Learning dominant physical processes with data-driven balance models

Jared L. Callaham[1*], J. Nathan Kutz[2], Bingni W. Brunton[3], and Steven L. Brunton[1]

[1] Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, United States

[2] Department of Applied Mathematics, University of Washington, Seattle, WA 98195, United States

[3] Department of Biology, University of Washington, Seattle, WA 98195, United States

Abstract

Throughout the history of science, physics-based modeling has relied on judiciously approximating observed dynamics as a balance between a few dominant processes. However, this traditional approach is mathematically cumbersome and only applies in asymptotic regimes where there is a strict separation of scales in the physics. Here, we automate and generalize this approach to non-asymptotic regimes by introducing the idea of an *equation space*, in which different local balances appear as distinct subspace clusters. Unsupervised learning can then automatically identify regions where groups of terms may be neglected. We show that our data-driven balance models successfully delineate dominant balance physics in a much richer class of systems. In particular, this approach uncovers key mechanistic models from the past hundred years in turbulence, nonlinear optics, geophysical fluids, and neuroscience.

Keywords– Physical modeling, machine learning, data-driven modeling, asymptotics, unsupervised learning, subspace clustering

1 Introduction

Across the engineering and physical sciences, decades of experimental and theoretical efforts have produced accurate and detailed physics-based models. The success of first principles modeling has resulted in governing equations describing a wide range of physics, including fluids, plasmas, combustion, atmospheric dynamics, electromagnetic waves, and quantum mechanics. However, it is well known that persistent behaviors are often determined by the balance of just a few dominant physical processes. This heuristic, which we refer to in general as *dominant balance*, has played a pivotal role in our study of systems as diverse as turbulence [1, 2], geophysical fluid dynamics [3–5], fiber optics [6, 7], and the earth's magnetic field [8]. It is also thought to play a role in the emerging fields of pattern formation [9–12], wrinkling [13], buckling [14], droplet formation [15, 16], electrospinning [17], and biofilm dynamics [18]. These balance relations, or order parameters [9], provide reduced-order mechanistic models to approximate the full complexity of the system with a tractable subset of the physics.

The success of dominant balance models is particularly evident in the field of fluid mechanics. The Navier-Stokes equations describe behavior across a tremendous range of scales, from water droplets to supersonic aircraft and hurricanes. Thus, much of our progress has required simplifying the physics with nondimensional parameters that determine which terms are important for a specific problem. Perhaps the most well known dimensionless quantity, the Reynolds number, describes the balance between inertial and viscous forces in a fluid. Other nondimensional numbers capture the relative importance of inertial and Coriolis forces (Rossby number), inertia and buoyancy (Froude number), and thermal diffusion and convection (Rayleigh number), among dozens of other possible effects. In many situations, the magnitude of these coefficients determine the important mechanisms at work in a flow; further, they determine which mechanisms

* Corresponding author (jc244@uw.edu)

Python code: github.com/jcallaham/dominant-balance

may be safely neglected. This approach has been especially important in making experimentally observable predictions for the profiles and scaling of wall turbulence [1, 19–25]. Similarly, in geophysical flows, balance arguments bypass the incredible complexity of the ocean and atmosphere to identify driving mechanisms such as geostrophy, the thermal wind, Ekman layers, and western boundary currents [3, 4]. Lighthill, one of the most influential fluid dynamicists of the 20th century, often relied on dominant balance arguments as physical motivation for his mathematical analyses [26–28]. Beyond fluid mechanics, asymptotic methods have been crucial in characterizing a diverse range of physical behavior.

More recently, modern developments in scientific computing have revolutionized our understanding of complex systems by enabling high-fidelity models that quantify multi-scale spatiotemporal interactions. At the same time, advanced tools from statistics have enabled the analysis of this increasing wealth of data. However, dominant balance models are still typically derived by hand using tedious scaling analysis or asymptotic expansions in limiting regimes. This severe restriction explains why such a powerful technique has not found even wider traction; many systems of practical or basic research interest lie between the extremes where scaling analysis can be unambiguously applied.

There is an exciting opportunity to leverage data-driven methods to identify dominant balance physics in these more challenging applications. Data-driven modeling is already driving changes in how we approach problems from control [29–32] to turbulence modeling [33] and forecasting [34, 35]. Indeed, some studies have addressed the dominant balance problem by using expert knowledge to design application-specific clustering algorithms, for example in a transitional boundary layer [36, 37] and stratified turbulence [38], in the latter case confirming the results of prior scaling analyses [39, 40]. Although these results are encouraging, to our knowledge the general challenge of identifying local dominant balance regimes from data remains open; our paper aims to address this gap.

In this work, we develop a generalized data-driven method to identify dominant balance regimes in complex physical systems. Beginning from the full evolution equations, we treat each term as a coordinate in an "equation space". Dominant balance relations have a natural geometric interpretation in this space, allowing a combination of unsupervised clustering and sparse approximation to automatically identify regions where groups of terms have negligible contributions to the local dynamics. We explore the proposed method on several systems, including a turbulent boundary layer (shown in Fig. 1), electromagnetic pulse propagation in an optical fiber, geostrophic balance in the Gulf of Mexico, and a biophysical model of a bursting neuron. In each case, we recover the expected balance relations from classical scaling analysis. The apparent ubiquity of the dominant balance phenomenon confirms a long-standing heuristic in physical sciences, while the ability to identify spatiotemporally local balance models via a data-driven approach opens new opportunities in a broad range of applications.

If successful, nonasymptotic data-driven methods could be used to better understand the behavior of more exotic dynamics such as non-Newtonian turbulence [41], hydrodynamic quantum analogues [42], and extreme event triggering [43], or to study important transitional behavior in cases where the asymptotics are already well known [44–48]. In the latter case, a clear understanding of the active mechanisms has proven crucial to successful control strategies [49, 50]. We may even be able to identify local dominant balance behavior in spatiotemporal systems without clear governing equations, such as neuroscience [51], epidemiology [52], ecology [53], active fluids [54–56], and schooling [57]. Automatic segmentation may also inform efficient numerical methods, in the vein of shock-capturing schemes [58], adaptive mesh refinement [59], or hybrid turbulence modeling [60]. It is our hope that this approach will shed light on more exotic physical processes that have remained elusive to traditional analysis.

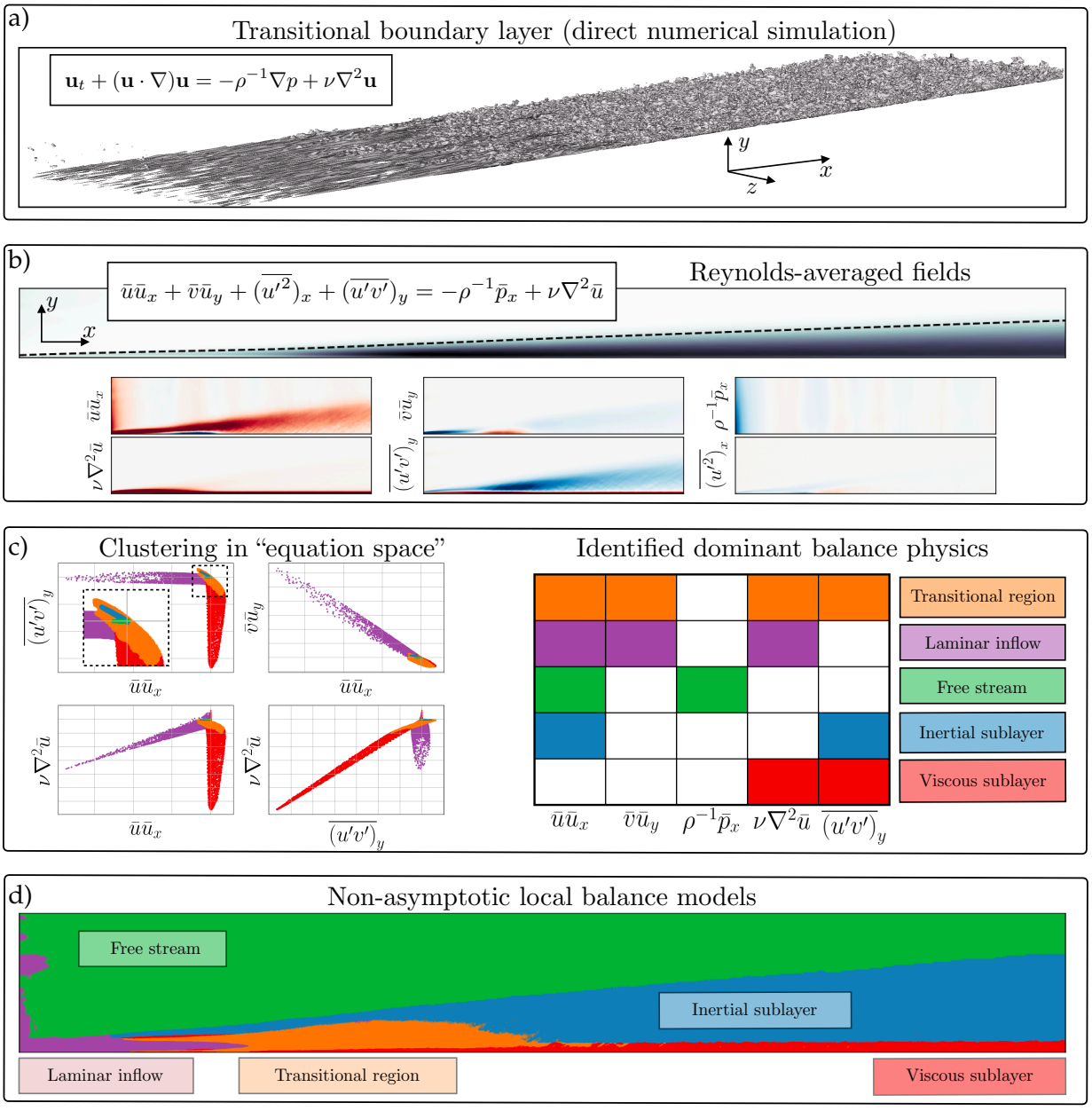

Figure 1: Schematic of the dominant balance identification procedure applied to a turbulent boundary layer. High-resolution direct numerical simulation results (a, visualized with a turbulent kinetic energy isosurface) are averaged to compute the Reynolds-averaged Navier-Stokes equations (b). The equation space representation of the field enables clustering and sparse approximation methods to extract the distinct geometrical structures in the six-dimensional space corresponding to dominant balance physics (c). Finally, the entire domain can be segmented according to these interpretable balance models, identifying distinct physical regimes (d). The equations and classical scaling analysis are discussed in Sec. 3.2.

2 Unsupervised dominant balance identification

In many fields of physics, painstaking analyses have produced models that are capable of describing a wide range of physical phenomena. However, it is well understood that the full complexity of such models is not always necessary to describe the local behavior of a system. We find that in many regimes the dynamics are governed by just a subset of the terms involved in the global description. For example, a general evolution equation for the field $u(x,t)$ on the domain $(x,t) \in \mathcal{D}$ can be written as

$$\mathcal{N}(u) = \sum_{i=1}^{K} f_i(u, u_x, u_{xx}, \dots, u_t, \dots) = 0. \tag{1}$$

Classically, this equation would be derived from fundamental physics (e.g. Maxwell's equations or the Navier-Stokes equations), but it could result from a model discovery procedure [61–63].

