

**Assessing the information content of complex flows**Lei Fang<sup>1</sup> and Nicholas T. Ouellette<sup>2,\*</sup><sup>1</sup>*Department of Civil and Environmental Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15620, USA*<sup>2</sup>*Department of Civil and Environmental Engineering, Stanford University, Stanford, California 94305, USA*

(Received 23 September 2020; accepted 22 January 2021; published 5 February 2021)

Complex dynamical systems can potentially contain a vast amount of information. Accurately assessing how much of this information must be captured to retain the essential physics is a key step for determining appropriate discretization for numerical simulation or measurement resolution for experiments. Using recent mathematical advances, we define spatiotemporally compact objects that we term dynamical linear neighborhoods (DLNs) that reduce the amount of information needed to capture the local dynamics in a well-defined way. By solving a set-cover problem, we show that we can compress the information in a full dynamical system into a smaller set of optimally influential DLNs. We demonstrate our techniques on experimental data from a laboratory quasi-two-dimensional turbulent flow. Our results have implications both for assessments of the fidelity of simulations or experiments and for the compression of large dynamical data sets.

DOI: [10.1103/PhysRevE.103.023301](https://doi.org/10.1103/PhysRevE.103.023301)**I. INTRODUCTION**

Spatially extended dynamical systems often contain huge amounts of information [1]. Since they are described by partial differential equations, the number of degrees of freedom is formally infinite. Nevertheless, in practice, most complex systems are well described by only a finite subset of these degrees of freedom; that is, a projection of the full dynamical system onto a lower-dimensional finite approximation can capture all of the relevant physics [2]. This observation underlies, for example, the numerical simulation of spatially extended systems on discrete grids or various mode decomposition schemes [3–5].

Accurate approximation, however, requires us to know how many degrees of freedom we need to retain before the physics of the discretized system differs appreciably from that of the original system. In some cases, physical reasoning can allow us to make estimates of the required number. In a turbulent fluid flow, for example, Kolmogorov's 1941 scaling theory [6] argues that the range of important length scales extends from the correlation length scale  $L$  down to the dissipation length scale  $\eta$ . The ratio of these scales is predicted to be  $L/\eta \sim \text{Re}^{3/4}$ , where  $\text{Re}$  is the turbulent Reynolds number. The number of degrees of freedom further scales as  $L^3$ , and therefore as  $\text{Re}^{9/4}$  [7]. Thus, classical theory allows us to make a prediction for how the number of degrees of freedom scales with a parameter that governs the complexity of the problem. Refinements of this estimate using, for instance, the multifractal formalism do not change the prediction of the number of degrees of freedom significantly [8]. For this reason, simulations of turbulence typically only resolve scales down to roughly  $\eta$  [7], and neglect the dynamics on finer scales.

Although the Kolmogorov estimate of the number of degrees of freedom is by and large considered to be successful, it nevertheless has several shortcomings. Importantly, it is only an estimate of the typical (in an average, order-of-magnitude sense) number of active degrees of freedom. Instantaneously, however, the actual number can be significantly larger [9,10]. Kolmogorov theory also only considers one aspect of the turbulence problem, namely the statistical properties of the (Eulerian) velocity field. If we are instead primarily interested in transport and Lagrangian trajectories, the number of required degrees of freedom is less clear due to questions related to sampling and dynamics [11–13]. And finally, Kolmogorov theory only applies to the flow of a homogeneous, constant-property fluid at asymptotically high Reynolds numbers. Thus, it cannot easily be used to estimate requirements for a vast range of other relevant flows including multiphase flow [14], transitional or intermediate-Reynolds-number flow [15], or stratified turbulence [16,17].

Here, we introduce a flexible technique for assessing the information content present in general dynamical flows. This method extends our previous work on a kind of coherent structure we termed linear neighborhoods [18] by augmenting them with a dynamical length scale determined using the recently developed stochastic sensitivity theory [12,13]. We define spatially extended regions we call dynamical linear neighborhoods (DLNs) that have the property that the flow field at any point inside in a specified time window can be computed up to a defined tolerance by simple linear extrapolation from a seed point. In an information-theoretic sense, then, a DLN can be seen as a single degree of freedom of the system. Computing the number of degrees of freedom of the system can then be reduced to the solution of a set-cover problem.

