Objective discovery of dominant dynamical processes with machine learning

Bryan Kaiser (bkaiser@lanl.gov)
Los Alamos National Laboratory https://orcid.org/0000-0002-4652-6935
Juan Saenz
Los Alamos National Laboratory
Maike Sonnewald
Princeton University
Daniel Livescu
Los Alamos National Laboratory

Physical Sciences - Article

Keywords: Dynamical Regime Identification, Verification Criterion, Unsupervised Learning Framework, ad hoc Conventional Analyses

Posted Date: July 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-745356/v1

Objective discovery of dominant dynamical processes with machine learning

LA-UR-21-27162

Bryan E. Kaiser[1*], Juan A. Saenz[1], Maike Sonnewald[2,3,4], and Daniel Livescu[5]

[1]Los Alamos National Laboratory, X-Computational Physics Division XCP-4, Los Alamos, NM 87545, USA

[2]Princeton University, Program in Atmospheric and Oceanic Sciences,Princeton, NJ 08540, USA

[3]NOAA/OAR Geophysical Fluid Dynamics Laboratory, Ocean and Cryosphere Division, Princeton, NJ 08540, USA

[4]University of Washington, School of Oceanography, Seattle, WA, USA

[5]Los Alamos National Laboratory, Computer Computational and Statistical Physics Division CCS-2, Los Alamos, NM 87545

[*] corresponding author e-mail: bkaiser@lanl.gov

Significant advances in the understanding and modeling of dynamical systems has been enabled by the identification of processes that locally and approximately dominate system behavior,[1] or dynamical regimes. The conventional regime identification method involves tedious and *ad hoc* parsing of data to judiciously obtain scales to ascertain which governing equation terms are dominant in each regime. Surprisingly, no objective and universally applicable criterion exists to robustly identify dynamical regimes in an unbiased manner, neither for conventional nor for machine learning-based methods of analysis. Here, we formally define dynamical regime identification as an optimization problem by using a verification criterion, and we show that an unsupervised learning framework can automatically and credibly identify regimes. This eliminates reliance upon conventional analyses, with vast potential to accelerate discovery.[2] Our verification criterion also enables unbiased comparison of regimes identified by different methods. In addition to diagnostic applications, the verification criterion and learning framework are immediately useful for data-driven dynamical process modeling,[3,4,5,6,7] and are relevant to researchers interested in the development of inherently interpretable methods[8] for scientific machine learning. Automation of this kind of approximate mechanistic analysis is necessary for scientists to gain new dynamical insights from increasingly large data streams.

Observations of dynamical systems often exhibit patterns of spatial and/or temporal sparsity within the terms of the relevant governing equations. A *dominant balance*[50] is a subset of governing equation terms which, on average, dominate the remaining equation terms by at least an order of magnitude. Observed dominant balances are often *non-asymptotic*, with no obvious parameter permitting a series expansion of the equation terms and guaranteeing that the error associated with each neglected term vanishes uniformly in the parameter limits. A *dynamical regime* is a dominant balance together with boundaries in space and/or time. Crucially, dominant balances are dominant *relative* to the magnitude of the negligible

terms within the same regime.

The importance of the relative nature of regime identification is best illustrated by d'Alembert's zero drag paradox,[10] which took over 150 years to be resolved by Prandtl.[11] The paradox emerged from the assumption of a global, rather than relative, threshold for the importance of frictional forces in fluid flow. Its resolution by Prandtl, who pointed out that frictional forces cannot be ignored in surface boundary layers, informs contemporary knowledge of aerodynamic stall and thus partially enables the high safety standards of commercial air travel we enjoy today. Prandtl used the conventional method for dynamical regime identification, which involves forming regime hypotheses from visual cues and dynamical intuition and then testing the hypotheses by comparison with the observed equation term magnitudes for each regime. The tediousness and *ad hoc* manner of this method of regime identification implies that a formal method is required to automatically identify regimes within large data sets, such as can be found in fields as diverse as nonlinear waves, plasma dynamics, earthquake dynamics, general relativity, quantum field theory, biochemical reaction-diffusion dynamics, fibrillation dynamics, epilepsy, and turbulent flows, fiber optics, biofilm dynamics, weather, and climate dynamics.[12,13,14,15,16]

In this Article, we propose an objective and robust method for identifying non-asymptotic regimes. We 1) formalize the regime identification problem as an optimization problem, 2) propose a verification criterion to quantify optimal solutions to this problem, 3) propose a custom dominant balance hypothesis selection algorithm, and 4) propose an unsupervised learning framework[17,18] for solving the problem, drawing upon recent successes of unsupervised learning as a tool for partitioning regimes[45,50] and upon the success of dimensionality reduction algorithms[20] at selecting dominant balance hypotheses.[50] The automation of regime identification is a necessary first step towards the development of a learning agent capable of developing parameterizations of chaotic dynamics by deploying the same logic system as human scientists: the scientific method. The goal of this approach is to develop scientific machine learning models that are as intelligible as they are predictive, attributes

that are not necessarily inversely proportional to one another despite the widely held beliefs to the contrary.[8]

Problem formulation

Given the array of data $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_N]$, consisting of N observations of the $D-$dimensioned vector of equation terms $\mathbf{e}_n$, we seek to label each observation with a $D-$dimensioned hypothesis vector $\mathbf{h}_n$, where $h_{ni} \in \{0, 1\}$ for each n^{th} observation of the i^{th} equation term. We assume that the equation is closed, $\sum_{i=1}^{D} e_{ni} = 0$, for all observations. The entire array of data is labeled by $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_N]$, and zeros in each hypothesis vector $\mathbf{h}_n$ indicate equation terms in $\mathbf{e}_n$ that are neglected. We choose a verfication criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$, such that the optimal fit hypotheses, $\mathbf{H}_{\text{opt}}$, can be obtained by varying the hypotheses $\mathbf{H}$ to find

$$\mathbf{H}_{\text{opt}} = \begin{cases} \underset{\mathbf{H}}{\text{argmax}}\, \mathcal{V}(\mathbf{E}, \mathbf{H}) & \text{if} \quad \max \mathcal{V}(\mathbf{E}, \mathbf{H}) > \mathcal{V}(\mathbf{E}, \mathbf{1}) \\ \mathbf{1} & \text{if} \quad \max \mathcal{V}(\mathbf{E}, \mathbf{H}) \leq \mathcal{V}(\mathbf{E}, \mathbf{1}) \end{cases}, \tag{1}$$

where $\mathbf{1}$ is an array of ones indicating all equation terms are retained for the entire data array. We use the notation conventions of Bishop,[21] where scalars are italicized, lower case bold represents one dimensional arrays, and upper case bold represents two or higher dimensional arrays.