Consider an "equation space" where each coordinate is defined by one of the K terms in Eq. (1). At each point (x,t) in space and time, each of the K terms f_i in the governing equations (1) may be evaluated at $u(x,t)$, resulting in a vector $\mathbf{f} \in \mathbb{R}^{\mathrm{K}}$:

$$\mathbf{f}(x,t) = \begin{bmatrix} f_1(u(x,t),\dots) & f_2(u(x,t),\dots) & \cdots & f_K(u(x,t),\dots) \end{bmatrix}^T. \tag{2}$$

By construction, $\mathbb{1}^T \mathbf{f}(x,t) = \mathcal{N}(u) = 0$ for all $(x,t) \in \mathcal{D}$. Simulated or measured field data is typically discretized, so the domain is approximated by N spacetime points: $\mathcal{D} \approx \left\{(x,t)^j \mid j = 1, 2, \dots, N\right\}$. The field at each of these points corresponds to a point in equation space.

We define a dominant balance regime as a region $\mathcal{R} \subset \mathcal{D}$ where the evolution equation is approximately satisfied by a subset of $p < K$ of the original terms in the equation; the remaining terms may be neglected. In this case $\mathbf{f}(x,t)$ will have near-zero entries corresponding to negligible terms when $(x,t) \in \mathcal{R}$. Geometrically, the field is approximately restricted to p of the original K dimensions of the equation space, resulting in a subspace that is aligned with the active p terms.

This geometric perspective on dominant balance physics leads naturally to segmentation via unsupervised clustering. For example, the Gaussian mixture model (GMM) framework learns a probabilistic model by assuming the data are generated from a mixture of Gaussian distributions with different means and covariances [64]. The learned covariances for each cluster can then be interpreted in terms of active and inactive terms in the evolution equation. The N spacetime points in $\mathcal{D}$ are used to train a mixture model; the algorithm treats points from a dominant balance regime as if they were generated from a distribution with near-zero variance in the directions corresponding to negligible terms. Data beyond the original inputs can efficiently be assigned to a balance model using the trained GMM.

In practice, there is no reason to expect the points will even approximate a mixture of Gaussian distributions. We therefore expect that the number of clusters required to capture all of the relevant physics will exceed the number of distinct balance regimes, resulting in redundant clusters. Furthermore, there is some ambiguity in the interpretation of "near-zero variance". We address both of these issues using sparse principal components analysis (SPCA) [65], which uses ℓ_1 regularization to extract a sparse approximation to the leading principal component. If a cluster describes a dominant balance regime, it should be well-described by its direction of maximum variance. Moreover, this leading principal component should have many near-zero entries. We apply SPCA to the set of points in each GMM cluster and take the active terms in the cluster to be those which correspond to nonzero entries in the sparse approximation to the leading principal component. The number of models can then be reduced by grouping clusters with the same set of active terms (or equivalently, the same sparsity pattern in the SPCA approximation).

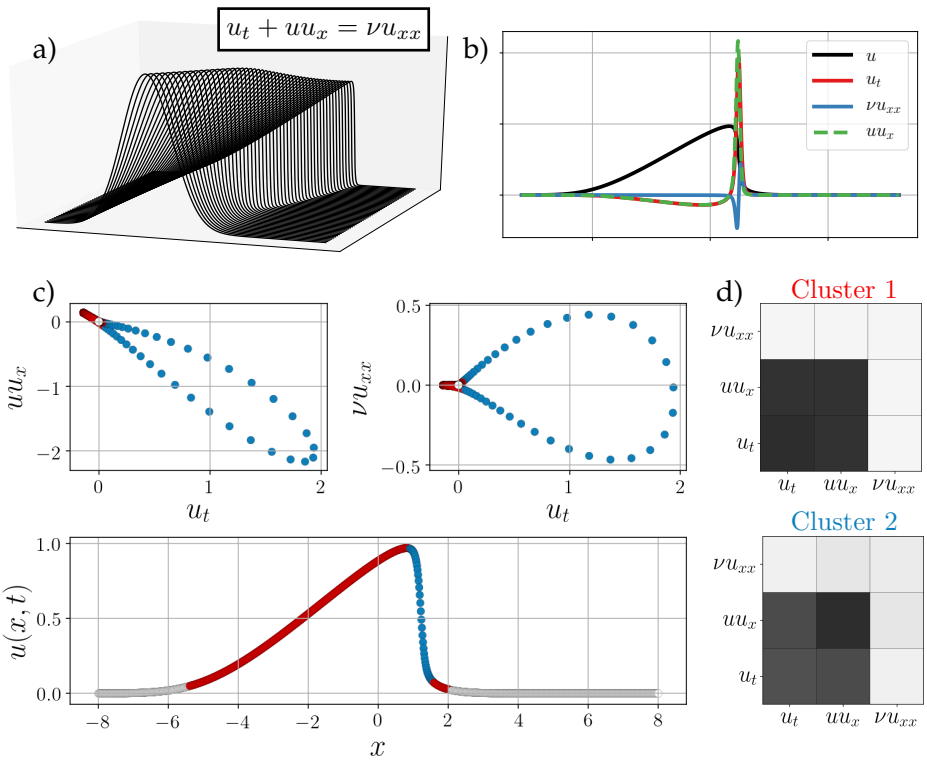

Figure 2: Example of dominant balance identification on the viscous Burgers' equation (a), with constituent terms shown in (b). The viscous term acts to diffuse sharp gradients and prevent formation of a discontinuous shock, but away from the shock front the dynamics are essentially inviscid. Away from the shock front, the field is approximately restricted to the $\nu u_{xx} = 0$ plane (c). This is reflected in the covariance matrices learned by the Gaussian mixture model (d).

Dominant balance identification can be seen as a localized active subspace analysis in equation space [66]. Rather than assuming that there is a global decomposition into approximately active and inactive subspaces, we simultaneously search for subspaces corresponding to different balance relations and the regions of the domain where the dynamics are well-described by this subspace.

For example, one of the simplest models that demonstrates dominant balance is the viscous Burgers' equation, shown in Fig. 2. Shocks form from the nonlinear advection and are dissipated by the viscous term. Away from the shock front, however, the gradients of the field are relatively weak, so viscosity does not contribute significantly to the dynamics. Figure 2 demonstrates the balance identification procedure applied to a snapshot of the viscous Burgers' equation example. Most of the field is classified into two clusters, corresponding to either no dynamics or an inviscid balance between acceleration and advection. Only a narrow slice along the shock front belongs to a cluster in which viscosity is active.

In simple cases, this two-step GMM-SPCA procedure might be replaced with a hard threshold; if a term exceeds some value ϵ it is "on". However, the proposed method offers two main advantages over thresholding. First, the idea of dominant balance has a natural geometric interpretation in equation space, thereby avoiding setting an arbitrary threshold for which diagnostics and interpretation may not be straightforward. Second, our method considers the *local, relative* importance of terms, whereas thresholding describes *global, absolute* importance. For example, this distinction

is significant in multiscale systems with some background process underlying intermittent bursts of activity. The intermittency is dominated by a balance between terms which may be much larger than the background process, although the dynamics during quiescent periods would be determined primarily by the background process. In this case an absolute thresholding method would either choose the background process to be always on or always off, whereas a relative approach recognizes that the dominant local balance simply changes during the intermittent activity. This is illustrated in Sec. 3.5, where we investigate a Hodgkin-Huxley-type model of spiking neuron, generalized to introduce multiscale bursting behavior.

3 Results

We now apply the dominant balance identification method to a range of physics with varying complexity: unsteady vortex shedding past a cylinder at Reynolds number 100; the mean field of a turbulent boundary layer; optical pulse propagation in supercontinuum generation; geostrophy in the Gulf of Mexico; and a Hodgkin-Huxley-type model of a biological neuron. Figure 3 shows a summary of the results, including slices of the equation space representations, identified balance models, and segmented fields. In each case, the results are consistent with classical scaling analyses and known physical behavior. Descriptions of the models and code used to generate this data are presented in Appendix A and are available online.

3.1 Flow past a circular cylinder at Re $= 100$

Governing equations and analytic scaling. Flow past a cylinder at moderate Reynolds number is a prototypical flow configuration for bluff body wakes. The wake transitions from steady laminar flow to periodic vortex shedding via a Hopf bifurcation at Re ≈ 47. The transition from linear instability to a stable limit cycle is itself a fascinating example of dominant balance in fluid mechanics and dynamical systems. The quadratic nonlinearity, initially inactive in the linear regime, mediates energy transfer between the mean flow and instability modes, deforming both until an energy balance is reached in the periodic limit cycle. This nonlinear stability mechanism was first described by Stuart and Landau [67, 68] and later employed for reduced-order modeling [69].

Even in the stable limit cycle, however, the local dynamics of the flow vary widely throughout the domain, highlighting mechanisms that give rise to von Kàrmàn-type vortex streets in a wide variety of flows. This unsteady, incompressible, viscous flow is governed by the two-dimensional Navier-Stokes equations:

$$\tilde{\mathbf{u}}_t + (\tilde{\mathbf{u}} \cdot \nabla)\tilde{\mathbf{u}} = -\frac{1}{\rho}\nabla\tilde{p} + \nu\nabla^2\tilde{\mathbf{u}}, \tag{3}$$

where $\tilde{\mathbf{u}}$ is the velocity field, $\tilde{p}$ is the pressure, ρ is the density, and ν is dynamic viscosity. Of course, these equations themselves involve some degree of approximation, ignoring effects such as compressibility and gravity, making use of the Newtonian form of the stress tensor, and assuming Fickian diffusion, though they have proven highly accurate when applied in the correct regime. Nevertheless, there are distinct regimes in this simple wake flow.