After defining DLNs, we demonstrate their computability and use by analyzing data acquired from a laboratory quasi-two-dimensional turbulent flow. We show that, as expected,

\*nto@stanford.edu

the number of DLNs required to span the system grows with Reynolds number. Additionally, we show that the flow field can be reconstructed with high accuracy using only the relatively small number of seed trajectories of the covering DLNs, suggesting potential ways to use this framework for data compression.

## II. THEORY

In previous work [18], we defined finite-time coherent structures by considering the behavior of the trajectories of fluid elements. These trajectories are the solutions of the ordinary differential equation

$$\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x}, t), \quad (1)$$

where  $\mathbf{u}$  is the Eulerian velocity field (that is, a solution of the Navier-Stokes equations). This equation admits the formal solution

$$\mathbf{F}_{t_0}^{t_0+T}(\mathbf{x}) = \mathbf{x}(t_0) + \int_{t_0}^{t_0+T} \mathbf{u}(\mathbf{F}_{t_0}^\tau(\mathbf{x}), \tau) d\tau \quad (2)$$

over the time span from  $t_0$  to  $t_0 + T$ , where  $\mathbf{F}_{t_0}^{t_0+T}(\mathbf{x})$  is the flow map that transforms the initial position of the fluid element  $\mathbf{x}(t_0)$  to its final position  $\mathbf{x}(t_0 + T)$  under the action of the velocity field  $\mathbf{u}$  and the integral should be interpreted as a path integral over the trajectory of the fluid element. If we now consider some other fluid element with initial position  $\mathbf{y}(t_0)$ , we could write an analogous solution to Eq. (1). However, if  $\mathbf{y}$  is close to  $\mathbf{x}$ , it would be reasonable to approximate its trajectory by linear extrapolation from the trajectory of  $\mathbf{x}$ . We define this linear approximation as

$$\begin{aligned} \mathbf{F}_{t_0}^{t_0+T}(\mathbf{y})_{\text{lin}} = & \mathbf{y}(t_0) + \int_{t_0}^{t_0+T} \left\{ \mathbf{u}(\mathbf{F}_{t_0}^\tau(\mathbf{x}), \tau) \right. \\ & \left. + \nabla \mathbf{u}(\mathbf{F}_{t_0}^\tau(\mathbf{x}), \tau) [\mathbf{F}_{t_0}^\tau(\mathbf{y}) - \mathbf{F}_{t_0}^\tau(\mathbf{x})] \right\} d\tau, \quad (3) \end{aligned}$$

where we have Taylor-expanded the velocity about the position of fluid element  $\mathbf{x}$  at time  $\tau$  and dropped higher-order terms. The error in this linear approximation is given by  $\|\mathbf{F}_{t_0}^{t_0+T}(\mathbf{y})_{\text{lin}} - \mathbf{F}_{t_0}^{t_0+T}(\mathbf{y})\|$ . Dimensionally, this error is a length scale that we denote as  $L_\ell$ , and it characterizes the distance between the estimated position of the fluid element at time  $t_0 + T$  and its true position. In our previous work [18], we defined the set of all  $\mathbf{y}$  for which  $L_\ell$  was smaller than some fixed threshold  $E$  as the linear neighborhood of the fluid element  $\mathbf{x}$ , because the behavior of these points could be captured up to a defined error tolerance by knowing the behavior of  $\mathbf{x}$  alone. However,  $E$  is a free parameter, and the linear neighborhood defined for any given fluid element depends on its value. One stringent choice for  $E$  would be choose the resolution  $L_R$  with which we know the velocity field. However, this choice is still arbitrary, and does not contain any information about the flow dynamics.

To bring a dynamical notion of the quality of the linear approximation into the picture, we turn to the recently developed stochastic sensitivity theory [12,13]. The central idea of this theory is to explicitly account for the unavoidable effects of error or uncertainty that arise in any practical setting on the solutions of Eq. (1), and in particular the way such uncertainty

will be amplified over time by the (nonlinear) dynamics. By modeling this process with a stochastic differential equation, stochastic sensitivity theory allows the calculation of a spatially resolved length scale  $L_S$  that gives a bound on the error we would expect in computing  $\mathbf{F}_{t_0}^{t_0+T}$  given only the dynamics and the accuracy of our computation scheme. Details of this calculation are given in Refs. [12,13]. If  $L_S$  is larger than  $L_R$ , the spatial resolution of the velocity field, the uncertainty in our computations will be dominated by the effective stochasticity arising from the system dynamics and we say that we are stochasticity limited [12]. Conversely, if  $L_R > L_S$ , poor resolution is the dominant source of uncertainty, and we say that we are resolution limited.