We propose Equation 1 as a definition of the dynamical regime identification problem, in which one seeks to partition the observations $\mathbf{E}$ into different regimes with different dominant balances, as labeled by $\mathbf{H}_{\text{opt}}$. The dominant balances within $\mathbf{H}_{\text{opt}}$ can be assigned by conventional methods,[22] or they can be assigned by using clustering algorithms to partition data into regimes and subsequently by using dimensionality reduction algorithms to select dominant balances for each regime,[50] as we shall describe below. The verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ defines the optimal regime hypotheses *and* it permits objective comparison of the optimal regime hypotheses from different methods.

Verification criterion

We seek an intelligible *verification criterion*, $\mathcal{V}(\mathbf{E}, \mathbf{H})$, where we define intelligible as both interpretable[8] *and* congruent with domain knowledge.[23] We propose a definition of optimal dominant balances as balances that satisfy two conditions for each regime,

1. the magnitude difference between the selected dominant terms and the negligible terms must be maximized;

2. the magnitude difference between the terms within the selected dominant set must be minimized.

If the first condition is not satisfied, then all equation terms should be retained, i.e., they are all equally dominant. We propose a verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ that is the sample-space weighted average of the *local magnitude score* for the n^{th} sample,

$$\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n) = \frac{\Gamma_n}{1 + \Omega_n} \in [0, 1], \tag{2}$$

where Γ_n is the normalized difference of the log of the smallest magnitude equation term of the set of terms that are considered dominant and the largest magnitude equation term of the set of terms that are considered negligible. We thus refer to Γ_n as the gap in magnitude and $\mathcal{M}_n$ as the local magnitude score. Ω_n is a penalty imposed by the difference between maximum and minimum magnitudes within the set of terms that are considered dominant. Definitions of Γ_n and Ω_n are provided in *Methods*. The score measures the consistence of local truncations of the equation with the observed magnitudes of equation terms.

We propose the weighted average of the score $\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)$, when averaged over N samples, as the verification criterion,

$$\mathcal{V}(\mathbf{E}, \mathbf{H}) = \frac{\sum_{n=1}^{N} w_n \cdot \mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)}{\sum_{n=1}^{N} w_n}, \tag{3}$$

where the array of weights $\mathbf{w} = [w_1, ..., w_N]$ are the discrete differentials of the observed

domain, e.g. space and/or time differentials. For example, if the N observations of data set $\mathbf{E}$ are of equation terms distributed across a two-dimensional space, then the verification criterion is the area-weighted average of all scores for each observation. Optimal verification criteria approach unity, $\mathcal{V}(\mathbf{E}, \mathbf{H}) \to 1$, to favor the identification of regimes that contain dominant balances that reflect the two conditions listed above.

Unsupervised learning framework

We propose an unsupervised machine learning framework[17,18] that automatically discovers regimes by using the verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ (Equation 3) to solve the problem defined by Equation 1. The framework is depicted in Figure 1, in which the regime identification problem is broken into partitioning, hypothesis selection, and hypothesis testing tasks. The left column outlines the conventional *ad hoc* method of regime identification, and the right column depicts our framework. Our framework intentionally emulates the scientific method: the hypothesized regimes $\mathbf{H}$ are tested by evaluating their fit to the equation data $\mathbf{E}$ by using the verification criterion.

The first task shown in Figure 1, row 1, is to partition $\mathbf{E}$ into different regimes. For humans, this task is often the mere act of visually recognizing the difference in dynamics from one sampled region to another. Sonnewald *et al.*[45] first suggested that the heuristic act of recognizing different regimes can be formulated as a partitioning problem that can be credibly solved using clustering algorithms. Clustering algorithms, a class of unsupervised machine learning algorithms that yield a finite set of categories according to similarities or relationships among its objects,[24,25] reveal underlying patterns of sparsity in the data. However, resulting clusters are sensitive to the choice of algorithm parameters,[44] and there is no definition of a cluster that is universal to all clustering algorithms.[27]

The second task, shown in the second row of Figure 1, is to select hypotheses $\mathbf{H}$ for all samples. Humans typically perform this task by estimating characteristic scales from observations and choosing a threshold for each regime by which some terms are deemed

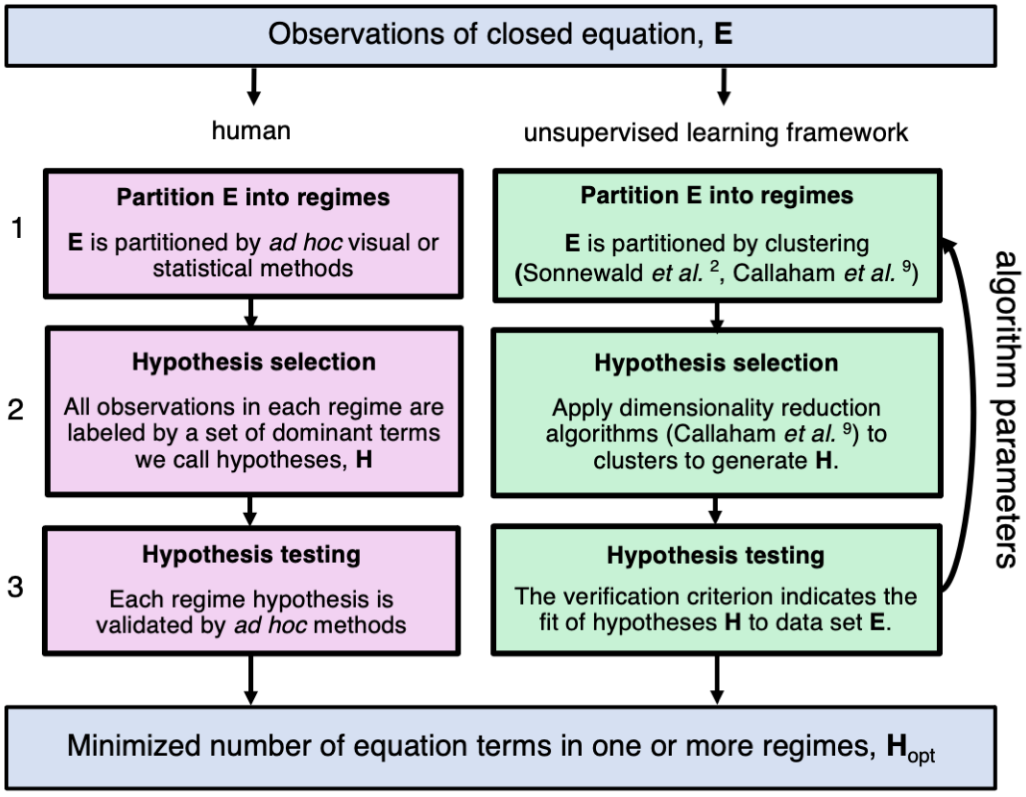

Figure 1: **Diagram of the dynamical regime problem**. Partitioning and empirical scaling analysis performed by a human (left column), and algorithms capable of performing said tasks (right column). The loop over algorithm parameters illustrates the procedure for obtaining $\mathbf{H}_{\text{opt}}$ in Equation 1.