For the wake behind a circular cylinder, the most relevant scales are the cylinder diameter L and free-stream velocity U. Dimensional analysis then suggests that

$$\tilde{\mathbf{u}} \sim U, \qquad \tilde{p} \sim \nu U^2, \qquad \nabla(\cdot) \sim \frac{1}{L}, \qquad \frac{\partial}{\partial t}(\cdot) \sim \frac{U}{L}.$$

Nondimensionalizing with respect to these scales, we find that the viscous term is smaller than the others by a factor of the Reynolds number, Re $= UL/\nu$, resulting in the familiar nondimensional

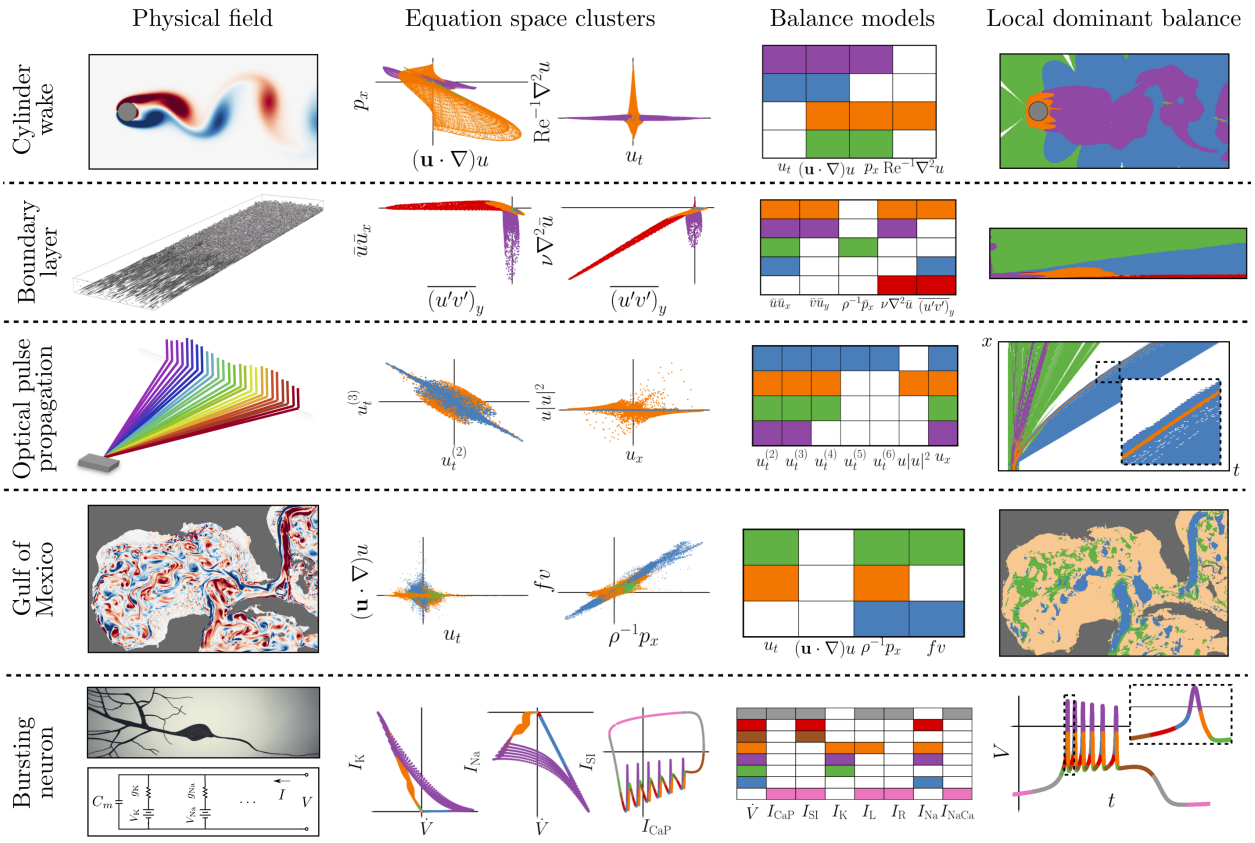

Figure 3: Dominant balance physics identified across a range of systems. For each case, a visualization of the system is shown on the left, followed by 2D views of the feature space colored by the identified balance relation, a key describing the active terms in each model, and the original field colored by the local balance. From top: a bluff body wake at moderate Reynolds number, a boundary layer in transition to turbulence, pulse propagation in an optical fiber, surface currents in the Gulf of Mexico, and a Hodgkins-Huxley model for an intrinsically bursting neuron.

form of the Navier-Stokes equations:

$$\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \frac{1}{\mathrm{Re}}\nabla^2\mathbf{u}. \tag{4}$$

The variables and operators have been nondimensionalized according to the previous scales. For even moderately large Reynolds numbers, we would expect the flow to behave in an approximately inviscid manner away from the cylinder. Thus, structures formed in the near-wake region will be advected downstream by the mean flow with only weak dissipation, as observed in the vortex street.

Near the cylinder, the no-slip boundary conditions due to viscosity change the behavior qualitatively. If we examine the flow at a point a distance $\delta \ll L$ from the wall, then δ is a more appropriate length scale for the gradients. However, since the near-wall flow varies on a similar timescale to the wake, suppose that U/L is still a good scale for the time derivative. The various terms then scale as

$$\tilde{\mathbf{u}}_t \sim \frac{U^2}{L}, \qquad (\tilde{\mathbf{u}}\cdot\nabla)\tilde{\mathbf{u}} \sim \frac{U^2}{\delta} \qquad -\frac{1}{\rho}\nabla\tilde{p} \sim \frac{U^2}{\delta}, \qquad \nu\nabla^2\tilde{\mathbf{u}} \sim \frac{\nu U}{\delta^2}.$$

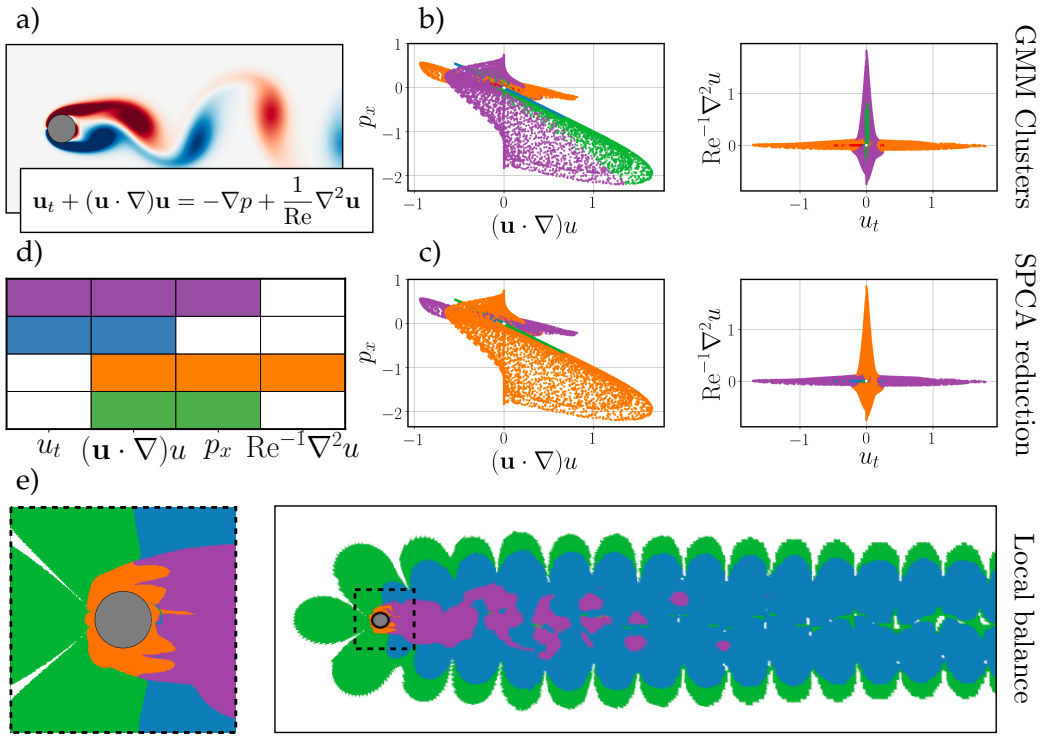

Figure 4: Vorticity snapshot for the wake behind a cylinder at $\mathrm{Re} = 100$ (a). A Gaussian mixture model (GMM) assigns field points to clusters by looking for groups with distinct mean and covariance (b). For instance, some clusters vary mainly in the acceleration-advection directions, while others vary principally in the viscous-advection directions. We would expect these to represent the far-field and boundary regions, respectively. This is confirmed by the sparse principal components analysis (SPCA) reduction, where clusters with significant nonzero variance in the same directions are grouped together (c). These directions can be interpreted as active terms in the balance relation (d). As anticipated, the region near the cylinder is dominated by a balance between viscosity and advection and pressure forces, while the far wake is approximately inviscid (e).

We find that the acceleration term is now smaller by a factor of δ/L, and expect the viscous term to be balanced by advection and the pressure gradient. The relatively strong gradients near the wall give rise to the vortex structures which characterize the wake.

Identified dominant balance. Figure 4 shows an example vorticity field along with views of the 4D equation space corresponding to Eq. (4). Although the method treats space and time equivalently, here we freeze time and explore a single snapshot; since the flow is periodic we expect the results to be representative. The visualization in equation space clearly reveals signatures of balance relations. One set of GMM clusters is nearly restricted to the the zero-viscosity plane, while another has reduced variance in the acceleration direction. The sparse approximations to the leading principal components of each cluster confirms this intuition; we use SPCA to construct balance models by grouping the Gaussian models with non-negligible variance in the same directions. As expected, the far wake is approximately inviscid, while the region near the cylinder is dominated by a balance between viscosity, pressure, and advection. This method also identifies other approximate regions, such as a low-pressure-gradient balance between acceleration and advection (blue), slowly varying potential flow (green), and a far-field region with near-zero dynamics (white).

3.2 Turbulent boundary layer

One of the major breakthroughs in the study of fluid mechanics in the 20th century was the development of boundary layer theory [1, 70]. In many practical applications fluids can be treated as inviscid, but close to solid boundaries strong velocity gradients lead to significant viscous forces. Prandtl showed in 1904 that careful scaling analysis applied to the governing Navier-Stokes equations reveals distinct regimes where the behavior of the fluid is essentially determined by a small subset of the full equations. In turn, these balance relations can be used to derive powerful scaling laws such as the so-called "law of the wall".

Although such analyses can be intractable for general turbulent flows, one of the most important canonical configurations is zero pressure gradient flow over a flat plate parallel to the free stream velocity. The zero pressure gradient ensures that the free-stream velocity is constant in the streamwise direction at large distances from the wall. This flow is statistically two-dimensional; the configuration does not vary in the cross-stream direction so the mean flow only varies in the streamwise and wall-normal directions.

Governing equations and analytic scaling. After performing the Reynolds decomposition of the variables into mean and fluctuating components, e.g. $u = \bar{u} + u'$, the mean flow is determined by the Reynolds-averaged Navier-Stokes (RANS) equations. For the streamwise mean velocity $\bar{u}$, the equation is

$$\bar{u}\bar{u}_x + \bar{v}\bar{u}_y = \rho^{-1}\bar{p}_x + \nu\nabla^2\bar{u} - (\overline{u'v'})_y - (\overline{u'^2})_x. \tag{5}$$

The terms on the left represent mean flow advection, while those on the right are the pressure gradient, viscosity, wall-normal Reynolds stress, and streamwise Reynolds stress, respectively.

One of the challenges in studying this flow is that there are multiple length scales. Following [74], we may consider a streamwise length scale L, a wall-normal length scale ℓ, and a viscous length scale $\eta = \nu/u_\tau$, where u_τ is the "friction velocity" associated with the shear stress at the wall.

