Predicting two-dimensional turbulence

R. T. Cerbus and W. I. Goldburg

Phys. Rev. E **91**, 043003 — Published 6 April 2015

DOI: 10.1103/PhysRevE.91.043003

Predicting 2D Turbulence

R.T. Cerbus[1, 2, *] and W.I. Goldburg[1]

[1]Department of Physics and Astronomy, University of Pittsburgh, 3941 O'Hara Street, Pittsburgh PA 15260
[2]Fluid Mechanics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan

Prediction is a fundamental objective of science. It is more difficult for chaotic and complex systems like turbulence. Here we use information theory to quantify spatial prediction using experimental data from a turbulent soap film. At high Reynolds number, *Re*, where a cascade exists, turbulence becomes easier to predict as the inertial range broadens. The development of a cascade at low *Re* is also detected.

I. INTRODUCTION

According to many textbooks, a hallmark of turbulence is its unpredictability [1, 2]. Here we address this issue using experimental data from a turbulent soap film. The starting point is Shannon's information theory [3–5], where in Neil Gershenfeld's words, "...information is what you don't already know" [6]. Our experiment conveys information about the physical state of the system.

The entropy from information theory [4] (and the Lyapunov exponents from dynamical systems [7]) is a measure of the limit on our ability to predict. However, the theory does not tell us how to make a prediction. To fully address the issue of turbulence's predictability, we need to make a choice about how to predict (described below) [8]. We are looking for the answer to simple questions: how difficult is it to predict turbulence and how does the difficulty depend on Reynolds number, *Re*?

Our main finding is that prediction is sensitive to whether a turbulent cascade is present or not. Turbulence becomes easier to predict and more predictable when a cascade develops and then more so as *Re* increases.

The turbulent cascade envisioned by Richardson and described mathematically by Kolmogorov is the prevalent picture of turbulence [9]. In this picture, energy (or enstrophy in two dimensions) is transported across scales from some injection scale until it reaches a dissipative scale and the cascade ends. A cascade exists in both three dimensional (3D) and two dimensional (2D) turbulence, which is studied here. The statistical structure of the cascade has important consequences for prediction.

The central quantity in information theory is the entropy density h [4]. It is the information we receive per measurement (which in this case refers to a single velocity value), after already having measured an infinite amount of previous data. A large h implies that one does not know what is coming, *i.e.* the system is unpredictable. The value of h gives the limit on one's ability to predict but does not indicate how to make a prediction.

We could also ask how much we already do know. This is the excess entropy E, which is the information about correlations in the system [10, 11]. It is the reduction of unpredictability. Accurate prediction requires an amount of information at least equal to E [12]. Although E further characterizes our ability to predict, we still must decide how to do so.

Now we must decide how to make a prediction. Our choice is to only use the information contained in prior measurements (data) and make a statistical model. There are no specific assumptions made about the system. (An interesting example of where a specific model is used can be found in Ref. [13].) The model consists of a set of states and the probabilities to transition between them. Here the states are simply the basis used to represent the data. The states could be the measured velocities themselves, *i.e.*, 10 cm/s, -23 cm/s, *etc.* There is more than one way to define which states to use and potential benefits from choosing them cleverly.

Following the work of Crutchfield [8, 12, 14], we choose the states such that we maximize our ability to predict (up to the limit set by h) and at the same time minimize how much information we need to do so. The amount of information then needed to make the prediction is called C, the statistical complexity [8]. That is why C is a measure of the difficulty in making a prediction. Given the definition of E above, it is clear that $C \geq E$, but the system-specific reasons can vary and are not always clear [8, 12].

More details on h, E and C can be found in the Appendices A-D. It should be mentioned here, however, that to calculate the probabilities necessary for estimating h, E and C, we must bin the data according to some rule. The main results we show are for a binary rule where we only distinguish between a velocity above and below the mean. Other bins were used with qualitatively the same results.

This study focuses on predicting the spatial variations of turbulence. A prediction in space means that given the velocity u at a point x, one anticipates the velocity at some other point r away. We are making predictions about the velocity fluctuations in space. Prediction is normally associated with time [15, 16], but there are several reasons for considering the spatial alternative.

We know that the temporal and spatial features of turbulence are distinct. The fundamental work of Kolmogorov dealt only with the spatial structure of turbulence [9, 17]. Kraichnan and others have also shown that

[*] Electronic address: rtc17@pitt.edu

many of the essential features of turbulence are retained if one throws away temporal correlations but keeps spatial ones [18–20]. Thus, a treatment of spatial prediction is arguably of more fundamental interest than temporal prediction, at least for turbulence.

For a specific application, consider airplane flight. The typical cruise speed of a Boeing 747 is $V \simeq 250$ m/s [21]. Contrast this with the rms velocity fluctuations σ of "strong" atmospheric turbulence $\sigma \simeq 7$ m/s [22]. Since $\sigma/V \simeq 0.03$ is small, one must use Taylor's frozen turbulence hypothesis when discussing the turbulence the airplane encounters [2, 23]. In other words, an airplane flies fast enough to sample only the spatial variations of turbulence. There is not enough time for the turbulent velocity field to evolve temporally.

While this is a study of 2D turbulence, the analysis is not specific to this system. Our work serves as an experimental test bed for these tools, which can be used generally for other complex systems.

II. EXAMPLE

As a simple illustration of these ideas, consider a coin flipping experiment where each subsequent flip will be the same as the previous one with probability $P \in [0, 1]$ [25]. This is the statistical model for, *e.g.*, correlated random walks [26].

