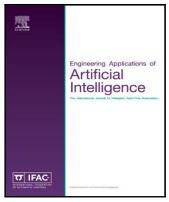
Engineering Applications of Artificial Intelligence 116 (2022) 105496

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Automated identification of dominant physical processes

Bryan E. Kaiser [a,*], Juan A. Saenz [a], Maike Sonnewald [b,c,d], Daniel Livescu [a]

[a] Los Alamos National Laboratory, Los Alamos, NM, USA
[b] Program in Atmospheric & Oceanic Sciences, Princeton University, Princeton, NJ, USA
[c] Ocean & Cryosphere Division, NOAA/OAR Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[d] School of Oceanography, University of Washington, Seattle, WA, USA

ARTICLE INFO

Keywords:
Dominant balance
Unsupervised machine learning
Clustering
Dynamical process
Nonlinear partial differential equations

ABSTRACT

The identification of processes that locally and approximately dominate dynamical system behavior has enabled significant advances in understanding and modeling nonlinear differential dynamical systems. Conventional methods of dominant process identification involve piecemeal and *ad hoc* (non-rigorous, informal) scaling analyses to identify dominant balances of governing equation terms and to delineate the spatiotemporal boundaries (boundaries in space and/or time) of each dominant balance. For the first time, we present an objective global measure of the fit of dominant balances to observations, which is desirable for automation, and was previously undefined. Furthermore, we propose a formal definition of the dominant balance identification problem in the form of an optimization problem. We show that the optimization can be performed by various machine learning algorithms, enabling the automatic identification of dominant balances. Our method is algorithm agnostic and it eliminates reliance upon expert knowledge to identify dominant balances which are not known beforehand.

1. Introduction

Observations of nonlinear dynamical systems can exhibit heterogeneous patterns of non-asymptotic dominant balances when subjected to asymmetric initial and/or boundary conditions. A dominant balance (Callaham et al., 2021) is a subset of governing equation terms which locally and statistically dominates the remaining equation terms by at least an order of magnitude. The dominant balances in observations of nonlinear systems are often non-asymptotic (Barenblatt, 1996). Non-asymptotic dominant balances do not permit equation truncation through formal methods with well defined convergence properties. Highly nonlinear differential dynamical systems with asymmetries imposed by initial and/or boundary conditions often exhibit multiple dominant balances delineated by boundaries in space and/or time.

One example of the importance of dominant balance identification is illustrated by d'Alembert's "zero drag" paradox (d'Alembert, 1752), which took over 150 years to be resolved by Prandtl (Prandtl, 1904). d'Alembert argued that, since frictional forces in fluid flow are very small, they can be neglected everywhere in the fluid. However, d'Alembert's argument meant that balls and cylinders flying through the air should experience zero drag. The paradox arose because frictional forces in fluid flows are often small, yet drag forces are virtually omnipresent in observations. The root of the paradox is the assumption of a global and absolute, rather than relative and local, threshold for the importance of frictional forces in fluid flow. Upon realizing these properties of dominant balances, Prandtl resolved the paradox by positing that the frictional terms in the fluid dynamical governing equations cannot be ignored within thin boundary layer regions on the surface of immersed objects (see Fig. 1). The dominant balances identified by Prandtl directly informed the development of aerodynamic stall prediction and, indeed, the whole field of aerodynamics.

Crucially, the equation terms that constitute dominant balances are dominant relative to the magnitude of the equation terms deemed negligible within the same dominant balance region. All equation terms in one dominant balance can be much smaller or larger than all equation terms in another dominant balance. Setting a global magnitude threshold on equation terms beforehand is problematic because dominant balances are useful as *localized* tools for diagnosing relevant dynamical processes. Dominant balance identification can aid the development of statistical models. This has been done in fields as diverse as nonlinear waves, plasma dynamics, earthquake dynamics, general relativity, quantum field theory, biochemical reaction–diffusion dynamics, fibrillation dynamics, epilepsy, turbulent flows, fiber optics, biofilm dynamics, weather, and climate dynamics (Strogatz, 1994; Blow and Wood, 1989; Seminara et al., 2012; Vallis, 2017; Peixoto and Oort, 1992).

* Corresponding author.
E-mail address: bkaiser@lanl.gov (B.E. Kaiser).

https://doi.org/10.1016/j.engappai.2022.105496
Received 16 June 2022; Received in revised form 20 September 2022; Accepted 28 September 2022
Available online xxxx
0952-1976/

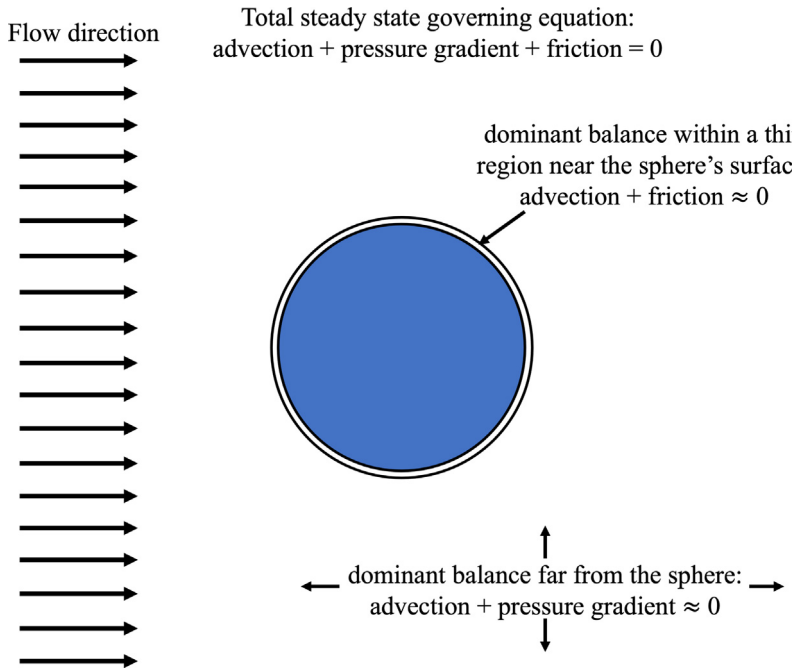


Fig. 1. Illustration of the resolution of d'Alembert's paradox. Prandtl observed that frictional terms, which may be negligibly small far from the surface of a body immersed in a fluid flow, are not negligibly small near the surface of the immersed body.

Conventional methods for dominant balance identification involve combining any available theoretical and/or empirical domain knowledge with estimates of characteristic scales to identify dominant balances. Callaham et al. (2021) proposed an unsupervised machine learning method that effectively automates the tedious, piecemeal, and *ad hoc* manner of conventional methods of dominant balance identification. However, to effectively use their method, the dominant balances must be known beforehand to choose the correct algorithm parameters. What if the dominant balances are not known beforehand? Here, we propose a definition of the dominant balance identification problem as an optimization problem. We then propose a simple algebraic verification criterion to generically define the optimal. By defining the problem and by proposing an optimization function consistent with previous balance identification methods, this study will enable robust and credible automation of dominant balance identification.