Beginning with the "outer" region of the boundary layer (where $y \gg \eta$), suppose the mean streamwise velocity $\bar{u}$ scales with the free stream U_∞, while the turbulent fluctuations u', v' scale with u_τ. As with the previous example, assume that the derivatives scale with the corresponding length scale, so that for instance $(\cdot)_y \sim 1/\ell$. For instance, the continuity equation $\bar{u}_x + \bar{v}_y = 0$ implies that $\bar{v} \sim U_\infty(\ell/L)$. By this reasoning typically we would expect the mean velocity gradient $\bar{u}_y$ to scale with U_∞/ℓ, but as argued in [74], the gradients in the outer part of the layer are much weaker than near the wall, and empirically a better estimate is $\bar{u}_y \sim u_\tau/\ell$. Then for the streamwise momentum equation we find

$$\bar{u}\bar{u}_x \sim \frac{U_\infty^2}{L}, \quad \bar{v}\bar{u}_y \sim \frac{u_\tau U_\infty}{L} \quad \nu\bar{u}_{xx} \sim \frac{\nu U_\infty}{L^2}, \quad \nu\bar{u}_{yy} \sim \frac{\nu u_\tau}{\ell^2}, \quad (\overline{u'v'})_y \sim \frac{u_\tau^2}{\ell}, \quad (\overline{u'^2})_x \sim \frac{u_\tau^2}{L},$$

and the pressure gradient is negligible by construction. Since $L \gg \ell$ we neglect the streamwise Reynolds stress compared to the wall-normal term. On the other hand, since $U_\infty \gg u_\tau$, we can assume the mean flow advection is dominated by the streamwise component $\bar{u}\bar{u}_x$. Finally, the viscous terms are smaller than the advection by a factor on the order of the Reynolds number $\mathrm{Re}_L = U_\infty L/\nu \gg 1$. The outer part of the boundary layer is then determined by an inertial balance between streamwise mean flow advection and wall-normal Reynolds stress:

$$(\overline{u'v'})_y = -\bar{u}\bar{u}_x. \tag{6}$$

However, this relation cannot describe the near-wall regime, where viscosity is known to be important. In this region we expect the wall-normal derivatives to scale with $(\cdot)_y \sim 1/\eta = u_\tau/\nu$.

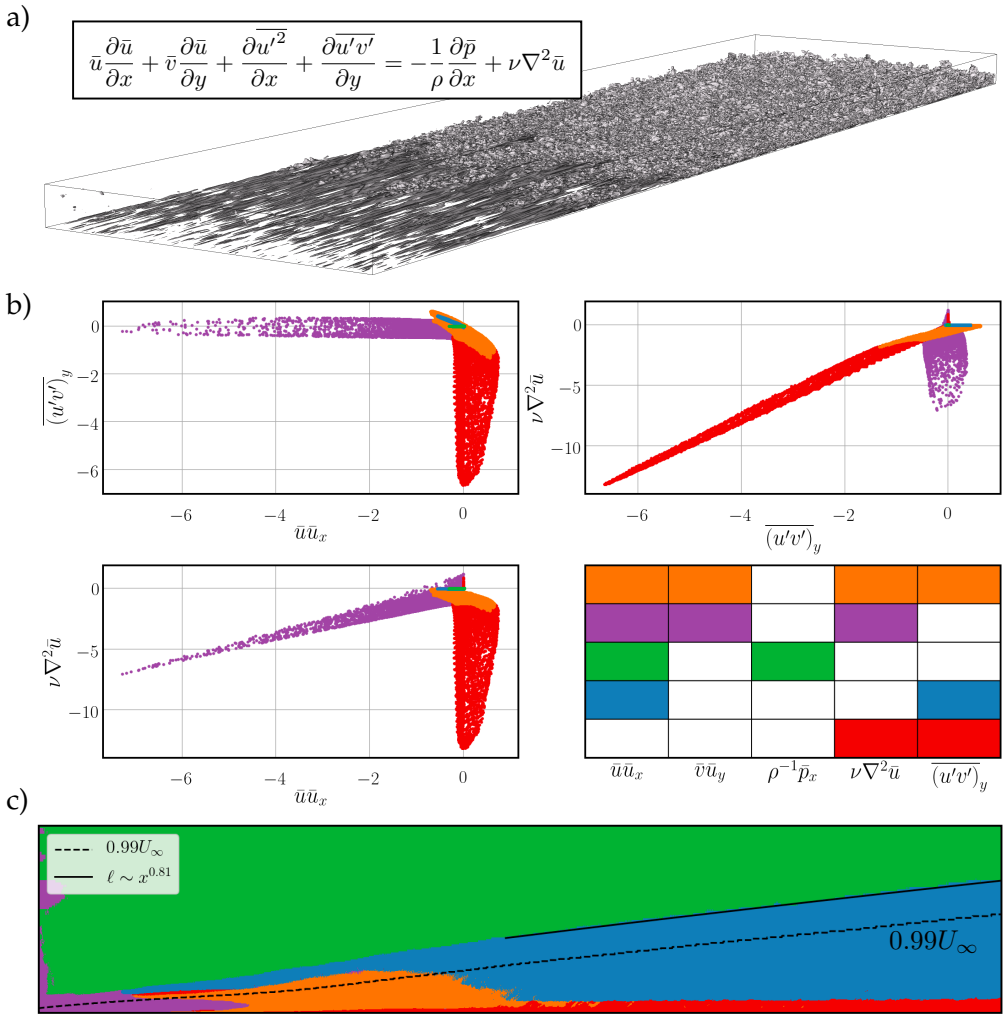

Figure 5: Direct numerical simulation (DNS) of a transitional boundary layer [36, 37, 71–73], visualized by contours of the turbulent kinetic energy (a). The Reynolds number based on free stream velocity and streamwise extent is $\mathrm{Re}_L = 192,000$. Active terms vary across the domain (b). The method recovers expected balance relations for the free-stream (green), the inertial sublayer (blue), and the viscous sublayer (red), along with a laminar region near the inlet (purple) and a transitional region (orange). The inertial sublayer follows the theoretically predicted power law (c). Boundary layer theory predicts that the length scale ℓ of the sublayer scales with $\ell \sim x^{4/5}$. As a rough criterion for the scale of the inertial balance model, we use the wall-normal coordinate at which the balance relation changes (solid line top), once the transitional region (purple) ends. A curve fit shows an approximate scaling of $\ell \sim x^{0.81}$.

As a consequence of the no-slip boundary conditions, in this region the free-stream velocity is not an appropriate scale for the streamwise component and we should instead use the friction velocity u_τ, so that

$$\bar{u}\bar{u}_x, \bar{v}\bar{u}_y \sim \frac{u_\tau^2}{L}, \quad \nu\bar{u}_{xx} \sim \left(\frac{\eta}{L}\right)\frac{u_\tau^2}{L}, \quad \nu\bar{u}_{yy} \sim \left(\frac{L}{\eta}\right)\frac{u_\tau^2}{L}, \quad (\overline{u'v'})_y \sim \left(\frac{L}{\eta}\right)\frac{u_\tau^2}{L}, \quad (\overline{u'^2})_x \sim \frac{u_\tau^2}{L}.$$

In this case the wall-normal Reynolds stress is larger than the mean flow advection by a factor of $L/\eta \gg 1$ and must instead be balanced by the viscosity. Therefore, in a thin viscous sublayer

near the wall the dominant balance is

$$(\overline{u'v'})_y = \nu \bar{u}_{yy}. \tag{7}$$

The overall picture is then that the Reynolds stress must be balanced by mean flow advection in the inertial sublayer and by viscosity in the near-wall region. Outside of the turbulent boundary layer the Reynolds stresses and mean wall-normal velocity are negligible, so small variations, for instance due to incompletely converged statistics, should be described by the balance $\bar{u}\bar{u}_x = -\rho^{-1}\bar{p}_x$. In a true zero pressure gradient flow both of these would be zero in the free stream.

Identified dominant balance. We investigate the dominant balance physics of transitional boundary layer data from a direct numerical simulation [36, 37, 73], openly available from the Johns Hopkins Turbulence Database [71, 72][1]. Figure 5 shows the equation space clusters and associated dominant balance models for the mean fields. As with the cylinder example, some sets of points have significantly reduced variance in certain directions of equation space, a strong signature of the dominant balance phenomenon.

The method identifies regions corresponding to the viscous sublayer (7), inertial sublayer (6), and slightly perturbed free stream. It also identifies a region near the inlet characterized by a lack of Reynolds stresses, suggesting the mean profile here should be consistent with the laminar solution. The boundaries between balance regimes need not be sharp, however, especially in a transitional flow. In this case a cluster containing all of the active terms in the zero-pressure-gradient flat plate turbulent boundary layer equation is identified between the laminar inflow region and fully developed turbulence downstream.

Equations (6) and (7) are a starting point for many of the results of boundary layer theory; from these a range of useful laws can be derived, such as the logarithmic mean velocity profile in the inertial sublayer. Although we ultimately hope that data-driven balance identification will open new avenues of analysis, we can also use established results to examine the validity of the proposed method.

For example, the dominant length scale ℓ in the inertial sublayer is expected to depend on the streamwise coordinate x via a power law $\ell \sim x^{4/5}$ [1]. It is not usually obvious how to extract a specific value of ℓ for which this scaling can be checked. However, as a rough proxy we may consider the wall-normal coordinate at which the dominant balance changes from that of the inertial sublayer to the free-stream. Figure 5 shows the growth of the inertial sublayer thickness according to this definition along with a power law fit with exponent 0.81, showing close agreement with the expected value of $4/5$. Although this evidence is somewhat circumstantial, it is at least suggestive that the balance model identification procedure reflects the underlying physics.

3.3 Optical pulse propagation

Another important example of dominant balance arises in nonlinear optics, where the interplay of an intensity dependent index of refraction with chromatic dispersion can generate localized optical solitons [75]. The derivation of the governing evolution equations of the electric field envelope from Maxwell's equations shows that for ultra-short pulses of light (e.g. a few femtoseconds), the time response of the polarization field can yield [76] a rich set of nonlinear dynamics.