negligible[42] for all samples within a regime. Callaham *et al.*[50] proposed sparse principal component analysis[29] (SPCA) for hypothesis selection because it labels features with small variances as negligible by performing a least absolute shrinkage and selection operator[30] (LASSO) regression on the principal axes from principal component analysis. This application of SPCA, or any other dimensionality reduction technique that pertains to convex data,[20] is geometrically and statistically consistent with EM clustering algorithms (e.g. $K-$means, Gaussian Mixture Models) because both algorithms assume convex, uni-modal, zero-skew data. We propose a custom algorithm for hypothesis selection, combinatorial hypothesis selection (CHS, see Methods), in which dominant balances are selected by calculating the magnitude score (Equation 2) for every possible dominant balance applied to the cluster-average equation terms and then selecting balance associated with the highest score.

The final task shown in Figure 1 is to measure the fit of hypotheses $\mathbf{H}$ to the data $\mathbf{E}$. This task was conventionally performed indirectly through *post hoc* validation of models constructed using relevant identified regimes. Crucially, the framework applies to any choice of clustering and hypothesis selection algorithms and, therefore, allows for objective evaluation and comparisons of different algorithms. We have formalized direct verification of hypotheses by defining the regime identification problem in Equation 1 and proposing a verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$.

The computational complexity of the framework depends on the algorithms chosen for clustering and hypothesis selection because the complexity of verification criterion is $\mathcal{O}(N)$. In *Methods* we demonstrate that the computation time of a single pass through the framework scales polynomially with sample size N for all combinations of a non-parameteric and a parametric clustering algorithm paired with SPCA hypothesis selection and CHS. However, practical application of the framework requires that the user search a subset of the potentially infinite range of possible algorithm parameters. Thus, familiarity with the chosen algorithms and the statistical properties of the data set will reduce the overall computational complexity and expedite dynamical regime discovery.

Examples

Global ocean barotropic vorticity Sonnewald *et al.*[45] used $K-$means clustering, manual hypothesis selection, and algorithm- and problem-specific verification criteria to discover new and canonical oceanic dynamical regimes. They used the time-mean vertically integrated barotropic vorticity equation,

$$\overbrace{\nabla \cdot (f\mathbf{U})}^{\substack{\text{advection of}\\ \text{planetary vorticity}}} = \overbrace{\frac{\nabla p_b \times \nabla H}{\rho}}^{\substack{\text{bottom}\\ \text{pressure torque}}} + \overbrace{\frac{\nabla \times \boldsymbol{\tau}}{\rho}}^{\substack{\text{wind \& bottom}\\ \text{stress curl}}} + \overbrace{\nabla \times \mathbf{A}}^{\substack{\text{nonlinear}\\ \text{torque}}} + \overbrace{\nabla \times \mathbf{B}}^{\substack{\text{diffusive}\\ \text{torque}}}, \tag{4}$$

which describes the balance of processes that control the rate of solid body rotation of a column of seawater (see *Methods* for more details).

Figures 2a) and 2b) show the optimal dominant balances for each regime and their spatial distributions, respectively, for $K-$means clustering and CHS, which are quantitatively similar and qualitatively consistent with the results of Sonnewald *et al.*[45] The optimal verification criterion, $\mathcal{V}(\mathbf{E},\mathbf{H}) = 0.90$, was evaluated at $K = 49$. This result is consistent with the range of prescribed clusters chosen by Sonnewald *et al.*[45] using information theoretic and a custom geographic convergence verification criteria. Figures 2c) and 2d) show the optimal results for HDBSCAN clustering and CHS, corresponding to a verification criterion of $\mathcal{V}(\mathbf{V},\mathbf{H}) = 0.87$. While the the $K-$means or HDBSCAN clustering results identify similar mid-latitude balances, the $K-$means results score higher and include nonlinear balances in expected locations such as the Gulf Stream on the United States eastern seaboard.

Tumor angiogenesis reaction-diffusion Anderson[46,32] calculated numerical solutions with different permutations of terms eliminated to identify dominant processes in tumor angiogenesis (the process by which tumors develop blood flow). We demonstrate that our framework directly identifies which terms are dominant without the need for multiple simulations. The tumor-induced angiogenesis model of Anderson *et al.*[46] is composed of conservation laws of three continuous variables (see *Methods*), where the endothelial-cell density per unit area (cells that rearrange and migrate from preexisting vasculature to form new capillaries), n, is governed by

$$\frac{\partial n}{\partial t} = \overbrace{d\nabla^2 n}^{\text{random motility}} - \overbrace{\nabla\cdot(\chi n\nabla c)}^{\text{chemotaxis}} - \overbrace{\nabla\cdot(\rho n\nabla f)}^{\text{haptotaxis}}, \tag{5}$$

$$= d\nabla^2 n - \chi n\nabla^2 c - \chi\nabla n\cdot\nabla c - n\nabla\chi\cdot\nabla c - \rho n\nabla^2 f - \rho\nabla n\cdot\nabla f, \tag{6}$$

where $\chi(c) = \chi_0/(1+\alpha_0 c)$. Figure 2e) shows the absolute time rate of change of cells as endothelial cell growth propagates towards the tumor. Figures 2f) and Figures 2g) show