If $P = 0.5$ we have the usual fair coin toss experiment, with $h = 1$ and $C = E = 0$, since this system is maximally uncertain but statistically simple to predict with no information being shared between the past and future. In this fully random case ($P = 0.5$) both 0 and 1 predict the same future, so they are combined into a single causal state. Of course, with only one causal state, $C = 0$ automatically (see Eq. D1).

Consider now a slight deviation of P from 0.5. Now $C = 1$ since we will always need to know 1 bit of information (the previous flip) to predict the future. We can also calculate h and E (see Appendices B and C), which are plotted together with C vs. P in Fig. 1. Since $P > 0.5$ means more predictable, it is clear that h should decrease with increasing P, while E should increase.

This example highlights the difference between E and C, the crypticity $\chi \equiv C - E$ [12, 48]. Here $C = E + h$, which is a unique feature of this system being first-order Markovian [10]. The extra information needed to predict beyond E is due to the randomness still intrinsic in the causal states themselves. There are many examples for which $C \neq E$ [12, 27], but this is not always so.

An important lesson we learn from this example is that h, E and C were all necessary to understand this system's behavior. For P only slightly different from 0.5, h and E will still suggest a nearly random system, much like a slightly biased coin. The fact that C is large and not 0 (its random value), shows that there are important correlations not present in a simple biased coin system. The system is both unpredictable (large h) and difficult to predict (large C). A similar result will be found for the low Reynolds number flow in Sec. IV.

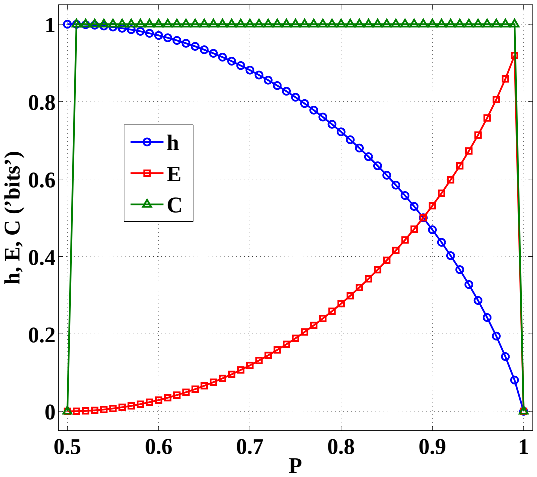

FIG. 1: Plot of the fundamental quantities h (◯), E (□) and C (△) for the simple example given here. Although h and E are continuous functions of P, C is not.

III. EXPERIMENTAL SETUP

Now consider a turbulent soap film, which is a good approximation to 2D turbulence since the film is only several μm thick [23, 28]. The soap solution is a mixture of Dawn (2%) detergent soap and water with 4 μm particles added for laser doppler velocimetry (LDV) measurements. Figure 2 contains a diagram of the experimental setup as well as thickness fluctuations visualized through thin film interference using a monochromatic light source. The thickness fluctuations act as a surrogate for velocity fluctuations [23, 28].

The soap film is suspended between two vertical blades. Nylon fishing wire connects the blades to the nozzle above and the weight below. The nozzle is connected by tubes to a valve and a top reservoir which is constantly replenished by a pump that brings the spent soap solution back up to the top reservoir. The flow is gravity-driven. Typical centerline speeds $\overline{u}$ are several hundred cm/s with rms fluctuations u' ranging roughly from 1 to 30 cm/s. The channel width w is usually several cm. The Reynolds number $Re = u'w/\nu$, where $\nu = 0.01$ cm^2/s is the kinematic viscosity, thus ranges from 10 to 10,000.

Turbulence is generated using several different protocols. We can (1) insert a row of rods (comb) perpendicular to the film, (2) replace on or both smooth walls with rough walls (saw blades) with the comb removed and possibly a rod inserted near the top [29], or (3) use a comb with smooth walls as in (1) but now very near the top of the soap film where the flow is still quite slow. The comb teeth are ~ 1 mm in diameter and several mm

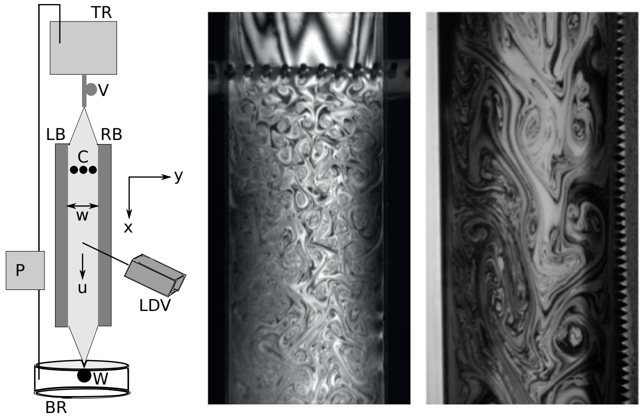

FIG. 2: Left: Experimental setup showing the reservoirs (TR, BR), pump (P), valve (V), comb (C), blades (LB, RB), LDV and weight (W). Middle: Fluctuations in film thickness from turbulent velocity fluctuations with smooth walls and a comb. Right: Thickness fluctuations with smooth and rough walls.

apart. The saw blade teeth are ~ 2 mm tall and wide.

When protocol (1) is used we almost always observe the direct enstrophy cascade [23, 28]. If procedure (2) is used, we can observe an inverse energy cascade [23, 28, 29], although this depends sensitively on the flux and w. When protocol (3) is used, we see no cascade at all.