Figure 6 shows an example of a process known as supercontinuum generation, in which nonlinear processes act on a localized pulse of light to generate a severe broadening of the optical spectrum. This is typically accomplished in microstructured optical fibers [77]. Thus an initial

[1] https://doi.org/10.7281/T17S7KX8

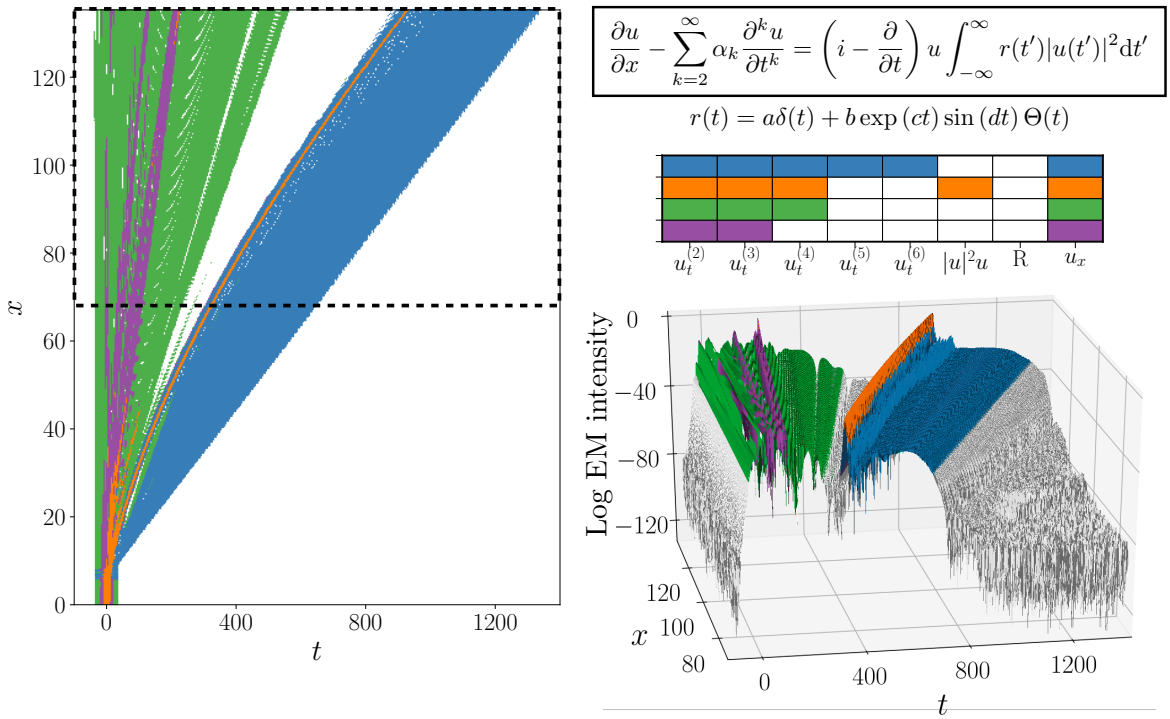

Figure 6: Identified balance models for the generalized nonlinear Schrödinger equation. The governing equations are derived from Maxwell's equations in 1D with a nonlinear time-delayed polarization response. Soliton propagation is understood to be maintained primarily by a balance between low-order dispersion and the cubic Kerr nonlinearity (delta-function component of the right-hand side integral) [7]. Although most of the field is identified with various linear dispersion relations, the strongest soliton is associated with cubic nonlinearity and dispersive terms through fourth order.

20-30 nanometer bandwidth can be stretched to hundreds of nanometers. The governing equation in this case is derived from Maxwell's wave equation in one dimension through the rotating wave approximation and the slowly varying envelope approximation [76]. The original PDE is linear and second order in a vacuum, but in order to handle complicated polarization responses in fibers the field is expanded about the frequency of the original pulse [6, 7]. This "center frequency" expansion leads to a Taylor series expansion of the linear polarization response, and the Raman convolution integral describing a time-delayed nonlinear response.

The resulting PDE, known as a generalized nonlinear Schrödinger equation (GNLSE) describes the evolution of the slowly varying complex envelope $u(x,t)$ of the pulse. When nondimensionalized with soliton scalings [7], the envelope equation is

$$\frac{\partial u}{\partial x} - \sum_{k=2}^{\infty} \alpha_k \frac{\partial^k u}{\partial t^k} = \left(i - \frac{\partial}{\partial t}\right) u \int_{-\infty}^{\infty} r(t')|u(t')|^2 \mathrm{d}t' \tag{8a}$$

$$r(t) = a\delta(t) + b\exp(ct)\sin(dt)\Theta(t). \tag{8b}$$

The various constants (α_k, a, b, c, d) describe the polarization response and are determined empirically.

Although the spectral domain is often of practical interest for studies of supercontinuum generation, in the time domain the pulse exhibits soliton behavior, as shown in figure 6. To leading

order, the soliton propagation is typically understood to be maintained by a balance between the second order dispersion and the instantaneous part of the nonlinear response, or intensity-dependent index of refraction. That is, evaluating the delta function component of the Raman kernel leads to the cubic Kerr nonlinearity. If only this cubic nonlinearity and second order dispersion are retained, equation (8a) is reduced to the usual nonlinear Schrödinger equation (NLS):

$$i\frac{\partial u}{\partial x} + \frac{\partial^2 u}{\partial t^2} + |u|^2 u = 0. \tag{9}$$

Figure 6 shows the balance models obtained through the unsupervised balance identification procedure applied to regions of the field where the intensity is within 40 dB of the peak. Most of the domain is associated with various linear dispersion relations, corresponding to different propagation speeds. Only a narrow region containing the strongest soliton is identified with the instantaneous nonlinear response, suggesting that a linear description is sufficient for much of the domain. The standard NLS equation is never identified, although the balance relation with cubic nonlinearity and fourth order dispersion (green) is consistent with standard truncation of the linear response at third or fourth order [6]. Interestingly, the full Raman time-delay response is never selected as an important term, although this is understood to be a critical mechanism for the initial scattering. Presumably the Gaussian mixture model approach is not sensitive enough to detect this, possibly due to the clearly invalid underlying assumption of normally distributed data.

3.4 Geostrophic balance in the Gulf of Mexico

Geophysical fluid dynamics is a particularly complex field; a full description of ocean dynamics for instance requires not only the Navier-Stokes equations on a rotating Earth with complicated bathymetry, but must also account for the effects of varying salinity, temperature, and pressure via a nonlinear equation of state. The ocean dynamics also couple to atmospheric and geological processes and solar forcing [3]. Scaling analyses have been remarkably successful; despite the complexity of the dynamics, in many cases it can be argued that greatly simplified versions of the governing equations are sufficient to describe the dominant motions.

Perhaps the most important model of this type is geostrophic balance. To a first approximation, the surface currents can be modeled with the 2D incompressible Navier-Stokes equations on a rotating sphere:

$$u_t + (\mathbf{u}\cdot\nabla)u + fv = -\frac{1}{\rho}p_x \tag{10a}$$

$$v_t + (\mathbf{u}\cdot\nabla)v - fu = -\frac{1}{\rho}p_y, \tag{10b}$$

where ρ is the density (in general a function of temperature, pressure, and salinity), and x and y are defined in the zonal and meridional directions, respectively. The Coriolis parameter f is given in terms of the Earth's angular velocity Ω and the latitude ϕ by $f = \Omega\sin\phi$. Note that this equation already includes some approximations. Compressibility, vertical motions, and both molecular and turbulent viscosities are all ignored in this model. Nevertheless, these equations are a standard starting point for many analyses of large scale ocean dynamics.

For flows with length scale L and velocity scale U, the relative importance of the Coriolis terms compared to the inertial terms is given by the Rossby number, $\mathrm{Ro} = U/fL$. In low Rossby number flows (relatively slow, large scale motions), the inertial terms become negligible and the dominant

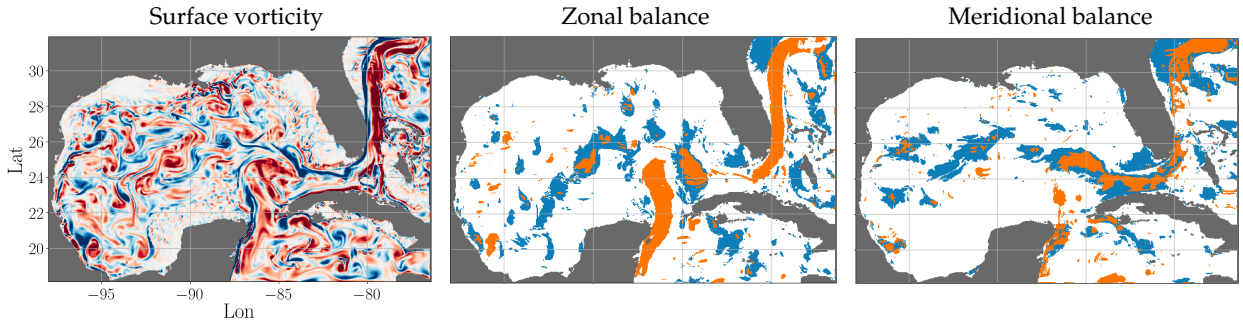

Figure 7: Surface vorticity in the Gulf of Mexico (left) along with identified balance models for zonal (middle) and meridional (right) dynamics. Orange regions are identified with the geostrophic balance, while the blue regions are time-varying in response to the Coriolis forces and regions in white are associated with the linearized rotating Navier-Stokes equations.

balance is between the Coriolis forces and pressure gradient forces:

$$+ fv = -\frac{1}{\rho} p_x \tag{11a}$$

$$- fu = -\frac{1}{\rho} p_y. \tag{11b}$$

This balance is thought to describe most approximately steady large scale currents [3].

We apply the unsupervised balance identification procedure to the high-resolution $1/25^\circ$ HYCOM reanalysis data for the Gulf of Mexico [78]. Figure 7 shows the regions corresponding to balance models for this data. The method identifies three regimes; geostrophic balance (orange), a balance between acceleration and Coriolis forces (blue), and the linearized rotating Navier-Stokes equations (white). The nonlinear advective term is not included in any of the models in this case, supporting the common use of linearized equations to study wavelike motions. Geostrophic balance is primarily identified in regions corresponding to slow, large scale motions: the southern end of the Gulf Stream and the relatively stable current between Cuba and the Yucatàn Peninsula.

Clearly the approximations in estimating gradients introduce significant error and variability into the balance identification procedure for this examples. However, the identified models are consistent with the expected behavior according to classical arguments. These results indicate some degree of robustness of the procedure and suggest that it may be applied to sufficiently clean experimental or data-assimilated observations.

3.5 Generalized Hodgkin-Huxley model of an intrinsically bursting neuron

Networks of biological neurons in an animal's nervous systems communicate with each other through the propagation of electrical potentials. These all-or-nothing events, known as *action potentials* or *spikes*, are large deviations from the membrane electrical potential at rest, as measured between the inside and outside of a neuron. Importantly, spikes can travel without significant degradation down the length of a neuron's long axon, which may be meters long.

The celebrated Hodgkin-Huxley model for spiking neurons reproduces an action potential through a balance of currents from multiple ions, each of which moves through the cell's membrane across specialized channels and pores at different phases of a spike [79]. These non-linear partial differential equations were the first detailed biophysical model to quantitively describe the

dynamic activity of neurons, and they underpin decades of ongoing attempts to understand more complex properties of neuronal electrical excitability [80].

The propagation of an action potential along an axon is well approximated by the cable equation of a cylinder of radius a,

$$C_M \frac{\partial V}{\partial t} = \frac{a}{2r_L} \frac{\partial^2 V}{\partial x^2} + \sum_j I_j, \tag{12}$$

where C_M is the membrane capacitance, r_L is the resistivity inside the cell, and I_j are each of the ionic currents in current per unit area due to the flow of ions into and out of the cell.

Hodgkin and Huxley originally modeled three (3) ionic currents: I_{Na} sodium, I_K potassium, and a leak I_L. The dynamics of V for a single action potential can then be expressed as a system of four (4) ordinary differential equations; the balance of currents in these equations reflect the biophysical mechanisms.

Adding more ionic currents and modeling the interactive balance of their dynamics produces more complex spiking behavior. In particular, here we consider a generalized Hodgkin-Huxley model with ten (10) currents that simulates the intrinsically bursting pattern of spikes observed in the R15 neuron of the sea slug *Aplysia* [81], as shown in Fig. 8. The R15 neuron has been used to study the mechanisms underlying intrinsic bursting, where several action potentials are generated in rapid succession interspersed with relative quiet with constant inputs. Under space-clamp conditions where an entire axon cable is considered to be spatially uniform, the equation describing the time-evolution of membrane voltage V under applied external input I_{stim} is

$$C_M \dot{V} = -\sum_j I_j + I_{\text{stim}}. \tag{13}$$

Specifically, the ionic currents I_j in our model are: I_{Na} the fast sodium Na^+ current; I_{Ca} the fast calcium Ca^{2+} current; I_K the delayed rectifier potassium current; I_{SI} the slow inward calcium current; I_{NS} the non-specific cation current; I_R the anomalous rectifier current; I_L the leakage rectifier current; I_{NaCa} the sodium-calcium exchanger current; I_{NaK} the sodium-potassium pump; I_{CaP} the calcium pump.