For our purposes here, these definitions provide a precise way of deciding when the linear approximation in Eq. (3) is acceptable: As long as  $L_\ell$  is less than the length scale associated with the dominant source of uncertainty, we cannot distinguish error due to linearization from error due to other effects. Thus, we define the dynamical linear neighborhood (DLN) of a fluid element  $\mathbf{x}$  to be the set of all  $\mathbf{y}$  for which  $L_\ell < \max(L_S, L_R)$ . If possible, one hopes to set  $L_R \leq L_S$  so that any uncertainty in the calculations is as small as possible. We term  $\mathbf{x}$  the center trajectory of the DLN, and stress again that, by construction, any trajectory  $\mathbf{y}$  within the DLN of  $\mathbf{x}$  is fully computable up to uncertainties inherent in the flow for the full time window over which the DLN is defined given information about  $\mathbf{x}$  alone.

Under this definition, every trajectory  $\mathbf{x}$  is the center trajectory for a DLN. If we consider a finite random sampling of trajectories, however, such as one might measure in a particle-tracking experiment, their DLNs will in general overlap significantly in some parts of the flow, while other parts of the domain may not be covered by any DLNs. This kind of random sampling of DLNs is thus not particularly informative. Instead, we seek here to determine a minimal set of DLNs that completely cover the flow domain. We can consider this set to be the most influential DLNs in the flow; and, since knowing only the center trajectories of this set of DLNs ought to allow us to recover the flow everywhere, the number of DLNs in this set can be taken to be a dynamically set estimate of the number of degrees of freedom in the flow.

We note that our method differs in spirit from traditional mode decomposition schemes such as proper orthogonal decomposition [3] or dynamic mode decomposition [5]. Such mode decompositions project the flow onto a set of ordered modes that are localized in, e.g., energy but not in space. In contrast, the DLNs in our method are spatially compact, and any gradients in DLN size across the flow domain can reveal local spatial scales in the flow as well.

## III. EXPERIMENTAL METHODS

To demonstrate our DLN approach, we analyze data obtained from experiments conducted in an electromagnetically driven thin-layer flow cell that produces nearly two-dimensional (2D) flow [19]. We describe this apparatus only briefly here, as we have given more details previously (see, e.g., Refs. [20,21]). The working fluid is a layer of

NaCl solution (14% NaCl by mass in de-ionized water) with lateral dimensions of  $86 \times 86 \text{ cm}^2$  and a depth of 0.5 cm. The fluid sits on a glass floor above a square grid of neodymium-iron-boron permanent magnets organized with their polarities alternating in stripes. The spacing of the magnets is  $L_m = 2.54 \text{ cm}$ , which sets the typical energy injection scale for the flow. We drive the flow by passing a dc current horizontally through the NaCl solution, and the interaction of this current with the magnetic field produces a Lorentz body force on the fluid. The flow remains 2D as long as the current is not too large [19]. The nondimensional strength of the forcing can be captured by the in-plane Reynolds number  $\text{Re} = u' L_m / \nu$ , where  $u'$  is the root-mean-square velocity and  $\nu$  is the kinematic viscosity [22]. We note that this Reynolds number does not directly capture either the width of the inverse energy cascade regime or the direct enstrophy cascade regime, but rather is indicative of the strength of the driving of the flow. Here, we consider Reynolds numbers ranging from 77 to 288. We vary the Reynolds number only by changing the electric current and not by changing any length scales. We seed the flow with fluorescent polystyrene microspheres  $51 \mu\text{m}$  in diameter. Above the NaCl solution, we float an additional 0.5-cm layer of fresh water both to remove any surface-tension-driven interactions among the tracer particles and to confine them to a single plane defined by the interface of the fresh water and NaCl solution. We image the tracer motion from above with a four-megapixel camera at 60 frames per second. Particle trajectories are extracted from the movies using a multiframe predictive tracking algorithm [23], and velocities are computed along the trajectories. We track roughly 30 000 particles per frame, and so can use the tracer velocities to construct well-resolved instantaneous Eulerian velocity fields. Finally, to remove noise and ensure the two-dimensionality of our data, we project the velocity fields onto a basis of streamfunction eigenmodes [19].