2. Problem formulation

Given the array of data $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_N]$, consisting of N observations of the D dimensional vector of equation terms $\mathbf{e}_n$, we seek to label each observation with a D dimensional hypothesis vector $\mathbf{h}_n$, where $h_{ni} \in \{0, 1\}$ for each nth observation of the ith 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 $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N]$, and zeros in each hypothesis vector $\mathbf{h}_n$ indicate equation terms in $\mathbf{e}_n$ that are neglected. We choose a verification 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 (Bishop, 2006), where scalars are italicized, lower case bold represents one dimensional arrays, and upper case bold represents two or higher dimensional arrays.

We propose Eq. (1) as a definition of the dominant balance identification problem, in which one seeks to partition the observations $\mathbf{E}$ into distinct regions each with different dominant balances, as labeled by $\mathbf{H}_{\text{opt}}$. The dominant balances within $\mathbf{H}_{\text{opt}}$ can be assigned by conventional *ad hoc* methods (Tennekes and Lumley, 1972), or they can be assigned by using clustering algorithms to partition data into distinct regions and subsequently by using dimensionality reduction algorithms to select dominant balances for each region (Callaham et al., 2021).

3. The local magnitude score

To define a verification criterion, we propose to define optimal dominant balances as balances that satisfy two conditions for each region,

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. These qualitative definitions are consistent with conventional *ad hoc* methods of scaling analysis (Zohuri, 2017).

The local order-of-magnitude score, $\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)$, hereafter the local magnitude score (LMS), pertains to a single observation of equation-space. It is a measure of the magnitude gap between dominant terms $\mathbf{h}_n \cdot \mathbf{e}_n$ and negligible terms $|\mathbf{h}_n - \mathbf{1}| \cdot \mathbf{e}_n$ (terms that are selected as dominant are labeled by $h_{ni} = 1$ and the neglected terms are labeled by $h_{ni} = 0$, for the nth observation and the ith equation term).

To define the LMS we must first define the selected and neglected sets of equation terms and their respective indices. Define $\mathrm{F} = \{1, \dots, D\}$ as the index set (Munkres, 2000) of the indices of the full set of equation terms in vector $\mathbf{e}_n$, such that

$$\mathbf{e}_n = \bigcup_{i \in \mathrm{F}} e_{ni}, \tag{2}$$

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 ×

number of equation terms]. The hypothesis vectors for each observation can be expressed as

$$\mathbf{h}_n = \bigcup_{i \in \mathrm{F}} h_{ni}, \tag{3}$$

where $\mathbf{h}_n$ is an indicator function (Cormen et al., 2009) 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

$$\mathrm{S}_n \subseteq \mathrm{F}. \tag{4}$$

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 nth observation is defined by set subtraction

$$\mathrm{R}_n = \mathrm{F} - \mathrm{S}_n, \tag{5}$$

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

$$\mathrm{R}_n \cap \mathrm{S}_n = \varnothing. \tag{6}$$

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 not a useful balance of terms (see Appendix A for further detail).

Let the arrays of selected and remainder equation terms from $\hat{\mathbf{e}}_n$ be $\mathbf{s}_n$ and $\mathbf{r}_n$, respectively. $\hat{\mathbf{e}}_n$ is the set of equation terms that are normalized such that the smallest equation term magnitude is unity,

$$\hat{\mathbf{e}}_n = \frac{\bigcup_{i \in \mathrm{F}_n} |e_{ni}|}{\min\left(\bigcup_{i \in \mathrm{F}} |e_{ni}|\right)} \geq 1, \tag{7}$$

where $\min(\bigcup_{i \in \mathrm{F}} |e_{ni}|) \neq 0$. If $\min(\bigcup_{i \in \mathrm{F}} |e_{ni}|) = 0$, then the minimum non-zero absolute valued element of $\mathbf{e}_n$ replaces the denominator in Eq. (7). The selected terms $\mathbf{s}_n$ (hypothesized as dominant) and remainder terms $\mathbf{r}_n$ (hypothesized as negligible) are defined as

$$\mathbf{s}_n = \bigcup_{i \in \mathrm{S}_n} \hat{e}_{ni}, \tag{8}$$

$$\mathbf{r}_n = \bigcup_{i \in \mathrm{R}_n} \hat{e}_{ni}, \tag{9}$$

respectively. Let the magnitude gap between the normalized subsets, Γ_n, be defined as a scalar for each nth observation,

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

The magnitude gap is normalized such that $\Gamma_n \in [0, 1]$ by imposing the floor condition (if $\Gamma_n < 0$ then $\Gamma_n = 0$) to correct for spurious large negative values of Γ_n that arise as $\min(\mathbf{s}_n) \to \max(\mathbf{r}_n)$. The behavior of Γ_n, defined in Eq. (10), as a function of the ratio $\min(\mathbf{s}_n)/\max(\mathbf{r}_n)$ is shown in Fig. 2, which shows that $\Gamma_n \to 1$ as the minimum magnitude of the selected subset approaches two orders of magnitude greater than the maximum of magnitude of the remainder subset.

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 a scalar penalty for the nth observation,

$$\Omega_n = \log_{10}(\max(\mathbf{s}_n)) - \log_{10}(\min(\mathbf{s}_n)) \in [0, \infty). \tag{11}$$

A base 10 logarithm is chosen for the penalty because it corresponds most directly to the notion of orders of magnitude. Ω_n is defined such that as $\mathrm{std}(\mathbf{s}_n) \to 0$, so does the penalty, $\Omega_n \to 0$.

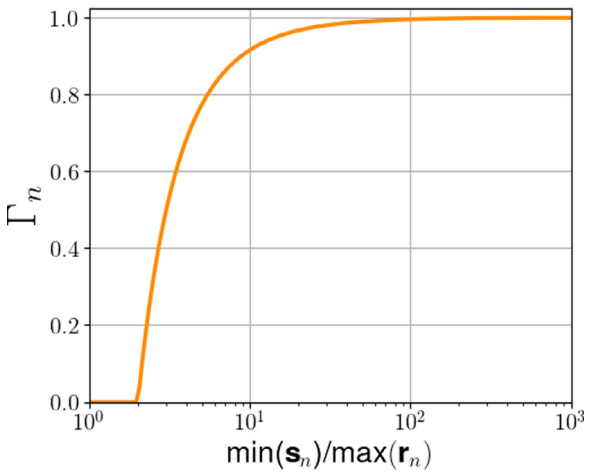


Fig. 2. The normalized magnitude gap between the selected and neglected equation terms. The convergence of the normalized magnitude gap of the nth observation, Γ_n, as a function of the number of orders of magnitude that separate the minimum magnitude term of the selected equation terms and the maximum magnitude remainder equation terms. $\Gamma_n \approx 1$ as at least two orders of magnitude separate the selected and remainder equation terms. If the selected equation terms are all the same magnitude then $\Omega_n = 0$ and $\mathcal{M}_n = \Gamma_n$. This figure indicates that we have formalized the notion of a dominant balance as bias towards the preferential selection of sets of terms that dominate neglected terms by at least two orders of magnitude.

