Inferring Kernel ϵ-Machines:
Discovering Structure in Complex Systems

Alexandra M. Jurgens[1, *] and Nicolas Brodu[1, †]

[1] *INRIA Bordeaux Sud Ouest, 33405 Talence Cedex, France*

(Dated: October 3, 2024)

Previously, we showed that computational mechanic's causal states—predictively-equivalent trajectory classes for a stochastic dynamical system—can be cast into a reproducing kernel Hilbert space. The result is a widely-applicable method that infers causal structure directly from very different kinds of observations and systems. Here, we expand this method to explicitly introduce the causal diffusion components it produces. These encode the kernel causal-state estimates as a set of coordinates in a reduced dimension space. We show how each component extracts predictive features from data and demonstrate their application on four examples: first, a simple pendulum—an exactly solvable system; second, a molecular-dynamic trajectory of n-butane—a high-dimensional system with a well-studied energy landscape; third, the monthly sunspot sequence—the longest-running available time series of direct observations; and fourth, multi-year observations of an active crop field—a set of heterogeneous observations of the same ecosystem taken for over a decade. In this way, we demonstrate that the empirical kernel causal-states algorithm robustly discovers predictive structures for systems with widely varying dimensionality and stochasticity.

CONTENTS

I. INTRODUCTION

In a broad sense the project of the natural sciences turns on the desire to uncover hidden structure in physical systems. But what, precisely, is meant by "hidden structure"? To be sure, we do not mean the most obvious form of structural analysis as performed in the engineering sciences. Rather, we wish to access something more abstract—a representation of our system that makes explicit the global patterns of behavior, and/or the patterns of behavior of distinguishable system components and the nature of how those components are related and organized.

Early work in nonlinear dynamical systems developed the field of attractor reconstruction from data using techniques like delay embeddings [1, 2]. In this kind of structural analysis we analyze the topology and geometry of the reconstructed attractor to better understand the dynamics of the system under study. For example, a system may have a set of behaviors (or patterns of behavior)

---

* alexandra.jurgens@inria.fr
† nicolas.brodu@inria.fr; corresponding author

it tends towards over time, visualized as trajectories approaching attracting points or limit cycles in phase space. Modern dynamical system reconstruction techniques are descendants of this kind of structural analysis [3, 4].

If we suppose the observed data are the composition of an observation function, applied to some underlying dynamical system at each instant of time, then we can also study how the observations change as the action of the underlying dynamics. This is the definition of the Koopman operator, acting on the observation function. Interestingly, this operator is linear, even though the original dynamical system is not. Given the advantages of linear algebra, a consequent body of research [5, 6] focuses on how to best express these operators. Algorithms exist to estimate them, provably consistently so in the limit of infinite data [7, 8]. This is generally accomplished by finding invariant subspaces using spectral decomposition approximations [9], effectively turning complicated dynamics into a simpler linear operations. However, the price to pay lies in the spectral basis functions, which become complex objects. These techniques can be extraordinarily powerful and flexible but struggle when it comes to the interpretability of these discovered latent spaces, which may be unintuitive. Alternatively, methods exist to model the non-linear evolution of the dynamical system. Recovering equations of motion from data is an ancient topic [10], but which is still active [4, 11–13]. The methods more closely related to our approach need to account for a stochastic component [14, 15].

A complementary perspective is offered by information-theoretic definitions of structure, which strive to describe systems by formalizing concepts like *irreducible information* and *complexity*. One such framework is offered by computational mechanics [16] which casts a system as a communication channel from its past behavior to its future behavior. Computational mechanics introduces the ϵ-*machine*—a model of a system built from the *causal states*, the set of predictively-equivalent distributions over the future given the past. In this setting, structural analysis of the system corresponds to describing the nature of its causal structure—plainly, what is the stochastic mapping between the system's past and future? We can easily construct a system where the answer is none at all—a coin flip for example. We can also imagine a system where every possible past uniquely determines the future. In general, we assume real systems lie somewhere in between these extremes.

Recently, computational mechanics was extended to marry these two styles of structural analysis in a novel way: a system's causal states may be consistently and uniquely embedded as points in a Hilbert space [17]. This allows us to construct the *kernel ϵ-machine*, a new geometric representation of the ϵ-machine wherein the causal states are points in a Hilbert space and the dynamics of the ϵ-machine may be represented by a stochastic model of their evolution. The previous contribution in this series introduced not only the kernel ϵ-machine but also a tractable algorithm for estimating empirical kernel causal

states from nearly arbitrary data [18]. The net result is that computational mechanics was extended to nearly arbitrary data types, including real data and with minimal constraints. The second major result was that we now have a presentation of the causal states imbued with not only information-theoretic notions of structure, but also geometric ones. This facilitates applying new, intuitive structural analysis of the causal states, including manifold learning techniques. This paper applies these techniques and the empirical kernel causal state algorithm to real-valued data for the first time.

