol{e}$ by direct minimization of its ℓ_1 norm. If the signal is additionally corrupted by dense low-amplitude noise as in Eq. (7), the problem becomes

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s},\boldsymbol{e}} \|\boldsymbol{s}\|_1 + \|\boldsymbol{e}\|_1 \text{ subject to } \left\| \begin{bmatrix} \boldsymbol{C\Psi} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{e} \end{bmatrix} - \boldsymbol{y} \right\|_2 < \epsilon. \tag{8}$$

We find, consistent with their results, that although the corruption $\boldsymbol{e}$ must be sparse in the sense that it has enough zero entries to enable identification via minimization of its ℓ_1 norm, it can

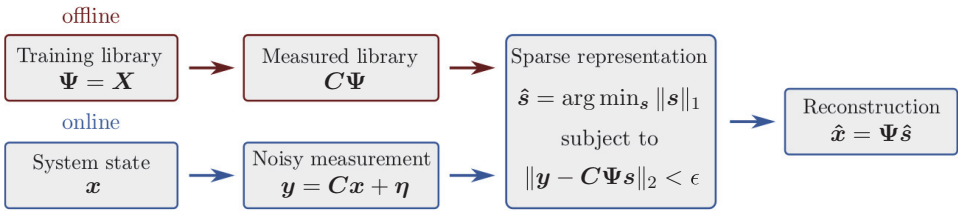

FIG. 2. Schematic of the sparse representation method shown graphically in Fig. 1. After constructing the library $\boldsymbol{C\Psi}$ in an offline step, the sparse coefficients $\hat{\boldsymbol{s}}$ consistent with measurements $\boldsymbol{y}$ are estimated. The full flow field can be reconstructed as a linear combination of the training examples in $\boldsymbol{\Psi}$.

actually consist of a substantial fraction of total measurements (see, e.g., Fig. 5). We demonstrate reconstruction from noisy measurements in Secs. V A and V B, but we relax the optimization problem even when noise is not explicitly added. Since Eq. (1) will generally not be exact, the relaxation ϵ plays a role similar to that of the regularization parameter λ in Eq. (4); larger relaxations ϵ allow sparser estimates $\hat{\mathbf{s}}$ which still satisfy the constraint in Eq. (7).

As mentioned earlier, there are many choices for the library of modes, and some may be more natural depending on the application. Fourier modes or wavelets may be useful in audio or image compression, and empirical POD modes are often used in fluids. To construct "tailored" libraries, such as POD modes, it is necessary to have a training set $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ of flow fields that contains representative examples. Such a training set can be obtained from simulations or experiments. Other libraries may be designed to be optimal in another sense. For instance, the K-SVD algorithm [66,67] iteratively constructs a library in which data should have a representation with some prescribed sparsity. GOBAL [68] is a similar method that enforces observability of the library modes. Wright *et al.* [12] simply use the training set as the library, so that the columns of $\boldsymbol{\Psi}$ are the prior observations.

This reconstruction framework enables many choices for the library and the optimization formulation. For example, gappy POD [26] solves the least-squares problem (3) with a library of POD modes. To identify the high-energy structures in the flow field and ensure that the problem remains underdetermined, the library of POD modes can be truncated. This may be done automatically, for instance with the hard threshold of Gavish and Donoho [110], although it is not clear in general what level of truncation is optimal for flow reconstruction.

B. Flow-field reconstruction from sparse representation

We now adapt the library-based reconstruction framework for fluid flow-field reconstruction. The procedure is shown schematically in Figs. 1 and 2. The signal $\boldsymbol{x} \in \mathbb{R}^n$ is the full discretized flow field and the measurement operator $\boldsymbol{C}$ still relates measurements $\boldsymbol{y}$ to the full field by $\boldsymbol{y} = \boldsymbol{Cx}$. Following the work of Wright *et al.* [12] in image analysis, we assume that the flow field has a sparse representation in a library of training examples $\boldsymbol{\Psi}$, as opposed to POD modes. That is, $\boldsymbol{x} = \boldsymbol{\Psi s}$ for some $\boldsymbol{s}$ with $\|\boldsymbol{s}\|_0 = K \ll n$. The coefficient vector $\boldsymbol{s}$ can be estimated by using one of the optimizations in Eqs. (5)–(8).

To improve the performance of this method for systems in fluid mechanics, we make several modifications. First, the empirical mean of the training set may be subtracted from all data. This is effective in cases where the mean represents a significant fraction of the energy in the data; for instance in sea surface temperature fields (Sec. V C). Second, the amplitudes of the reconstructed flow fields can be rescaled so that the total energy of the flow is equal to that computed from the training data. This rescaling is useful for reconstruction from noisy measurements; flow fields reconstructed with Eq. (7) tend to have lower amplitudes than the true field, since high levels of noise can be consistent with qualitatively accurate fields of reduced amplitude. In this work we only use this modification in Sec. V A.

Third, we develop a method of localized reconstruction for complex fluid flows. In the reconstruction process above, it is assumed that the test field is a simple linear combination of global fields in the library. However, for flows with coherent structures at multiple spatial scales, it may be prohibitively expensive to collect enough data to have representative examples of all likely global flow fields. Said another way, for multiscale flows, it is difficult to collect enough data for the library to converge to a statistical stationary distribution. Fortunately, if we decompose the global domain into local patches, each patch may be much lower rank, enabling a local sparse representation. To facilitate localized reconstruction, we introduce local kernels $\boldsymbol{\Phi}_j$, $j = 1, 2, \ldots, k$ that restrict the measurement $\boldsymbol{y}_j = \boldsymbol{C}\boldsymbol{\Phi}_j\boldsymbol{x}$ and reconstruction to the jth local region:

$$\hat{\boldsymbol{s}}_j = \arg\min_{\boldsymbol{s}_j} \|\boldsymbol{s}_j\|_q \text{ subject to } \|\boldsymbol{y}_j - \boldsymbol{C}\boldsymbol{\Phi}_j\boldsymbol{\Psi}\boldsymbol{s}_j\|_2 < \epsilon. \tag{9}$$

These decoupled optimization problems lead to compact local estimates $\hat{\boldsymbol{x}}_j = \boldsymbol{\Phi}_j\boldsymbol{\Psi}\hat{\boldsymbol{s}}_j$, with a full state estimate, $\hat{\boldsymbol{x}} = \sum_j \hat{\boldsymbol{x}}_j$, that is globally valid.

Finally, we define a metric to compare the quality of various reconstructions. The normalized root-mean-square residual of the difference of the reconstruction $\hat{\boldsymbol{x}}$ and the test field $\boldsymbol{x}$ is

$$\text{error} = \frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2}{\|\boldsymbol{x}\|_2}. \tag{10}$$

If the empirical mean $\bar{\boldsymbol{x}}$ is subtracted from both $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, then the appropriate metric is

$$\text{error} = \frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2}{\|\boldsymbol{x} + \bar{\boldsymbol{x}}\|_2}. \tag{11}$$

At this point, it is important to summarize some of the main assumptions that sparse representation relies on. First, as with all reconstruction methods based on a tailored library, we assume that the flow is statistically stationary, so that a sufficiently large library based on training data can generalize to future states. We then assume that the training set is comprehensive enough that the observed states are well-approximated by a linear combination of library elements. This is a related requirement, but while the former is a property of the flow, the latter is a property of the training data itself. All reconstruction methods further rely on the measurements containing sufficient information to accurately identify the coefficients $\hat{\boldsymbol{s}}$. Just as reconstruction is impossible, even with relatively dense sampling, for a flow state that is essentially orthogonal to the library, we cannot realistically expect to accurately reconstruct a turbulent channel flow from one point measurement, no matter how extensive the library. Finally, sparse representation assumes that a flow field of interest may be expressed as a linear combination of a small number of other flow fields in the training library. This assumption holds for simple flows, such as periodic vortex shedding at low Reynolds number, but may be unjustified for complex, multiscale flows unless a staggering amount of data is available. We examine the implications of these assumptions in the following sections.

1. Algorithm

The complete flow-field reconstruction based on sparse representation is as follows:

(1) Compute the library $\boldsymbol{\Psi} \in \mathbb{R}^{n\times r}$. The library may be given by the unmodified training data $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, although we also investigate reconstruction using POD modes and a K-SVD library. Optionally, the empirical mean flow field $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ may be subtracted from $\boldsymbol{X}$.

(2) Take measurements $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{\eta}$ of the flow field $\boldsymbol{x}$ by using the measurement operator $\boldsymbol{C} \in \mathbb{R}^{p\times n}$ with noise $\boldsymbol{\eta}$. For the examples below, the measurement matrix $\boldsymbol{C}$ consists of rows of the identity matrix corresponding to measured locations in the discretized field.

(3) Solve the appropriate optimization problem in Eqs. (5)–(8) for the coefficient vector $\hat{\boldsymbol{s}}$. If the coefficients are found by minimizing the ℓ_1 norm, we refer to this as sparse representation.

(4) Reconstruct the estimated flow field with $\hat{\boldsymbol{x}} = \boldsymbol{\Psi}\hat{\boldsymbol{s}}$. Optionally, rescale the estimated field to have the same variance as the training fields; this is helpful for very noisy measurements.

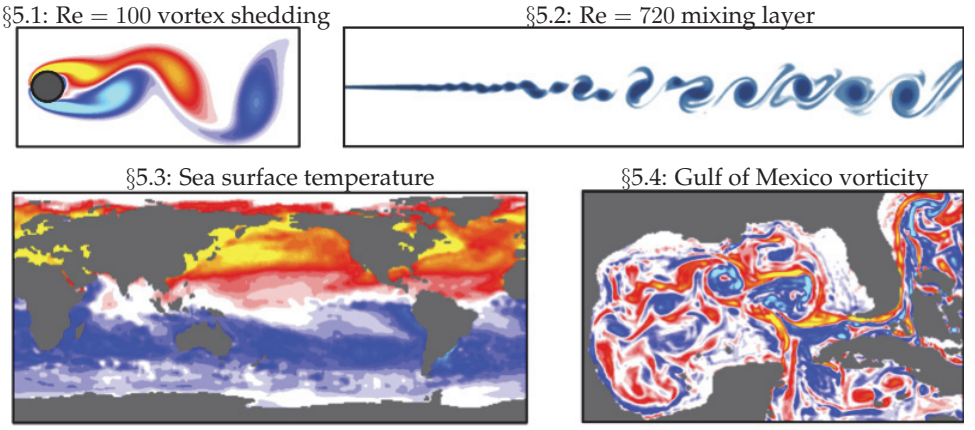

FIG. 3. Example flow fields from the data sets which we investigate with the sparse representation-based reconstruction method. We study two canonical flows (periodic vortex shedding past a cylinder at Re = 100 and a mixing layer at Re = 720) and two geophysical data sets (sea surface temperature and Gulf of Mexico vorticity fields).

For dictionary learning with K-SVD, we use KSVD-BOX v13. We solve the pursuit problem in Eq. (6) with OMP-BOX v10 [111]. To solve the convex optimization problems (7)–(9), we use the CVX Matlab package [112,113]. The complexity of sparse approximation grows with the number of measurements and the number of modes in the dictionary but does not directly depend on the number of points in the original discretized field.

IV. FLOW CONFIGURATIONS

Here we describe the fluid flows explored in this work and the methods used to obtain the data. We apply flow-field reconstruction to four data sets of increasing complexity, shown in Fig. 3: vortex shedding past a cylinder at Re = 100, a mixing layer at Re = 720, observations of sea surface temperature, and sea surface vorticity in the Gulf of Mexico.

A. Periodic vortex shedding

The first test case is given by the two-dimensional fluid flow past a circular cylinder at Reynolds number 100, which is characterized by periodic, laminar vortex shedding. This flow is a canonical benchmark system, although it is considerably simpler than most flows of practical interest.

Our data were generated from direct numerical simulation of the incompressible Navier-Stokes equations using the immersed boundary projection method [114,115]. The computational domain consists of four nested grids with the smallest grid covering a domain of 9×4 cylinder diameters and the largest grid covering a domain of 72×32 diameters. The resolution for each grid is 450×200 (50 points per cylinder diameter) and the simulation uses a time step of $\Delta t = 0.02$ time units that are nondimensionalized by the free-stream velocity and the cylinder diameter. We collect 151 post-transient snapshots, corresponding to five periods of vortex shedding, with each snapshot separated by $10\Delta t$. The Reynolds number for this flow, based on the cylinder diameter and free-stream velocity, is $\mathrm{Re} = DU_\infty/\nu = 100$, where D is the diameter, U_∞ is the free-stream velocity, and ν is the kinematic viscosity. The training set consists of the first 32 snapshots, which spans one full period. We analyze the vorticity field, although the method could be applied to velocity, pressure, scalar concentrations, or any other field variables of interest. The mean vorticity field is included in visualizations, although analyses and error calculations are performed after subtracting the empirical mean of the training data.