Finally, the LMS for the nth sample, is given by

$$\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n) = \frac{\Gamma_n}{1 + \Omega_n} \in [0, 1]. \tag{12}$$

The score measures the consistency of truncations of the equation with the average observed magnitudes of equation terms for the nth observation. While Eq. (12) defines the optimization problem in terms of a single variable, a two-variable optimization (i.e., the maximization of Γ_n and minimization of Ω_n) is a viable alternative methodology to the uni-variate optimization presented here.

3.1. Local magnitude score example

To understand the LMS, consider a simple problem in which one needs to decide which equation terms to keep and which to neglect. Consider the one dimensional form of the heat equation with two additional terms (i.e. advection and source),

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial x^2} - u \frac{\partial T}{\partial x} + \lambda(T - T_0) = 0. \tag{13}$$

Now assume that after scale analysis (Zohuri, 2017) for a given problem of interest we find that the terms in Eq. (13) scale as

$$\mathcal{O}\left(\frac{\partial T}{\partial t}\right) \sim 1, \tag{14}$$

$$\mathcal{O}\left(\kappa \frac{\partial^2 T}{\partial x^2}\right) \sim \epsilon, \tag{15}$$

$$\mathcal{O}\left(u \frac{\partial T}{\partial x}\right) \sim 1, \tag{16}$$

$$\mathcal{O}\left(\lambda(T - T_0)\right) \sim \epsilon, \tag{17}$$

where we will consider both the case where ϵ is a small parameter and the case where ϵ is a large parameter. We can rewrite Eq. (13) in terms of the scale analysis

$$1 - \epsilon - 1 + \epsilon = 0. \tag{18}$$

There are two dominant balances possible for Eq. (18), corresponding to the small parameter case $\epsilon \ll 1$ and to the large parameter case $\epsilon \gg 1$. If there is only a single observation of the terms of Eq. (18), then $N = 1$ and it can be expressed in equation data array as

$$\mathbf{e}_1 = [1, \epsilon, -1, -\epsilon], \tag{19}$$

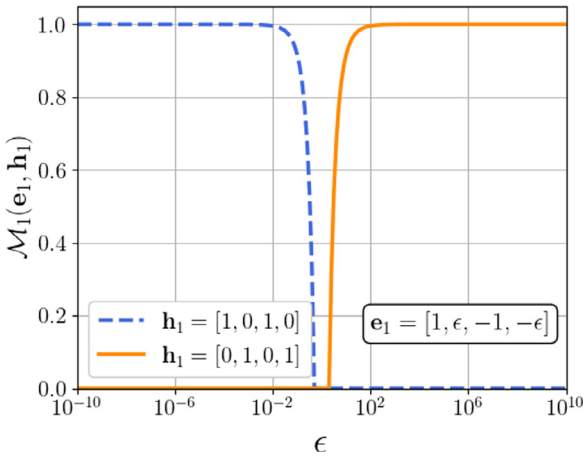


Fig. 3. LMS example. The LMS for the hypotheses $\mathbf{h}_1 = [1, 0, 1, 0]$ and $\mathbf{h}_1 = [0, 1, 0, 1]$ when applied to Eq. (19) as ϵ is varied from $\epsilon \ll 1$ to $\epsilon \gg 1$. A score of unity is awarded if the dominant terms dominate by two orders of magnitude. In this example there is only one sample, thus the average of a single data point is its original value and therefore $\mathcal{V}(\mathbf{E}, \mathbf{H}) = \mathcal{M}_1(\mathbf{e}_1, \mathbf{h}_1)$. This figure indicates that we have constructed the LMS to preferentially select sets of terms that dominate neglected terms by at least two orders of magnitude regardless of the signs of equation terms.

and therefore $\mathbf{E} = \mathbf{e}_1$ and $\mathbf{H} = \mathbf{h}_1$. If we chose a brute force search of all possible dominant balance hypotheses, then there are $2^D - D - 1 = 11$ choices because each hypothesis is a permutation of two types. A dominant balance of one term is not meaningful (see Appendix A). A dominant balance of all governing equation terms is trivial and conceptually equivalent a dominant balance of no governing equation terms. The possible dominant balance hypotheses are

$$\text{all possible hypotheses} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{20}$$

The score of all of these hypotheses are approximately zero for all magnitudes of ϵ except for $\mathbf{h}_1 = [1, 0, 1, 0]$ and $\mathbf{h}_1 = [0, 1, 0, 1]$, which represent the dominant terms when ϵ is relatively small and large, respectively, shown in Fig. 3. Note that in this example $\Omega_1 = 0$ and therefore $\mathcal{M}_1 = \Gamma_1$. The score rapidly converges to unity if the scale separation between dominant and neglected terms is larger than two orders of magnitude.

4. A verification criterion

We propose the weighted average of $\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)$, when averaged over N samples, as a tenable verification criterion in Eq. (1),

$$\mathcal{V}(\mathbf{E}, \mathbf{H}) = \frac{\sum_{n=1}^{N} w_n \cdot \mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)}{\sum_{n=1}^{N} w_n}, \tag{21}$$

where the array of weights $\mathbf{w} = [w_1, \ldots, w_N]$ are the discrete differentials of the observed domain, e.g. space and/or time differentials. If $N = 1$, then $\mathcal{V}(\mathbf{E}, \mathbf{H}) = \mathcal{M}_n(\mathbf{e}_1, \mathbf{h}_1)$. For example, if N observations of data set $\mathbf{E}$ are equation terms distributed across a one-dimensional space that evolve in time, then the verification criterion is the weighted average of all scores where each nth weight is product of the time step and grid spacing for the nth observation, e.g. $w_n = (\Delta t \cdot \Delta x)_n$. The score is designed such that the optimal is unity.

The verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ is one possible objective function that defines optimal dominant balances in Eq. (1). While our choice of verification criterion is ultimately subjective, we note that (a) our choice is consistent with domain knowledge as we state in the two conditions above and show in *Examples*, and (b) it permits objective comparison of dominant balances identified by different methods for labeling equation data with $\mathbf{H}_{\text{opt}}$. Other definitions of the verification criterion are possible and encouraged; the goal is to formally identify dominant balances by solving Eq. (1).