We note that the principles we outline here, of describing and predicting the state-space structure of a complex system, is a well-developed tradition in time-series forecasting methods and the study of chaos in dynamical systems [19] and which is closely related to the subfield of pattern formation [20–22]. This said, the prediction of trajectories in state space is a separate topic. Compared to the methods mentioned above for recovering dynamical systems equations, in our case the consistency of the evolution in both data and causal states spaces also needs to be ensured. We address these issues in a sequel.

Section II reviews the necessary aspects of computational mechanics theory. Section III then reviews the theoretical construction of kernel causal states, as well as the empirical kernel causal state algorithm as introduced previously and the use of a diffusion mapping technique for nonlinear dimension reduction. Finally, we present the results of the empirical kernel causal state algorithm applied to two simulated examples in Section IV and two real data examples in Section V.

II. COMPUTATIONAL MECHANICS

Traditionally, the basic objects under study in computational mechanics are conditional distributions over realizations of stochastic processes. In particular, we generally are interested in building the ϵ-*machine*: a stochastic process's minimal optimally predictive model [23]. The states of the ϵ-machine are called the *causal states*, which are sets of past observations of a process that induce the same distribution over futures.

We take a *stochastic process* $\mathcal{P}$ to consist of a τ-indexed random variable Z defined on the measurable space $(\mathcal{Z}^\tau, \Sigma, \mu)$. Specific realizations of random variables are denoted by use of the lower case and time indexing is done by the use of subscripts. For example, we write $Z_t = z_t$ to say that $z \in \mathcal{Z}$ is the specific value of Z at time t. The dynamic of the stochastic process is given by the shift operator, also called the *translation operator*, which is an operator $\tau : \mathcal{Z}^\tau \to \mathcal{Z}^\tau$ that maps t to $t+1$: $\tau z_t = z_{t+1}$. It also acts on the measure: $(\tau\mu)(E) = \mu(\tau^{-1}E)$ for $E \in \Sigma$. Temporal blocks of the process are given by $\{z_{a < t \leq b} : a < b; a, b \in \tau\}$.

Computational mechanics is primarily concerned with the prediction of the future of a system given observations

of its past. To be technical about these terms, we refer to the observation at $t = 0$ as the *present*. We call $Z_{t \leq 0}$ the *past X* and $Z_{t > 0}$ the *future Y*. Using distinct symbols for the past and future is both for readability and to match notation from our previous work [18]. We assume that the past and future are defined on their own measurable spaces: $\left(\mathcal{Z}^{\tau \leq 0}, \Sigma_X, \mu_X \right)$ and $\left(\mathcal{Z}^{\tau > 0}, \Sigma_Y, \mu_Y \right)$, respectively. A time subscript on a past or future represents the time at which that past ends or that future begins, as appropriate. So, a specific realization of the past at time τ is written $x_\tau = z_{t \leq \tau}$. A corresponding specific realization of the future is $y_\tau = z_{t > \tau}$. Thus, at each time step an instance of the process z splits into paired instances of a past and a future: $z = (x_\tau, y_\tau) \ \forall \tau \in \tau$.

For the theoretical development of the causal states we make two assumptions about the stochastic process. We assume the process $\mathcal{P}$ is conditionally stationary, which is to say that the conditional distribution over futures given a specific past $\Pr(Y \mid X = x)$ is the same for all $t \in \mathbb{Z}$ up to a null-measure set. For this reason we may drop the time index t when it is unnecessary. Conditional stationarity is a weaker condition than requiring $\mathcal{P}$ be stationary, which would require $\tau \mu = \mu$. We also typically assume that the measure is ergodic, which is to say that all shift-invariant sets $I \in \mathcal{Z}^{\mathbb{Z}}$ are either $\mu(I) = 1$ or $\mu(I) = 0$. Practically speaking, the empirical kernel causal state algorithm as described in Section III B and Section III C may be applied to data even when these conditions are not met (or unable to be tested for), but the strict guarantees of convergence of the causal states may not hold. Since real-world applications are already working with finite data and other practical limitations, this is not typically of great concern.

We will, however, enforce two very minor constraints. First, the temporal indices τ are taken to be the integers $\mathbb{Z}$, which is to say that the process is *discrete-time*. Second, $\mathcal{Z}$ is taken to be compact—typically, either a discrete finite set or a closed interval in $\mathbb{R}$ (or a Cartesian product of a finite number of these, which is sufficient to ensure that $\mathcal{Z}$ is compact by Tychonoff's theorem). Note that we do not constrain $\mathcal{Z}$ to be real-valued—some or all data may be symbolic. Indeed, in the classic setting for computational mechanics [23] $\mathcal{Z}$ is a discrete and finite alphabet, typically using only two symbols $\{0, 1\}$, pasts and futures are semi-infinite strings and the temporal blocks become words. Loosening of these constraints is possible even in the theoretical realm [17], although this will not be discussed here.