B. Mixing layer

As a more complex example, we consider a two-dimensional, compressible mixing layer at Reynolds number

$$\mathrm{Re} = \frac{\Delta U\, \delta}{\nu} = 720,$$

where δ is the initial vorticity thickness, ΔU is the velocity difference across the layer, and ν is the kinematic viscosity. Stanley and Sarkar [116] showed that two-dimensional numerical simulations in this regime reproduce the flow structures observed in three-dimensional experiments.

We generated this data set by direct numerical simulation of the compressible Navier-Stokes equations using a finite-volume, fifth-order WENO scheme [117]. The spatial coordinates are normalized by the vorticity thickness at the inlet. The velocities are normalized by the speed of sound of the fluid far from the mixing region. The Mach numbers of the high- and low-stream velocity are 0.5 and 0.25, respectively. The computational domain is $x \in [0, 800]$ and $y \in [-200, 200]$. The flow is forced at the inflow boundary at its most unstable fundamental frequency, and its subharmonic. Nonreflective boundary conditions are implemented on the other boundaries. The grid is smoothly stretched away from the mixing region to the nonreflective boundaries to prevent contamination by reflections. The grid in the mixing region is uniform with $\Delta x = 0.08$ and $\Delta y = 0.02$, respectively. After removing the transient portion of the simulation, we collect 2400 snapshots of the vorticity field in a window of $(x, y) \in (0, 128) \times (-12, 12)$, separated by nondimensional time steps of $\Delta t = 0.5827$. We compute the normal vorticity from these velocity fields both for ease of visualization and for the importance of vorticity in identifying dynamically significant coherent structures [118].

Forcing at the inlet excites instability waves, which roll up into vortices and convect downstream. These vortices pair and eventually merge into successively larger vortices. This process contributes to the linear growth of the mixing layer [119], and at higher Reynolds number, the turbulent mixing layer is dominated by the linear growth of the coherent structures [120]. These structures play a significant role in mixing, transport, and entrainment in turbulent shear flows [118]. Therefore, we study this laminar mixing layer as a representative case to assess the ability of sparse representation to generalize to other shear flow configurations.

C. Sea surface temperature field

Real-world data are rarely as well-behaved as the numerical solutions of canonical flows. However, it is these flows, with limited training data, multiscale dynamics, and unmodeled coupling to external systems, where the ability to infer the structure of the field would be most useful. To this end we explore reconstruction methods with the NOAA Optimum Interpolation Sea Surface Temperature (SST) V2 data set. Due to seasonal fluctuations, the SST field exhibits strongly periodic structure, although complex ocean dynamics still lead to rich flow phenomena. Flow data are available on a weekly basis on a one degree grid, and it is produced by combining local and satellite temperature observations. We use all available data at the time of analysis (1914 weeks, spanning October 1981 to June 2018). Our training set consists of the first 20 years of data (1040 weeks, spanning 1981–2001). We calculate a long-term annual mean field and subtract the mean for all analyses, since the mean accounts for the majority of the spatial structure of the field and is therefore uninformative with respect to the performance of reconstruction methods.

D. Gulf of Mexico surface vorticity

Finally, we consider the Gulf of Mexico surface velocity estimates from the HYbrid Coordinate Ocean Model (HYCOM) group. This data-assimilative model synthesizes remotely sensed and *in situ* measurements on a hybrid coordinate system. We use daily $1/12.5^\circ$-resolution data from 1992–2018 (9268 snapshots, combined from HYCOM experiment numbers 19.0, 19.1, 90.9, 91.1, and

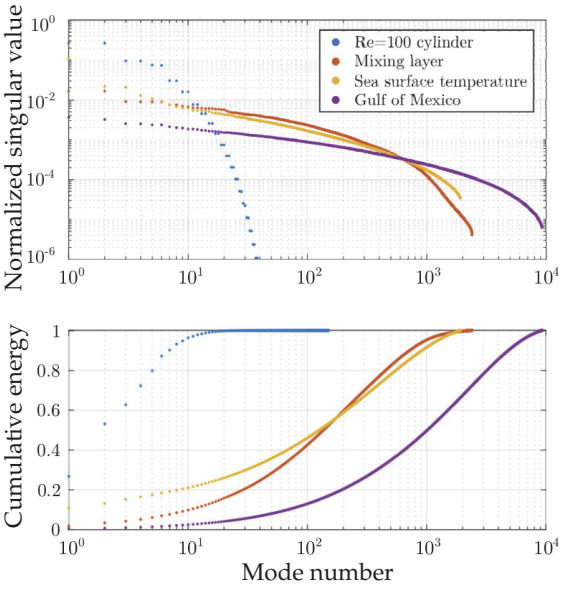

FIG. 4. Singular value spectra for the flows studied in this work. Each is normalized by the sum of singular values for that flow. The normalized cumulative sum of the singular values (bottom) represents the energy captured in the dominant POD modes. The rate of convergence gives an indication of the complexity of the flow. The singular values for vortex shedding past a cylinder (blue) converge quickly, whereas the Gulf of Mexico vorticity data (purple) has a long tail. The sea surface temperature (yellow) and mixing layer vorticity (red) are of intermediate complexity.

91.2). The training set consists of 8341 snapshots, or approximately 90% of the total data, with the remainder withheld for independent validation. As with the mixing layer and cylinder, we compute vorticity from 2D velocity measurements, although the methods are readily applied to any quantity of interest. We analyze the fluctuating vorticity fields relative to the empirical mean of the training set, but include the mean flow in visualizations.

Figure 4 shows the singular value spectra, equivalent to the POD eigenspectra, of the four data sets. This offers a rough comparison of the complexity of the flows. The low-dimensional dynamics of the flow behind a cylinder is clear from the sharp decay of singular values; most of the energy is contained within the first twenty POD modes. On the other hand, the spectrum for the Gulf of Mexico data converges slowly, indicating complex multiscale dynamics. The difficulties with this data set are intuitive: by restricting our view to the Gulf of Mexico we study a flow with an unmodeled coupling to a much larger chaotic system. The mixing layer and sea surface temperature data exhibit intermediate complexity. For the mixing layer, the flow near the inlet is approximately periodic, and the behavior becomes more complex as the flow evolves downstream. Similarly, the SST fields show strong seasonal fluctuations with perturbations.

V. RESULTS

We now investigate flow-field reconstruction from limited measurements using the four data sets described in Sec. IV, which span a range of physical scales and complexity. For the two numerically generated flows, we study the impact of measurement noise on reconstruction accuracy and find improved robustness with sparse representation. In the mixing layer, we demonstrate the advantages of sparse representation by analyzing the flow in windows to enable higher levels of sparsity. Finally, we demonstrate the proposed method on two geophysical data sets: global sea

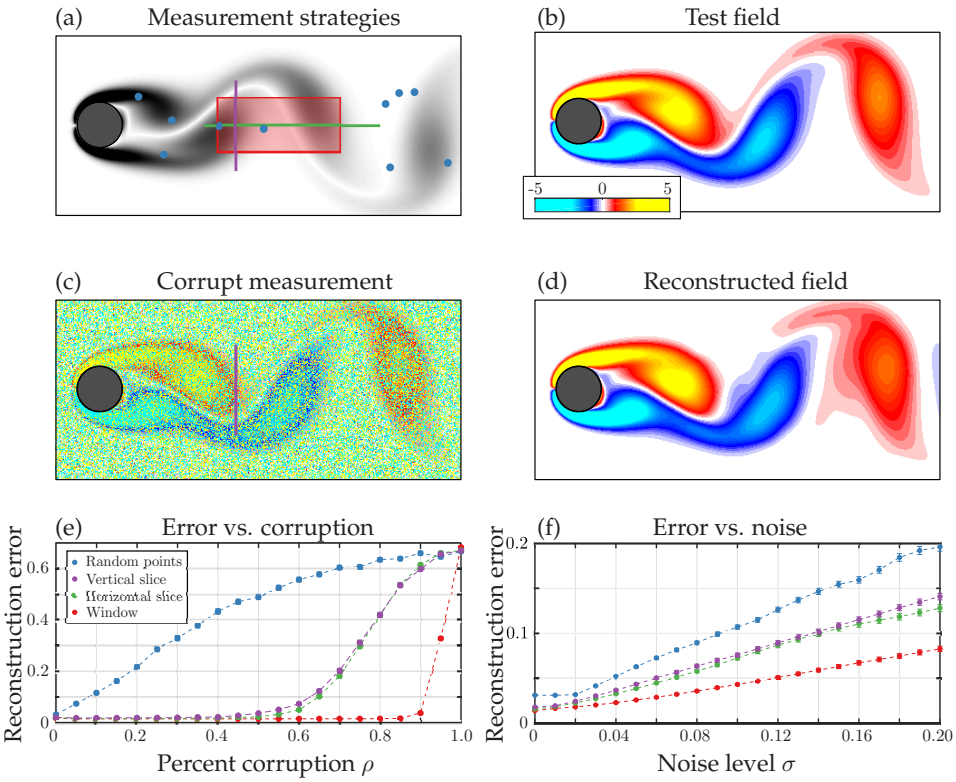

FIG. 5. Sparse reconstruction in a library of training data accurately recovers the flow past a cylinder in the presence of noise and corruption using a variety of measurement strategies. (a) Illustration of different measurements: random points (blue), vertical and horizontal slices (purple and green, respectively), and a window (red). (b) Flow field from the test data set. (c) Example flow snapshot with corruption in 70% of grid locations. The vertical stripe shows the measurement location. (d) Reconstructed flow field from sparse representation using corrupted measurements shown in panel (c). (e) Normalized reconstruction error with increasing percentage of grossly corrupted grid points (see Sec. III A). Colors correspond to the measurements in panel (a) and error bars show standard error as obtained by simulating 10 different realizations of the Gaussian noise for the 119 test fields. (f) Normalized reconstruction error with increasing levels of dense, normally distributed noise. In all cases, more measurements result in better performance. A contour plot of reconstruction errors is shown in Fig. 18 in Appendix C.

surface temperature and Gulf of Mexico surface vorticity. In all cases, we compare the performance of sparse representation to other library-based methods, including gappy POD.

A. Periodic vortex shedding

Figure 5 demonstrates reconstruction of the flow past a cylinder from various measurements with increasing levels of noise and corruption. In particular, we consider sparse representation in a library of the training data and find that this flow can be accurately reconstructed, even in the presence of significant noise. We also investigate various measurement strategies, including random point measurements, a "window" inspired by PIV-type measurement, and continuous "slices" in both vertical and horizontal orientations. For example, Fig. 5 demonstrates recovery of the entire field using Eq. (8) from a cross-stream slice measurement where 70% of the measured points are corrupted by replacing the observed value with a uniform random value on the range of observed vorticity. In all cases, the measurement strategies with larger numbers of observations (e.g., the red

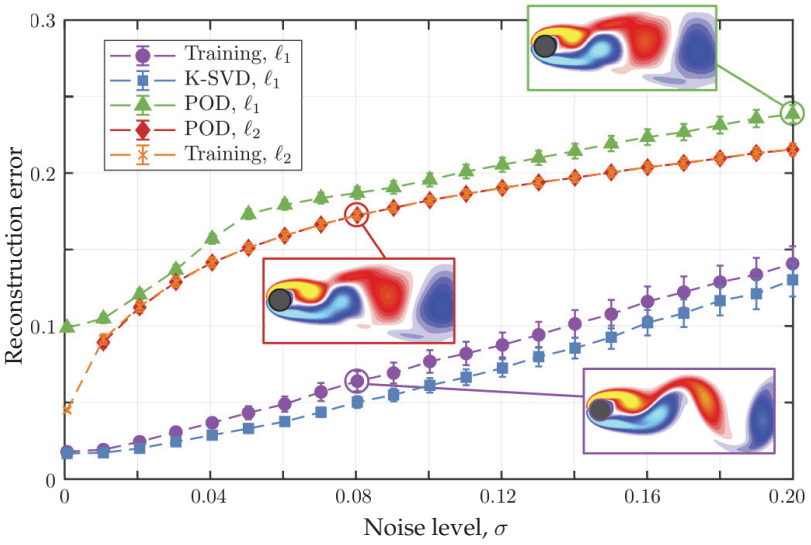

FIG. 6. Comparison of reconstruction with different libraries and norms from a noisy vertical measurement slice (Fig. 5, purple). The horizontal axis is the level of dense Gaussian noise, the vertical axis is the normalized residual error in the reconstruction (10), and the error bars indicate standard error on the mean residual. Insets show typical examples of reconstruction. Over this range of noise, sparse representation with the training library (red) shows an average 35% improvement over the more standard gappy POD (purple).

window) exhibit more robust reconstruction performance. In addition, for corrupt measurements, there is a *phase change* observed at a critical corruption density, consistent with the wider sparse representation literature. It is not surprising that sparse representation is effective for this simple example, since the flow is periodic and patterns observed in the training data generalize to the test fields. In fact, for this reason, we do not have a truly independent set of flow fields on which we validate the ability of these methods to generalize beyond training data. However, these results are encouraging, because sparse representation exhibits accurate and robust reconstruction across a variety of physical measurement configurations and noise intensity.