5. Unsupervised learning framework

We propose an unsupervised machine learning framework (Kohavi and John, 1997; Dy and Brodley, 2004) that automatically discovers dominant balances by using the verification criterion $\mathcal{V}(\mathbf{E}, \mathbf{H})$ (Eq. (21)) to solve the problem defined by Eq. (1). The framework is depicted in Fig. 4. The dominant balance identification problem is broken into partitioning, hypothesis selection, and hypothesis testing tasks. The left column outlines the conventional *ad hoc* method of dominant balance identification, and the right column depicts our framework. Our framework intentionally emulates the scientific method: the hypothesized dominant balances $\mathbf{H}$ are tested by evaluating their fit to the equation data $\mathbf{E}$ by using the verification criterion.

The first task shown in Fig. 4, row A, is to partition $\mathbf{E}$ into different dominant balance regions. 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. (2019) first suggested that the heuristic act of recognizing different dominant balance regions can be formulated as a partitioning problem that can be credibly solved using clustering algorithms. They are a class of unsupervised machine learning algorithms that yield a finite set of categories according to similarities or relationships among its objects (MacQueen et al., 1967; Hartigan, 1975). Clustering reveals underlying patterns of sparsity in the data. However, the resulting clusters are sensitive to the choice of algorithm parameters (Pedregosa et al., 2011). In addition, no definition of a cluster that is universal to all clustering algorithms exists (Estivill-Castro, 2002).

The second task, shown in the row B of Fig. 4, 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 dominant balance by which some terms are deemed negligible (Zohuri, 2017) for all samples within a region. Callaham et al. (2021) proposed sparse principal component analysis (Zou et al., 2006) (SPCA) for hypothesis selection. SPCA labels features with small variances as negligible. This is achieved through the application of least absolute shrinkage and selection operator (Tibshirani, 1996) (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 (Van Der Maaten et al., 2009), is geometrically and statistically consistent with expectation–maximization clustering algorithms (e.g. K–means, Gaussian Mixture Models). Both algorithms assume convex, uni-modal, zero-skew data.

We propose a simple hypothesis selection algorithm, that we will refer to as the combinatorial hypothesis selection (CHS) algorithm, for equations with less than eight terms $D \leq 8$, because of the computational complexity of brute force combinatorial guessing. The advantage of choosing CHS is that it contains no parameters that require tuning, such as the LASSO regression coefficient that must be chosen for SPCA. 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 (Eq. (12)) for all possible hypotheses and then selecting the hypothesis that is awarded the highest score.

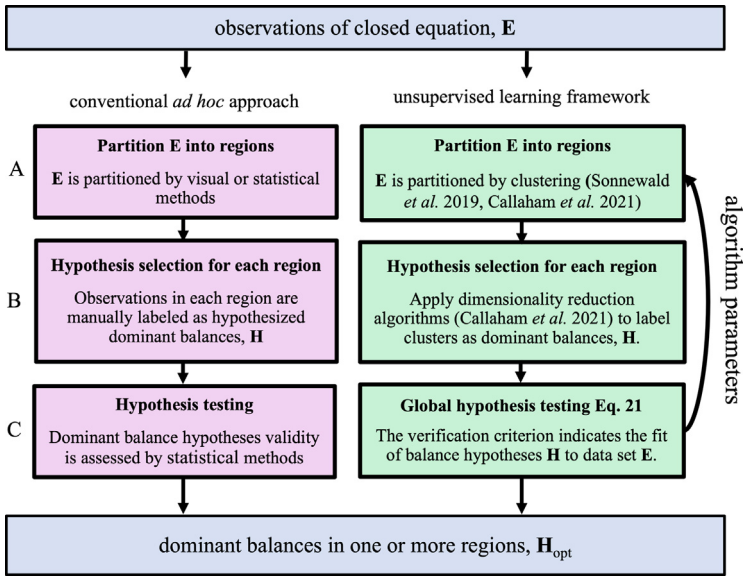


Fig. 4. The dominant balance identification problem. Partitioning and empirical scaling analysis performed by a human (left column), and algorithms capable of performing said tasks (right column). The loop over algorithm parameters illustrates the procedure for obtaining $\mathbf{H}_{\text{opt}}$ in Eq. (1). Sonnewald et al. (2019) utilize $K-$means clustering to divide oceanic vorticity into regions with distinct dominant balances. Callaham et al. (2021) demonstrated how clustering can be used to discover regions of dominant balances and sparse principal component analysis can then be used to label each region with a dominant balance.

Table 1
Synopsis of automated dominant balance identification by Sonnewald et al. (2019) and by Callaham et al. (2021).

Study	Dynamics	Clustering	Hypothesis selection	Verification
Sonnewald et al. (2019)	Global ocean vorticity	K-means	None	(1) Robustness of identified ocean regions (2) Information criteria convergence
Callaham et al. (2021)	(1) Turbulent boundary layer (2) Optical pulse propagation (3) Regional ocean vorticity (4) Bursting neurons (5) Rotating detonations	Gaussian Mixture Model	SPCA	None

Eq. (12) can be applied to a single data sample or to a weighted average of samples by using Eq. (21). The exponential time complexity limits the feasibility of computing CHS to equations with relatively few terms, as is shown in the *Synthetic data* example below.

The final task shown in Fig. 4, row C, is to measure the fit of hypotheses **H** to the data **E**. This task was conventionally performed indirectly through *post hoc* validation of models constructed using relevant identified dominant balances. Crucially, the framework applies to any choice of clustering and hypothesis selection algorithms. This allows for objective evaluation and comparisons of different algorithms. We have formalized direct verification of hypotheses by defining the dominant balance identification problem in Eq. (1) and proposing a verification criterion $\mathcal{V}(\mathbf{E},\mathbf{H})$. Table 1 shows the components of the dominant balance identification problem that were performed in previous studies.

The computational complexity of the framework depends on the algorithms chosen for clustering and on the hypothesis selection because the complexity of verification criterion is $\mathcal{O}(N)$. The computation time of a single pass through the framework scales polynomially with sample size N for all combinations of a non-parametric 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 dominant balance discovery.

5.1. Synthetic data example

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. Fig. 5a shows the synthetic data e_{ni} consisting of $D = 8$ equation terms, featuring two dominant balance regions in which dominant terms have magnitudes of $\mathcal{O}(10)$ and negligible terms have magnitudes of $\mathcal{O}(10^{-1})$. The dominant balance regions 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 dominant balance regions are prescribed by the Heaviside step function $\mathcal{H}$, such that:

$$\mathbf{e}_i(x,y) = (-1)^i \eta(y)(\lambda \mathcal{H}(\phi) + \beta), \tag{22}$$

$$\eta(y) = \eta_0 \sin(\omega y), \tag{23}$$

$$\phi = \begin{cases} y - 0.5 & \text{if} \quad 0 < i < D/2 \\ 0.5 - y & \text{if} \quad D/2 \leq i < \infty \end{cases}, \tag{24}$$

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 nth observation of the ith feature.