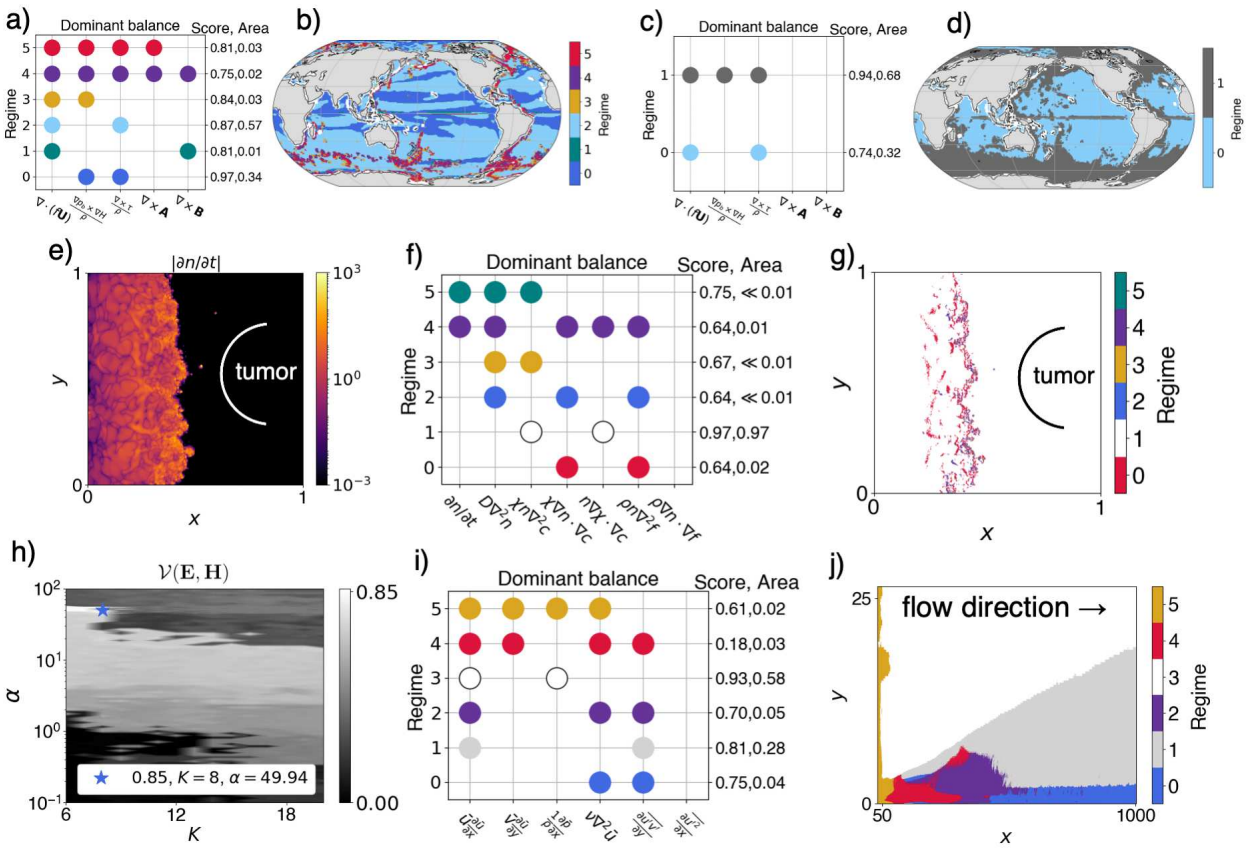

Figure 2: **Examples of framework-identified regimes in two-dimensional spatial domains**. a) to d) show the dominant balances and spatial distributions of regimes in ocean vorticity determined by $K-$means (a,b) and HDBSCAN clustering (c,d) with CHS. e) to g) show endothelial cell growth rates, dominant balances, and spatial distributions of regimes, respectively, determined by $K-$means clustering with CHS. h) to j) show the optimization over the verification criterion, dominant balances, and spatial distributions of regimes, respectively, determined by GMM clustering with SPCA hypothesis selection for a spatially developing turbulent boundary layer.

the dominant balances and spatial distributions of the optimal regimes identified by using $K-$means clustering and CHS. The optimal verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H}) = 0.96$ occurred at $K = 9$. The results from pnly one simulation suggest that the fastest cell growth is a residual of a dominant chemotactic-haptotatic balance, $\chi \nabla n \cdot \nabla c \sim \rho n \nabla^2 f$, (cluster 0, red) in the regions of tissue.

Turbulent boundary layers Canonical turbulent boundary layer dynamical regimes,[22] previously identified by Callaham *et al.*[50] using Gaussian Mixture Model (GMM) clustering

with SPCA hypothesis selection and no quantitative verification criteria, can be identified automatically. Turbulent boundary layers (TBLs) develop as a high-speed flow blows over non-deformable surfaces. The equation that governs the velocity in the direction of the mean flow, u, is

$$\overbrace{\overline{u}\frac{\partial \overline{u}}{\partial x}+\overline{v}\frac{\partial \overline{u}}{\partial y}}^{\substack{\text{mean}\\ \text{momentum flux}\\ \text{divergence}}} = \overbrace{-\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x}}^{\substack{\text{mean}\\ \text{pressure}\\ \text{gradient}}} + \overbrace{\nu\nabla^2\overline{u}}^{\substack{\text{mean}\\ \text{momentum}\\ \text{diffusion}}} - \overbrace{\frac{\partial \overline{u'v'}}{\partial y}-\frac{\partial \overline{u'^2}}{\partial x}}^{\substack{\text{turbulent}\\ \text{momentum flux}\\ \text{divergence}}}, \tag{7}$$

where the velocity and pressure fields (u, v, p) have been decomposed into mean and fluctuating components denoted by overbars and primes, respectively. The x direction points in the downwind direction, and the y direction points in the direction normal to the surface.

Figures 2h) shows the framework optimization over LASSO regression coefficient α and prescribed number of clusters K with the optimal verification criterion of $\mathcal{V}(\mathbf{E}, \mathbf{H}) = 0.85$ for $K = 8$ and $\alpha = 49.94$. The optimal regimes are shown in Figures 2i) and 2j). The regimes are consistent with the results of Callaham *et al.*[50] and with domain knowledge,[33] but notably our framework required no fluid dynamical knowledge.

Conclusions

We, for the first time, have formalized the dynamical regime identification problem by defining it as the maximization of a verification criterion (Equations 1-3). Our formalism is independent of the method by which the optimization problem is solved, thus transforming a previously *ad hoc* method of dynamical analysis into a method for objective regime identification. This formalism is the foundation of our unsupervised learning framework, in which equation data is partitioned by clustering algorithms,[50,45] labeled as dominant balances by dimensionality reduction algorithms,[50] and the fit of the labels to the data is evaluated by the verification criterion. This framework is repeated over a user-specified range of algorithm parameters to find the optimal regimes as defined by the highest verification criterion. We show that our framework yields results consistent with domain knowledge and previous stud-

ies,[50,45] and we emphasize that the framework is broadly applicable to chaotic systems. We anticipate that this work could dramatically expedite the discovery of unknown regimes in new data and accelerate efforts in data-driven dynamical process modeling.[3,4,5,6,7,34,?] This work can be seen as a first step in a new paradigm in the development of machine learning for scientific applications in which the algorithms are inherently intelligible because they are constructed in a manner that explicitly incorporates the scientific method.