With the wealth of potential reconstruction techniques within the library-based optimization framework, it is interesting to explore the relative resilience to noise of different choices of the library and regularizing norm. Figure 6 shows a comparison of reconstruction accuracy with increasing noise level for several of these combinations. We find that over a wide range of Gaussian noise levels, the sparse reconstruction (ℓ_1 norm) with either the training library or a K-SVD library outperforms POD-based methods. The slope of the ℓ_2-based methods are smaller than those of the ℓ_1-based methods, so they will eventually achieve lower relative error, although at such large levels of noise it is unlikely that any method will result in a useful reconstruction. The poor performance of ℓ_1 optimization with a POD library indicates that the POD basis does not admit a sparse representation; empirical POD modes are eigenfunctions of a time-averaged correlation matrix, so that the energy in any particular flow field is distributed across modes. Thus, although the POD basis is optimal in the sense that it offers the best global reconstruction for a given number, this does not translate into optimality for the problem of reconstruction from limited measurements. In contrast, the K-SVD library is designed to admit a sparse representation of the data, and it is not surprising that this library results in the best performance.[2] However, sparse representation in a

[2]K-SVD allows for tuning several parameters; although our chosen values work well, these are not necessarily optimal.

library of the training data exhibits similar performance and benefits from simple implementation and interpretable results. These results reinforce the importance of sparse representation with respect to robustness to noise, a quality which has made ℓ_1 regularization a popular tool to prevent overfitting in parameter estimation [84].

It is not surprising that sparse representation is so effective on this example, since the flow is low-rank, periodic, and does not exhibit multiscale phenomena. For a periodic flow, sparse representation reduces to choosing the single example with the correct phase from the library. In contrast, each flow field is a dense linear combination of POD modes, so that this library does not admit a sparse representation. Reconstruction from limited, noisy measurements via sparse representation in a training library can therefore provide a robust alternative to ℓ_2-based methods. In addition, the set of sparse coefficients can also be used for robust estimation integral quantities such as the lift coefficient, provided the fields in the training set are labeled with the corresponding quantity of interest, as demonstrated in Appendix C.

B. Mixing layer

The downstream evolution of the mixing layer leads to globally aperiodic dynamics, so we cannot expect to exactly reproduce an arbitrary flow field with a single example from the training library, as was the case for the periodic vortex shedding behind a cylinder. However, we find that highly sparse representations still lead to accurate flow-field estimates. This suggest that the library of mixing flow fields generalize to new flows that are not in the training set, allowing a sparse representation.

Figure 7 demonstrates reconstruction of the normal vorticity of the mixing layer from spatially downsampled measurements, given by 10:1 downsampling of the original data in the middle region containing the mixing layer. Since noise is added to the measurements, the sparse representation is found using Eq. (7) with an optimally chosen value of ϵ (see Appendix B). This may be thought of as a *super-resolution* problem [121,122], where low-resolution measurement data is synthesized into a higher-resolution field based on a high-resolution library. We compare reconstruction via sparse representation in the training library to ℓ_2 minimization in a truncated library of POD modes ($r = 50$). Both suffer from some degree of overfitting, since the specific global arrangement of vortices in the test data is likely not observed in the training data. Still, sparse representation builds a reasonable picture of the early perturbations and later large-scale vortical structure, whereas both are barely identifiable in the POD reconstruction.

The relatively limited accuracy of sparse representation in this case suggests that the global test field is not a sparse linear combination of examples in the training set, presumably because all possible arrangements of vortices and their phases have not been observed. Compared to vortex shedding past a cylinder, this flow exhibits more complex, multiscale dynamics that are driven by the successive vortex pairing process. The local measurements may not be informative or correlated with the global structure of the flow field, which is an implicit assumption of the global library-based estimation. In such cases, where the global domain is larger than the decorrelation length, it may be helpful to assume that measurements inform only the local spatial region of the flow. Thus, we apply the localized reconstruction process introduced in Sec. III B. We divide the full flow field into ten windows that grow linearly in the streamwise direction, consistent with the streamwise dynamical scaling, and solve the local sparse representation problems independently. Reconstructions are then formed from sparse combinations of the windowed training fields.

Figure 7(g) compares the relative sparsity of the global and windowed sparse representations, given by the fraction of total nonzero coefficients across all independent windowed optimization problems. The local representations are more sparse and have higher fidelity than the global reconstruction. This suggests a connection between multiscale features of the flow and the sparsity of representation. By restricting the scope of the reconstruction problem, we simplify the effective dynamics, and this is reflected in the order of magnitude difference in the sparsity of representation. These results suggest that the proposed flow-field reconstruction method may generalize well to spatially complex flow fields. Note that, when averaged across the test data, the various

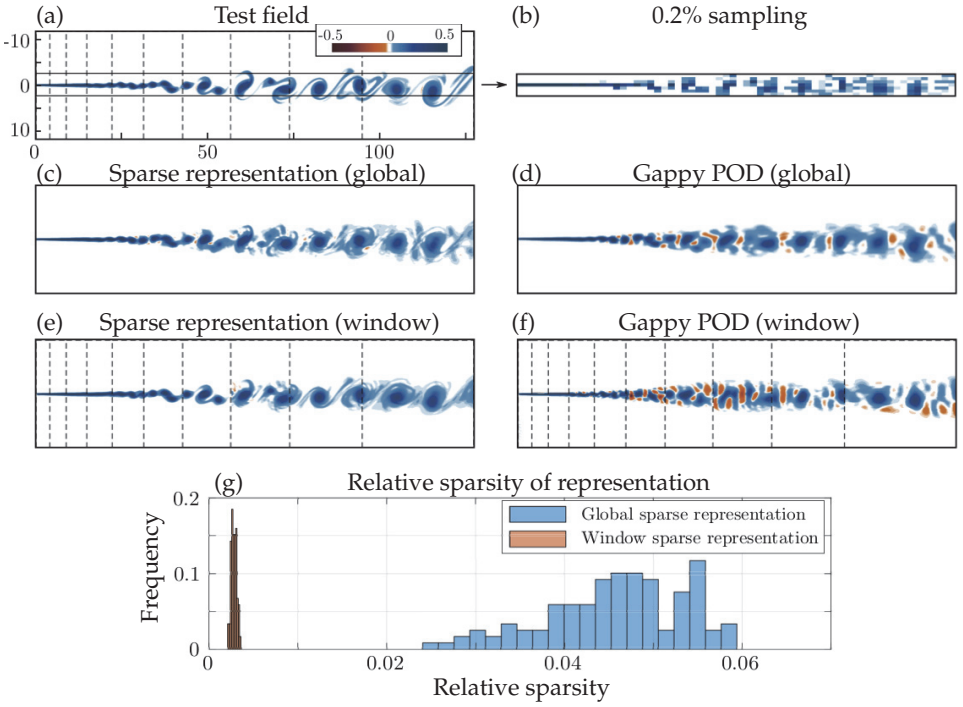

FIG. 7. Global reconstruction of a mixing layer vorticity field via super-resolution. (a) The test field is (b) measured with 10:1 downsampling and reconstructed with both (c) sparse representation in a training library and (d) least-squares regression in a POD library. Although both methods overfit the data, sparse representation captures the large-scale structures more effectively than least-squares POD. (e), (f) Separating the domain into windows that grow linearly in the streamwise direction, consistent with the flow dynamics, leads to a more realistic reconstruction from sparse representation, although gappy POD does not improve with this approach. (g) The relative sparsity, given by the fraction of nonzero coefficients, of the global and windowed sparse representations shows that windowing enables improved sparsity. A color map of reconstruction errors is shown in Fig. 19 in Appendix C.

reconstruction results are comparable in an ℓ_2 error metric ($\sim$0.4), even though the fields obtained via sparse representations exhibit more qualitatively accurate flow structures. Gappy POD performs relatively well in this case because there are more samples ($p = 448$) than POD modes in the truncated basis ($r = 50$). The optimal performance of gappy POD tends to be in the oversampled regime; see Appendix B for a discussion and demonstration on the sea surface temperature data set. However, the ℓ_2 metric is likely not ideal for measuring differences in convecting flow structures, because shifting the exact test field by a couple of pixels will result in an error that is comparable to the gappy POD field.

Figure 8 demonstrates reconstruction from ten noisy point measurements evenly spaced along the centerline of each window. In this case, sparse representation significantly outperforms gappy POD, although when averaged across the test data, the global sparse reconstruction is more accurate, in an ℓ_2 sense, than the windowed estimate. As discussed further in Sec. VI and shown in Fig. 12, the dynamics have longer timescales in the downstream windows, because the flow evolves from a linear instability wave to complex vortex interactions. Thus, the training data may not include enough representative examples of downstream behavior to admit a sparse representation. Indeed, the local reconstructions in the first seven windows are highly accurate and the errors in the total flow field estimate are largely due to the final three windows.

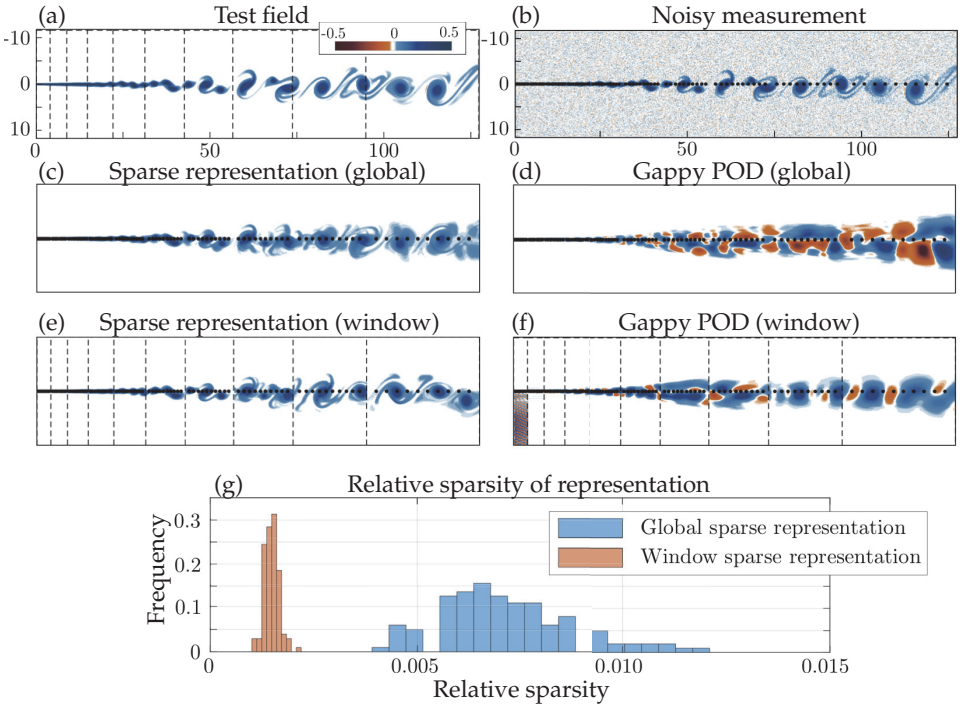

FIG. 8. Local mixing layer vorticity field reconstruction. As in Fig. 7, we construct windows that grow linearly in the spanwise direction and collect ten point measurements at noise level $\sigma = 0.3$ (white and black dots) from the mixing layer centerline in each window (separated by dashed lines). (e), (f) Local reconstructions are computed based on a windowed library of training examples. The normalized residual error averaged across the test set is (c) 0.49 for the global sparse reconstruction, (e) 0.58 for the windowed sparse reconstruction, (d) 1.06 for global gappy POD reconstruction, and (f) 1.05 for the local POD estimate. Although the windowed sparse representation has larger global errors than the global estimation, the local reconstruction is more accurate in the windows closer to the inlet, since the timescales of the flow are shorter there and the training set is more likely to contain examples of similar fields [see also Fig. 12(b)]. The relative sparsity of the global and windowed sparse representations in panel (g) shows improved sparsity with windowing.

C. Sea surface temperature field

The global sea surface temperature (SST) data set represents a flow field that is strongly driven by periodic seasonal forcing but also deviates from oscillatory behavior under the influence of complex oceanographic and environmental processes. For this reason, the SST data may be viewed as a problem of intermediate difficulty for the algorithm, with complexity somewhere between the Re = 100 flow past a cylinder from Sec. V A and the strongly aperiodic Gulf of Mexico data in Sec. V D. Figure 9 shows a comparison of mean-subtracted temperature field reconstructions from randomly located point measurements restricted to the mid-latitude region from 50° S to 50° N. We reconstruct the field with sparse representation in a training library and compare the result to gappy POD with the same measurements in a heavily truncated library ($r = 2$) and with many more measurements in a library of $r = 50$ POD modes. Both methods perform similarly in the error metric given by Eq. (10); average reconstruction errors across the test data are within one standard error of one another.