The type of cascade is identified by calculating the one-dimensional velocity energy spectrum $\mathcal{E}(k)$, where $\frac{1}{2}u'^2 = \int_0^\infty \mathcal{E}(k)dk$. For the enstrophy cascade, $\mathcal{E}(k) \propto k^{-3}$ and for the energy cascade $\mathcal{E}(k) \propto k^{-5/3}$ [23, 28]. A number of measurements were taken above the blades where the flow is slower. For protocol (3), $\mathcal{E}(k)$ is flat and so apparently there is no cascade, although the flow is not laminar ($u' \neq 0$). See Fig. 3 for some representative spectra. In Fig. 4 the data for $Re < 100$ have a flat $\mathcal{E}(k)$.

In all cases, we measure the longitudinal (streamwise) velocity component at the horizontal center of the channel. The data rate is $\simeq 5000$ Hz and the time series typically had more than 10^6 data points. For this system the time series is really a spatial series by virtue of Taylor's frozen turbulence hypothesis [2, 9, 23, 28]. This means that the spatial variations are swept through the LDV's measuring point by the mean flow so quickly that it is as if the LDV were scanning a frozen-in-time velocity field. This distinction between spatial and temporal is essential, as discussed above and in Ref. [24].

IV. RESULTS

The quantities C, E and h are plotted vs. Re in Fig. 4. The data are roughly divided in Re into no-cascade (flat $\mathcal{E}(k)$ for $Re < 100$) and cascade (power law $\mathcal{E}(k)$ for $Re > 100$) regimes. Although C and E intersect at finite $Re \simeq 7000$ in Fig. 4, this meeting point depends on the analysis. In order to calculate probabilities from continuous data, one must bin the measurements. For different binning protocols we find a different meeting point. However, the Re-dependent behavior of h, E and C discussed below is the same. See Appendices A and D for more details on the treatment of the data.

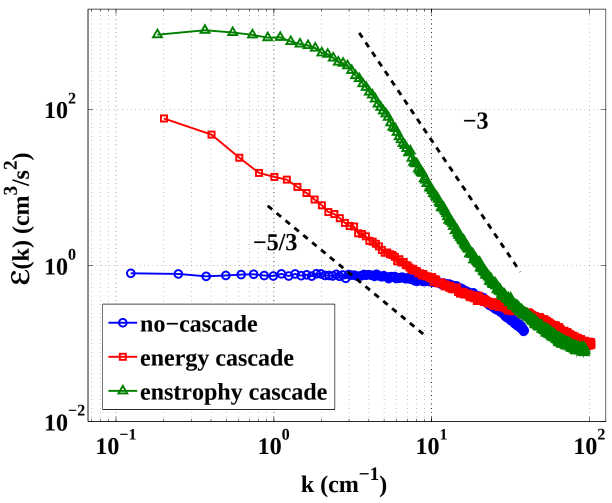

FIG. 3: Representative one-dimensional energy spectra in a log-log plot of $\mathcal{E}(k)$ vs. k. The enstrophy cascade (△) has a slope close to -3 while the energy cascade (□) has a slope close to -5/3. The flat curve (◯) has no cascade.

A. Cascade Turbulence

Now consider the behavior of h, E and C in the "cascade regime" of Fig. 4, $Re > 100$. At these values of Re, $\mathcal{E}(k)$ shows power law scaling as in Fig. 3. Both energy and enstrophy cascade data are present. We see from Fig. 4 that the unpredictability (h) is decreasing, the amount of information needed to predict (C) is also decreasing, while information about correlations (E) is increasing (all logarithmically). The opposite trend in Re for E and C is noteworthy. It is surprising that the behavior of h, E and C for $Re > 100$ does not depend on which cascade is present, only on whether or not there is a cascade at all.

The increase of E with Re can be understood from the traditional view that as Re increases, the "inertial range" of correlated scales broadens [9]. The increase in correlations across spatial scales is reflected by an increase in E. We can go further to suggest a connection between E and the broadness of the inertial range. Dimensional arguments suggest that the turbulent degrees of freedom go as $N \propto Re$ for the enstrophy cascade and $N \propto Re^{3/2}$ for the inverse energy cascade. In the 3D energy cascade, $N \propto Re^{9/4}$ [30]. Thus the behavior $E \propto \log_2 Re$ in Fig. 4 indicates that E is a logarithmic measure of the extent

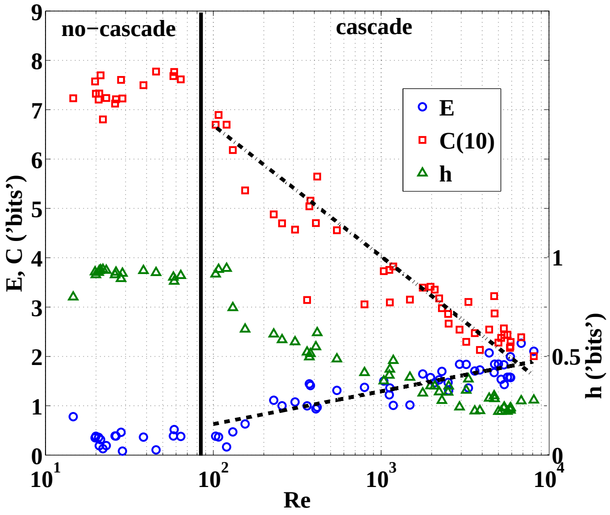

FIG. 4: The statistical complexity C (□), excess entropy E (○) and entropy density h (△) as functions of Re for binarized ($A = 2$) data (see Appendix A for details on binning). We plot h on a different scale for better visibility. The maximum value of h here is $\log_2 2 = 1$, which the no-cascade data for $Re < 100$ approach very closely. Here $L = 10$ and we used our MATLAB program with the χ^2 test to calculate C (see Appendix D for details). The lines are not fits to the data but are meant to suggest the behavior of C and E as functions of Re. For the cascade region, C and h are decreasing functions of Re while E increases. The vertical line separates the data according to whether there is a cascade or not.

of the inertial range.