The workhorse of computational mechanics is the predictive equivalence relation. By predictive equivalence we mean that two specific pasts x, x' correspond to the same distribution over futures Y. This induces an equivalence relation $\epsilon[\cdot]$ that maps pasts to sets of pasts via:

$$\epsilon[x] = \left\{ x' \in \mathcal{Z}^{\mathbb{N}_0} : \Pr(Y \mid X = x') = \Pr(Y \mid X = x) \right\} . \tag{1}$$

If x and x' are predictively equivalent, they belong to the set of pasts induced by their equivalence class. These sets of pasts and their associated distributions over futures are what we want to find: the elements of the range of the equivalence mapping are exactly the causal states of the process. The causal state random variable is denoted $\mathcal{S}$ and its realization is written σ. The set of causal states of a process is denoted $\boldsymbol{\mathcal{S}}$. (To be more precise, $\boldsymbol{\mathcal{S}}$ is the closure of the set of causal states, where the closure is taken under convergence in distribution [17].)

The cardinality of $\boldsymbol{\mathcal{S}}$ depends on the process under study. On the one hand—the extreme of utter unpredictability—a sequence of independent identically-distributed variables trivially is comprised of a single causal state, because every past x will induce an identical distribution over futures. At the opposite extreme of exact predictability, on the other hand, a discrete period-k process (e.g. *ababab...* has $k = 2$) has exactly k causal states. In general, the structure of $\boldsymbol{\mathcal{S}}$ is highly variable: the set may be finite or infinite, have countable or uncountable cardinality, may be well-ordered, or fractal.

The ϵ-*machine* of a process is the causal state set $\boldsymbol{\mathcal{S}}$ together with the transition dynamic over the causal states induced by the shift operator of the process. When both $\boldsymbol{\mathcal{S}}$ and $\mathcal{Z}$ are discrete and finite, the dynamic over the causal states may be simply expressed with a set of symbol-labeled transition matrices $\left\{ T_{\sigma, \sigma'}^{(z)} = \Pr(z, \sigma' \mid \sigma) : \sigma, \sigma' \in \boldsymbol{\mathcal{S}}, z \in \mathcal{Z} \right\}$. When $\boldsymbol{\mathcal{S}}$ is finite, then, the ϵ-machine takes on the familiar form of an *edge-emitting finite-state hidden Markov model*.[1] When $\boldsymbol{\mathcal{S}}$ is infinite but $\mathcal{Z}$ is discrete and finite, the transition dynamic may described with an *iterated function system* [24, 25]. When $\mathcal{Z}$ becomes continuous, the process is *hidden semi-Markov* [26] and the transition dynamic over causal states becomes a general inhomogeneous stochastic difference equation: the probability of state-to-state transitions depends both on the state before the transition and the value $z \in \mathcal{Z}$ inducing the transition.

By construction, the ϵ-machine is the minimal optimally predictive model of a process. Additionally, ϵ-machines are used to directly calculate in closed form a wide variety of relevant information theoretic properties, such as the Shannon entropy rate and the statistical complexity, a measure of the process's complexity [16]. In addition, being built constructively from conditional distributions over futures means that ϵ-machines are invariant under bijective transformations—even nonlinear ones—of the observation space. That is to say, a physical process measured under a change of reference frame will still induce the same equivalence classes, making the ϵ-machine an intrinsic property of $\mathcal{P}$ robust to changes in measurement. These advantages make the ϵ-machine theoretically

---

[1] With the additional constraint that the symbol-labeled transitions are *unifilar*, which is to say that the current state of the process σ_t and the next observed symbol z_{t+1} uniquely determines the next state σ'_{t+1}.
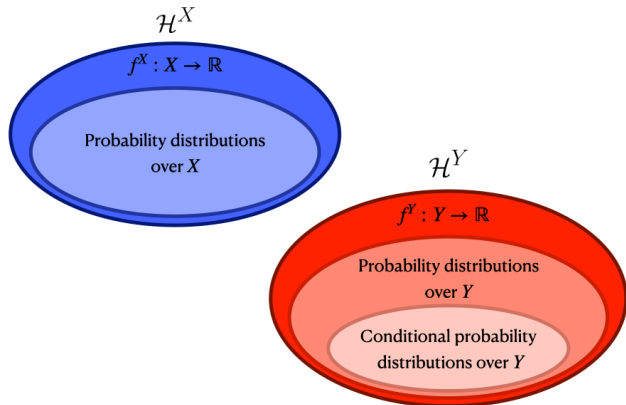
FIG. 1. Probability distributions over pasts X and conditional distributions over futures Y can be represented as points in the reproducing kernel Hilbert spaces $\mathcal{H}^X$ and $\mathcal{H}^Y$, respectively.

compelling, but the exactness of the theoretical construction have made ϵ-machines difficult to work with from an inference perspective.

The recently developed kernel causal state method [18] is, to our knowledge, the first causal state inference algorithm that is practicable for nearly arbitrary data. It is also well-grounded: proof that causal states are always well defined and empirical observation-based approximations will converge has been recently been extended to continuous-valued processes under the mild conditions listed above [17]. This vastly expands the practicality of computational mechanics to new classes of processes.

III. KERNEL HILBERT ϵ-MACHINES

Recent work on the topology and geometry within computational mechanics showed that the causal states can be naturally embedded into a Hilbert space using an appropriate kernel [17]. This connects computational mechanics to the well-studied arena of reproducing kernel Hilbert spaces (RKHS), setting the stage for the *kernel ϵ-machine* (KϵM)—an ϵ-machine representation that allows for tractable analysis of more complex processes than previously possible [18].