Fluctuations in the sea surface temperature field are dominated by seasonal oscillations, so that two POD modes capture a surprising amount of structure. In the absence of measurement noise, gappy POD has proven effective in estimating flow fields that can be represented accurately in terms

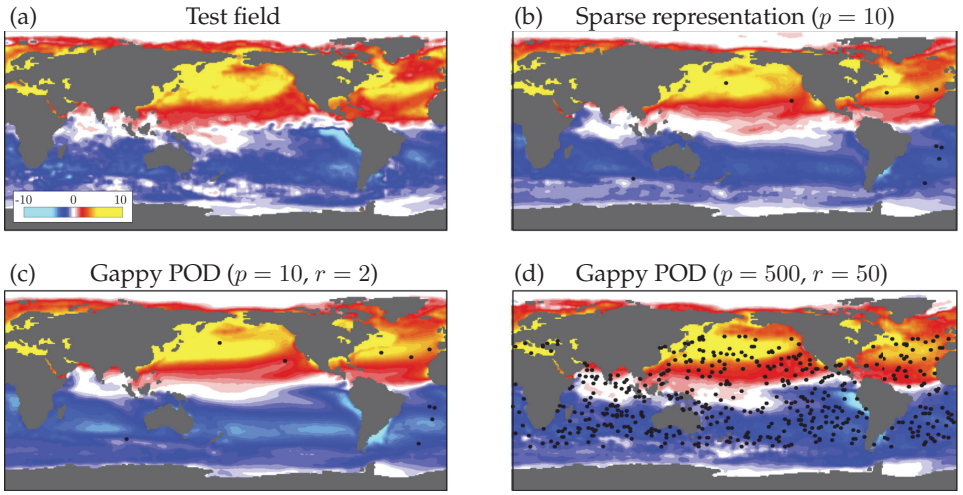

FIG. 9. (a) Example reconstruction of a sea surface temperature field using (b) sparse representation in a training library and gappy POD with (c) $p = 10$ and (d) $p = 500$ measurements. The POD libraries were truncated at $r = 2$ and $r = 50$ modes respectively, which were the empirically determined optimal values. Errors in all three estimates are around 0.30. The long-term annual mean temperature field has been subtracted to highlight variations in the data. A color map of reconstruction errors is shown in Fig. 20 in Appendix C.

only a few modes [29]. The fact that a flow as apparently complex as in Fig. 9(a) can be accurately reconstructed with either method from as few as 10 random point measurements is a reflection of the underlying low-dimensional structure of this data set.

D. Gulf of Mexico surface vorticity

The final test flow is the HYCOM Gulf of Mexico ocean velocity data, which poses the greatest challenge for reconstruction. On the timescale of the available data, the flow is not statistically stationary and, given the spatial complexity of the flow, accurate reconstruction requires significantly more measurements than the other fields considered in this work. Figure 10 shows a typical example of field reconstruction from $p = 4000$ random point measurements, which account for about 8.5% of all grid locations. Global reconstructions from both sparse representation in the training set and gappy POD with a truncated library of $r = 500$ modes successfully reconstruct much of the large-scale structure in the field and have comparable reconstruction errors, although the sparse representation estimate is contaminated by nonphysical high-frequency fluctuations.

As with the mixing layer, the test field cannot be accurately represented as a sparse combination of training examples because of the complexity of the flow and the fact that the training data do not fully generalize to the test set. The sparsest representation identified by Eq. (7) contains $K = 3995$ (around 48%) nonzero coefficients [Fig. 10(c)], around 400 times as many as the sea surface temperature field in Fig. 9(b). However, we achieve a more accurate reconstruction through the local kernel approach outlined in Sec. III B and the Appendix. By separating the reconstruction problem into localized kernels, we can stitch these local reconstructions together to obtain a global field that more accurately captures the large-scale vortical structures in the test field. This is intuitive from a measurement perspective, since the localization essentially relaxes the optimization constraints so that the sparse representation need only be consistent with *local* measurements. Seeking a global reconstruction that matches all measurements simultaneously is overly restrictive for a flow in which spatial correlations decay rapidly. As with the mixing layer, sparsity appears to be a hallmark of complexity; on average, each kernel representation contains only $K \approx 106$ ($\sim$1%) nonzero

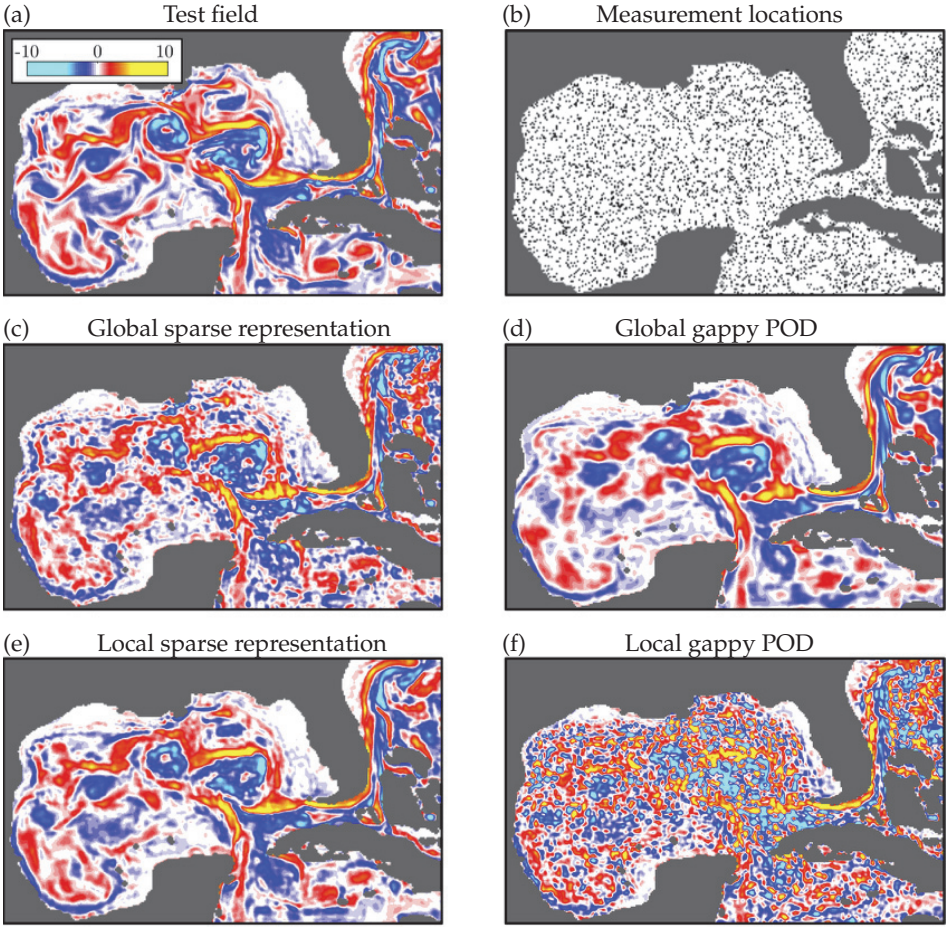

FIG. 10. (a) Reconstruction of Gulf of Mexico vorticity field from (b) $p = 4000$ random points. (c) The global sparse reconstruction from a training library suffers from overfitting, since there is not a highly sparse global representation of this flow field in the training set. (d) Gappy POD on the global field is comparable to global sparse representation; residual errors in both cases are $\sim$0.60. (e) Sparse reconstruction from the same measurements but using the local kernel method described in Sec. III B with $k = 96$ equally spaced kernels, enables *locally* sparse representations that combine to form a significantly more accurate global estimate, with a reconstruction error of 0.37. (f) The local least-squares POD still suffers from high-frequency overfitting. A color map of reconstruction errors is shown in Fig. 21 in Appendix C.

coefficients. This suggests that features in local patches of the flow may closely resemble those present in the training data, even if the global flow field does not.

Figure 11 demonstrates reconstruction from uniformly downsampled measurements. The test field is sampled at a 5:1 ratio in the ocean region ($p = 1261$ points) and reconstructed with gappy POD and local sparse representation, using the same parameters as in Fig. 10. Again, reconstruction from a local sparse representation is more accurate than gappy POD, suggesting that the method may be useful for interpolating low-resolution sensor data.

VI. DISCUSSION

This work demonstrates the enhanced robustness and accuracy of flow-field reconstruction by sparse representation in a library of training data, and we have explored this approach on a range

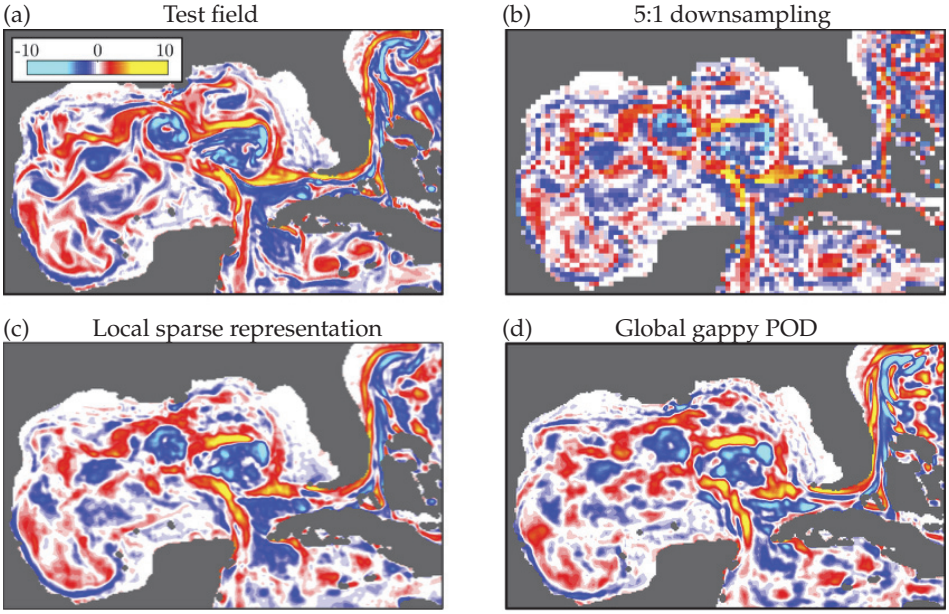

FIG. 11. (a) Reconstruction of Gulf of Mexico vorticity field from (b) uniform 5:1 downsampling. (c) Local sparse representation with $k = 96$ equally spaced kernels yields a 26% improvement in reconstruction error over (d) gappy POD in a library truncated to $r = 500$ modes. Global sparse representation and local gappy POD both perform worse than these methods.

of example flow fields of increasing complexity. We also discuss potential limitations of sparse representation, along with proposed methodological extensions and improvements. The success of sparse representation depends on the availability of both an extensive library that contains representative examples of relevant flow structures and sufficiently rich sensor information to infer which of these structures are active. Both of these requirements are related to the flow physics and spatiotemporal scales of the particular system under consideration.

First, the library must contain a sufficiently extensive collection of example flow fields, so that a new flow field may be approximated by a sparse combination of these examples. Even for aperiodic flows, this may be satisfied if the training set contains a long enough flow history. To quantify if the training library is sufficiently complete to generalize to a new test field, we compute the residual error obtained by approximating a test field by orthogonal projection onto the training library. If the test field is well approximated in the training library, the residual is small, and if the test field has new structures that are not observed in the training data, there will be a large residual. Figure 12 shows the residual error for each of the four flow examples from Sec. V as a function of the length of the training data; the orthogonal projection is obtained by computing the POD subspace for the given library with m training examples. Even with very few examples in the library, the flow past a cylinder generalizes to the test data, since the flow is periodic. However, the mixing layer and Gulf of Mexico vorticity data have relatively large generalization error, even for large training libraries, indicating that there are new structures that have not been observed in the training data. With enough training data, it is conceivable that the generalization error can be controlled for these flows, although this may be prohibitively expensive in terms of data collection and processing. Instead, decomposing the flow domain into local patches results in considerably improved library generalization [see Fig. 12(b)], meaning that fewer training data are required for an accurate representation of a new flow field in the library. As shown in Figs. 7, 8, and 10, the local patches also admit a *sparser* representation in the library, resulting in more accurate and robust flow reconstructions. The improved performance of a local sparse representation is intuitive because

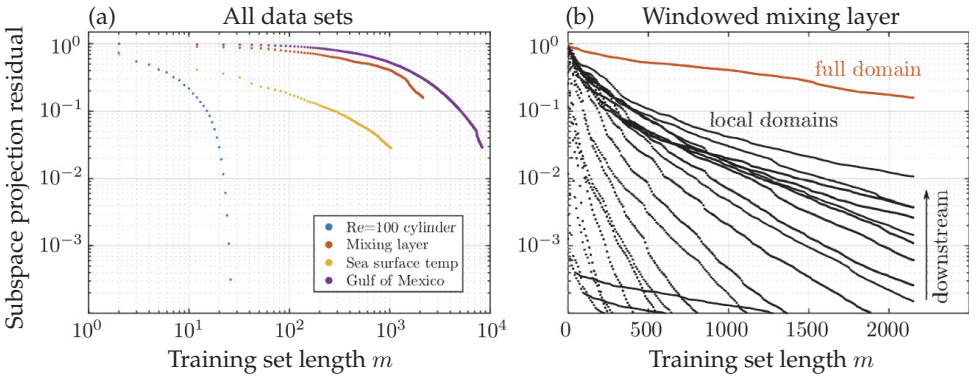

FIG. 12. Residual error in projection of test data onto the linear subspace spanned by POD modes. As more data are added to the training set (horizontal axis), arbitrary fields from the test set are more likely to be in the span of the training data. This residual represents a lower bound on the error in a global reconstruction based on this linear subspace. (a) Comparison of all flows studied in this work. (b) Subspace projection residuals for the windowing scheme shown in Figs. 7 and 8. Even for windows encompassing the complex vortex pairing behavior downstream, the test data are approximately within the span of the subspace, although the timescales of convergence are much longer.

decomposing the spatial domain makes it more likely to find similar local flow structures in the training data. Thus, the generalization error of the training library provides a useful diagnostic to quantify the expected performance of sparse representation.