An interpretation of the behavior of C is also suggested by the traditional picture of 2D turbulence [23, 28]. As Re grows, the inertial range broadens, and more of the velocity fluctuations come under the governance of the cascade. Thus, the randomness h will decrease, and because the cascade's structure is dominating, our prediction cost C decreases. This is the result of the general principle that patterns help us to predict [14]. Here the pattern is the cascade's structure.

Turbulence has traditionally been thought of as unpredictable [1, 2], an idea we have tested here with h, E and C. We see that the spatial predictability of (2D) turbulence is a function of Re. As Re increases we can predict further and more easily. This is in stark contrast to turbulence's increasing temporal unpredictability with Re, at least as evidence by numerical work [15, 16]. This reiterates the important difference between time and space in turbulence, which is of fundamental interest and practical importance (recall the airplane).

B. Transition to Cascade Turbulence

Next consider the region of Fig. 4 labeled "no-cascade". The absence of a cascade is evidenced by a lack of power law scaling in $\mathcal{E}(k)$ as in Fig. 3. Here h, E and C are relatively constant with respect to Re. It is notable that h is very near to the random (white noise) value of $\log_2 2 = 1$, which is nothing like laminar flow where $h = 0$. When a cascade emerges at $Re \simeq 100$, all three quantities begin to change noticeably. This change in behavior is decidedly different from the laminar to turbulent transition which only involves the onset of fluctuations [1, 30].

Simulations of 3D turbulence have shown that statistics of the velocity derivatives are gaussian (or subgaussian) up until a small value of the Reynolds number [31, 32]. Below this value of Reynolds number, there is a "regime which is a complex time-dependent flow rather than a turbulent one." They observe a transition similar to the one described here, evidenced primarily by non-gaussian velocity derivative statistics. (Recall that non-gaussian statistics are a general feature of fully developed turbulence [33].)

We can also use a more traditional tool from turbulence, the correlation function $c(r) \equiv \langle u(x)u(x+r)\rangle_x/u'^2$ plotted in Fig. 5 [9]. $c(r)$ has typically been thought of as a tool for determining the range of length scales over which u is correlated. $c(r)$ is telling us that for small $Re \leq 100$, the range of scales over which u is correlated is very small.

Figures 3 and 5 both indicate that for $Re \leq 100$ the data is like white noise. The values of $h \simeq 1$ and $E \simeq 0$ in Fig. 4 reinforce this interpretation. On the other hand, if the fluctuations were truly like white noise, then C should also be zero in this regime, which it is not. Recall that in the simple example from Sec. II, C is large when h and E are close to their random values. The data are nearly random but have an explicit albeit short dependence on the past which drives C from zero to its maximum value. If we were to only look at h (or E), we would miss that there is nontrivial (non-random) behavior for low Re.

We have yet to understand why self-similar turbulence emerges from this "complex, time-dependent flow" [31]. One sees from another nonlinear system, Rayleigh-Benard convection, that there is a lot to be learned even at modest levels of excitation [34].

The traditional approaches to the laminar-turbulent transition deal with instabilities of the laminar flow [1, 35]. Whether it is the quasi-periodicity of Landau [30] or the nonperiodicity of Ruelle and Takens [36], none of these approaches deal with the development of a Richardson or Kolmogorov cascade [37]. And yet a cascade is always present in "fully-developed turbulence" [9, 17]. How does this cascade emerge? New approaches and models are necessary to understand how cascade behavior develops out of a "complex, time-dependent flow" [31]. Since this development is clearly visible in Fig. 4, an information theory approach seems promising.

We further suggest an information-theoretic indicator of a cascade. Based on the above arguments, large E and $1/C$ should both indicate a well-developed cascade. With that in mind, we can also consider the "predictive

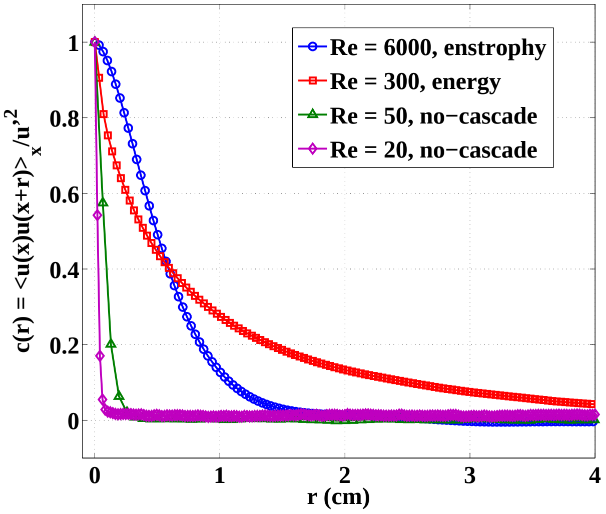

FIG. 5: The velocity autocorrelation function $c(r)$ plotted *vs.* r for several values of Re. For small Re, $c(r)$ quickly decays to zero, indicating little correlation in the velocity u. For larger Re, where Fig. 3 indicates spatial structure, there is a wider range of correlated scales. The $Re = 300$ curve has a longer correlation length L than the higher $Re = 6000$ curve presumably because this lower Re curve corresponds to an inverse energy cascade. The inverse energy cascade is supposed to involve larger length scales than the enstrophy cascade [23, 28]. Here L is defined as the distance at which $c(r)$ decays to $1/e$.