#### IV. RESULTS AND DISCUSSION

From these data, we seek to find the most influential DLNs described above. To do so, we observe that this problem is equivalent to the well documented set-cover problem in the computer-science literature [24]. Here, we use an iterative greedy algorithm to solve the set-cover problem. At each step of the iteration, we add to our set the DLN that contains the largest number of uncovered trajectories until either all trajectories are covered or the coverage does not improve. In Figs. 1(a) and 1(c), we show the set of most influential DLNs and corresponding center trajectories for two flows with different Reynolds numbers. Note that we plot these DLNs at the initial time  $t_0$  even though in general they are three-dimensional spacetime objects [18]. In particular, note that as the center trajectories move in the flow, the associated DLNs will move with them, though they will not necessarily appreciably change their size. The corresponding Eulerian flow fields at  $t_0$  are shown in Figs. 1(b) and 1(d). For smaller Reynolds numbers, we observe that the DLNs are relatively large and fairly isotropic. As the Reynolds number increases, the DLNs shrink, as shown in Fig. 2 where we plot the probability density functions of the area of each DLN. They

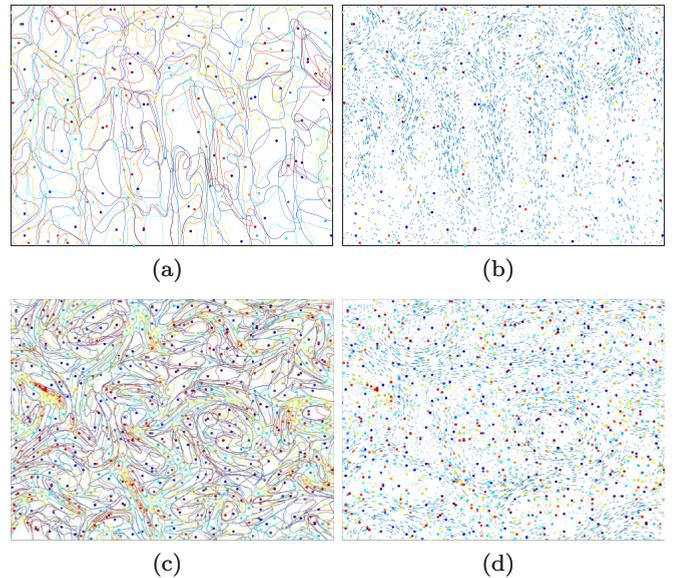


FIG. 1. (a) Most influential DLNs as determined by our greedy algorithm for a flow with  $\text{Re} = 77$ . Here,  $T = 2 \text{ s}$ , corresponding to one eddy turnover time for the largest Reynolds number in our data set ( $\text{Re} = 288$ ). (b) Corresponding downsampled velocity field (only 85% of the velocity vectors are shown, for clarity) at the initial time  $t_0$ . The locations of the center trajectories of the most influential DLNs shown in (a) are marked. (c), (d) The same quantities as in (a) and (b) but for a flow with  $\text{Re} = 234$ .

also begin to exhibit anisotropy indicative of the strong local heterogeneity of the turbulent flow. Additionally, as we would expect, as the Reynolds number increases, more DLNs are needed to cover the whole domain. As shown in Fig. 3, we find that the number of DLNs required scales as  $\text{Re}^{3/2}$ . Because the size of the domain does not change, this scaling suggests that the typical linear size of a DLN (estimated as the square root of its area) should vary as  $\text{Re}^{-3/4}$ , which is precisely what we find. Although it would be satisfying to compare this

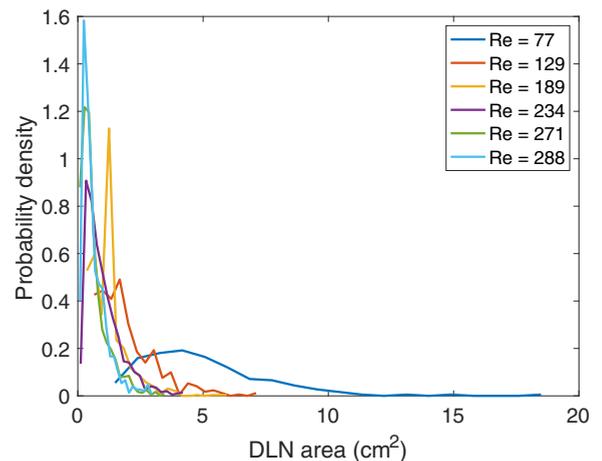


FIG. 2. Probability density functions of the area of individual DLNs for our six different Reynolds numbers. As  $\text{Re}$  grows, the typical area of a DLN shrinks.