Throughout these examples, we find that least-squares solutions generally overfit to noisy sensor measurements, resulting in nonphysical high-frequency fluctuations, as in Figs. 7 and 10. In contrast, sparse representation in a library of training examples results in robust and accurate flow reconstruction, preventing overfitting and ensuring that that the unmeasured regions of the field are consistent with prior knowledge. If a sufficiently sparse approximation is not possible, however, the method will not reliably produce a field that is qualitatively similar to actual observations. However, even if a sparse representation of the entire flow field does not exist, Figs. 8(b), 8(c), and 10(d) demonstrate that locally sparse representations can be used for globally accurate reconstruction.

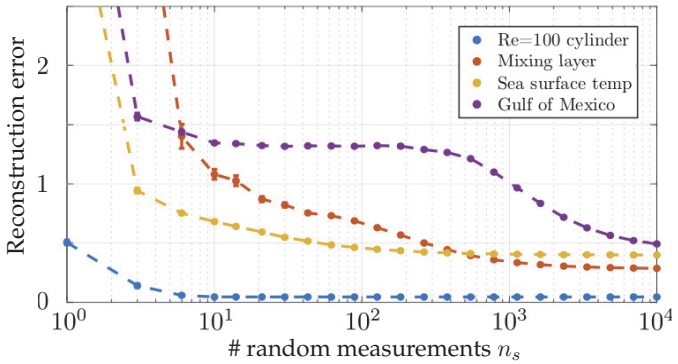

FIG. 13. Normalized residual error of sparse representation-based reconstructions with increasing number of random point measurements. We use the global reconstruction method for the periodic vortex shedding past a cylinder (blue), mixing layer (red), sea surface temperature fields (yellow), and Gulf of Mexico (purple) flow fields. For computational efficiency we estimate coefficient vectors $\hat{\boldsymbol{s}}$ using orthogonal matching pursuit (OMP) with empirical estimates of the sparsity K. Accurate reconstruction requires both rich training data and sufficient measurement information; the latter condition can vary widely depending on the flow.

A second condition for successful sparse flow reconstruction, even with a rich enough library, is for the measurements to provide sufficient information to correctly identify the sparse library coefficients. For example, it is unrealistic to hope that a single point measurement can be used to reconstruct a highly turbulent flow field, no matter how comprehensive the training library. In the Bayesian perspective, even perfect knowledge of the prior distribution is not enough for an accurate estimation, unless it is sufficiently conditioned on measurement information. Figure 13 shows the accuracy of the sparse representation method versus the number of random point measurements for each example. In each case, there is a rough number of sensors where the error sharply decreases: $p = 10$ for the flow past a cylinder and sea surface temperature fields, $p = 100$ for the mixing layer, and $p = 4000$ for the Gulf of Mexico data, which roughly correspond to the number of measurements used in Sec. V. Further increasing the number of sensors results in a minimum error plateau, which is defined by the generalization error of the library, as described above. It is important to note that it may be possible to reconstruct the flow field with less error from fewer measurements by leveraging additional knowledge about the flow; for example, from a reduced-order model [63] or via time-delay embedding [123].

VII. CONCLUSION

In this study, we develop a method for flow-field reconstruction based on sparse representation in a library of examples. This method builds on prior work in library-based reconstruction and sparse representation; in particular the sparse representation for the classification algorithm [12,13]. We apply this method to several example flows, ranging from simple canonical flows to challenging geophysical data sets, and demonstrate improved accuracy and robustness to noise and corruption compared with typical least-squares flow reconstruction, provided that the library is sufficiently rich and the measurements are sufficiently informative.

This work suggests several directions of ongoing research to refine the method for practical applications. For example, the requirement that the flow be statistically stationary appears to preclude application in flows with changing operating conditions. However, the demonstration of the method on the Gulf of Mexico vorticity field suggests that the stationarity requirement may be violated, provided the library is rich enough. A more thorough investigation of parameter variation, for example to reconstruct flow past a cylinder over a range of Reynolds numbers, would be an interesting future direction; sparse representation has already been shown to identify different dynamical regimes in a classification context [13]. In addition, sparse representation fundamentally relies on a linear embedding, with the associated limitations. It is unclear how to naturally extend these approaches to nonlinear embeddings, although this is an interesting avenue of ongoing work. However, the localized methods of Secs. V B and V D provide a principled approach to lessen the burden on the required training data.

Although Figs. 12 and 13 give rough metrics for sufficiency of the training data and measurement information, more quantitative and principled criteria for both requirements would be useful to determine *a priori* which flows are good candidates for this method, how much training data are required, and which sensor configurations will sufficiently inform the structure of the flow field. The performance of this method may also be improved with more sophisticated library learning methods [66–68] or optimal sensor placement strategies [29,30,65,124], both of which are active areas of research.

The localized reconstruction procedure presented here, while effective, has also not been optimized. Continuing to develop this procedure is a promising future direction because it offers a route to circumvent some apparent limitations of library-based reconstruction [see, e.g., Fig. 4(b)]. For example, in convection-dominated flows or flows with continuous symmetries, one might envision a "universal" library constructed by convolving a kernel with the flow fields, in the spirit of recent work on convolutional neural networks [69].

Furthermore, there are many ways to formulate and solve the sparse optimization problem (7), and alternative approaches, such as sequentially thresholded least squares [86] or pursuit

algorithms, may outperform ℓ_1 regression. In addition, the sparse optimization procedure may be too computationally expensive for some real-time control applications, motivating ongoing work to improve algorithmic efficiency; fortunately, the timescales of the geophysical flows investigated in this work are slow compared with the sparse optimization. Finally, most flows of interest are three dimensional, and it will be important to demonstrate this method on three-dimensional flows. While it is straightforward to generalize this method to three-dimensional fields, e.g., by the same vectorization approach used for two-dimensional flows, the matrices representing discretized 3D flows can become very large. The main computational cost is due to the optimization problem, which scales with the number of measurements and not the dimensionality of the discretized flow field; however, the number of measurements necessary to sufficiently inform the structure of a complex three-dimensional flow field may lead to prohibitively expensive computations.

In the examples considered here, we have shown several advantages which make sparse representation an attractive candidate for flow-field reconstruction. In particular, we find that reconstruction from sparse representation in a library of training examples is robust against noise and leads to physical flow-field estimates. We have shown that this framework can be modified to handle dense sensor noise and gross measurement corruption. The method can also be extended to complex flow fields by decomposing the spatial domain and seeking localized sparse representations. With this flexibility, sparse representation may provide a powerful tool for estimating complex flow fields in a range of applications.

The immersed boundary projection layer solver used to generate the cylinder wake data is publicly available [125]; the data are available in [data Ref. 126]. The NOAA OISST v2 data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA [data Ref. 127] (accessed 2018). HYCOM data are available in [data Ref. 128].

ACKNOWLEDGMENTS

We gratefully acknowledge funding support from the Air Force Office of Scientific Research (FA9550-18-1-0200 and FA9550-16-1-0650) and the Army Research Office (W911NF-19-1-0045). K.M. gratefully acknowledges support by the Washington Research Foundation, the Gordon and Betty Moore Foundation (Award No. 2013-10-29), the Alfred P. Sloan Foundation (Award No. 3835), and the University of Washington eScience Institute. Special thanks to Nathan Kutz for sharing his insights into the extreme utility of sparsity promoting methods for complex physical systems. We also thank Bing Brunton, Ben Erichson, and Lionel Mathelin for valuable discussions on sparsity and flow estimation. Funding for the development of HYCOM was provided by the National Ocean Partnership Program and the Office of Naval Research. Data assimilative products using HYCOM are funded by the U.S. Navy.

APPENDIX A: LOCAL RECONSTRUCTION METHOD

Here, we provide details on the kernel-based localized reconstruction method applied to the mixing layer and Gulf of Mexico vorticity fields; for example, to produce the estimates in Figs. 10(e) and 10(f). As described in Sec. III B, we construct a global flow-field estimate as a weighted superposition of local reconstructions in a decomposed domain, which admits sparser representations in the training library.

We introduce compact overlapping kernels $\boldsymbol{\Phi}_j$, $j = 1, 2, \ldots, k$ normalized so that at each grid location $\mathbf{r}$, $\sum_j \boldsymbol{\Phi}_j(\mathbf{r}) = 1$. These kernels separate the global estimation problem into k local problems:

$$\hat{\boldsymbol{s}}_j = \arg\min_{\boldsymbol{s}_j} \|\boldsymbol{s}_j\|_q \text{ subject to } \|\boldsymbol{y}_j - \boldsymbol{C}\boldsymbol{\Phi}_j\boldsymbol{\Psi}\boldsymbol{s}_j\|_2 < \epsilon.$$

Since the kernels are normalized, the local flow-field estimates $\hat{\boldsymbol{x}}_j = \boldsymbol{\Phi}_j\boldsymbol{\Psi}\hat{\boldsymbol{s}}_j$ can be combined to form a global estimate $\hat{\boldsymbol{x}} = \sum_j \hat{\boldsymbol{x}}_j$. The simplest such kernels are the windows used for the mixing

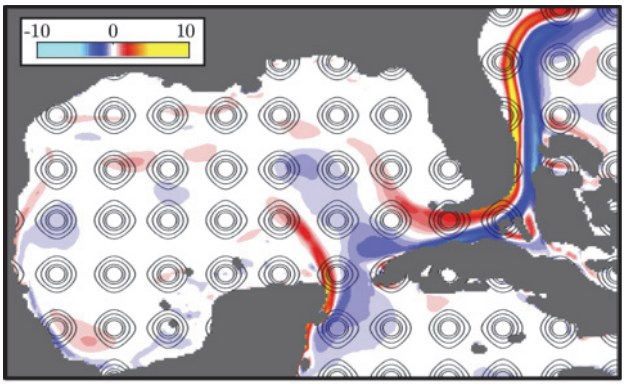

FIG. 14. Locations of kernel centers for the local reconstructions in Figs. 10(e) and 10(f), shown superimposed on the empirical mean vorticity field. The contours shown are for Gaussian kernels prior to normalization.

layer reconstructions in Figs. 8(c) and 8(d), where each kernel has the value 1 within its window and 0 outside.

For the Gulf of Mexico data we define 96 points $\mathbf{r}_1,\ \mathbf{r}_2, \ldots,\ \mathbf{r}_j,\ \mathbf{r}_{96}$ as the kernel centers on a uniform 12×8 grid covering the spatial domain, as in Fig. 14. Each kernel $\mathbf{\Phi}_j$ is constructed with a radial Gaussian function of the distance from each grid location to the kernel center:

$$\mathbf{\Phi}_j(\mathbf{r}) = \frac{1}{N(\mathbf{r})} e^{|\mathbf{r}-\mathbf{r_j}|^2/\sigma^2},$$

where the width σ is half the longitudinal distance between successive kernel centers. Values below 10^{-2} are set to zero, and the normalization factors $N(\mathbf{r})$ are then calculated as the sum of the

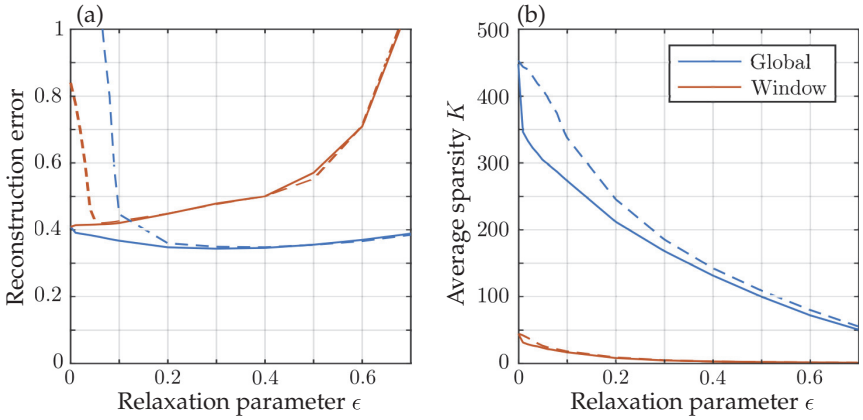

FIG. 15. The sparse representation problem in Eq. (7) allows for choosing the relaxation parameter ϵ. (a) Average reconstruction error across the test set vs ϵ for the downsampled mixing layer (see Fig. 7) from clean and noisy measurements (solid and dashed lines, respectively). (b) Sparsity of representation for the same reconstruction problem vs ϵ. Relaxing the problem improves reconstruction accuracy, especially when the measurements are noisy, but increasing ϵ too far can allow for solutions which are inconsistent with measurements.