Figs. 5b,c,d show the results using $K-$means clustering and SPCA hypothesis selection. Fig. 5b 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

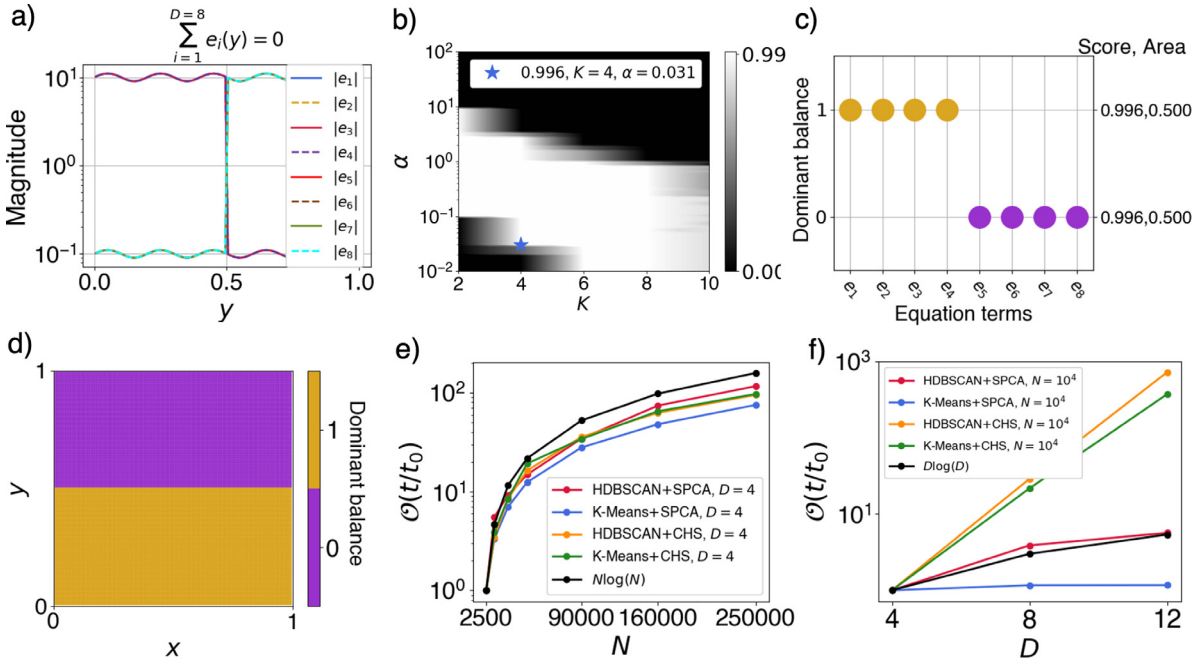


Fig. 5. Synthetic data example. a The synthetic data equation term magnitudes for all y at a fixed x. **b** The map of the verification criterion as the number of prescribed clusters for $K-$means and the LASSO regression coefficient α for hypothesis selection by SPCA are varied. The blue star corresponds to the optimal verification criteria. **c** The optimal dominant balances. **d** The spatial distribution of the optimal dominant balances. **e** The variation in wall time as a function of sample size for different algorithms. **f** The variation in wall time as a function of the number of equation terms for different algorithms.

optimal is marked with the blue star, $\mathcal{V}(\mathbf{E},\mathbf{H}) = 0.996$, though much of the white band in Fig. 5b corresponds to equivalently optimal results. Fig. 5c shows the optimal dominant balances, and Fig. 5d shows the spatial distribution of the optimal dominant balances. The algorithm parameter ranges were specified as follows: for $K-$means clustering the number of prescribed clusters K was specified as $K = \{2,\ldots,10\}$. The other hyperparameters were the default choices as provided by SciKit Learn (Pedregosa et al., 2011). 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}. Identical optimal balances 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 balances are robust because the magnitude separation between dominant and negligible terms is at least two orders of magnitude and the spatial boundary between the dominant balance regions is 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 (Mahajan et al., 2012), $D = 2$. 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 (the continuous LASSO regression coefficient, α); therefore, we recommend CHS for equations with fewer terms than 8. For example, if the number of equation terms is less than 8 and $K-$means is the chosen clustering algorithm, then the user need only optimize the verification criterion over the number of clusters K instead of performing a multi-variate optimization over both K and the LASSO regression coefficient α for SPCA, as shown in Fig. 5b. Since α is a continuous variable, there are an infinite number of discrete choices of α within any given range. Therefore, one can minimize the *total* wall time elapsed (e.g., the total time elapsed for all α and K in Fig. 5b) by using CHS instead and therefore only performing the optimization over K. However, since CHS is combinatorially complex in D we recommend SPCA for governing equations with many terms (say, $D > 8$) and CHS for governing equations with fewer terms (say, $D \leq 8$) because it is parameter free.

The average wall times elapsed for the framework computations are shown as a function of the number of samples, N, in Fig. 5e and as a function of the number of equation terms, D, in Fig. 5f. The wall times in Fig. 5e are normalized by the wall time to compute the $N = 2500$ case for each algorithm and the wall times in Fig. 5f are similarly normalized by the wall time to compute the $D = 4$ cases. Each result (each data point in Fig. 5e and 5f) was computed on a single 2.6 GHz Intel Xeon E5-2660 v3 processor. Each point represents the average wall time for one pass through the framework (Fig. 4). Fig. 5e shows that the computation time scales polynomially with sample size N for all algorithm choices. Fig. 5f shows that CHS becomes prohibitively complex with increasing number of equation terms because its complexity scales with $\mathcal{O}(2^D)$. The time complexity behavior shown in Figs. 5e and 5f informs our is less than 8.

5.2. Global ocean barotropic vorticity example

Sonnewald et al. (2019) used $K-$means clustering, manual hypothesis selection, and algorithm- and problem-specific verification criteria to discover new and canonical oceanic dominant balances. We use the vorticity data of Sonnewald et al. (2019), who computed a 20-year mean of the Estimating the Circulation and Climate of the Ocean (ECCO) ocean state estimate (Forget et al., 2015; Wunsch and Heimbach, 2013; Anon, 2017a,b), version 4 release 2, at 1° resolution to calculate terms of the vertically integrated barotropic vorticity equation.