This section first reviews how causal states naturally form a Hilbert space. It then recounts the algorithm for empirical kernel causal state construction from data. Finally, it concludes with a walk through the diffusion mapping algorithm, which finds a low-dimensional embedding of the empirically estimated kernel causal state set.

A. Causal states form a Hilbert space

Section II defined $\mathcal{S}$ as the (closure of the) set of the causal states of a process. Now, let $\mathbb{S}$ be the vector space

of signed measures generated by the closed span of $\mathcal{S}$. This is the smallest vector space that contains all the causal states. The dimensionality of $\mathbb{S}$ may be much less than the cardinality of the causal state set $\mathcal{S}$. In general, for any process generated by a hidden Markov model, $\dim \mathbb{S}$ will be finite but $\mathcal{S}$ may be uncountably infinite [17, 25].

The space $\mathbb{S}$ is a subspace of the space of probability measures over half-infinite sequences $\mathbb{P}\left(\mathcal{Z}^{\mathbb{N}}\right)$. By the Riesz representation theorem, given a symmetric positive-definite kernel $k : \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \to \mathbb{R}$, we may embed any measure $\mu \in \mathbb{P}\left(\mathcal{Z}^{\mathbb{N}}\right)$ into a Hilbert space by writing down the function:

$$f_\mu = \int_{\mathcal{Z}^{\mathbb{N}}} k\left(y, \cdot\right) d\mu\left(y\right) \ , \tag{2}$$

which belongs to the reproducing kernel Hilbert space $\mathcal{H}_k$. (The dot notation indicates a free argument so that $k\left(y, \cdot\right) \in \mathcal{H}_k \ \forall \, y \in \mathcal{Z}^{\mathbb{N}}$.) This Hilbert space is generated by the kernel k and equipped with the inner product between measures:

$$
\begin{aligned}
\langle f_\mu, g_\nu \rangle_k &= \int \int k(y, y') \, d\mu(y) \, d\nu(y') \\
&= \int f_\mu(y') \, d\nu(y') \ .
\end{aligned}
\tag{3}
$$

Since $\mathbb{S} \subseteq \mathbb{P}\left(\mathcal{Z}^{\mathbb{N}}\right)$, this allows us to embed the causal states into a reproducing kernel Hilbert space. Let $\mu_{\epsilon[x]}$ be the conditional measure over the set of pasts $\epsilon\left[x\right] \subseteq \mathcal{Z}^{\mathbb{N}_0}$ induced by the equivalence mapping Eq. (1) on an arbitrary past x. This conditional measure is well-defined and approximations from empirical observation will converge in distribution to $\mu_{\epsilon[x]}$ for processes meeting the conditions laid out in Section II [17, 27]. Then the *kernel causal state* $\sigma_{k,\epsilon[x]}$ is given by:

$$\sigma_{k,\epsilon[x]} = \int_{\mathcal{Z}^{\mathbb{N}}} k\left(y, \cdot\right) d\mu_{\epsilon[x]}(y). \tag{4}$$

When the kernel is *characteristic* this embedding is unique [28]. When the kernel is *universal*, the kernel-induced inner product norm metrizes convergence in distribution in the space of measures, which is to say that $\| f_{\mu_n} - f_\mu \|_k \to 0$ implies $\int f d\mu_n \to \int f \mu$ for $f \in C\left(\mathcal{Z}^{\mathbb{Z}}\right)$ [17, 29]. It has been shown that universal kernels are also characteristic. In practice, finding a universal kernel typically means working with kernels derived from a radial basis function, such as Gaussian kernels.

It should be emphasized that the choice of kernel imposes a specific geometry onto the kernel causal state set $\mathcal{S}_k$. This is, in a sense, the entire point, as imbuing the causal state set with a geometry allows us to tractably work with the ϵ-machines of systems with causal state sets too numerous and/or complex to write down explicitly. However, we wish for this geometry to be in some sense "natural" or, at the very least, consistent with both the

topology of convergence in distribution of the causal states and the product topology of the sequence space $\mathcal{Z}^{\mathbb{Z}}$. The choice of kernel and implications for the imposed geometry will be further discussed shortly and in greater detail in Appendix B.

B. Empirical kernel causal state algorithm

In practice, it can be challenging to explicitly build the equivalence classes $\epsilon[x]$ defined in Eq. (1) or to write down in closed form the conditional measures $\mu_{\epsilon[x]}$ for the vast majority of processes. With this motivation, our previous paper [18] introduced a method to construct the empirical kernel causal states of a process given a set of observations of total length T. These empirical kernel causal states are guaranteed to converge in distribution to the kernel causal states in the limit of $T \to \infty$, although the nature of this convergence may be complicated to determine [17, 30–32].

We construct not just one but two Hilbert spaces—one $\mathcal{H}^X$ over pasts and the other $\mathcal{H}^Y$ over futures. This requires constructing two kernels—k^X and k^Y, respectively. Kernel evaluations are performed on sequences (specifically, pasts or futures). To conserve the underlying product topology of sequence space, the kernel is constructed to compare sequences site-wise, with a damping parameter for indices receding further into the past or future, respectively. That is to say, when two sequences match indices sufficiently far into the past (future), those sequences become arbitrarily close under the Hilbert space metric.