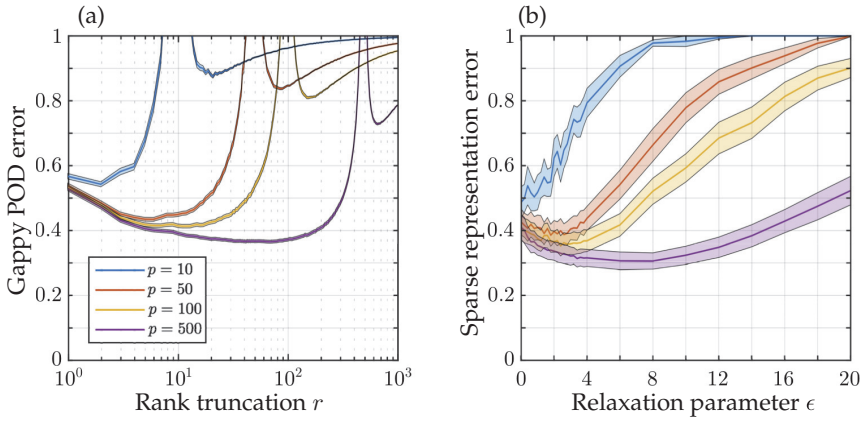

FIG. 16. Parameter tuning for sea surface temperature field estimation. (a) Rank truncation of gappy POD library. Sharp peaks correspond to $p = r$, when the number of point measurements equals the number of retained modes and the matrix $\boldsymbol{C\Psi}$ becomes square. The optimal truncation appears to depend on the number of measurements but is always in the oversampled case $r < p$. (b) Relaxation of the sparse representation problem. As with the mixing layer, some relaxation of the constraint can improve accuracy, but an overly relaxed optimization problem leads to inconsistent estimates. For both plots, shaded bands indicate standard error on the mean.

unweighted values of all kernels

$$N(\mathbf{r}) = \sum_{j=1}^{k} \boldsymbol{\Phi}_j(\mathbf{r}),$$

so that the resulting estimates may be combined in a weighted average.

Computationally, each location $\mathbf{r}$ is a point on the grid, and so just as the flow-field snapshots are arranged into column vectors, the corresponding values of $\boldsymbol{\Phi}_j$ are the entries in a sparse diagonal matrix. The product $\boldsymbol{\Phi}_j\mathbf{x}$ of a kernel with a discretized flow field is then nonzero only in the region surrounding the kernel center $\mathbf{r}_j$.

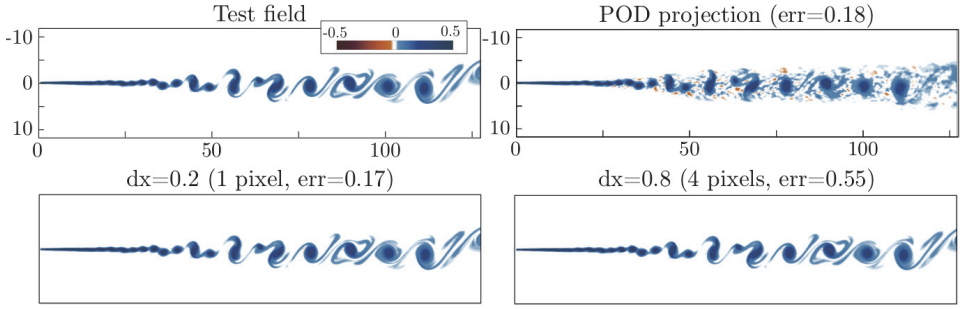

FIG. 17. Several perspectives on the global ℓ_2 error metric. The projection onto the POD basis is the optimal representation in the linear span of the training data. However, shifting the field by one grid step in the streamwise direction generates an ℓ_2 error nearly the same as the POD projection. A shift of four steps results in more than 50% error, although the fields capture exactly the same coherent structures. Whether it is more important to capture global energy content or estimate physically accurate structures may depend on the application.

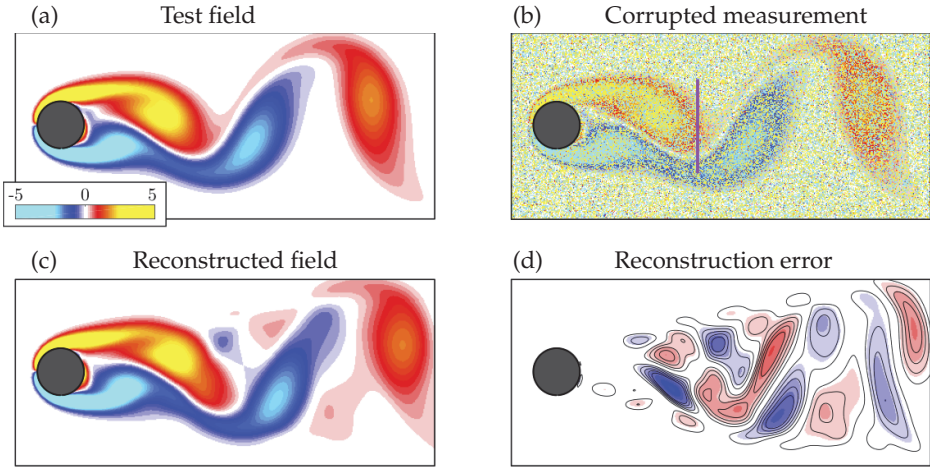

FIG. 18. Visualization of reconstruction error for the cylinder wake from corrupted measurements (see Fig. 5). (a) Test snapshot. (b) Example flow snapshot with corruption in 70% of grid locations. The vertical stripe shows the measurement location. (c) Reconstructed flow field from sparse representation using corrupted measurements shown in panel (b). (d) Errors in the flow-field estimate of panel (c), shown on the same scale as the original field. The contours are on intervals of 0.1 between -0.5 and 0.5.

APPENDIX B: PARAMETER TUNING

Machine learning methods typically allow for tuning some set of parameters for optimal performance. For example, increasing the regularization parameter λ in Eq. (4) can prevent overfitting to noisy data, but beyond some optimal value the solution is prone to bias. For the sparse representation

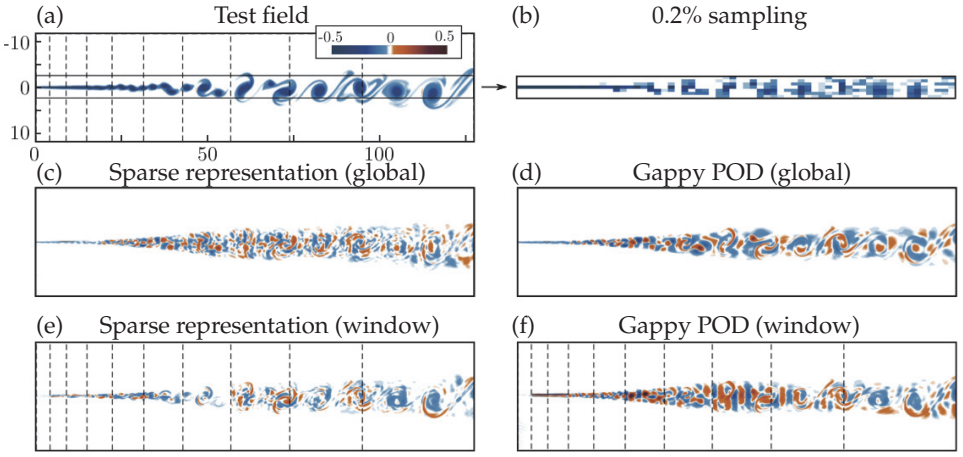

FIG. 19. Error in the reconstruction of a mixing layer vorticity field via super-resolution (see Fig. 7). (a) The test field is measured with (b) 10:1 downsampling and reconstructed with both sparse representations in (c) a training library and (d) least-squares regression in a POD library. The global reconstructions are compared with separating the domain into windows that grow linearly in the streamwise direction and reconstructing with (e) sparse representation and (f) gappy POD. On average, the windowing approach does not lead to more accurate reconstructions in an ℓ_2 sense than global sparse representation, although the error structures are more localized.

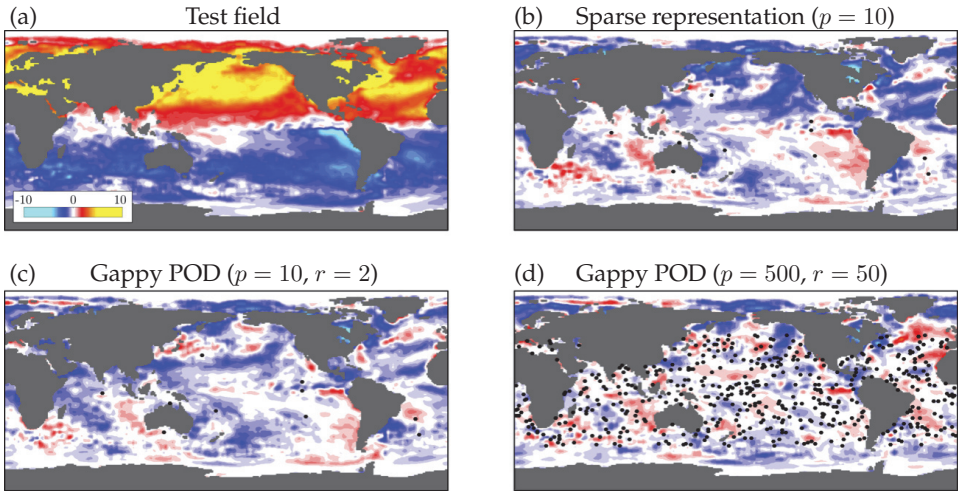

FIG. 20. (a) Visualization of errors in estimates of sea surface temperature field using (b) sparse representation in a training library and gappy POD with (c) $p = 10$ and (d) $p = 500$ measurements. The POD libraries were truncated to $r = 2$ and $r = 50$ modes respectively, which were empirically determined to be optimal. Global ℓ_2 errors in all three estimates are around 0.30.

problem (7) which is central to our proposed method, the only free parameter is the relaxation ϵ. A larger value of ϵ allows for a larger difference between the observations and the estimated flow field, which may be useful for instance if the measurements are noisy or if the training data may not generalize well. In these cases, relaxing the constraint can lead to a sparse solution and accurate estimation.

The sparse representation method can be sensitive to the choice of ϵ. For example, Fig. 15(a) shows the error in reconstructing a downsampled test field (see Fig. 7) with increasing relaxation. For clean data and global reconstruction (solid blue lines), the method is only weakly dependent on the choice of ϵ, whereas a careful selection of this parameter is important for reconstruction from windows (red lines) and if measurements are noisy (dashed lines). For the windowed reconstruction, error increases sharply beyond an optimal value of ϵ; Fig. 7(b) suggests that this may be because overly relaxed constraints allow the solution to be highly sparse but generally inconsistent with observations. We expect that an appropriate choice of ϵ will in general also depend on both the magnitude of observed fluctuations and the number of measurements.

Gappy POD can also be tuned via truncation of the POD library. In this case the optimization problem is typically the overdetermined case (3), although if there are fewer measurements than modes in the library, the least-squares solution to Eq. (5) with $q = 2$ is an analogous estimation method. We find that gappy POD performs best in the oversampled case $r < p$. That is, the library is truncated to contain fewer modes than measurements. For example, gappy POD has comparable accuracy to sparse representation for the sea surface temperature field estimates in Fig. 9, but this accuracy is sensitive to the truncation. Figure 16(a) shows gappy POD reconstruction error vs rank truncation r averaged across the test data for various numbers of random point measurements p. The error peaks sharply when the measured library matrices $\boldsymbol{C\Psi}$ become nearly square, but for any value of p the optimal truncation is $r < p$.