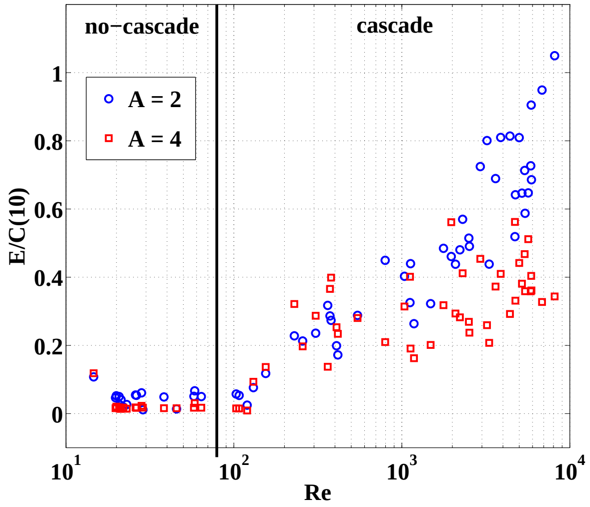

FIG. 6: The predictive efficiency E/C plotted vs. Re using the same data as in Fig. 4 as well as a quaternary partition $A = 4$ with partition walls placed symmetrically with respect to the mean (see Appendix A for details on binning). We used $L = 10$ for both partitions (see Appendix D). Here we find that E/C is increasing only after a cascade develops.

efficiency" E/C [38], which is an increasing function of Re, as shown in Fig. 6 for two different binning protocols. The ratio E/C tells us the fraction of the information needed to predict C that is due to correlations E. It is nearly zero when no cascade is present and grows smoothly after one has emerged. This shows that E/C is a nice tool for studying the transition to cascade turbulence.

Besides this cascade transition, the laminar to fluctuation transition is also of interest. Unfortunately, we are not able to access a truly laminar regime with our apparatus. For laminar flow and this geometry, $h = E = C = 0$ [8]. Looking at Fig. 4, and with the reasonable assumption that h and C are continuous functions of Re, one expects a local maximum in C and h at some low value of Re. This maximum would correspond to a special transition in the evolution of the flow between laminar and turbulent behavior. The observation of this maximum requires a different experimental setup.

V. CONCLUSION

The approach here is not limited to incompressible Navier-Stoke's turbulence. In fact it is useful for any nonlinear system, even those for which one does not know the equations of motion. When we think of turbulence in terms of information and prediction, we can make new distinctions and draw new insights. We have been able to highlight a cascade transition and have seen that spatially, turbulence is becoming easier to predict statistically as Re increases. As for our airplane, Figs. 4 and 6 bring bittersweet news. Although its passengers will certainly experience a rougher flight as Re increases, at least they won't be as surprised.

We would like to thank D. P. Feldman for explaining several concepts to us and for making his excellent lecture notes available online. C. J. Ellison was kind enough to explain some of the finer points of the formalism to us. We are also indebted to M. Bandi for providing numerous suggestions and insights. The criticisms and suggestions from several referees have also been beneficial. This work is supported by NSF Grant No. 1044105 and by the Okinawa Institute of Science and Technology (OIST). R.T.C. is also supported by a Mellon Fellowship through the University of Pittsburgh.

Appendix A: Data

The approach used here is data driven. We are given a data stream and use it to say something about the system that made it. The main assumption is that the system is stationary [4, 8]. We don't appeal to the Navier-Stoke's equation or any of Kolmogorov's universality assumptions [9, 17]. This method is generally applicable to many types of systems.

The formalism is now introduced. In the discussion that follows an uppercase U denotes the data (the random variable, the message) with possible velocity values $\mathcal{U}$ and the lowercase u denotes a particular member of that set. We can also consider groups of length L denoted by the set $\mathcal{U}^L$ and its particular members u^L. We are interested in treating a group because of the correlations that may exist between its members. Overhead arrows indicate a direction in the 1D data set relative to an arbitrary reference point x. For example, $\overrightarrow{U^L}$ refers to any block of data of size L taken to the right of x. For example, if $L = 3$, then a particular block $\overrightarrow{u^3}$ is as below

$$...u_{x-\Delta x}, u_x, \overrightarrow{u_{x+\Delta x}, u_{x+2\Delta x}, u_{x+3\Delta x}}, u_{x+4\Delta x}, ...$$

where Δx is the spatial resolution. If no L is mentioned, the block is (semi-)infinite.

Let U be a velocity component in the soap film, which is characterized by the experimental probability distribution $P(U)$. The focus is on the information shared between different directions $\overleftarrow{U}$ and $\overrightarrow{U}$ relative to the arbitrary point x [8, 39]. If we had data with explicit time dependence, we would talk about the past, future and present [8].

In order to use this formalism with turbulence, the continuous experimental data must be converted to symbols [40]. A partition is defined which assigns data values in specific ranges to unique symbols [11, 40]. This is usually referred to as binning the data. All experiments of continuous systems do this because of limited resolution ϵ. There are numerous previous studies where even binarizing a turbulent velocity signal has given more insight than traditional techniques [24, 40–42].

In this work we primarily use a binary partition (alphabet size $A = 2$) with the single partition wall located at the mean velocity. This smaller alphabet allows us to use a larger L with confidence and so cover a wider range of length scales in our analysis. Just as with h in Ref. [24], we have found that the general behavior of C and E with respect to Re is independent of the partition size; partitions of sizes $A = 4$, 8 gave similar results. Here the choice was made to use the same alphabet size A for all Re. This was done so that all data, if random, would have the same maximum value of $h = \log_2 A$. Thus, all data are treated at the same level of description. Of course, there are alternative choices for setting the partition size.

Appendix B: Entropy density h

We have already spoken of the entropy density h as a measure of unpredictability. The definition of entropy we are most familiar with is [3, 4]

$$H(U) = -\sum_{u \in U} p(u) \log_2 p(u), \tag{B1}$$

with units of "bits". This is the unpredictability of single data points given no immediate knowledge of any previous data points. An example of this would be estimating the unpredictability of letters in the English language based solely on the frequency of the letters and not on words.