This introduces two categories of meta-parameter: first, the specifics of the site-wise comparison; second, those related to the temporal damping and aggregation over time. In the first category is the choice of metric (e.g. Euclidean metric for real-valued data, or the discrete metric for symbolic data) as well as kernel width or radius. In the second category we consider the damping method and strength (e.g., exponential or power law) as well as the method of aggregating time indices (e.g., kernel products or kernel sums). We will not go into further depth here on the nature of these parameters (see Appendix B for more details) but list them to emphasize that the construction of the kernel—and thus, imbuing of the causal state set with a specific geometry—must be done carefully and with consideration to the underlying system.

Once the kernel is constructed, though, the empirical estimation of the kernel causal states is a relatively straightforward four-step algorithm, as depicted in Fig. 2:

1. Assuming that Z is vector-valued and constructed from D measurements m^i: $z_t = \{m_t^1, m_t^2, \ldots, m_t^D\}$, determine an appropriate history length L_p and future length L_f. The measurements m^i may include heterogeneous data types, see Section II. Pasts and
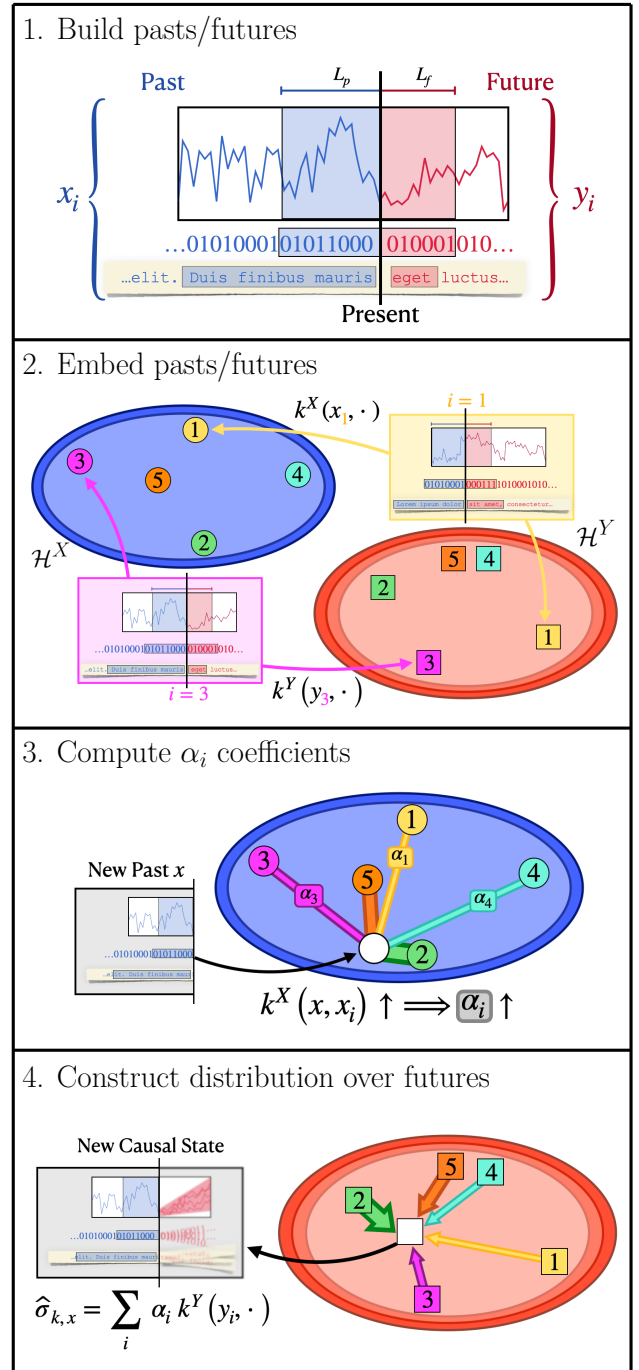

FIG. 2. The empirical kernel causal state algorithm. Step 1: Build a library of past sequences x_i and future sequences y_i from observed data. Step 2: Construct an appropriate kernel for pasts k^X and a kernel for futures k^Y. Embed pasts (circles) and futures (squares) into their respective Hilbert spaces, $\mathcal{H}^X$ (blue) and $\mathcal{H}^Y$ (red). Step 3: Calculate the empirical kernel causal state embedding coefficients α_i for a new past x by comparing x to each embedded past x_i. More similar pasts will have larger embedding coefficients. Step 4: Construct the distribution over futures for x using the calculated α_i to weight the sum over the associated futures y_i.

futures are the finite-length temporal blocks

$$x_t = \left(z_{t-L_p+1}, \ldots, z_{t-1}, z_t\right) \quad \text{and} \tag{5}$$

$$y_t = \left(z_{t+1}, z_{t+2}, \ldots, z_{t+L_f}\right) , \tag{6}$$

respectively. We then have N pasts $\boldsymbol{\mathcal{X}} = \{x_1, x_2, \ldots, x_N\}$ and futures $\boldsymbol{\mathcal{Y}} = \{y_1, y_2, \ldots, y_N\}$ where $N = T - L_p - L_f + 1$.