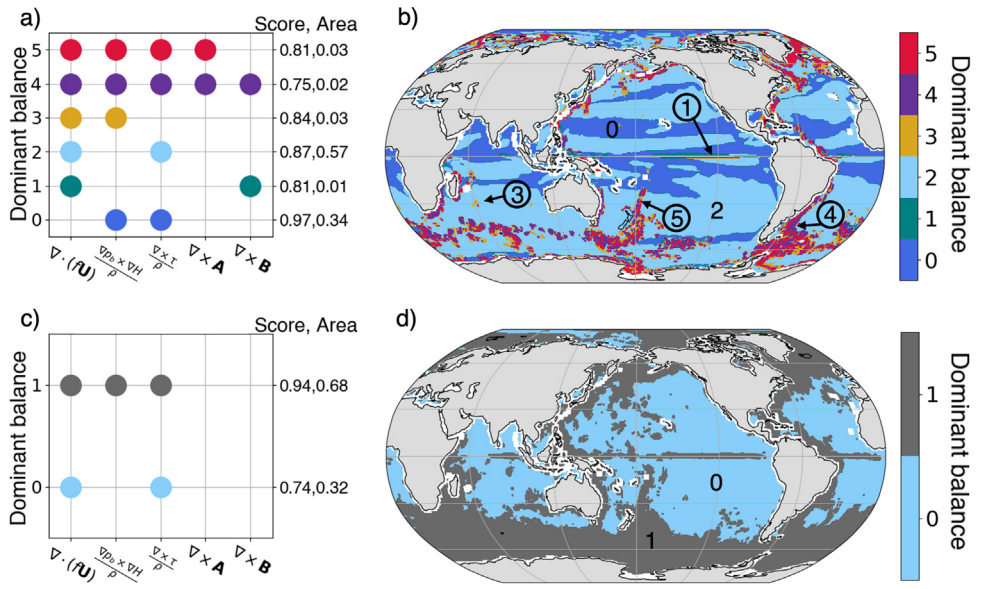


Fig. 6. Oceanic barotropic vorticity dominant balance examples. a The optimal dominant balances found by $K-$means and CHS. **b** The spatial distributions of the optimal dominant balances found by $K-$means and CHS. **c** The optimal dominant balances found by HDBSCAN and CHS. **d** The spatial distributions of the optimal dominant balances found by HDBSCAN and CHS. The optimal verification criterion for the $K-$means clustering approach is $\mathcal{V}(\mathbf{E},\mathbf{H}) = 0.90$ while the optimal verification criterion for the HDBSCAN clustering approach is $\mathcal{V}(\mathbf{V},\mathbf{H}) = 0.87$. Comparison of the optimal verification criteria indicates that the $K-$means clustering algorithm fits the data better than the HDBSCAN clustering algorithm for the oceanic barotropic vorticity data set.

The time-mean vertically-integrated barotropic vorticity equation,

$$\overbrace{\nabla\cdot(f\mathbf{U})}^{\text{advection of planetary vorticity}} = \overbrace{\frac{\nabla p_b\times\nabla H}{\rho}}^{\text{bottom pressure torque}} + \overbrace{\frac{\nabla\times\boldsymbol{\tau}}{\rho}}^{\text{wind \& bottom stress curl}} + \overbrace{\nabla\times\mathbf{A}}^{\text{nonlinear torque}} + \overbrace{\nabla\times\mathbf{B}}^{\text{diffusive torque}}, \tag{25}$$

describes the balance of processes that control the rate of solid body rotation of a column of seawater.

Figs. 6a and 6b show the optimal dominant balances 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. (2019). The differences can be attributed to the selection of the optimal number of clusters as $K = 49$ as opposed to $K = 50$, algorithm stochasticity, and different standardization methods (see Appendix C). 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. (2019), using information theoretic and a custom geographic convergence verification criteria. Figs. 6c,d 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 $K-$means and 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.

5.3. Nonlinear diffusion in tumor-induced angiogenesis example

Anderson and Chaplain (1998), Anderson et al. (2000) 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 and Chaplain (1998) is composed of conservation laws of three continuous variables, where the endothelial-cell density per unit area (cells that rearrange and migrate from preexisting vasculature to form new capillaries), n, is governed by

$$\frac{\partial n}{\partial t} = \overbrace{d\nabla^2 n}^{\text{random motility}} - \overbrace{\nabla\cdot(\chi n\nabla c)}^{\text{chemotaxis}} - \overbrace{\nabla\cdot(\rho n\nabla f)}^{\text{haptotaxis}}, \tag{26}$$

$$= d\nabla^2 n - \chi n\nabla^2 c - \chi\nabla n\cdot\nabla c - n\nabla\chi\cdot\nabla c - \rho n\nabla^2 f - \rho\nabla n\cdot\nabla f, \tag{27}$$

where $\chi(c) = \chi_0/(1+\alpha_0 c)$. Fig. 7a shows the absolute time rate of change of cells as endothelial cell growth propagates towards the tumor. Figs. 7b,c show the optimal dominant balances and their spatial distributions 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 only 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.

5.4. Spatially-developing turbulent boundary layer example

Canonical turbulent boundary layer dominant balances (Tennekes and Lumley, 1972), previously identified by Callaham et al. (2021) using Gaussian Mixture Model (GMM) clustering with SPCA hypothesis selection and no quantitative verification criteria, can be identified automatically using the present method. 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

$$\overbrace{\overline{u}\frac{\partial\overline{u}}{\partial x}+\overline{v}\frac{\partial\overline{u}}{\partial y}}^{\substack{\text{mean}\\\text{momentum flux divergence}}} = \overbrace{-\frac{1}{\rho}\frac{\partial\overline{p}}{\partial x}}^{\substack{\text{mean}\\\text{pressure gradient}}} + \overbrace{\nu\nabla^2\overline{u}}^{\substack{\text{mean}\\\text{momentum diffusion}}} - \overbrace{\frac{\partial\overline{u'v'}}{\partial y}-\frac{\partial\overline{u'^2}}{\partial x}}^{\substack{\text{turbulent}\\\text{momentum flux divergence}}}, \tag{28}$$

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.

Fig. 8a 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 dominant balances are shown in Fig. 8b,c. The dominant balances are consistent with the results of Callaham et al. (2021) and with domain knowledge (Schetz and Bowersox, 2011), but notably our framework required no fluid dynamical knowledge.