Our dominant balance approach identifies several interpretable regimes of physics in the generalized Hodgkin-Huxley model that are largely consistent with known biophysics. The addition of a set of calcium-dependent currents underly the slower oscillations between quiescence and excitable bursting, as evident in the slower limit cycle. Notably, in these clusters, colored pink and gray in Fig. 8, the balance of ions is dominated by terms with strong calcium dependence (I_{CaP}, I_{SI}, and I_{NaCa}). In contrast, the time-course of V at each fast spike is dominated by voltage-gated ionic currents. In Fig. 8, the rising part of each spike is mediated by activation of sodium channels, and the inward I_{SI} and I_{Na} increase V (red and blue). V reaches peak voltage as the sodium channels inactivate and delayed rectifier potassium channels I_K activate (purple). The exit of potassium from the cell decreases V back towards the resting potential.

There are three currents that have not been identified to belong to any cluster: the fast calcium current, sodium-potassium pump, and the non-specific cation current. Although these are dynamically important for the model, they are relatively small compared to the other terms ($\mathcal{O}(0.1-1)$ compared to $\mathcal{O}(100)$ for the spiking dynamics) and so they don't appear to participate in any of the local dominant balance relationships identified by this method. This is a similar situation to the Raman time-delay nonlinearity in the optical pulse propagation example (Sec. 3.3) and the nonlinear advection in the Gulf of Mexico (Sec. 3.4). In all of these cases, the influence of the neglected terms appears to be of a more subtle nature than the dominant balance physics we explore in this work.

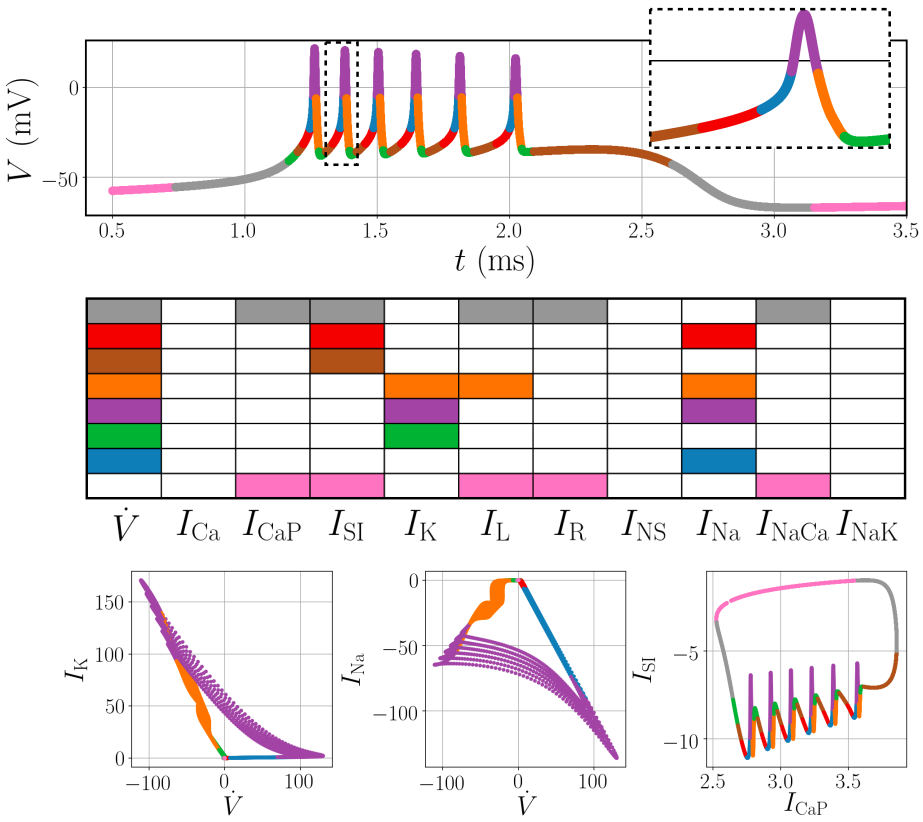

Figure 8: Generalized Hodgkins-Huxley model for an intrinsically bursting neuron. Dynamics in quiescent periods are characterized by currents related to calcium concentration (pink and gray), while the spiking dynamics are dominated by the classic sodium-potassium cycle.

4 Discussion

In one guise or another, dominant balance analysis has played a major role in the development of our understanding of many complex systems. In this paper we have proposed a method of identifying dominant balance regimes in an unsupervised manner directly from data. This approach leverages our understanding of the full physical complexity in the form of governing equations, but by using simple clustering and sparse approximation methods we avoid any *a priori* assumptions about balance relations. Nevertheless, in contexts ranging from fluid turbulence to nonlinear optics the method recovers classical dominant balance relationships.

The critical step in this process is the "equation space" perspective described in Sec. 2. By considering each term in the governing equation to describe a direction in this space, the dominant balance relations naturally manifest via restriction to sparse subspaces, i.e. dramatic reductions in variance in directions corresponding to negligible terms. This enables the Gaussian mixture models to identify clusters with variance in different directions, and the sparse principal components analysis to extract sparse subspaces by finding directions with significantly nonzero variance. These machine learning tools are therefore applied in a targeted and clearly motivated context, but the equation space perspective necessarily ties the output to underlying physics.

The method as presented here is perhaps the simplest version possible of this type of analysis. As such, there are clear opportunities for further refinement. For example, the Gaussian mixture model analysis is built on the assumption of normally distributed data. There is no reason to think

that the equation space representation of physical fields would be normally distributed, which may limit the sensitivity of the method. Other methods such as spectral clustering or a custom, =physically motivated algorithm may be more effective at segmenting this type of data.

On the other hand, the method can be sensitive to computation of the various terms in the equation, especially gradients. When possible, the terms were extracted directly from the numerical solvers, although this may present a challenge for noisy experimental data. One way to address this could be a reanalysis-type smoothing procedure, as was used by the HYCOM group to generate the Gulf of Mexico data. Similar data-assimilation approaches have been successful at resolving mean profiles of turbulent flows from limited experimental data [82, 83].

When properly developed and validated, the ability to automatically extract balance relations from data has exciting potential applications. For instance, identifying regions of flow fields where viscosity is important could be a principled way to inform schemes such as adaptive mesh refinement [59] or hybrid turbulence modeling [60, 84]; currently regions are typically chosen using heuristics or expert knowledge. An understanding of balance relations could even potentially be used to develop novel control strategies. By designing or actuating with the goal of manipulating which regimes are active, such an approach might be used to achieve drag reduction or mixing enhancement.

More generally, dominant balance analysis has historically been a critical tool for understanding local physical behavior in complex systems. To date we have only been able to apply these methods to systems for which the governing equations are well-understood and which admit an asymptotic scaling analysis. Generalizing this analytic approach with data-driven dominant balance identification could allow application of this powerful perspective to complex geometries, non-asymptotic regimes, and even systems for which the governing equations are unknown.

However, as with all applications of machine learning and data science methods to physical systems, a critical step in application to any system will be careful validation that the balance identification procedure reproduces the expected results. The dominant balance modeling approach described here is designed to build on, rather than circumvent, physical expertise. The study of dominant balance regimes has been foundational to our understanding of many complex systems; we hope that data-driven methods can integrate with this legacy to enable even wider applicability.

Acknowledgements

JLC acknowledges support from the NDSEG fellowship. JNK acknowledges support from the Air Force Office of Scientific Research (AFOSR FA9550-17-1-0329). BWB acknowledges support by the Washington Research Foundation. SLB acknowledges funding support from the Air Force Office of Scientific Research (AFOSR FA9550-18-1-0200) and the Army Research Office (ARO W911NF-19-1-0045). The authors also acknowledge support from the Defense Advanced Research Projects Agency (DARPA PA-18-01-FP-125). We also thank Jean-Christophe Loiseau, Lionel Mathelin, and Kunihiko Taira for valuable discussions about the implications and applications for dominant balance identification.

Appendix A: Data provenance

Direct numerical simulation of flow past a circular cylinder. We simulate this configuration at Re $= 100$ with unsteady incompressible DNS using the open source spectral element solver Nek5000 [85]. The domain consisted of 17,432 seventh order spectral elements on $x, y \in (-20, 50) \times (-20, 20)$, refined close to a cylinder of unit diameter centered at the origin. Diffusive terms are

integrated with third order backwards differentiation, while convective terms are advanced with a third order extrapolation. The results of this simulation have been validated against those of the immersed boundary projection method [86] by comparing aerodynamic coefficients and vortex shedding frequency. We extract the vorticity field and spatial terms in equation (4) directly from the solver for further analysis. Time derivatives for dominant balance identification were estimated with a second order central difference.

Direct numerical simulation of a transitional boundary layer. To study dominant balance physics in the turbulent boundary layer, we use the transitional DNS by Lee and Zaki [36, 37, 73], openly available from the Johns Hopkins Turbulence Database [71, 72][2]. The full computational domain consists of a long flat plate with an elliptical leading edge. The extent of the domain (in units defined by the plate half-thickness) is $(x, y, z) \in (1040, 40, 240)$ with periodic boundary conditions in the spanwise (z) direction, discretized to $(Nx, Ny, Nz) = (4097, 257, 2049)$. Since the configuration of interest is a zero pressure gradient flat plate boundary layer, the DNS results are only saved once the flow passes the elliptical leading edge ($x > 30.2185$). The inflow consists of small amplitude free-stream turbulence superimposed on a uniform streamwise velocity U_∞ incident on the plate. The interactions of these perturbations with the laminar boundary layer cause a downstream transition to turbulence [73].

Since we are interested here in the mean momentum balance, we only use the 2D mean field (also available from JHTDB), which was computed from 4701 data snapshots once the flow reached a statistically stationary state. Without direct access to the gradients, we compute the constituent terms of the RANS equations with second-order accurate finite differences, as shown in Fig. 1b. Although some of these fields show small fluctuations, the overall smoothness suggests the statistics are approximately converged.

Supercontinuum generation in photonic crystal fiber. The generalized nonlinear Schrödinger equation (GNLSE), nondimensionalized with soliton scaling [7], is given by Eq. (8a). The various constants describe the polarization response and are determined empirically. In this case we use the values described by Dudley *et al* for photonic crystal fiber [77]. We also use the split-step spectral method and initial conditions described in these works to simulate the pulse propagation[3].

Surface currents in the Gulf of Mexico. We study the high-resolution $1/25°$ HYCOM reanalysis data for the Gulf of Mexico [78]. We use data from only the first field in the data set, corresponding to January 1993. Data-assimilated fields are available for the 2D velocity components, sea surface temperature, salinity, and sea surface height; vorticity is shown in Fig. 7.