(Note that we allow $L_p \neq L_f$. See Appendix B for discussion of measurement-dependent history-/future lengths.)

2. Generate two reproducing Hilbert spaces $\mathcal{H}^X$ and $\mathcal{H}^Y$ using two symmetric positive-definite kernels: $k^X(x, \cdot) : \boldsymbol{\mathcal{X}} \to \mathcal{H}^X$ and $k^Y(y, \cdot) : \boldsymbol{\mathcal{Y}} \to \mathcal{H}^Y$, respectively. We embed each past x_i into $\mathcal{H}^X$ and future y_i into $\mathcal{H}^Y$ with the appropriate kernel:

$$\widetilde{x}_{i,k} = k^X(x_i, \cdot) ,$$
$$\widetilde{y}_{i,k} = k^Y(y_i, \cdot) .$$

This results in N many paired points in each Hilbert space.

3. Construct the conditional distribution over futures for a specific past x by determining the relative similarity of x with each x_i and assigning an appropriate coefficient α_i. One scheme is given by the regularized estimator from [31]:

$$\alpha(x) = \text{inv}\left(G^X + \gamma I_N\right) \boldsymbol{K}(x) , \tag{7}$$

where $G^X_{ij} = k^X(x_i, x_j)$ is the Gram matrix over past sequences, γ is a small regularization value, and $\boldsymbol{K}(x)$ is a column vector such that $\boldsymbol{K}_i(x) = k^X(x, x_i)$. More elaborate coefficient schemes can be found in [32].

4. Build the empirical kernel causal state $\widehat{\sigma}_x$ using the coefficients calculated in step 3 [31, Eq. 11]:

$$\widehat{\sigma}_x = \sum_i^N \alpha_i(x) k^Y(y_i, \cdot) . \tag{8}$$

The distribution over futures is, in effect, a weighted sum over all previously observed future sequence embeddings, where the weight schema takes into account the relative similarity of all other observed past sequences to the associated past x.

Steps 3 and 4 are repeated for all past sequences x, resulting in a set of empirical kernel causal states $\widehat{\boldsymbol{\mathcal{S}}} = \{\widehat{\sigma}_x : x \in \boldsymbol{\mathcal{X}}\}$. Past sequences that induce the same distribution over futures will map to the same point in $\mathcal{H}^Y$ in the limit of infinite data, reproducing the action of the equivalence class mapping in Eq. (1). $\widehat{\boldsymbol{\mathcal{S}}}$ is not guaranteed to have any specific geometry beyond lying inside $\mathcal{H}^Y$. Algorithm step 3 makes the implicit assumption that similar pasts yield similar distributions over future. However,

we wish to impress upon the reader that this does not induce a continuity assumption in the sequence space. For example, consider two past trajectories x and x' on opposite sides of a boundary between basins of attraction. The embedded points $\widetilde{x}_k$ and $\widetilde{x}'_k$ may become very close in $\mathcal{H}^X$, especially with finite-length histories L_p and temporal damping. However, if both sides of the basin are equivalently well-sampled and k^X is adequately discerning, then the values of each α_i will differ slightly for x and x', with each past preferring its side of the basin. Assuming k^Y is also sufficiently discerning, the difference in coefficients is then amplified in step 4 so that the $x \to \sigma_{k,x}$ mapping accurately reflects the discontinuity in the data space.

Notice that both k^X and k^Y must be correctly parametrized to handle discontinuities in the $x \to \sigma_{k,x}$ mapping. This is feasible in principle up to any chosen resolution provided enough data. An example of a closely related situation is given in Section IV A. The important point is that no additional constraint on the structure of the empirical kernel causal state set $\widehat{\boldsymbol{\mathcal{S}}}$ is imposed by the algorithm in and of itself. In practice though, the ability to discriminate fine structures (e.g., fractal boundaries) at any given resolution depends on both careful parametrization and the availability of sufficient data.

C. Assigning coordinates to causal states

Section III B's algorithm embeds the kernel causal states using N coefficients α_i. The causal states, though, are a property of the process, defined irrespective of the number of observations. Ideally, this should be reflected in the way their empirical estimates are represented. In particular, when the process is generated by an ordinary or a stochastic differential equation, whose phase space is described by M parameters, then each point in that phase space is its own causal state [18]. For these broad classes of processes, the true causal state set $\boldsymbol{\mathcal{S}}$ thus has infinite cardinality, but it can be fully indexed by a fixed number M of parameters, independent of N. An estimate of M can then be recovered from data [18]. For generic hidden Markov processes, M can be arbitrarily high [33], although it is in general finite as noted in Section III A. When estimating the causal states from measurements of a physical process, we do not have access to the microstates and their internal degrees of freedom. Yet, if a small number of parameters can encode causal states with high accuracy, even imperfectly, they still hold most of the predictive power at the data scale at which they are computed. Our previous work [18] proposed to encode the empirical kernel causal states via a small number of coordinates using a diffusion mapping [34].