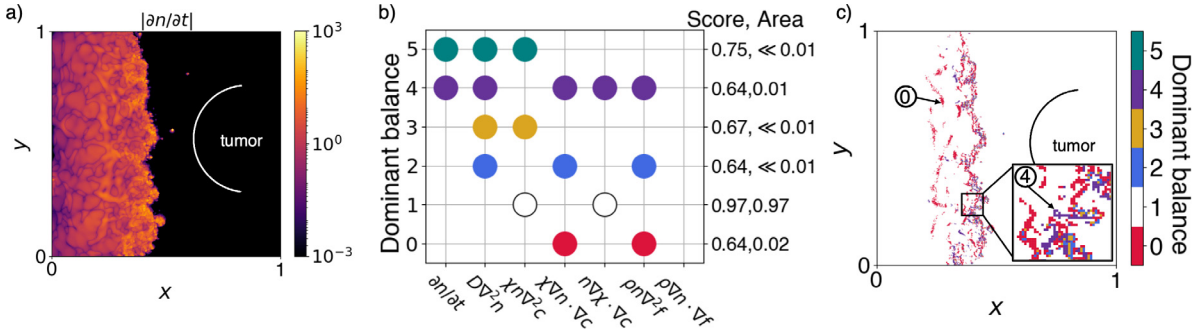


Fig. 7. Tumor-induced angiogenesis dominant balance example. a Endothelial cell growth rates. **b** Optimal dominant balances. **c** The spatial distributions of the optimal dominant balances. Balances are identified by $K-$means clustering with CHS. The results suggest that the balances 0 and 4 dominate the propagation of endothelial cell growth from left to right, and that the terms $\chi\nabla^2 c$ and $\rho\nabla n \cdot \nabla f$ are possibly negligible. The RHS y axis of **b** quantifies the areal percentages of balances 2,3, and 5 as less than 1%.

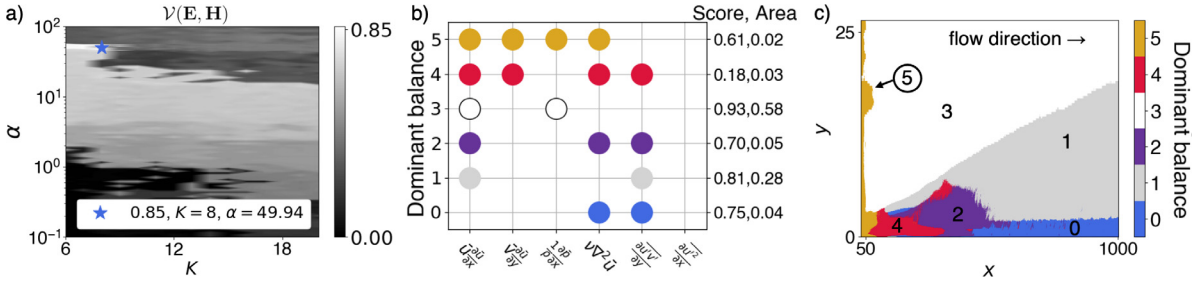


Fig. 8. Turbulent boundary layer dominant balance example. a The map of the verification criterion as a function of algorithm parameters K (number of clusters prescribed for GMM clustering) and α (LASSO regression coefficient for hypothesis selection by SPCA). The blue star indicates the optimal verification criterion. **b** The optimal dominant balances. **c** The spatial distributions of the optimal dominant balances. The optimal balances and their spatial distributions agree with those identified by Callaham et al. (2021) as well as domain knowledge, but here the algorithm parameters K and α are chosen by automatically selecting the parameters corresponding to the optimal verification criterion. .

6. Conclusions

Our proposed formal definition of the dominant balance identification problem is defined by the global maximization of a verification criterion over all equation term data (Eqs. (1) through (21)). Our formalism is independent of the method by which the optimization problem is solved, thus permitting objective comparison of different methods of balance identification and transforming previously *ad hoc* and piecemeal analyses into an objective framework for dominant balance identification.

We show that our framework yields results consistent with domain knowledge and previous studies (Callaham et al., 2021; Sonnewald et al., 2019). We emphasize that the framework is broadly relevant to analyses of chaotic systems. We note that the verification criterion (Eq. (21)) could be used as a loss function for a supervised learning approach to the dominant balance identification problem as defined by Eq. (1), where a neural network could simultaneously perform tasks A, B, and C in Fig. 4. We also note that two dimensional example data sets were chosen here for illustrative purposes only. The presented framework readily applies to arbitrary dimensional data and/or the temporal dimension. We anticipate that this work could dramatically expedite the discovery of unknown dominant balances in new data and accelerate efforts in data-driven dynamical process modeling (Rudy et al., 2017; Raissi, 2018; Rackauckas et al., 2020; Reichstein et al., 2019; Schneider et al., 2017).

Code availability

Code for dominant balance identification and the tumor-induced angiogenesis solver are available at https://github.com/bekaiser.

CRediT authorship contribution statement

Bryan E. Kaiser: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Visualization. **Juan A. Saenz:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing, Supervision. **Maike Sonnewald:** Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. **Daniel Livescu:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing, Supervision, Project administration, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was performed under the auspices of DOE. Financial support comes partly from Los Alamos National Laboratory (LANL), USA, Laboratory Directed Research and Development (LDRD), USA project "Machine Learning for Turbulence", 20180059DR. This work is approved for unlimited release, LA-UR-22-20935. 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. M.S. acknowledges funding from Cooperative Institute for Modeling the Earth System, Princeton University, USA, 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.

Appendix A. The ambiguity of a single-term dominant balance

Consider the equation

$$a - \sum_{i=1}^{N_i} b_i = 0, \tag{A.1}$$

where $a \gg 1$, $b_i \ll 1$, and N_i is a sufficiently large number such that the equation is satisfied. If we seek to identify the dominant balances of this equation, there is just a single term, thus the dominant balance is

$$a \approx 0, \tag{A.2}$$

because term a is not balanced by any other term but rather the summation of a large set of small terms. Therefore, a single term dominant balance is not useful in dominant balance identification. For this reason, we maintain that length($\mathbf{s}_n$)≥ 2. If this condition is imposed the resulting LMS score forEq. (A.1) will be low, $\mathcal{M} \approx 0$.

Appendix B. Score relationship to Buckingham Π theorem

The LMS is (a) invariant to the magnitude of the equation vector $\mathbf{e}_n$ and (b) invariant to the sign of the elements of the equation vector, thus

$$\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n) = \mathcal{M}_n(\pm c\mathbf{e}_n, \mathbf{h}_n), \tag{B.1}$$

where c is a positive scalar constant and the subscript denotes the nth example. 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 select dominant Π groups. Buckingham Π theorem is a formal method for the identification of the minimum number of non-dimensional parameters that describe a dynamical system. If the governing equation(s) for the system is(are) known, then it can be shown that the Π groups are consistent with the non-dimensional equation coefficients (Zohuri, 2017).

Appendix C. Standardization for clustering

All equation data were standardized prior to clustering using RobustScaler.fit() from Scikit Learn library (Pedregosa et al., 2011).

Appendix D. Global ocean barotropic vorticity

In Eq. (25) 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.