Main references

[1] Barenblatt, G. I. *Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics*. 14 (1996).

[2] Sonnewald, M. *et al.* Bridging observation, theory and numerical simulation of the ocean using machine learning. *arXiv preprint, arXiv:2104.12506* (2021).

[3] Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Science Advances* **3**, e1602614 (2017).

[4] Raissi, M. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research* **19**, 932–955 (2018).

[5] Rackauckas, C. *et al.* Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385* (2020).

[6] Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).

[7] Schneider, T., Lan, S., Stuart, A. & Teixeira, J. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters* **44**, 12–396 (2017).

[8] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

[9] Callaham, J. L., Koch, J. V., Brunton, B. W., Kutz, J. N. & Brunton, S. L. Learning dominant physical processes with data-driven balance models. *Nature communications* **12**, 1–10 (2021).

[10] d'Alembert, J. L. R. *Essai d'une Nouvelle Théorie de la Rrésistance des Fluides* (1752).

[11] Prandtl, L. Über flussigkeitsbewegung bei sehr kleiner reibung. *Verhandl. III, Internat. Math.-Kong., Heidelberg, Teubner, Leipzig, 1904* 484–491 (1904).

[12] Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (1994).

[13] Blow, K. J. & Wood, D. Theoretical description of transient stimulated raman scattering in optical fibers. *IEEE Journal of Quantum Electronics* **25**, 2665–2673 (1989).

[14] Seminara, A. *et al.* Osmotic spreading of bacillus subtilis biofilms driven by an extracellular matrix. *Proceedings of the National Academy of Sciences* **109**, 1116–1121 (2012).

[15] Vallis, G. K. *Atmospheric and Oceanic Fluid Dynamics* (2017).

[16] Peixoto, J. P. & Oort, A. H. Physics of climate (1992).

[17] Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artificial intelligence* **97**, 273–324 (1997).

[18] Dy, J. G. & Brodley, C. E. Feature selection for unsupervised learning. *Journal of machine learning research* **5**, 845–889 (2004).

[19] Sonnewald, M., Wunsch, C. & Heimbach, P. Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science* **6**, 784–794 (2019).

[20] Van Der Maaten, L., Postma, E. & Van den Herik, J. Dimensionality reduction: a comparative. *J Mach Learn Res* **10**, 13 (2009).

[21] Bishop, C. M. *Pattern Recognition and Machine Learning* (2006).

[22] Tennekes, H. & Lumley, J. *A First Course in Turbulence* (1972).

[23] Wang, J., Oh, J., Wang, H. & Wiens, J. Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2417–2426 (2018).

[24] MacQueen, J. *et al.* Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 281–297 (Oakland, CA, USA, 1967).

[25] Hartigan, J. A. *Clustering algorithms* (John Wiley & Sons, Inc., 1975).

[26] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[27] Estivill-Castro, V. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter* **4**, 65–75 (2002).

[28] Zohuri, B. *Dimensional Analysis beyond the Pi Theorem* (2017).

[29] Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics* **15**, 265–286 (2006).

[30] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).

[31] Anderson, A. R. & Chaplain, M. A. J. Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bulletin of mathematical biology* **60**, 857–899 (1998).

[32] Anderson, A. R., Chaplain, M. A., Newman, E. L., Steele, R. J. & Thompson, A. M. Mathematical modelling of tumour invasion and metastasis. *Computational and mathematical methods in medicine* **2**, 129–154 (2000).

[33] Schetz, J. A. & Bowersox, R. D. *Boundary Layer Analysis* (2011).

[34] Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *science* **324**, 81–85 (2009).

Methods

Data Availability Statement

The tumor angiogenesis reaction-diffusion and synthetic datasets generated during and analysed during the current study are available from the corresponding author on reasonable request. The Estimating the Circulation and Climate of the Ocean (ECCO) ocean state estimate (https://data.nas.nasa.gov/ecco/), version 4 release 2,[35,36,37,38] was used to estimate the global ocean vorticity budget. The spatially-developing turbulent boundary layer data is available in the Johns Hopkins University turbulence database (http://turbulence.pha.jhu.edu) in the developing boundary layer repository.[39]

Code Availability Statement

The tumor angiogenesis reaction-diffusion code is publicly available at github.com/bekaiser/tumor_dynamics. Pending permission from Los Alamos National Laboratory, the unsupervised learning framework and synthetic data generation codes will be publicly available at github.com/bekaiser.

An intelligible verification criterion

The local order-of-magnitude score, $\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)$, hereafter the local magnitude score, pertaining to a single observation, measures the magnitude gap between dominant terms $\mathbf{h}_n \cdot \mathbf{e}_n$ and negligible terms $|\mathbf{h}_n - \mathbf{1}| \cdot \mathbf{e}_n$ in a single observation $\mathbf{e}_n$ (recall that the terms that are selected as dominant are labeled by $h_{ni} = 1$, and the neglected terms are labeled by $h_{ni} = 0$). Define $\mathrm{F} = \{1, ..., D\}$ as the *index set*[40] of the indices of the full set equation terms in vector $\mathbf{e}_n$, such that

$$\mathbf{e}_n = \bigcup_{i \in \mathrm{F}} e_{ni}, \tag{8}$$

for observation n such that $1 \leq n \leq N$.

We refer to the binary sets that represent the dominant terms as *hypotheses* because they represent informal equation truncations that are not guaranteed to have asymptotic properties. The hypotheses for the entire data set $\mathbf{E}$ form an array, $\mathbf{H}$, which has the same dimensions as $\mathbf{E}$, [number of samples $\times$ number of equation terms]. The hypothesis vectors for each observation can be expressed as

$$\mathbf{h}_n = \bigcup_{i \in \mathrm{F}} h_{ni}, \tag{9}$$

where $\mathbf{h}_n$ is an indicator function[41] that consists entirely of ones and zeros, which represent selected dominant terms and negligible terms, respectively.