Figure 16(b) shows reconstruction error vs relaxation parameter ϵ for sparse representation-based estimates of the sea surface temperature fields for varying number of point measurements. The points are chosen at random with a new realization for each field in the test set. In this case we do not artificially introduce noise, although an appropriate choice of ϵ still improves the accuracy of the sparse representation method.

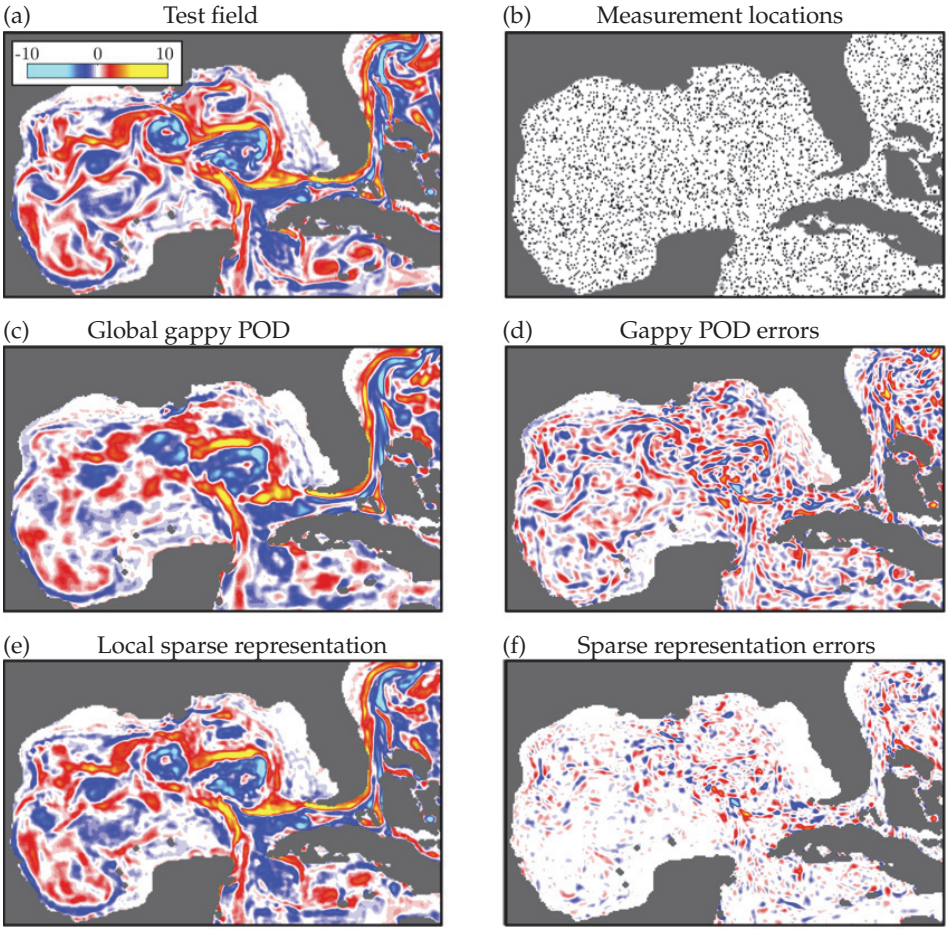

FIG. 21. (a) Reconstruction of Gulf of Mexico vorticity field from (b) $p = 4000$ random points. (c), (d) Although gappy POD resolves the large-scale structures, (e), (f) the localized version of sparse representation represents a significant reduction in error. Errors are on the same color scale as the fields. The global ℓ_2 error in the gappy POD estimate is 0.60, while the sparse representation error is 0.37. Parameters for the estimation are the same as described for Fig. 10.

APPENDIX C: RECONSTRUCTION ERRORS AND INTEGRAL QUANTITIES

Although the global ℓ_2 error as defined by Eq. (10) is a convenient metric for the reconstruction accuracy, it is an incomplete description. For example, Fig. 17 explores shifting a mixing layer vorticity field by a single pixel, which results in an ℓ_2 error comparable to the optimal representation in the training set (via projection onto the POD basis).

It may therefore be helpful to visualize the errors in reconstruction for the various flows investigated here. This Appendix includes plots of the reconstruction error for the cylinder wake (Fig. 18, c.f. Fig. 5) mixing layer (Fig. 19, c.f. Fig. 7), sea surface temperature (Fig. 20, c.f. Fig. 9), and Gulf of Mexico vorticity (Fig. 21, c.f. Fig. 10).

The sparse coefficient vector can also be used to estimate integral quantities such as the lift coefficient. This may be thought of as a "labeled" problem, where each field in the training set is augmented with the quantity of interest. The estimate of this quantity for the test field is then given by the product of the sparse coefficient vector with the vector of labels. For example, Fig. 22

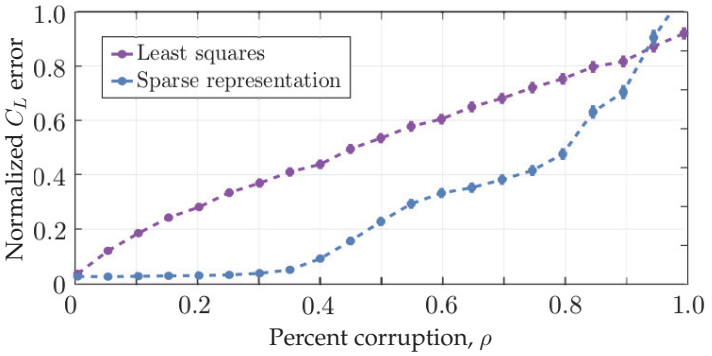

FIG. 22. Estimation of the lift coefficient C_L for the $\mathrm{Re} = 100$ cylinder wake from the horizontal slice measurement (green line in Fig. 5). The vertical axis shows the estimation error normalized by the rms lift coefficient over the vortex shedding period. The sparse representation method described in Appendix C outperforms least squares estimation over a wide range of sparse corruption levels.

compares the lift coefficient estimate for the cylinder wake using this approach against the standard least squares prediction. The measurements are taken from the "horizontal slice" [green line in Fig. 5(a)] with noise level $\sigma = 0.01$ and varying levels of sparse corruption. Sparse representation outperforms least squares over a wide range of corruption levels.

[1] B. R. Noack, M. Morzynski, and G. Tadmor, *Reduced-Order Modelling for Flow Control* (Springer, Berlin, 2011), Vol. 528.
[2] R. King, *Active Flow and Combustion Control 2014* (Springer, Cham, 2014), Vol. 127.
[3] S. L. Brunton and B. R. Noack, Closed-loop turbulence control: Progress and challenges, Appl. Mech. Rev. **67**, 050801 (2015).
[4] D. Sipp and P. J. Schmid, Linear closed-loop control of fluid instabilities and noise-induced perturbations: A review of approaches and tools, Appl. Mech. Rev. **68**, 020801 (2016).
[5] J. Pfeiffer, Closed-loop active flow control for road vehicles under unsteady cross-wind conditions, Ph.D. thesis, Technische Universität Berlin, 2016.
[6] B. Strom, S. L. Brunton, and B. Polagye, Intracycle angular velocity control of cross-flow turbines, Nat. Energy **2**, 17103 (2017).
[7] R. K. Maurya, *Characteristics and Control of Low Temperature Combustion Engines* (Springer, Cham, 2017), Chap. 9, pp. 483–510.
[8] A. Yakhot, T. Anor, and G. E. Karniadakis, A reconstruction method for gappy and noisy arterial flow data, IEEE Trans. Med. Imaging **26**, 1681 (2008).
[9] S. Sankaran, M. E. Moghadam, A. M. Kahn, E. E. Tseng, J. M. Guccione, and A. L. Marsden, Patient-specific multiscale modeling of blood flow for coronary artery bypass graft surgery, Ann. Biomed. Eng. **40**, 2228 (2012).
[10] M. D. Graziano, M. D'Errico, and G. Rufino, Ship heading and velocity analysis by wake detection in SAR images, Acta Astronaut. **128**, 72 (2016).
[11] E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability* (Cambridge University Press, 2003).
[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210 (2009).
[13] I. Bright, G. Lin, and J. N. Kutz, Compressive sensing based machine learning strategy for characterizing the flow around a cylinder with limited pressure measurements, Phys. Fluids **25**, 127102 (2013).

[14] S. L. Brunton, J. H. Tu, I. Bright, and J. N. Kutz, Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems, J. Appl. Dyn. Syst. **13**, 1716 (2014).
[15] I. Bright, G. Lin, and J. N. Kutz, Classification of spatiotemporal data via asynchronous sparse sampling, Multiscale Model. Simul. **14**, 823 (2016).
[16] B. Kramer, P. Grover, P. Boufounos, S. Nabi, and M. Benosman, Sparse sensing and DMD-based identification of flow regimes and bifurcations in complex flows, J. Appl. Dyn. Syst. **16**, 1164 (2017).
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (MIT Press, 2012), pp. 1097–1105.
[18] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).
[19] J. Yu and J. S. Hesthaven, Flowfield reconstruction method using artificial neural network, AIAA J. **57**, 482 (2019).
[20] L. Sirovich, Turbulence and the dynamics of coherent structures. I - Coherent structures. II - Symmetries and transformations. III - Dynamics and scaling, Q. Appl. Math. **45**, 561 (1987).
[21] G. Berkooz, P. Holmes, and P. L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, Annu. Rev. Fluid Mech. **25**, 539 (1993).
[22] P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech. **656**, 5 (2010).
[23] C. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, Spectral analysis of nonlinear flows, J. Fluid Mech. **641**, 115 (2009).
[24] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, On dynamic mode decomposition: Theory and applications, J. Comput. Dynam. **1**, 391 (2014).
[25] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems* (SIAM, 2016).
[26] R. Everson and L. Sirovich, Kauhunen-LoÈve procedure for gappy data, J. Opt. Soc. Am. A **12**, 1657 (1995).
[27] T. Bui-Thanh, M. Damodaran, and K. Willcox, Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition, AIAA J. **42**, 1505 (2004).
[28] D. Venturi and G. E. Karniadakis, Gappy data and reconstruction procedures for flow past a cylinder, J. Fluid Mech. **519**, 315 (2004).
[29] K. Willcox, Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition, Comput. Fluids **35**, 208 (2006).
[30] B. Yildrim, C. Chryssostomidis, and G. E. Karniadakis, Efficient sensor placement for ocean measurements using low-dimensional concepts, Ocean Model. **27**, 160 (2009).
[31] B. Podvin, Y. Fraigneau, F. Lusseyran, and P. Gougat, A reconstruction method for the flow past an open cavity, J. Fluids Eng. **128**, 531 (2005).
[32] R. Adrian, On the role of conditional averages in turbulence theory, *Proceedings of the 4th Biennial Symposium on Turbulence in Liquids* (Science Press, Princeton, 1975), pp. 323–332.
[33] J. P. Bonnet, D. R. Cole, J. Delville, M. N. Glauser, and L. S. Ukeiley, Stochastic estimation and proper orthogonal decomposition: Complementary techniques for identifying structure, Exp. Fluids **17**, 307 (1994).
[34] D. Ewing and J. Citriniti, Examination of a LSE/POD complementary technique using single and multi-time information in the axisymmetric shear layer, in *Proceedings of the IUTAM Symposium on Simulation and Identification of Organized Structures in Flows* (Springer, Dordrecht, 1999).
[35] C. E. Tinney, F. Coiffet, J. Delville, A. M. Hall, P. Jordan, and M. N. Glauser, On spectral linear stochastic estimation, Exp. Fluids **41**, 763 (2006).
[36] V. Durgesh and J. W. Naughton, Multi-time-delay LSE-POD complementary approach applied to unsteady high-Reynolds-number near wake flow, Exp. Fluids **49**, 571 (2010).
[37] L. Ukeiley, N. Murray, Q. Song, and L. Cattafesta, Dynamic surface pressure based estimation for flow control, *IUTAM Symposium on Flow Control and MEMS* (Springer, Dordrecht, 2008), pp. 183–189.
[38] R. J. Adrian, Conditional eddies in isotropic turbulence, Phys. Fluids **22**, 2065 (1979).
[39] T. C. Tung and R. J. Adrian, Higher-order estimates of conditional eddies in isotropic turbulence, Phys. Fluids **23**, 1469 (1980).