We must therefore estimate time derivatives and both velocity and pressure gradients to compute the terms in Eqns. (10a) and (10b). Since this information is not directly accessible from the model (as for the numerical examples), we use finite differences to estimate the velocity derivatives. The pressure field itself is also not available; as a rough estimate we use the residuals of the left-hand side of Eqns. (10a) and (10b) in place of pressure gradients. We also assume constant density throughout the field. Finally, since this field is two-dimensional but the terms in each evolution equation represent the same physics, we simply stack the features for each velocity component into a single $(2N \times 4)$ matrix with columns corresponding to acceleration, convection, Coriolis forces, and the pressure gradient. Although these are strong assumptions and approxi-

[2] https://doi.org/10.7281/T17S7KX8

[3] MATLAB code freely available at http://www.scgbook.info/

mations, we would expect them to only make the dominant balance identification problem more difficult, since they represent attempts to deal with limited information about the system.

Generalized Hodgkins-Huxley model of a bursting neuron. A full set of model equations, including biophysical parameters, follow [81] and are given in the simulation code. Briefly, gating variables following Hodgkin-Huxley form are described by solutions to differential equations of the general form $\dot{z} = (z_\infty - z)/\tau_z$, where z_∞ are the steady-state values and τ_z are the time constants associated with the gating variable z. To produce the data used in our analysis, this system of ordinary differential equations was integrated numerically in MATLAB using ode15.

Appendix B: Parameter tuning

The proposed method was designed to minimize the number of hyper-parameters that need to be tuned. However, there are two important parameters that must be selected: the number of clusters for the Gaussian mixture model (GMM), and the ℓ_1 regularization for sparse principal components analysis (SPCA).

Since the data is not actually drawn from a mixture of Gaussian distributions it can be difficult to make a principled choice for the number of GMM clusters. Intuitively, if there are too few clusters the GMM procedure cannot be expected to capture all of the distinct directions of variance in the equation space. The secondary SPCA reduction makes the method somewhat robust to this parameter; the final balance models tend to be similar provided that there are enough clusters. However, if there are too many clusters, the constituent distributions of the mixture model may not contain enough points to be dominated by a single principal component.

The ℓ_1-regularization for SPCA is somewhat easier to choose with a simple model selection procedure. A larger regularization value tends to yield more sparsity in the leading principal component, corresponding to neglected terms in the cluster. We define the residual for a given regularization value as the ℓ_2-norm of the neglected terms across all clusters. For example, if SPCA with a regularization of 0.1 yields a principal component with a zero in the direction corresponding to viscosity for one of the clusters, the SPCA residual for 0.1 in that cluster would be the magnitude of the viscous terms in that cluster. Sweeping a range of regularization values yields a Pareto-type curve showing the tradeoff of sparsity against descriptiveness.

This metric offers a guideline for choosing an appropriate regularization, although there is still some flexibility in the specific value. As Fig. 9 shows, tuning the regularization differently yields a different set of balance models. As with many model selection procedures, a different value may be selected depending on the desired level of descriptiveness and parsimony. Based on physical considerations, in this work we looked for regularizations that resulted in a diversity of balance relations with 2-3 active terms each (middle panel of Fig. 9).

Appendix C: Model uncertainty

The idea of dominant balance is not necessarily clearly defined outside of asymptotic regimes; strictly speaking, all terms in a model are likely to have some nonzero contribution throughout the domain of interest. Considering for example the cylinder wake, clearly the boundary layer is not steady, nor is the far-field region actually inviscid.

Fortunately, since GMM is a probabilistic clustering method it comes with a natural notion of uncertainty. The clustering procedure assigns to each point a probability of belonging to each cluster. We can propagate this through the SPCA reduction by summing the probabilities that each point in the field belongs to one of the clusters that reduces to the same balance model. This

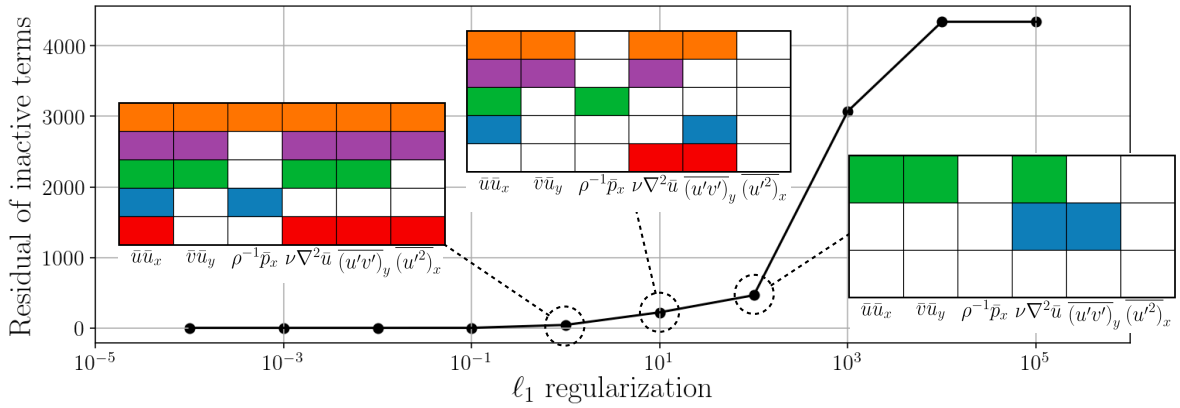

Figure 9: Model selection procedure used to choose a sparse regularization value for the principal components analysis, demonstrated on the turbulent boundary layer example. Although there is some flexibility depending on the desired accuracy and simplicity in the specific application, the residual of neglected terms suggests a range of appropriate values. In this work we chose regularizations that were as sparse as possible but spanned most of the original terms in the equation and had relatively small residuals (middle panel). Often this led to a set of balance relations, each with 2-3 terms, which collectively captured much of the richness of the full system.

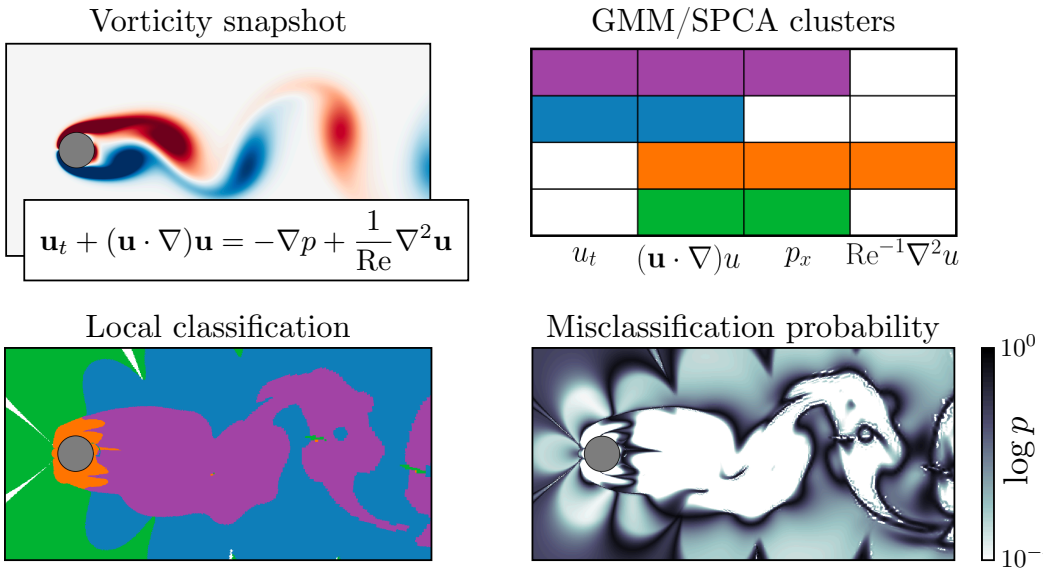

Figure 10: Uncertainty estimation for the dominant balance identification procedure. The Gaussian mixture model clusters points in the domain by assigning a probability of belonging to each Gaussian distribution. Summing the probabilities that each point belongs to a GMM cluster which SPCA reduces to the same balance model gives an overall estimate of the uncertainty associated with the identified dominant balance.

results in an estimate of the probability of misclassification of each point, as shown in Fig. 10. As expected, this measure generally becomes large in transitional regions. However, keeping their approximate nature in mind, the balance models offer a principled and intuitive segmentation of the domain according to the dominant physics.

References

[1] H. Schlichting. *Boundary-Layer Theory*. McGraw-Hill, 1955.

[2] K. R. Sreenivasan. Fluid turbulence. *Reviews of Modern Physics*, 71(2):S383, 1999.

[3] A. Gill. *Atmosphere-Ocean Dynamics*. Academic Press, 1982.

[4] M. J. Lighthill. Dynamics of rotating fluids: a survey. *Journal of Fluid Mechanics*, 26:411–431, 1966.

[5] A. J. Majda. *Introduction to PDEs and waves for the atmosphere and ocean*. American Mathematical Society New York, 2003.

[6] K. J. Blow and D. Wood. Theoretical description of transient stimulated Raman scattering in optical fibers. *IEEE Journal of Quantum Electronics*, 25, 1989.

[7] L. F. Mollenauer and J. P. Gordon. *Solitons in Optical Fibers: Fundamentals and Applications*. Elsevier, 2006.

[8] U. R. Christensen and J. Aubert. Scaling properties of convection-driven dynamos in rotating spherical shells and application to planetary magnetic fields. *Geophys. J. Int.*, 166(1):97–114, 2006.

[9] M. C. Cross and P. C. Hohenberg. Pattern formation outside of equilibrium. *Reviews of modern physics*, 65(3):851, 1993.

[10] S. W. Morris, E. Bodenschatz, D. S. Cannell, and G. Ahlers. Spiral defect chaos in large aspect ratio Rayleigh-Bènard convection. *Physical Review Letters*, 71(13):2026, 1993.

[11] E. Bodenschatz, W. Pesch, and G. Ahlers. Recent developments in Rayleigh-Bènard convection. *Annual Review of Fluid Mechanics*, 32(1):709–778, 2000.

[12] B. Grzybowski, H. A. Stone, and G. M. Whitesides. Dynamic self-assembly of magnetized, millimetre-sized objects rotating at a liquid-air interface. *Nature*, 405:1033–1036, 2000.

[13] E. Cerda and L. Mahadevan. Geometry and physics of wrinkling. *Phys. Rev. Letters*, 90(7):074302, 2003.

[14] N. Tsapsis, E. R. Dufresne, S. S. Sinha, C. S. Riera, J. W. Hutchingson, L. Mahadevan, and D. A. Weitz. Onset of buckling in drying droplets of colloidal suspensions. *Phys. Rev. Letters*, 94(1):018302, 2005.

[15] X. D. Shi, M. P. Brenner, and S. R. Nagel. A cascade of structure in a drop falling from a faucet. *Science*, 265(5169):219–222, 1994.

[16] A. S. Utada, A. Fernandez-Nieves, H. A. Stone, and D. A. Weitz. Dripping to jetting transitions in coflowing liquid streams. *Physical Review Letters*, 99(9):094502, 2007.