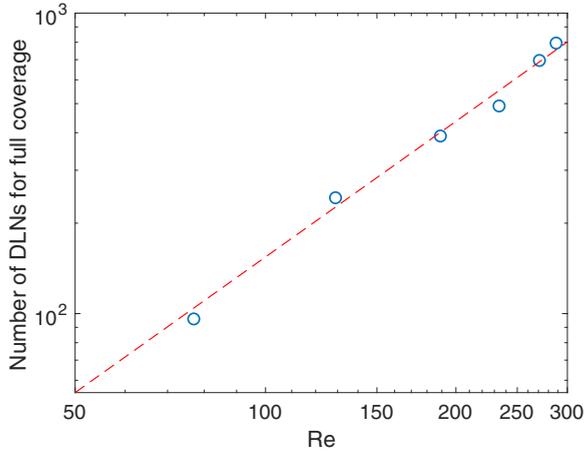
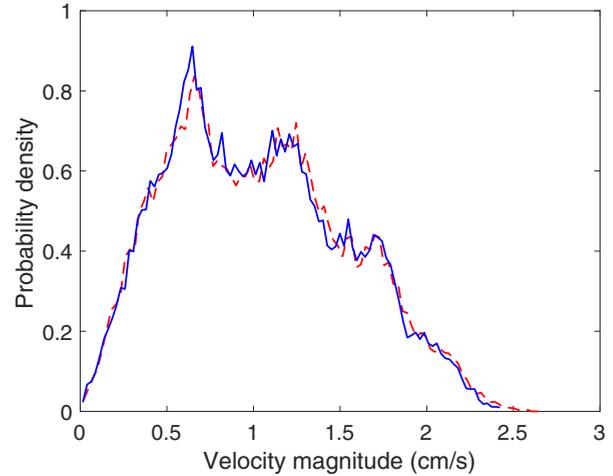


FIG. 3. The number of DLNs required to cover the full domain as a function of the flow Reynolds number. The dashed line is a  $Re^{3/2}$  power law.

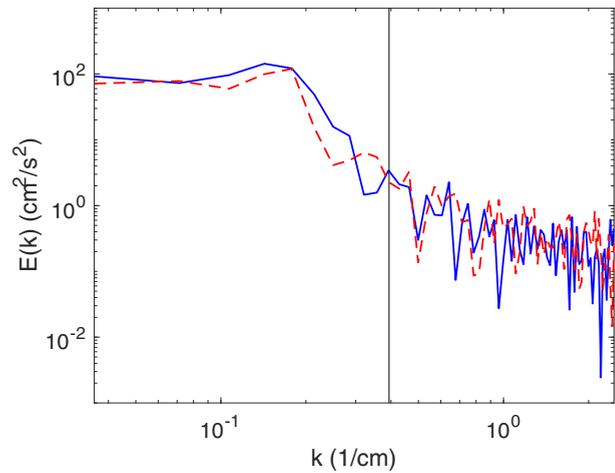
scaling to a theoretical prediction, it is not clear how to do so. To change the Reynolds number in our system, we change only the strength of the forcing; how this affects the widths of the energy and enstrophy cascades in two-dimensional turbulence, however, is nontrivial to determine. Thus, for the present we leave this scaling as an empirical finding.

Showing that the number of DLNs required to cover the flow domain increases with Reynolds number supports our hypothesis that they are good choices for the relevant degrees of freedom in the flow, suggesting that they contain most of the relevant information. A more stringent test of this conjecture is to attempt to reconstruct the flow field knowing only the DLNs and the center trajectories. To do so, we followed our definition of DLNs and linearly extrapolated the velocity known at each center trajectory of the set of influential DLNs over its full associated DLN. In Fig. 4, we compare the results of this reconstruction with the full measured field, both statistically [Figs. 4(a) and 4(b)] and instantaneously [Figs. 4(c) and 4(d)], at a Reynolds number of 216. Even though the reconstructed field uses only 5.7% of the measured trajectories, the accuracy of the reconstructed fields is excellent. This result again underscores our claim that the DLNs accurately capture the essential information contained in the flow: the quality of the reconstruction suggests that the additional information measured in the experiment is redundant.