The indices of elements in $\mathbf{e}_n$ that are selected as dominant terms by the hypothesis $\mathbf{h}_n$ form the selection index set S_n, where

$$\mathrm{S}_n \subseteq \mathrm{F}. \tag{10}$$

The number of selected elements may vary for each observation n, and if $\mathrm{S}_n = \mathrm{F}$ then $\mathbf{h}_n = \mathbf{1}$ and no equation terms are neglected. It follows that the remainder index set R_n for the n^{th} observation is defined by set subtraction

$$\mathrm{R}_n = \mathrm{F} - \mathrm{S}_n, \tag{11}$$

and, therefore, the remainder index set and selected index set are non-overlapping,

$$\mathrm{R}_n \cap \mathrm{S}_n = \varnothing. \tag{12}$$

Thus the cardinality, or size, of the selected index set and remainder index set are $2 \leq \mathrm{card}(\mathrm{S}_n) \leq D$ and $0 \leq \mathrm{card}(\mathrm{R}_n) \leq D - 2$, respectively. The lower bound of two selected terms is not necessary nor required; we impose it because a dominant balance of just one

term is conceptually ambiguous. Let the arrays of selected and remainder equation terms from $\mathbf{e}_n$ be $\mathbf{s}_n$ and $\mathbf{r}_n$, respectively. $\mathbf{s}_n$ and $\mathbf{r}_n$ are normalized by the smallest element of $\mathbf{e}_n$ and defined as

$$\mathbf{s}_n = \frac{\bigcup_{i \in \mathrm{S}_n} |e_{ni}|}{\min\left(\bigcup_{i \in \mathrm{F}} |e_{ni}|\right)}, \tag{13}$$

$$\mathbf{r}_n = \frac{\bigcup_{i \in \mathrm{R}_n} |e_{ni}|}{\min\left(\bigcup_{i \in \mathrm{F}} |e_{ni}|\right)}, \tag{14}$$

respectively. If $\min(\bigcup_{i \in \mathrm{F}} |e_{ni}|) = 0$, then the minimum non-zero absolute valued element of $\mathbf{e}_n$ replaces the denominators in Equations 13 and 14. Let the relative magnitude gap between the normalized subsets, Γ, be defined as a scalar for each n^{th} observation:

$$\Gamma_n = \begin{cases} \frac{\log_{10}(\min(\mathbf{s}_n) - \max(\mathbf{r}_n))}{\log_{10}(\min(\mathbf{s}_n) + \max(\mathbf{r}_n))} & \text{if} \quad \min(\mathbf{s}_n) > \max(\mathbf{r}_n) \\ 0 & \text{if} \quad \min(\mathbf{s}_n) \leq \max(\mathbf{r}_n) \end{cases} . \tag{15}$$

The magnitude gap Γ is normalized such that $\Gamma \in [0, 1]$, by imposing the floor condition *if* $\Gamma < 0$ *then* $\Gamma = 0$ to correct for spurious large amplitude negative values of Γ that arise as $\min(\mathbf{s}_n) \to \max(\mathbf{r}_n)$. $\Gamma \to 1$ as the number of orders of magnitude between the element with the minimum absolute value of the selected subset and element with the maximum absolute value of the remainder subset approaches infinity, and if $\Gamma = 1$ then the hypothesis that the terms in the selected subset dominate the terms in the remainder subset is exact (for any numerical implementation the optimal is limited by machine precision, so $\Gamma \approx 1$). If the magnitude difference between the two subsets vanishes, then $\Gamma \to 0$, and if a remainder subset term exceeds the absolute magnitude of the selected subset, then the hypothesis does not represent a dominant balance, and $\Gamma = 0$ is prescribed.

Since the goal is to choose the selected subset, $\mathbf{s}_n$, such that it corresponds to the dominant terms, the feature magnitudes of the selected subset should be approximately the same. Otherwise, the smallest magnitude term(s) in the selected subset should be removed from

that subset and added to the remainder subset. To penalize large absolute magnitude differences within the selected subset, we introduce the scalar penalty for the n^{th} observation,

$$\Omega_n = \log_{10}(\max(\mathbf{s}_n)) - \log_{10}(\min(\mathbf{s}_n)) \in [0, \infty). \tag{16}$$

A base 10 logarithm is chosen for the penalty because it corresponds most directly to the notion of orders of magnitude. If $\Omega_n \to 0$, the absolute magnitudes of the selected subset terms approach uniformity.

Scaling and the local magnitude score

The local magnitude score is a) invariant to the magnitude of the feature vector $\mathbf{x}_n$ and b) invariant to the sign of the elements of the feature vector,

$$\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n) = \mathcal{M}_n(\pm c\mathbf{e}_n, \mathbf{h}_n), \tag{17}$$

where c is a positive scalar constant. Therefore, the score is invariant to the choice of dimensional or non-dimensional equations, and, equivalently, it can be applied to Buckingham Π theorem to identify dominant Π groups. Readers unfamiliar with the Buckingham Π theorem and how it pertains to the non-dimensionalization of partial differential equations are referred to Zohuri.[42]

Combinatorial hypothesis selection

We propose a simple hypothesis selection algorithm that we will refer to as the combinatorial hypothesis selection (CHS) algorithm. Since the number of all possible hypotheses for an equation is a permutation of two types (0 or 1) with repetition allowed, the number of possible hypotheses is $\mathcal{O}(2^D)$. If the number of equation terms, D, is not large, then hypotheses can be feasibly generated by calculating the magnitude score (Equation 2) for all possible hypotheses and then selecting the hypothesis that is awarded the highest score. Equation 2

can be applied to a single data sample or to an average of samples. The exponential time complexity limits the feasibility of computing CHS to equations with relatively few terms.

Synthetic data for complexity analyses

Consider a two-dimensional array of data with an even number of equation terms, where half of the terms are two orders of magnitude larger in one half of the domain and vice versa, with no variability in the x direction. Figure 3a), in *Extended data* shows the synthetic data e_{ni} consisting of $D = 8$ equation terms, featuring two regimes in which dominant terms have magnitudes of $\mathcal{O}(10)$ and negligible terms have magnitudes of $\mathcal{O}(10^{-1})$. The regimes are separated by a discontinuity at $y = 0.5$. Multiplicative sinusoidal noise is added to give the two regions variance that is proportional to 10% of the signal amplitude in each region. The two regions can be considered dynamical regimes: in each, half of the terms dominate the other equation terms by two orders of magnitude. The regions of dominant terms are prescribed by the Heaviside step function $\mathcal{H}$, such that:

$$\mathbf{e}_i(x, y) = (-1)^i \eta(y)(\lambda \mathcal{H}(\phi) + \beta), \tag{18}$$

$$\eta(y) = \eta_0 \sin(\omega y), \tag{19}$$

$$\phi = \begin{cases} y - 0.5 & \forall i < D/2 \\ 0.5 - y & \forall i \geq D/2 \end{cases}, \tag{20}$$

where x and y are spatial coordinates. The equation closes exactly for all N samples, $\sum_{i=1}^{D} e_{ni} = 0$, and the prescribed coefficients are $\lambda = 10^1$, $\beta = 10^{-1}$, $\eta_0 = 10^{-1}$, and $\omega = 10\pi$, Once again, e_{ni} is the n^{th} observation of the i^{th} feature.