Consider two examples. First look at a random string of 1s and 0s where $p(0) = p(1) = 0.5$. Here $H = 1$ is the maximum possible value. Next consider a periodic string such as "...0101...". Here again $p(0) = p(1) = 0.5$, and so here also $H = 1$. However, something is wrong since a periodic string should be perfectly predictable.

Since this definition of unpredictability misses any structure or correlations extending across scales, it is generalized to the block entropies [10, 11]

$$H_L = H(U^L) = -\sum_{u^L \in U^L} p(u^L) \log_2 p(u^L). \tag{B2}$$

This is the unpredictability of blocks of data. Of course, if we want to go back to looking at the unpredictability of a single data point, we can manipulate the H_L. The unpredictability of a single data point knowing L immediately previous data points is

$$h_L = H_{L+1} - H_L. \tag{B3}$$

The L-dependence is inconvenient, but if we make L large enough h_L will become L-independent (for most systems) [10, 11]. We are now ready to introduce the entropy density

$$h = \lim_{L \to \infty} h_L = H(\overrightarrow{U}^1 | \overleftarrow{U}) \tag{B4}$$

with an equivalent definition in terms of the conditional entropy [4]. This says explicitly how unpredictable a single data point is given all previous ones.

To further develop intuition for how h is associated with unpredictability, recall the Lyapunov exponents [7]. If a system is chaotic, its largest Lyapunov exponent λ is greater than 0 [7]. If our measurement has a resolution of ϵ and we enforce a tolerance of Δ, then our system is typically predictable up to a distance of $\frac{\log_2(\Delta/\epsilon)}{\lambda}$. Consider an information approach to the same problem. We choose to (or are forced to) have a particular partition size ϵ. This will correspond to $A = \frac{\max(U) - \min(U)}{\epsilon}$. Our maximum possible uncertainty in bits is $\log_2 A$. It will take $n = \frac{\log_2 A}{h}$ steps into the future to add up to this uncertainty and beyond this our data stream is unpredictable.

We estimate h using the limit of h_L from Eq. B3 in Eq. B4, as discussed in Ref. [24] and elsewhere [10, 11]. The undersampling bias in the $H(U^L)$ is corrected using Grassberger's method [11], although this did not affect the value of h very much. The h_L typically reached h at $L \simeq 10$.

Appendix C: Excess entropy E

While h tells us about the unpredictability of $\overrightarrow{U}^1$ given $\overleftarrow{U}$, we may also want to know how much we actually learned about $\overrightarrow{U}$ from $\overleftarrow{U}$. This is the excess entropy E. It is in some sense the opposite of unpredictability. E doesn't ask how much information we get from $\overrightarrow{U}$ upon measuring, but how much we don't get. We already know it. Stated mathematically [10, 11]:

$$E = H(\overrightarrow{U}) - H(\overrightarrow{U}|\overleftarrow{U}) \equiv I(\overrightarrow{U};\overleftarrow{U}) \qquad \text{(C1)}$$

where $I(\overrightarrow{U};\overleftarrow{U})$ is the mutual information shared between $\overrightarrow{U}$ and $\overleftarrow{U}$ [4].

This E is the information we got from $\overleftarrow{U}$ that reduces unpredictability. However, just like h, this is a statistical statement that doesn't tell us how to use that information. E does provide us with a lower bound on the amount of information needed to make predictions, since we need to account for all correlations. No matter how it's done, E bits will be necessary [10], otherwise we ignore some structure in the system.

An alternative expression is used to estimate E [10]:

$$E = \sum_{L=1}^{\infty}(h_L - h) \qquad \text{(C2)}$$

This calculation uses essentially the same quantities involved in estimating h. It turns out that for many chaotic systems, $h_L - h \propto 2^{-\gamma L}$ (γ is some constant independent of L) [10]. This empirical relationship has been shown to improve the estimation of E [10]. This expression will be used when possible.

Appendix D: Crutchfield complexity C

We now come to prediction using a statistical model. We must determine a set of special states called causal states S [8]. These will make up a minimal representation of our system for predictive purposes. In other words, we are trying to build the simplest possible statistical model of our data. For more details see Ref. [14]. There Shalizi *et al.* show that within the information theory framework, the approach described below is maximally predictive with a minimal amount of information needed.

A statistical model consists of a set of states and the transition probabilities between them. To determine S consider all unique blocks of data U^L. One would like to make L large to capture as many correlations as possible, but the finite amount of data means only finite L can be statistically reliable. For our data, $L \simeq 10$ is a good compromise. This L is also chosen because it is the value of L at which h_L typically reached h.

We now calculate the conditional probability $p(\overrightarrow{U}^L|\overleftarrow{u}^L)$ that any particular block $\overleftarrow{u}^L$ will give rise to any other block of the same length. If the conditional probability distributions conditioned on two blocks are the same, they are indistinguishable from a statistically predictive point of view. Thus block 1 and block 2 are equivalent, $u_1^L \sim u_2^L$, if $p(\overrightarrow{U}^L|\overleftarrow{u}_1^L) = p(\overrightarrow{U}^L|\overleftarrow{u}_2^L)$. This process incorporates pattern recognition by construction, which is why C was originally introduced as a complexity quantifier [8, 43].