This method embeds $\widehat{\boldsymbol{\mathcal{S}}}$ into $\mathbb{R}^M$ using coordinates constructed from the eigenspectra of a normalized proximity matrix on the empirical kernel causal states. This matrix is often called the *diffusion operator* because when the underlying data is drawn from a manifold the ma-

FIG. 3. Relation between the causal diffusion components ψ—the eigenvectors of the diffusion matrix $M^{\mathcal{S}_k}$—and the coordinates that encode the empirical kernel causal states.

trix approximates the Laplace-Beltrami operator on that manifold. This connection gives an appealing physical interpretation to the eigenspectra but it should be emphasized that causal states are not guaranteed or even expected to be drawn from a manifold.

A major advantage of this method is that we can construct the diffusion operator using the same kernel over futures used to construct the kernel causal states. The inner product on kernel causal states is given by Eq. (3). To calculate the inner product on the empirical kernel causal states we use the embedding coefficients from Section III B:

$$\langle \widehat{\sigma}_x, \widehat{\sigma}_{x'} \rangle = \sum_l \sum_m \alpha_l(x)\alpha_m(x')k^Y(y_l, y_m) \ . \quad (9)$$

With this, we define our kernel causal state proximity matrix as $G_{ij}^{\mathcal{S}_k} = \langle \widehat{\sigma}_{x_i}, \widehat{\sigma}_{x_j} \rangle$. We then normalize $G^{\mathcal{S}_k}$ to eliminate the influence of the sampling density, resulting in a nonsymmetric row-stochastic matrix $M^{\mathcal{S}_k}$. (See [34, 35] for further details.) Intuitively, $M_{ij}^{\mathcal{S}_k}$ gives the probability that a virtual point at $\widehat{\sigma}_{x_i}$ would travel to $\widehat{\sigma}_{x_j}$ under the action of purely isotropic diffusion acting with respect to the proximity set by the kernel inner product.

We note a potential point of confusion here: This is *not* the true ϵ-machine dynamic over the causal states referenced in Section II. Rather, this diffusion is virtual and only used here as a non-Euclidean measure of the proximity of the points. This virtual diffusion could, however, be used as a mathematical foundation with respect to which to quantify properties of the ϵ-machine dynamic. Indeed, the diffusion distance is invariant under permutation of the data indices, hence can only reflect static properties of $\widehat{\mathcal{S}}$.

Finally, we perform a spectral decomposition on $M^{\mathcal{S}_k}$. We call the right eigenvectors ψ the *causal diffusion components*.

Note that since $M^{\mathcal{S}_k}$ is row-stochastic, $\lambda_0 = 1$ and ψ_0 is constant. We normalize the eigenvectors so that $\psi_{0,i} = 1$

for all i, and omit it in the definition of the causal diffusion components. The associated left eigenvector ϕ_0 gives the diffusion-induced density at each sample $\widehat{\sigma}_{x_i}$.[2]

We use the coefficients of these eigenvectors as shown in Fig. 3 to assign coordinates to each $\widehat{\sigma}_{x_i}$. In this way we embed the empirical kernel causal state estimates in $\mathbb{R}^M$:

$$\widehat{\sigma}_{x_i}|_M \equiv (\psi_{1,i}, \ldots, \psi_{M,i}) \ . \quad (10)$$

The diffusion distance is recovered by scaling these coordinates by the corresponding eigenvalues of $M^{\mathcal{S}_k}$:

$$d\left(\widehat{\sigma}_{x_i}|_M, \widehat{\sigma}_{x_l}|_M\right) = \sum_{j=1}^M \lambda_j^2 \left(\psi_{j,i} - \psi_{j,l}\right)^2 \ . \quad (11)$$

This distance matches that of the virtual diffusion process corresponding to $M^{\mathcal{S}_k}$; i.e., how "long" it takes to "diffuse" between two given empirical kernel causal states. This is not the same as the distance in $\mathcal{H}^Y$. However, the relation between both can be inferred by expressing the spectral decomposition of $M^{\mathcal{S}_k}$ as that of an operator in $\mathcal{H}^Y$. (See also [34].)

Note that we retained only $M < N$ components. This is justified for the cases in this section's introduction where M can be formally inferred. More generally, and for natural processes especially, we cannot state in advance how many components are needed to encode the causal states up to a prescribed accuracy. When a spectral gap exists we identify M by the eigenvalue decay profile; see examples in our previous work [18] and in the next sections. Otherwise, we can set M so that the residual distance $d\left(\widehat{\sigma}_{x_i}|_M, \widehat{\sigma}_{x_i}|_{N-1}\right)$, averaged over all samples, remains below a given threshold.

Intuitively, each added component captures more predictive information—that not previously contained in the others. This is similar to how principal components progressively capture variance in a data set, but for predictive information instead of variance and using a strongly nonlinear transformation of the data. To see this, recall that eigenvalues $1 \geq \lambda_j \geq 0$ are sorted in decreasing magnitude and that Eq. (11) specifies which proportion of the distance between the empirical kernel causal states is captured by each coordinate. However, the empirical kernel causal states themselves are embeddings of conditional distributions over futures. Hence, refining their proximity also means, in some sense, refining the ability to discriminate between their predictions. This notion aligns with other embeddings of causal states [24, 36], but with the additional property of ordering the coordinates.