Figures 3b), 3c), and 3d) show the results using $K-$means clustering and SPCA hypothesis selection. Figure 3b) shows the variation of the verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ with α, the LASSO regression coefficient for SPCA, and K, the prescribed number of clusters for $K-$means clustering. The optimal is marked with the blue star, $\mathcal{V}(\mathbf{E}, \mathbf{H}) = 0.996$, though

much of the white band in Figure 3b) corresponds to equivalently optimal results. Figure 3c) shows the dominant balances of the optimal regimes and Figure 3d) shows the spatial distribution of the optimal regimes. Identical optimal regimes were identified by using $K-$means clustering with CHS, by using Hierarchical Density-Based Scan (HDBSCAN) clustering and SPCA hypothesis selection, and HDBSCAN and CHS. The optimal regimes are robust because the magnitude separation between dominant and negligible terms is at least two orders of magnitude everywhere and spatial boundaries of the regimes are discontinuous.

While comprehensive complexity analyses are beyond the scope of this Article, we can infer some general properties of the framework's time complexity. Exhaustive searches over algorithm parameters may very well be NP-hard. The search over K, the prescribed number of clusters for $K-$means, to minimize the sum of the square of the Euclidean distance of each data point to its nearest center is NP-hard even for just two equation terms,[43] $D = 2$. CHS is prohibitively complex at large numbers of equation terms D because its complexity scales with the number of possible dominant balances, $\mathcal{O}(2^D)$. However, SPCA hypothesis selection adds an additional parameter for optimization; therefore, we recommend CHS for equations with fewer terms than 10.

The average wall time elapsed for the framework computations over algorithm parameter ranges are shown in Figure 3. The algorithm parameter ranges were specfied as follows: for $K-$means clustering the number of prescribed clusters K was specified as $K = \{2, ..., 10\}$ and the other hyperparameters were the default choices as provided by SciKit Learn.[44] For HDBSCAN clustering the prescribed minimum number of samples for a cluster was specified as 100 samples, and the minimum cluster size was varying from 2000 samples to 3000 samples. For hypothesis selection by SPCA, the LASSO regression coefficient was varied between 10^{-2} and 10^2. Figures 3e) and 3f) show the time complexity with respect to number of samples N and number of equation terms D, respectively. Each point represents the average wall time for one pass through the framework (Figure 1). Figure 3e) shows that the computation time scales polynomially with sample size N for all algorithm choices. Figure

3f) shows that CHS becomes prohibitively complex with increasing number of equation terms because its complexity scales with $\mathcal{O}(2^D)$. However, SPCA adds an additional parameter for optimization; therefore, we recommend choosing CHS if the number of equation terms is less than 10.

Global ocean barotropic vorticity

We use the vorticity data processing method of Sonnewald *et al.*,[45] who computed a 20-year mean of the ECCO data at 1° resolution to calculate terms of the vertically integrated barotropic vorticity equation. The variables in Equation 4 are: f is the Coriolis parameter, $\mathbf{U}$ is the vertically integrated horizontal velocity, p_b is the bottom pressure, H is the depth, ρ is a reference density, $\boldsymbol{\tau}$ represents surface stress, ∇ is applied only to the horizontal coordinates, $\mathbf{A}$ contains nonlinear horizontal momentum fluxes, and $\mathbf{B}$ contains linear horizontal diffusive fluxes.

Tumor angiogenesis reaction-diffusion

We solve the non-dimensional, tumor-induced angiogenesis governing equations of Anderson *et al.*[46] The tumor angiogenic factor concentration, c (chemicals secreted by the tumor that promote angiogenesis), and the fibronectin concentration, f (macromolecules that are secreted by n and stimulate the directional migration of n), are governed by

$$\frac{\partial f}{\partial t} = \beta n - \gamma n f, \tag{21}$$

$$\frac{\partial c}{\partial t} = -\eta c n. \tag{22}$$

Endothelial cell migration up the fibronectin concentration gradient is termed haptotaxis,[47,48] while endothelial cell migration up the gradient of tumor angiogenic factor concentration is termed chemotaxis.[49]

In Figure 2c), the tumor is located at $x, y = 1, 0.5$, and the endothelial cell growth

is propagating in the positive x direction towards the tumor. Figures 2d) and 2d) show the regime distributions and dominant balances, respectively, for the optimal results for $K-$means clustering and CHS hypothesis selection.

We numerically solve the same problem as Anderson *et al.*,[46] with the exception that 1% amplitude red noise was added to the initial c and f fields to provide additional variability for illustrative purposes. A second-order accurate finite difference code was used to calculate each term in the expanded form of the endothelial cell density equation, such that $\mathbf{E}$ is composed of observations of the terms in Equation 6). We employ the same boundary conditions, initial conditions, and constant coefficients (d, α_0, χ_0, ρ, β, γ, and η) as[46] at double the resolution. Second-order finite differences were employed for spatial derivatives, and 4th-order adaptive Runge-Kutta was employed for the temporal evolution. No flux boundary conditions were applied to all four boundaries of the square domain:

$$\mathbf{n} \cdot (d\nabla n - \chi(c) n \nabla c - \rho n \nabla f) = 0, \tag{23}$$

where $\mathbf{n}$ is the unit normal vector to the boundaries. The initial conditions, for a circular tumor (TAF distribution) some distance from three clusters of endothelial cells, are:

$$c(x, y, 0) = \begin{cases} 1, & 0 \leq r \leq 0.1 \\ \frac{(\nu - r)^2}{\nu - r_0}, & 0.1 < r \leq 1 \end{cases}, \tag{24}$$

where $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}$.

$$f(x, y, 0) = k \mathrm{e}^{-\frac{x^2}{\epsilon_1}}, \tag{25}$$

$$n(x, y, 0) = \mathrm{e}^{-\frac{x^2}{\epsilon_2}} \sin^2(6\pi y), \tag{26}$$

where $\nu = (\sqrt{5} - 0.1)/(\sqrt{5} - 1)$, $r_0 = 0.1$, $x_0 = 1$, $y_0 = 1/2$, $k = 0.75$, $\epsilon_1 = 0.45$, $\epsilon_2 = 0.001$.

The constant coefficients were specified as $d = 0.00035$, $\alpha_0 = 0.6$, $\chi_0 = 0.38$, $\rho = 0.34$, $\beta = 0.05$, $\gamma = 0.1$, and $\eta = 0.1$.