[17] S. V. Fridrikh, J. H. Yu, M. P. Brenner, and G. C. Rutledge. Controlling the fiber diameter during electrospinning. *Physical Review Letters*, 90(14):144502, 2003.

[18] A. Seminara, T. E. Angelini, J. N. Wilking, H. Vlamakis, S. Ebrahim, R. Kolter, D. A. Weitz, and M. P. Brenner. Osmotic spreading of bacillus subtilis biofilms driven by an extracellular matrix. *Proceedings of the National Academy of Sciences*, 109(4):1116–1121, 2012.

[19] G. I. Barrenblatt, A. J. Chorin, O. H. Hald, and V. M. Prostokishin. Structure of the zero-pressure-gradient turbulent boundary layer. *Proceedings of the National Academy of Sciences*, 94:7817–7819, 1997.

[20] M. V. Zagarola and A. J. Smits. Mean-flow scaling of turbulent pipe flow. *J. Fluid Mech.*, 373:33–79, 1998.

[21] J. F. Morrison, B. J. McKeon, W. Jiang, and A. J. Smits. Scaling of the streamwise velocity component in turbulent pipe flow. *Journal of Fluid Mechanics*, 508:99–131, 2004.

[22] T. B. Nickels, I. Marusic, S. Hafez, and M. S. Chong. Evidence of the k_1^{-1} law in a high-Reynolds-number turbulent boundary layer. *Physical Review Letters*, 95:074501, 2005.

[23] I. Marusic, R. Mathis, and N. Hutchins. Predictive model for wall-bounded turbulent flow. *Science*, 329(5988):193–196, 2010.

[24] A. J. Smits, B. J. McKeon, and I. Marusic. High-Reynolds number wall turbulence. *Annual Review of Fluid Mechanics*, 43, 2011.

[25] I. Marusic, J. P. Monty, M. Hultmark, and A. J. Smits. On the logarithmic region in wall turbulence. *Journal of Fluid Mechanics*, 716:R3, 2013.

[26] M. J. Lighthill. Contributions to the theory of heat transfer through a laminar boundary layer. *Proc. R. Soc. A*, 202:359–377, 1950.

[27] M. J. Lighthill. On sound generated aerodynamically. Part I. *Proc. R. Soc. A*, 211:564–587, 1952.

[28] M. J. Lighthill. *Surveys in Mechanics*, chapter Viscosity effects in sound waves of finite amplitude, pages 250–351. Cambridge University Press, 1956.

[29] M. Pastoor, L. Henning, B. R. Noack, R. King, and G. Tadmor. Feedback shear layer control for bluff body drag reduction. *Journal of Fluid Mechanics*, 608:161–196, 2008.

[30] S. L. Brunton and B. R. Noack. Closed-loop turbulence control: Progress and challenges. *Applied Mechanics Review*, 67(5), 2015.

[31] S. Verma, G. Novati, and P. Koumoutsakos. Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Nat. Acad. Sci.*, 115(23):5849–5854, 2018.

[32] C.-A. Yeh and K. Taira. Resolvent-analysis-based design of airfoil separation control. *Journal of Fluid Mechanics*, 867:572–610, 2019.

[33] K. Duraisamy, G. Iaccarino, and H. Xiao. Turbulence modeling in the age of data. *Annual Reviews of Fluid Mechanics*, 51:357–377, 2019.

[34] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet. The analog data assimilation. *Monthly Weather Review*, 145(10):4093–4107, 2017.

[35] P. R. Vlachas, J. Pathak, B. R. Hunt, T. P. Sapsis, M. Girvan, E. Ott, and P. Koumoutsakos. Forecasting of spatio-temporal chaotic dynamics with recurrent neural networks: a comparative study of reservoir computing and backpropagation algorithms. *arXiv:1910.05266v1*, 2019.

[36] J. Lee and T. A. Zaki. Detection algorithm for turbulent interfaces and large-scale structures in intermittent flows. *Computers & Fluids*, 175:142–158, 2018.

[37] Z. Wu, J. Lee, C. Meneveau, and T. Zaki. Application of a self-organizing map to identify the turbulent-boundary-layer interface in a transitional flow. *Physical Review Fluids*, 4:023902, 2019.

[38] G. D. Portwood, S. M. de Bruyn Kops, J. R. Taylor, H. Salehipour, and C. P. Caulfield. Robust identification of dynamically distinct regions in stratified turbulence. *J. Fluid Mech.*, 807(R2), 2016.

[39] J. J. Riley and S. M. de Bruyn Kops. Dynamics of turbulence strongly influenced by buoyancy. *Physics of Fluids*, 15(7), 2003.

[40] G. Brethouwer, P. Billant, E. Lindborg, and J.-M. Chomaz. Scaling analysis and simulation of strongly stratified turbulent flows. *Journal of Fluid Mechanics*, 585:343–368, 2007.

[41] D. Samanta, Y. Dubief, M. Holzner, C. Schäfer, A. N. Morozov, C. Wagner, and B. Hof. Elasto-inertial turbulence. *Proceedings of the National Academy of Sciences*, 110(26):10557–10562, 2013.

[42] J. W. M. Bush. Pilot-wave hydrodynamics. *Annual Review of Fluid Mechanics*, 47:269–292, 2015.

[43] W. Cousins and T. P. Sapsis. Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Physica D: Nonlinear Phenomena*, 280:48–58, 2014.

[44] B. Hof, A. Juel, and T. Mullin. Scaling of the turbulence transition threshold in a pipe. *Physical Review Letters*, 91(24):244502, 2003.

[45] B. Hof, C. W. H. van Doorne, J. Westerweel, F. T. M. Nieuwstadt, H. Faisst, B. Eckhardt, H. Wedin, R. R. Kerswell, and F. Waleffe. Experimental observation of nonlinear traveling waves in turbulent pipe flow. *Science*, 305(5960):1594–1598, 2004.

[46] B. Eckhardt, T. M. Schneider, B. Hof, and J. Westerweel. Turbulence transition in pipe flow. *Annual Review of Fluid Mechanics*, 39:447–468, 2007.

[47] K. Avila, D. Moxey, A. de Lozar, M. Avila, D. Barkley, and B. Hof. The onset of turbulence in pipe flow. *Science*, 333(6039):192–196, 2011.

[48] D. Barkley, B. Song, V. Mukund, G. Lemoult, M. Avila, and B. Hof. The rise of fully turbulent flow. *Nature*, 526(7574):550, 2015.

[49] Y. Du and G. E. Karniadakis. Suppressing wall turbulence by means of a transverse traveling wave. *Science*, 288(5469):1230–1234, 2000.

[50] B. Hof, A. de Lozar, M. Avila, X. Tu, and T. M. Schneider. Eliminating turbulence in spatially intermittent flows. *Science*, 327(5972):1491–1494, 2010.

[51] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz. Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of Neuroscience Methods*, 258:1–15, 2016.

[52] J. L. Proctor and P. A. Eckhoff. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International Health*, 7:139–145, 2015.

[53] S. A. Levin and R. T. Paine. Disturbance, patch formation, and community structure. *Proceedings of the National Academy of Sciences*, 71(7):2744–2747, 1974.

[54] C. Dombrowski, L. Cisneros, S. Chatkaew, R. E. Goldstein, and J. O. Kessler. Self-concentration and large-scale coherence in bacterial dynamics. *Physical Review Letters*, 93(9):098103, 2004.

[55] H. H. Wensink, J. Dunkel, S. Heidenreich, K. Drescher, R. E. Goldstein, H. Löwen, and J. M. Yeomans. Meso-scale turbulence in living fluids. *Proc. Nat. Acad. Sci.*, 109(36):14308–14313, 2012.

[56] T. Gao, R. Blackwell, M. A. Glaser, M. D. Betterton, and M. J. Shelley. Multiscale polar theory of microtubule and motor-protein assemblies. *Physical Review Letters*, 114(4):048101, 2015.

[57] A. Becker, H. Masoud, J. Newbolt, M. Shelley, and L. Ristroph. Hydrodynamic schooling of flapping swimmers. *Nature Communications*, 6:8514, 2015.

[58] C. W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77(2):439–471, 1988.

[59] M. J. Berger and J. Oliger. Adaptive mesh refinement for hyperbolic partial differential equations. *Journal of Computational Physics*, 1984.

[60] P. R. Spalart. Detached-eddy simulation. *Annual Review of Fluid Mechanics*, 2009.

[61] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, April 2009.

[62] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Nat. Acad. Sci.*, 113(15):3932–3937, 2016.

[63] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4), 2017.

[64] C. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[65] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 2012.

[66] P. G. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: applications to Kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–1524, 2014.

[67] J. T. Stuart. On the non-linear mechanics of hydrodynamic stability. *J. Fluid Mech.*, 4(1):1–21, 1958.

[68] L. D. Landau and E. M. Lifshitz. *Fluid Mechanics*. Elsevier, 1959.

[69] 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.*, 497:335–363, 2003.

[70] S. B. Pope. *Turbulent Flows*. Cambridge University Press, 2000.

[71] E. Perlman, R. Burns, Y. Li, and C. Meneveau. Data exploration of turbulence simulations using a database cluster. In *Supercomputing SC07*. IEEE, 2007.

[72] Y. Li, E. Perlman, M. Wan, Y. Yang, R. Burns, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence*, 9(31), 2008.

[73] T. A. Zaki. From streaks and spots and on to turbulence: exploring the dynamics of boundary layer transition. *Flow, turbulence, and combustion*, 91(3):451–473, 2013.

[74] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge Monographs on Mechanics, 1996.

[75] G. Agrawal. *Nonlinear Fiber Optics, 6th Ed.* Academic Press, 2019.

[76] J. N. Kutz and E. Farnum. Solitons and ultra-short optical waves: the short-pulse equation versus the nonlinear Schrödinger equation. *Edited by H. E. Hernández-Figueroa, E. Recami*, page 148, 2014.

[77] J. M. Dudley and J. R. Taylor. *Supercontinuum generation in optical fibers*. Cambridge, 2010.

[78] HYCOM + NCODA global $1/25^\circ$ reanalysis (Expt. 50.1). https://hycom.org/dataserver, 2019.

[79] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[80] G Bard Ermentrout and David H Terman. *Mathematical foundations of neuroscience*, volume 35. Springer Science & Business Media, 2010.

[81] CC Canavier, JW Clark, and JH Byrne. Simulation of the bursting activity of neuron r15 in aplysia: role of ionic currents, calcium balance, and modulatory transmitters. *J. Neurophys.*, 66(6):2107–2124, 1991.

[82] 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. *Experiments in Fluids*, 58(5), 2017.

[83] A. F. C. da Silva and T. Colonius. Ensemble-based state estimator for aerodynamic flows. *AIAA Journal*, 56(7), 2018.

[84] B. Chaouat. The state of the art of hybrid RANS/LES modeling for the simulation of turbulent flows. *Flow, turbulence, and combustion*, 99(2), 2017.

[85] P. F. Fischer, J. W. Lottes, and S. G. Kerkemeir. Nek5000 web pages. http://nek5000.mcs.anl.gov, 2008.

[86] K. Taira and T. Colonius. The immersed boundary method: A projection approach. *Journal of Computational Physics*, 225:2118–2137, 2007.