All equivalent blocks are then combined and organized into a set of predictive causal states S. For example, suppose there are only three states u_1, u_2, and u_3 (forget about L here). If $p(\overrightarrow{U}|\overleftarrow{u}_1) = p(\overrightarrow{U}|\overleftarrow{u}_2) \neq p(\overrightarrow{U}|\overleftarrow{u}_3)$, then $u_1 \sim u_2 \nsim u_3$ and we have two causal states $s_1 = (u_1, u_2)$ and $s_2 = (u_3)$. Refer back to the example in Sec. II. It is apparent that if $P = 0.5$ (or 1) there is only one causal state, but if $P \neq 0.5$ (or 1), there are two causal states.

The Shannon information (entropy) contained in S is the statistical complexity [8, 44]

$$C = H[S] = -\sum_s p(s)\log_2 p(s). \qquad \text{(D1)}$$

This is the total amount of information needed to statistically reproduce the data, as we shall soon see.

Here is how this prediction work in practice: we find the causal states S as just described and so we also have the transition probabilities between the states S. Start out in some state u belonging to a particular s. Determine the next s' statistically using the known transition probabilities $p(s'|s)$ (the $'$ means the next step). Then determine a particular u' belonging to this s' according to $p(u'|s')$. This is symbolically represented by

$$u \xrightarrow{u\in s} s \xrightarrow{p(s'|s)} s' \xrightarrow{p(u'|s')} u'.$$

Then repeat. In this way the data is reproduced in a statistical sense. In summary, we can write down the probability of any u starting from any other u. This is statistical prediction.

We needed to know an amount of information $C = H[S]$ to carry out the above prediction program. That is, we need to ask (on average) C "yes" or "no" questions in order to find the current state of the system, and then predict from there. By design, this connects with the system's predictability, since organizing the message's parts into causal states will affect the value of C.

We can appreciate the distinction between C and h by considering an unbiased coin flip. The system is maximally unpredictable with $h = 1$, since one has no clue as to what will come next. In contrast, $C = 0$ since no information is needed for statistical prediction. There is only one causal state. This may strike the readers as strange, since random data is supposedly impossible to predict. This is only true if we insist on a prediction that has absolute certainty. Here we are predicting statistically.

When actually handling real data to identify S, one must deal with imperfections. These may be due to external noise or the finiteness of the amount of data. Regardless of the origin, one must set some sensible threshold to determine if two conditional probability distributions are the same, since they will never be identical. An example of some conditional probability distributions is shown in Fig. 7. Two of the distributions are similar, indicating that the two states belong to the same causal state. The third distribution is entirely different. The task is to choose a sensible metric to make this distinction objectively.

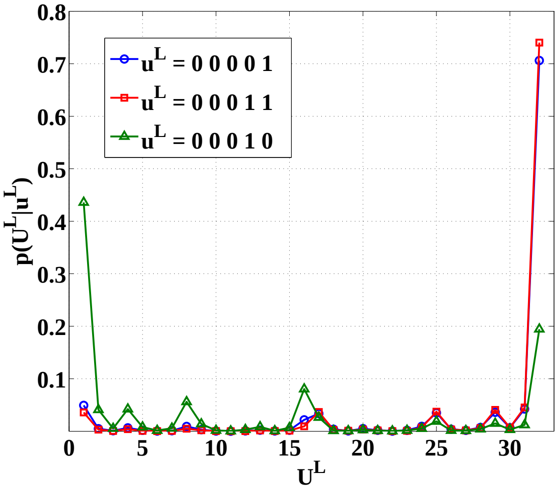

FIG. 7: An example of three conditional pdfs used to determine the causal states. The data used here is binarized turbulence data with $L = 5$ (giving a total of 32 possible states) and $Re = 3300$ (◯ = 00001, + = 00011, △ = 00010). The horizontal axis features all the possible future states while the vertical axis is the conditional probability that given a certain past state, any of those possible future states will occur. Here the distribution for states ◯ and □ appear similar while that for state △ is quite different.

We wrote a MATLAB program that uses a χ^2 test to compare conditional probability distributions [45]. We use a 0.95 confidence level, but the results are not sensitive to this choice. Results from our method are in good agreement with another frequently used algorithm [46, 47]. In the end, of course, the choice has an element of subjectivity to it.

Note that alternative expressions for h and E are [8, 27]

$$h = H[\overrightarrow{U}^1 | \overleftarrow{S}] \tag{D2}$$

and

$$E = I[\overrightarrow{S}; \overleftarrow{S}] = H[\overrightarrow{S}] - H[\overrightarrow{S}; \overleftarrow{S}] = C - H[\overrightarrow{S}; \overleftarrow{S}]. \tag{D3}$$

Equations D2 and D3 say that the causal states serve as a sufficient representation. Equation D2 also serves as a check on our determination of S by comparing h calculated with Eq. D2 with our previous method from Eqs. B3 and B4. From Eq. D3 we see that C may be different from E. Actually, it can be shown that $C \geq E$. The difference between these two has various interpretations.

The interpretation of Crutchfield and coworkers is that a system may have some "hidden" information, or crypticity $\chi = C - E$ [27, 48]. One might think that looking at the correlations in the infinite $\overleftarrow{U}$ would be enough to know how to predict, but we actually need a little more χ for prediction. This still comes from the data ($\overleftarrow{U}$) but one needed to build this statistical model to get it out. Wiesner and coworkers have interpreted χ as the information erased at each step in the system's evolution [38]. If we were to simulate this system on a computer, $k_B T \chi$ (where k_B = Boltzmann's constant and T is the computer's temperature) would be the minimum thermodynamic cost. This is an extension of Landauer's work on computation. He was the first to suggest that the erasure of information has a thermodynamic cost [49].