Appendix E. Tumor-induced angiogenesis simulation

We use the non-dimensional, tumor-induced angiogenesis governing equations of Anderson and Chaplain (1998). 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

$$\frac{\partial f}{\partial t} = \beta n - \gamma n f, \tag{E.1}$$

$$\frac{\partial c}{\partial t} = -\eta c n, \tag{E.2}$$

Endothelial cell migration up the fibronectin concentration gradient is termed haptotaxis (Carter, 1965, 1967), while endothelial cell migration up the gradient of tumor angiogenic factor concentration is termed chemotaxis (Sholley et al., 1984).

In Fig. 7a, 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. Fig. 7b and Fig. 7c show the dominant balances and their spatial distributions, respectively, for the optimal results for $K-$means clustering and CHS hypothesis selection.

We numerically solve the same problem as Anderson and Chaplain (1998), 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 Eq. (27). We employ the same boundary conditions, initial conditions, and constant coefficients (d, α_0, χ_0, ρ, β, γ, and η) as Anderson and Chaplain (1998) 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:

$$\mathbf{n} \cdot (d\nabla n - \chi(c) n \nabla c - \rho n \nabla f) = 0, \tag{E.3}$$

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:

$$c(x, y, 0) = \begin{cases} 1, & 0 \leq r \leq 0.1 \\ \frac{(\nu - r)^2}{\nu - r_0}, & 0.1 < r \leq 1, \end{cases} \tag{E.4}$$

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

$$f(x, y, 0) = k \mathrm{e}^{-\frac{x^2}{\epsilon_1}}, \tag{E.5}$$

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

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 $\eta = 0.1$.

Appendix F. Spatially developing turbulent boundary layer

We use the same data set as Callaham et al. (2021), namely the turbulent boundary layer direct numerical simulation data available in the Johns Hopkins University turbulence database (Zaki, 2013). 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 $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$.

References

Anderson, A.R., Chaplain, M.A.J., 1998. Continuous and discrete mathematical models of tumor-induced angiogenesis. Bull. Math. Biol. 60 (5), 857–899.

Anderson, A.R., Chaplain, M.A., Newman, E.L., Steele, R.J., Thompson, A.M., 2000. Mathematical modelling of tumour invasion and metastasis. Comput. Math. Methods Med. 2 (2), 129–154.

Anon, ECCO Consortium, 2017a. A Twenty-Year Dynamical Oceanic Climatology: 1994–2013. Part 1: Active Scalar Fields: Temperature, Salinity, Dynamic Topography, Mixed-Layer Depth, Bottom Pressure.

Anon, ECCO Consortium, 2017b. A Twenty-Year Dynamical Oceanic Climatology: 1994–2013. Part 2: Velocities, Property Transports, Meteorological Variables, Mixing Coefficients.

Barenblatt, G.I., 1996. Scaling, Self-Similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics. (14).

Bishop, C.M., 2006. Pattern Recognition and Machine Learning.

Blow, K.J., Wood, D., 1989. Theoretical description of transient stimulated Raman scattering in optical fibers. IEEE J. Quantum Electron. 25 (12), 2665–2673.

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

Carter, S.B., 1965. Principles of cell motility: the direction of cell movement and cancer invasion. Nature 208 (5016), 1183–1187.

Carter, S.B., 1967. Haptotaxis and the mechanism of cell motility. Nature 213 (5073), 256–260.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms.

d'Alembert, J.L.R., 1752. Essai d'une Nouvelle Théorie de la Rrésistance des Fluides.

Dy, J.G., Brodley, C.E., 2004. Feature selection for unsupervised learning. J. Mach. Learn. Res. 5 (Aug), 845–889.

Estivill-Castro, V., 2002. Why so many clustering algorithms: a position paper. ACM SIGKDD Explor. Newsl. 4 (1), 65–75.

Forget, G., Campin, J.-M., Heimbach, P., Hill, C.N., Ponte, R.M., Wunsch, C., 2015. ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. Copernicus GmbH.

Hartigan, J.A., 1975. Clustering Algorithms. John Wiley & Sons, Inc..

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97 (1–2), 273–324.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1, (14), Oakland, CA, USA, pp. 281–297.

Mahajan, M., Nimbhorkar, P., Varadarajan, K., 2012. The planar k-means problem is NP-hard. Theoret. Comput. Sci. 442, 13–21.

Munkres, J.R., 2000. Topology.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Peixoto, J.P., Oort, A.H., 1992. Physics of climate. American Institute of Physics, New York, NY (United States).

Prandtl, L., 1904. Über Flussigkeitsbewegung bei sehr kleiner Reibung. In: Verhandl. III, Internat. Math.-Kong., Heidelberg, Teubner, Leipzig, 1904. pp. 484–491.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A., 2020. Universal differential equations for scientific machine learning. arXiv preprint arXiv:2001.04385.

Raissi, M., 2018. Deep hidden physics models: Deep learning of nonlinear partial differential equations. J. Mach. Learn. Res. 19 (1), 932–955.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204.

Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N., 2017. Data-driven discovery of partial differential equations. Sci. Adv. 3 (4), e1602614.

Schetz, J.A., Bowersox, R.D., 2011. Boundary Layer Analysis.

Schneider, T., Lan, S., Stuart, A., Teixeira, J., 2017. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. Geophys. Res. Lett. 44 (24), 12–396.

Seminara, A., Angelini, T.E., Wilking, J.N., Vlamakis, H., Ebrahim, S., Kolter, R., Weitz, D.A., Brenner, M.P., 2012. Osmotic spreading of Bacillus subtilis biofilms driven by an extracellular matrix. Proc. Natl. Acad. Sci. 109 (4), 1116–1121.

Sholley, M., Ferguson, G., Seibel, H., Montour, J., Wilson, J., 1984. Mechanisms of neovascularization. Vascular sprouting can occur without proliferation of endothelial cells. Lab. Invest. 51 (6), 624–634.

Sonnewald, M., Wunsch, C., Heimbach, P., 2019. Unsupervised learning reveals geography of global ocean dynamical regions. Earth Space Sci. 6 (5), 784–794.

Strogatz, S.H., 1994. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.

Tennekes, H., Lumley, J., 1972. A First Course in Turbulence.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Vallis, G.K., 2017. Atmospheric and Oceanic Fluid Dynamics.

Van Der Maaten, L., Postma, E., Van den Herik, J., 2009. Dimensionality reduction: a comparative. J. Mach. Learn. Res. 10 (66–71), 13.

Wunsch, C., Heimbach, P., 2013. Dynamically and kinematically consistent global ocean circulation and ice state estimates. In: International Geophysics. vol. 103, Elsevier, pp. 553–579.

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

Zohuri, B., 2017. Dimensional Analysis beyond the Pi Theorem.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Statist. 15 (2), 265–286.