Turbulent boundary layers

We use the same turbulence simulation data set and data processing method as Callaham *et al.*,[50] where ρ and ν are constants that represent the fluid density and kinematic viscosity, respectively. The overbar averaging operator represents averaging over the spanwise direction as well as averaging over time, and the diffusion operator is defined as $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$.

Methods references

[35] Forget, G. *et al.* Ecco version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation (2015).

[36] Wunsch, C. & Heimbach, P. Dynamically and kinematically consistent global ocean circulation and ice state estimates. In *International Geophysics*, vol. 103, 553–579 (Elsevier, 2013).

[37] ECCO Consortium. A twenty-year dynamical oceanic climatology: 1994-2013. part 1: Active scalar fields: Temperature, salinity, dynamic topography, mixed-layer depth, bottom pressure. (2017).

[38] ECCO Consortium. A twenty-year dynamical oceanic climatology: 1994-2013. part 2: Velocities, property transports, meteorological variables, mixing coefficients (2017).

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

[40] Munkres, J. R. *Topology* (2000).

[41] Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms* (2009).

[42] Zohuri, B. *Dimensional Analysis beyond the Pi Theorem* (2017).

[43] Mahajan, M., Nimbhorkar, P. & Varadarajan, K. The planar k-means problem is np-hard. *Theoretical Computer Science* **442**, 13–21 (2012).

[44] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[45] Sonnewald, M., Wunsch, C. & Heimbach, P. Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science* **6**, 784–794 (2019).

[46] Anderson, A. R. & Chaplain, M. A. J. Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bulletin of mathematical biology* **60**, 857–899 (1998).

[47] Carter, S. B. Principles of cell motility: the direction of cell movement and cancer invasion. *Nature* **208**, 1183–1187 (1965).

[48] Carter, S. B. Haptotaxis and the mechanism of cell motility. *Nature* **213**, 256–260 (1967).

[49] Sholley, M., Ferguson, G., Seibel, H., Montour, J. & Wilson, J. Mechanisms of neovascularization. vascular sprouting can occur without proliferation of endothelial cells. *Laboratory investigation; a journal of technical methods and pathology* **51**, 624–634 (1984).

[50] Callaham, J. L., Koch, J. V., Brunton, B. W., Kutz, J. N. & Brunton, S. L. Learning dominant physical processes with data-driven balance models. *Nature communications* **12**, 1–10 (2021).

Acknowledgements

This work was performed under the auspices of DOE. Financial support comes partly from Los Alamos National Laboratory (LANL), Laboratory Directed Research and Development (LDRD) project "Machine Learning for Turbulence," 20180059DR. LANL, an affirmative action/equal opportunity employer, is managed by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract 89233218CNA000001. Computational resources were provided by the Institutional Computing (IC) program at LANL.

MS acknowledges funding from Cooperative Institute for Modeling the Earth System, Princeton University, under Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce.

Author contributions

B.K. drafted the manuscript, contributed to the design of the framework, wrote all software, and performed all example analyses. J.S. substantively revised the manuscript, contributed to the design of the framework, and contributed to the problem formulation. M.S. substantively revised the manuscript, contributed to the software implementation of the framework, and contributed the vorticity data set. D.L. revised the manuscript, contributed to the design of the framework, and substantively contributed to the problem formulation.

Competing interest declaration

The authors declare no competing interests.

Extended data

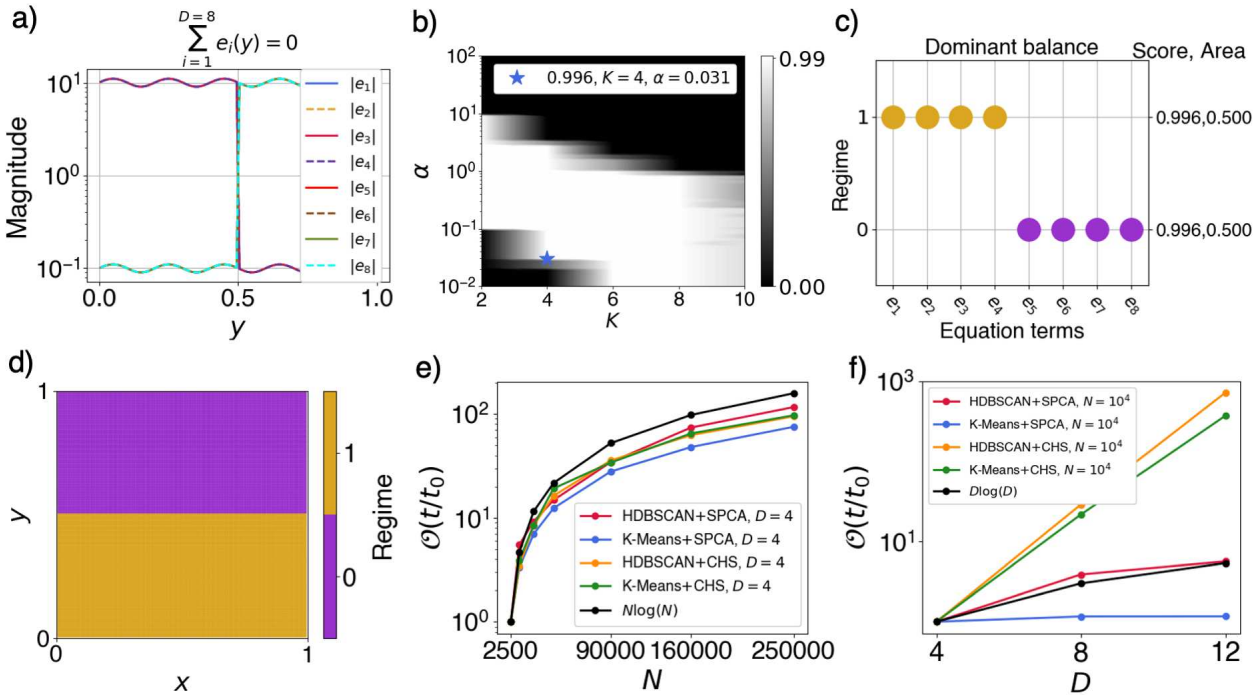

Figure 3: **Synthetic data and computational complexity.** The synthetic data for all y at constant x is shown in a). b) shows the variation global magnitude score as the number of prescribed clusters for $K-$means and the LASSO regression coefficient α for SPCA are varied. c) and d) show the dominant balances and their spatial distribution for the optimal results that occur in the white band in b) which correspond to the verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H}) = 0.996$. e) and f) show the variation in wall time as a function of sample size and number of equation terms, respectively, for difference algorithm choices.