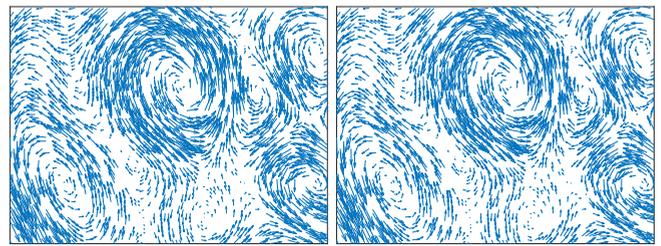
Finally, we provide evidence that the information reduction provided by our DLN approach is nontrivial by comparing the quality of the DLN-based reconstruction to a simpler reconstruction scheme. Any high-quality experiment or numerical simulation will in general over-resolve the flow field to ensure that no essential physics is missed; thus, one might argue that it is not surprising that we can reconstruct the flow field well from a small subset of the measured information. To test this hypothesis, we used a simple data-reduction scheme. First, we chose a random set of measured trajectories distributed roughly uniformly over the flow domain. Triangulating these points using a Delaunay triangulation allowed us to associate a polygon with each point, at which point we could



(a)



(b)



(c)

(d)

FIG. 4. Quality of a reconstruction of the flow field using a finite set of DLNs containing only 5.7% of the measured trajectories for flow with  $Re = 216$ . (a) Probability density functions of the velocity magnitude for both the original measurement (solid line) and the reconstruction (dashed line). (b) Energy spectra of the original measurement (solid line) and the reconstruction (dashed line). The vertical line shows the energy injection scale. (c) A section of the original velocity field (downsampled by 50% for clarity). (d) The same section of the reconstructed velocity field with vectors plotted at the same points as in (c).

use the same velocity field reconstruction scheme we used for the DLNs. To vary the number of DLNs used to reconstruct the field, we simply multiplied  $L_\ell$  in the DLN definition by a

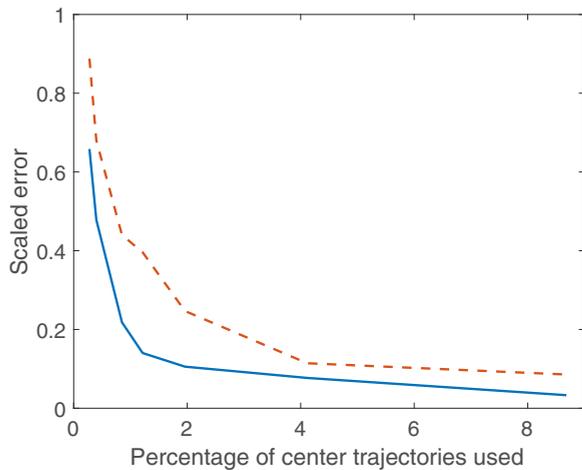


FIG. 5. Comparison of the quality of our DLN-based reconstruction scheme (solid line) with one based on a random sampling of the original velocity field (dashed line; see text for details). The two methods are compared by measuring the mean of the magnitude of the difference between the measured velocity and the reconstructed velocity at every point in the flow field, scaled by the root-mean-square velocity. This error is plotted against the number of trajectories used for the reconstruction, scaled by the total number of measured trajectories. Regardless of the number of trajectories used, the DLN-based scheme always outperforms the random scheme.

factor, essentially requiring a looser or tighter bound on the error in the linear approximation. As shown in Fig. 5, this

method does provide a reasonably good reconstruction of the velocity field, given that the experimentally measured velocity fields are indeed over-resolved. However, for any number of seed trajectories chosen, a DLN-based reconstruction always outperforms the simpler scheme, potentially by large factors.

## V. CONCLUSIONS

To summarize, we have defined spatially compact structures (DLNs) that, by construction, allow one to reconstruct the flow field inside as well as is dynamically possible using only the information from a single center trajectory. We showed that not only can these regions be computed in a real flow, but that by solving a set-cover problem we can isolate the most influential DLNs that then give us insight into the number of dynamically relevant degrees of freedom in the flow. Using these influential DLNs, we show that we can very accurately reconstruct the flow field using less information than would be required by a simpler, random compression scheme. These results suggest potential avenues both for thinking about new adaptive mesh refinement schemes for numerical simulations and for data compression schemes to allow easier data sharing.

## ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation under Grant No. CMMI-1563489.

- [1] D. A. Egolf, I. V. Melnikov, W. Pesch, and R. E. Ecke, Mechanisms of extensive spatiotemporal chaos in Rayleigh-Bénard convection, *Nature (London)* **404**, 733 (2000).
- [2] L. Sirovich, B. W. Knight, and J. D. Rodríguez, Optimal low-dimensional dynamical approximations, *Q. J. Appl. Math.* **48**, 535 (1990).
- [3] G. Berkooz, P. Holmes, and J. L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, *Annu. Rev. Fluid Mech.* **25**, 539 (1993).
- [4] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, Spectral analysis of nonlinear flows, *J. Fluid Mech.* **641**, 115 (2009).
- [5] P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, *J. Fluid Mech.* **656**, 5 (2010).
- [6] A. N. Kolmogorov, The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers, *Dokl. Akad. Nauk SSSR* **30**, 301 (1941).
- [7] S. B. Pope, *Turbulent Flows* (Cambridge University Press, Cambridge, UK, 2001).
- [8] G. Paladin and A. Vulpiani, Degrees of freedom of turbulence, *Phys. Rev. A* **35**, 1971 (1987).
- [9] Y. Kaneda and T. Ishihara, High-resolution direct numerical simulation of turbulence, *J. Turbul.* **7**, N20 (2006).
- [10] T. Ishihara, K. Morishita, M. Yokokawa, A. Uno, and Y. Kaneda, Energy spectrum in high-resolution direct numerical simulations of turbulence, *Phys. Rev. Fluids* **1**, 082403(R) (2016).
- [11] S. Wiggins, The dynamical systems approach to Lagrangian transport in oceanic flows, *Annu. Rev. Fluid Mech.* **37**, 295 (2005).
- [12] L. Fang, S. Balasuriya, and N. T. Ouellette, Disentangling resolution, precision, and inherent stochasticity in nonlinear systems, *Phys. Rev. Research* **2**, 023343 (2020).
- [13] S. Balasuriya, Stochastic sensitivity: A computable Lagrangian uncertainty measure for unsteady flows, *SIAM Rev.* **62**, 781 (2020).
- [14] S. Balachandar and J. K. Eaton, Turbulent dispersed multiphase flow, *Annu. Rev. Fluid Mech.* **42**, 111 (2010).
- [15] U. Piomelli, T. A. Zang, C. G. Speziale, and M. Y. Hussaini, On the large-eddy simulation of transitional wall-bounded flows, *Phys. Fluids A* **2**, 257 (1990).
- [16] G. N. Ivey and J. Imberger, On the nature of turbulence in a stratified fluid. Part I: The energetics of mixing, *J. Phys. Oceanogr.* **21**, 650 (1991).
- [17] S. K. Venayagamoorthy and J. R. Koseff, On the flux Richardson number in stably stratified turbulence, *J. Fluid Mech.* **798**, R1 (2016).
- [18] L. Fang, S. Balasuriya, and N. T. Ouellette, Local linearity, coherent structures, and scale-to-scale coupling in turbulent flow, *Phys. Rev. Fluids* **4**, 014501 (2019).
- [19] D. H. Kelley and N. T. Ouellette, Onset of three-dimensionality in electromagnetically driven thin-layer flows, *Phys. Fluids* **23**, 045103 (2011).
- [20] L. Fang and N. T. Ouellette, Multiple stages of decay in two-dimensional turbulence, *Phys. Fluids* **29**, 111105 (2017).

- [21] Y. Liao and N. T. Ouellette, Spatial structure of spectral transport in two-dimensional flow, *J. Fluid Mech.* **725**, 281 (2013).
- [22] D. H. Kelley and N. T. Ouellette, Using particle tracking to measure flow instabilities in an undergraduate laboratory experiment, *Am. J. Phys.* **79**, 267 (2011).
- [23] N. T. Ouellette, H. Xu, and E. Bodenschatz, A quantitative study of three-dimensional Lagrangian particle tracking algorithms, *Exp. Fluids* **40**, 301 (2006).
- [24] V. Chvatal, A greedy heuristic for the set-covering problem, *Math. Oper. Res.* **4**, 233 (1979).