[40] Y. G. Guezennec, Stochastic estimation of coherent structures in turbulent boundary layers, Phys. Fluids A **1**, 1054 (1989).
[41] A. M. Naguib, C. E. Wark, and O. Juckenhöfel, Stochastic estimation and flow sources associated with surface pressure events in a turbulent boundary layer, Phys. Fluids **13**, 2611 (2001).
[42] D. R. Cole and M. N. Glauser, Applications of stochastic estimation in the axisymmetric sudden expansion, Phys. Fluids **10**, 2941 (1998).
[43] J. A. Taylor and M. N. Glauser, Towards practical flow sensing and control via POD and LSE based low-dimensional tools, J. Fluids Eng. **126**, 337 (2004).
[44] L. M. Hudy and A. Naguib, Stochastic estimation of a separated-flow field using wall-pressure-array measurements, Phys. Fluids **19**, 024103 (2007).
[45] N. E. Murray and L. S. Ukeiley, Estimation of the flowfield from surface pressure measurements in an open cavity, AIAA J. **41**, 969 (2003).
[46] N. E. Murray and L. S. Ukeiley, An application of gappy pod, Exp. Fluids **42**, 79 (2007).
[47] J. T. Pinier, J. M. Ausseur, M. N. Glauser, and H. Higuchi, Proportional closed-loop feedback control of flow separation, AIAA J. **45**, 181 (2007).
[48] J. H. Tu, J. Griffin, A. Hart, C. W. Rowley, L. N. Cattafesta, and L. S. Ukeiley, Integration of non-time-resolved PIV and time-resolved velocity point sensors for dynamic estimation of velocity fields, Exp. Fluids **54**, 1429 (2013).
[49] A. Surana and A. Banaszuk, Linear observer synthesis for nonlinear systems using Koopman Operator framework, IFAC-PapersOnLine **49**, 716 (2016).
[50] P. Holmes, J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monographs on Mechanics (Cambridge University Press, 1996).
[51] B. R. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, and F. Thiele, A hierarchy of low-dimensional models for the transient and post-transient cylinder wake, J. Fluid Mech. **497**, 335 (2003).
[52] M. Buffoni, S. Camarri, A. Iollo, E. Lombardi, and M. V. Salvetti, A non-linear observer for unsteady three-dimensional flows, J. Comput. Phys. **227**, 2626 (2008).
[53] J.-C. Loiseau and S. L. Brunton, Constrained sparse Galerkin regression, J. Fluid Mech. **838**, 42 (2018).
[54] J.-C. Loiseau, B. R. Noack, and S. L. Brunton, Sparse reduced-order modeling: Sensor-based dynamics to full-state estimation, J. Fluid Mech. **844**, 454 (2018).
[55] T. Suzuki, Reduced-order Kalman-filtered hybrid simulation combining particle tracking velocimetry and direct numerical simulation, J. Fluid Mech. **709**, 249 (2012).
[56] D. Foures, N. Dovetta, D. Sipp, and P. J. Schmid, A data-assimilation method for Reynolds-averaged Navier-Stokes-driven mean flow reconstruction, J. Fluid Mech. **759**, 404 (2014).
[57] S. Symon, N. Dovetta, B. J. McKeon, D. Sipp, and P. J. Schmid, Data assimilation of mean velocity from 2D PIV measurements of flow over an idealized airfoil, Exp. Fluids **58**, 61 (2017).
[58] B. Combés, D. Heitz, A. Guilbert, and E. Mémin, A particle filter to reconstruct a free-surface flow from a depth camera, Fluid Dynam. Res. **47**, 051404 (2015).
[59] R. Kikuchi, T. Misaka, and S. Obayashi, Assessment of probability density function based on POD reduced-order model for ensemble-based data assimilation, Fluid Dyn. Res. **47**, 051403 (2015).
[60] V. Mons, J.-C. Chassaing, T. Gomez, and P. Sagaut, Reconstruction of unsteady viscous flows using data assimilation schemes, J. Comput. Phys. **316**, 255 (2016).
[61] A. F. C. da Silva and T. Colonius, Ensemble-based state estimator for aerodynamic flows, AIAA J. **56**, 2568 (2018).
[62] Andre F. C. da Silva, An EnKF-based flow state estimator for aerodynamic problems, Ph.D. thesis, California Institute of Technology, 2019.
[63] C. W. Rowley and S. T. M. Dawson, Model reduction for flow analysis and control, Annu. Rev. Fluid Mech. **49**, 387 (2017).
[64] K. Taira, S. L. Brunton, S. T. M. Dawson, C. W. Rowley, T. Colonius, B. J. McKeon, O. T. Schmidt, S. Gordeyev, V. Theofilis, and L. S. Ukeiley, Modal analysis of fluid flows: An overview, AIAA J. **55**, 4013 (2017).
[65] K. Manohar, B. W. Brunton, J. Nathan Kutz, and S. L. Brunton, Data-driven sparse sensor placement, IEEE Control Syst. Mag. **38**, 63 (2018).

[66] M. Aharon, M. Elad, and A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. **54**, 4311 (2006).
[67] M. Elad and M. Aharon, Image denoising via learned dictionaries and sparse representation, in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2006).
[68] L. Mathelin, K. Kasper, and H. Abou-Kandil, Observable dictionary learning for high-dimensional statistical inference, Arch. Comput. Methods Eng. **25**, 103 (2018).
[69] K. Fukami, K. Fukagata, and K. Taira, Super-resolution reconstruction of turbulent flows with machine learning, J. Fluid Mech. **870**, 106 (2019).
[70] N. B. Erichson, L. Mathelin, Z. Yao, S. L. Brunton, M. W. Mahoney, and J. N. Kutz, Shallow learning for fluid flow reconstruction with limited sensors and limited data, arXiv:1902.07358.
[71] J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann, The shifted proper orthogonal decomposition: A mode decomposition for multiple transport phenomena, SIAM J. Sci. Comput. **40**, A1322 (2018).
[72] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2019).
[73] M. Milano and P. Koumoutsakos, Neural network modeling for near wall turbulent flow, J. Comput. Phys. **182**, 1 (2002).
[74] C. Wehmeyer and F. Noé, Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics, J. Chem. Phys. **148**, 1 (2018).
[75] F. J. Gonzalez and M. Balajewicz, Learning low-dimensional feature dynamics using deep convolutional recurrent autoencoders, arXiv:1808.01346.
[76] K. Lee and K. Carlberg, Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders, arXiv:1812.08373.
[77] B. Lusch, J. Nathan Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nat. Commun. **9**, 4950 (2018).
[78] K. Champion, B. Lusch, J. Nathan Kutz, and Steven L. Brunton, Data-driven discovery of coordinates and governing equations, arXiv:1904.02107.
[79] Z. J. Zhang and K. Duraisamy, Machine learning methods for data-driven turbulence modeling, in *Proceedings of the 22nd AIAA Computational Fluid Dynamics Conference, AIAA, Dallas* (AIAA, Reston, VA, 2015), p. 2460.
[80] J. Ling, A. Kurzawski, and J. Templeton, Reynolds averaged turbulence modeling using deep neural networks with embedded invariance, J. Fluid Mech. **807**, 155 (2016).
[81] J. N. Kutz, Deep learning in fluid dynamics, J. Fluid Mech. **814**, 1 (2017).
[82] K. Duraisamy, G. Iaccarino, and H. Xiao, Turbulence modeling in the age of data, Ann. Rev. Fluid Mech. **51**, 357 (2019).
[83] S. Mallat, Understanding deep convolutional networks, Philos. Trans. R. Soc., A **374**, 20150203 (2016).
[84] H. Xu, C. Caramanis, and S. Mannor, Robust regression and LASSO, IEEE Trans. Inf. Theory **56**, 3561 (2010).
[85] J. N. Kutz, *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data* (Oxford University Press, 2013).
[86] P. Zheng, T. Askham, S. L. Brunton, J. Nathan Kutz, and A. Y. Aravkin, A unified framework for sparse relaxed regularized regression: SR3, IEEE Access **7**, 1404 (2018).
[87] S. G. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. **41**, 3397 (1993).
[88] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers* (IEEE, New York, 1993), pp. 40–44.
[89] J. A. Tropp and A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory **53**, 4655 (2007).
[90] D. Needell and J. A. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. **26**, 301 (2009).
[91] D. Donoho, For most large underdetermined systems of linear equations the minimal ℓ_1 norm near solution approximates the sparsest solution, Commun. Pure Appl. Math. **59**, 907 (2006).

[92] E. Candès and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inf. Theory **52**, 5406 (2006).
[93] D. L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory **52**, 1289 (2006).
[94] E. Candès, Compressive sampling, in *Proc. International Congress of Mathematicians* (European Mathematical Society, Zurich, 2006).
[95] R. G. Baraniuk, Compressive sensing, IEEE Signal Processing Magazine **24**, 118 (2007).
[96] X. Huang, Compressive sensing and reconstruction in measurements with an aerospace application, AIAA J. **51**, 1011 (2013).
[97] J. H. Tu, C. W. Rowley, J. N. Kutz, and J. K. Shang, Spectral analysis of fluid flows using sub-Nyquist-rate PIV data, Exp. Fluids **55**, 1805 (2014).
[98] J.-L. Bourgiuignon, J. A. Tropp, A. S. Sharma, and B. J. McKeon, Compact representation of wall-bounded turbulence using compressive sampling, Phys. Fluids **26**, 015109 (2014).
[99] Z. Bai, T. Wimalajeewa, Z. Berger, G. Wang, M. Glauser, and P. K. Varshney, Low-dimensional approach for reconstruction of airfoil data via compressive sensing, AIAA J. **53**, 920 (2015).
[100] M. Liu, D. Zhang, and D. Shen, Ensemble sparse classification of Alzheimer's disease, NeuroImage **60**, 1106 (2012).
[101] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert, Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling, NeuroImage **76**, 11 (2013).
[102] N. Nabizadeh and M. Kubat, Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features, Comput. Electr. Eng. **45**, 286 (2015).
[103] Y. Panagakis, C. Kotropoulos, and G. R. Arce, Music genre classification using locality preserving non-negative tensor factorization and sparse representations, in *10th International Society for Music Information Retrieval Conference* (ISMIR, Canada, 2009).
[104] M. Esfahanian, H. Zhuang, and N. Erdol, A new method for detection of North Atlantic right whale up-calls, J. Acoust. Soc. Am. **136**, 2073 (2014).
[105] E. N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, J. Atmos. Sci. **26**, 636 (1969).
[106] Z. Zhao and D. Giannakis, Analog forecasting with dynamics-adapted kernels, Nonlinearity **29**, 2888 (2016).
[107] R. Lguensat and P. Tandeo, The analog data assimilation, Mon. Weather Rev. **145**, 4093 (2017).
[108] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. Roy. Stat. Soc. B **58**, 267 (1996).
[109] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2009).
[110] M. Gavish and D. L. Donoho, The optimal hard threshold for singular values is $4/\sqrt{3}$, IEEE Trans. Inf. Theory **60**, 5040 (2014).
[111] R. Rubenstein, M. Zibulevsky, and M. Elad, Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit, Technical report, Techicon (2008).
[112] M. C. Grant and S. P. Boyd, Graph implementations for nonsmooth convex programs, in *Recent Advances in Learning and Control*, edited by V. D. Blondel, S. P. Boyd, and H. Kimura (Springer, London, 2008), pp. 95–110.
[113] M. C. Grant and S. P. Boyd, CVX: Matlab Software for Disciplined Convex Programming, http://cvxr.com/cvx (2013).
[114] K. Taira and T. Colonius, The immersed boundary method: A projection approach, J. Comput. Phys. **225**, 2118 (2007).
[115] T. Colonius and K. Taira, A fast immersed boundary method using a nullspace approach and multi-domain far-field boundary conditions, Comput. Methods Appl. Mech. Eng. **197**, 2131 (2008).
[116] S. Stanley and S. Sarkar, Simulations of spatially developing two-dimensional shear layers and jets, Theor. Comput. Fluid Dyn. **9**, 121 (1997).
[117] V. Coralic and T. Colonius, Finite-volume WENO scheme for viscous compressible mulitcomponent flows, J. Comput. Phys. **274**, 95 (2014).
[118] A. K. M. F. Hussain, Role of coherent structures in turbulent shear flows, Proc. Indian Acad. Sci. **4**, 129 (1981).

[119] C. D. Winant and F. K. Browand, Vortex pairing: The mechanism of turbulent mixing-layer growth at moderate Reynolds number, J. Fluid Mech. **63**, 237 (1974).
[120] G. L. Brown and A. Roshko, On density effects and large structures in turbulent mixing layers, J. Fluid Mech. **64**, 775 (1974).
[121] J. Yang, J. Wright, T. S. Huang, and Y. Ma, Image super-resolution via sparse representation, IEEE transactions on image processing **19**, 2861 (2010).
[122] W. T. Freeman, T. R. Jones, and E. C. Pasztor, Example-based super-resolution, IEEE Comput. Graph. Appl. **22**, 56 (2002).
[123] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. Nathan Kutz, Chaos as an intermittently forced linear system, Nat. Commun. **8**, 1 (2017).
[124] V. Mons, J.-C. Chassaing, and P. Sagaut, Optimal sensor placement for variational data assimilation of unsteady flows past a rotationally oscillating cylinder, J. Fluid Mech. **823**, 230 (2017).
[125] https://github.com/crowley/ibpm.
[126] http://www.dmdbook.com.
[127] https://www.esrl.noaa.gov/psd/.
[128] http://hycom.org.