[1] D. J. Tritton, *Physical Fluid Dynamics* (Oxford University Press, USA, 1988)

[2] H. Tennekes, J. L. Lumley *A First Course in Turbulence* (MIT Press, Cambridge, Mass., 1972)

[3] C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1964)

[4] T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)

[5] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1962)

[6] N. Gershenfeld, *The Physics of Information Technology* (Cambridge University Press, Cambridge, 2000)

[7] G. L. Baker, J. P. Gollub, *Chaotic Dynamics: An Introduction* (Cambridge U. Press, 2nd Ed., Cambridge, 1996)

[8] J. P. Crutchfield, Nat. Phys. **8**, 17 (2012)

[9] P. A. Davidson, *Turbulence: An Introduction for Scientists and Engineers* (Oxford University Press, Oxford, 2004)

[10] J. P. Crutchfield, D. P. Feldman, Chaos **13**, 25 (2003)

[11] T. Schürmann, P. Grassberger, Chaos **6**, 414 (1996)

[12] C. J. Ellison, J. R. Mahoney, J. P. Crutchfield, J. Stat. Phys. **136**, 1005 (2009)

[13] G. Boffetta, A. Celani, A. Crisanti, A. Vulpiani, Phys. Fl. **9**, 724 (1996)

[14] C. R. Shalizi, J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001)

[15] E. Aurell, G. Boffetta, A. Crisanti, G. Paladin, A. Vulpiani, Phys. Rev. E **53**, 2337 (1996)

[16] C. E. Leith, R. H. Kraichnan, J. Atmos. Sci. **29**, 1041 (1972)

[17] A. N. Kolmogorov, "The local structure of turbulence in incompressible viscous fluids for very large Reynolds numbers," Dokl. Akad. Nauk. SSSR **30**, 299 (1941) (Proc. R. Soc. Lond. A **434** (reprinted))

[18] R. H. Kraichnan, Phys. Rev. Lett. **72**, 1016 (1994)

[19] B. I. Shraiman, E. D. Siggia, Nature **405**, 639 (2000)

[20] G. Falkovich, K. Gawedzki, M. Vergassola, Rev. Mod. Phys. **73**, 913 (2001)

[21] Boeing, http://www.boeing.com/boeing/commercial/... ...747family/pf/pf_400_prod.page, 4 July, 2014

[22] J. D. McMinn, AIAA Guidance Navigation and Control Conference, AIAA973532, (1997)

[23] H. Kellay, W. I. Goldburg, Rep. Prog. Phys. **65**, 845-894 (2002)

[24] R. T. Cerbus, W. I. Goldburg, Phys. Rev. E **88**, 053012 (2013)

[25] R. Quax, A. Appolloni, P. M. A. Sloot, Eur. Phys. J. Spec. Top. **222**, 1389 (2013)

[26] E. A. Codling, M. J. Plank, S. Benhamou, J. R. Soc. Interface **5**, 813 (2008)

[27] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, Phys. Rev. Lett. **103**, 094101 (2009)

[28] G. Boffetta, R. Ecke, Ann. Rev. Fluid Mech. **44**, 427 (2012)

[29] H. Kellay, T. Tran, W. I. Goldburg, N. Goldenfeld, G. Gioia, P. Chakraborty, Phys. Rev. Lett. **109**, 254502 (2012)

[30] L. D. Landau, E. M. Lifshitz, *Fluid Mechanics* (Butterworth-Heinemann, 2nd Ed., Oxford, 1987)

[31] J. Schumacher, K. R. Sreenivasan, V. Yakhot, New. J. Phys. **9**, 89 (2007)

[32] J. Schumacher, J. D. Scheel, D. Krasnov, D. A. Donzis, V. Yakhot, K. R. Sreenivasan, PNAS **111**, 10961 (2014)

[33] K. R. Sreenivasan, R. A. Antonia, Ann. Rev. Fl. Mech. **29**, 435 (1997)

[34] L. P. Kadanoff, Phys. Today **54**, 34 (2001)

[35] A. Brandstäter, J. Swift, H. L. Swinney, A. Wolf, J. D. Farmer, E. Jen, J. P. Crutchfield, Phys. Rev. Lett. **51**, 1442 (1983)

[36] D. Ruelle, F. Takens, Commun. Math. Phys. **20**, 167 (1971)

[37] H. L. Swinney, J. P. Gollub, Phys. Today **31**, 41 (1978)

[38] K. Wiesner, M. Gu, E. Rieper, V. Vedral, Proc. Roy. Soc. A **468**, 4058 (2012)

[39] J. P. Crutchfield, D. P. Feldman, Phys. Rev. E **55**, R1239 (1997)

[40] C. S. Daw, C. E. A. Finney, E. R. Tracy, Rev. Sci. Instrum. **74**, 915 (2002)

[41] A. J. Palmer, C. W. Fairall, W. A. Brewer, IEEE Trans. Geo. Remote Sensing **38**, 2056 (2000)

[42] M. Lehrman, A. B. Rechester, Phys. Rev. Lett. **87**, 164501 (2001)

[43] D. P. Feldman, J. P. Crutchfield, Phys. Lett. A **238**, 244 (1998)

[44] J. P. Crutchfield, K. Young, Phys. Rev. Lett. **63**, 105 (1989)

[45] E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses, 3rd Ed.* (Springer, New York, 2005)

[46] C. R. Shalizi, K. L. Shalizi, J. P. Crutchfield, arXiv:cs/0210025v3 [cs.LG]

[47] C. R. Shalizi, K. L. Shalizi, *An Algorithm for Building Markov Models from Time Series*, http://vserver1.cscs.lsa.umich.edu/∼crshalizi/CSSR/, 15 May 2013.

[48] J. R. Mahoney, C. J. Ellison, R. G. James, J. P. Crutchfield, Chaos **21**, 037112 (2011)

[49] R. Landauer, Phys. Lett. A **217**, 188 (1996)