Another interpretation is to recall that the diffusion map is parametrized to approximate a Laplace operator. In this perspective, the eigenvectors represent the vibrating modes of a manifold enclosing the empirical kernel

---

[2] Note that $1 = \phi_0^T \psi_0 = \sum_i \phi_0$ with this normalization of ψ_0, matching the density interpretation.
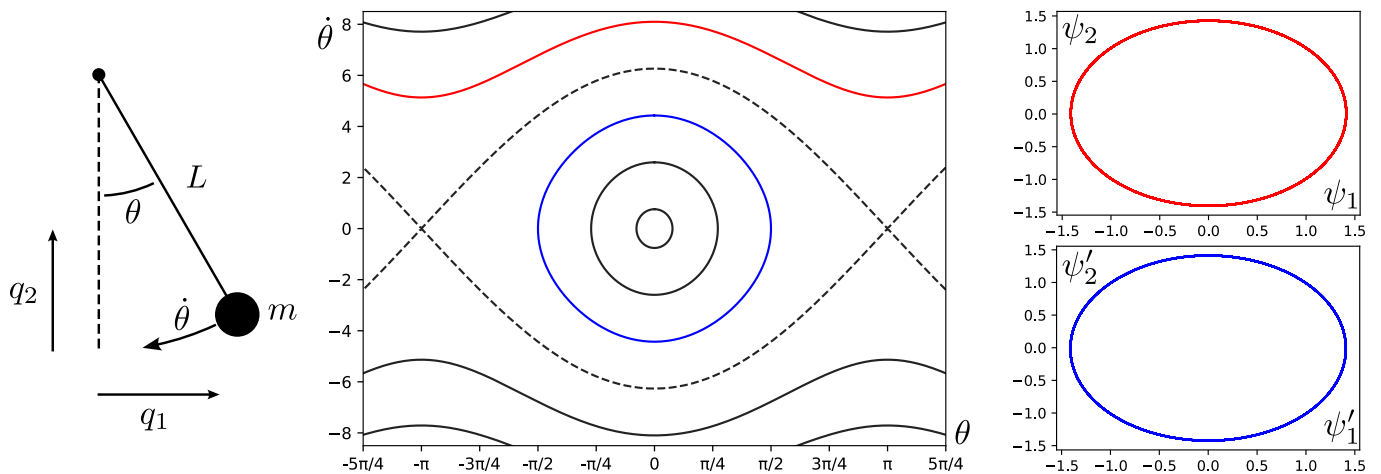
FIG. 4. Left: The simple pendulum with angular velocity $\dot{\theta}$ and angle from the vertical θ. Middle: Pendulum trajectories in the angle and angular velocity $\left(\theta, \dot{\theta}\right)$ phase space at different energy levels. The dotted line is the separatrix, for which the total energy $E = E_{\text{sep}}$. The ellipsoid trajectories in the middle represent trajectories with $E < E_{\text{sep}}$ and the trajectories above and below the separatrix have $E > E_{\text{sep}}$. The red and blue trajectories were used as input to the empirical kernel causal state algorithm. Right: Each resultant color coded set of empirical kernel causal states plotted using their first two diffusion coordinates. Eigenvalues and spectral gaps are given in Fig. 5.

causal states, with the appropriate eigenvalues associated to frequencies. Components of the diffusion map with $j > M$ are then assimilated to high frequency noise that can be safely discarded. More generally, the connection between diffusion maps and harmonic analysis has been well detailed in [37, 38]. This connection opens up the possibility to compare the empirical kernel causal states of a process estimated at multiple scales, which we keep for future work.

Another point of view on choosing M is to imagine the signals as measurements of some unknown physical process we seek to model. It is often not desirable to embed all observed fluctuations as there may be sensor noise or additional microscopic phenomena that perturb observations of the underlying system. In this view, lower frequency components capture the essence of the dynamics and additional components capture finer and finer details. The goal is to retain sufficient components to reflect the system's physics, but not those of irrelevant external factors.

Taking the last three sections together, the empirical kernel causal states and the causal diffusion components define the structural model class of *kernel ϵ-machines.* The members of this model class inherit the desirable properties of their analytical brethren—namely optimality, minimality, and uniqueness. We explicitly left out discussing the dynamic over the causal states of kernel ϵ-machines for our purposes here, as our focus is on reporting the successful empirical approximation of the kernel causal states $\mathcal{S}_k$ for a series of examples using both synthetic and real data. We leave to the future the task of inferring the kernel causal state dynamic for the examples reported here.

IV. SIMULATED EXAMPLES

This section presents two examples of the empirical kernel causal state algorithm applied to simulated data. The first system is the simple pendulum, a low-dimensional deterministic nonlinear system for which the ϵ-machine can be solved exactly. We walk through finding the causal state set and kernel causal state set analytically and compare these results to the empirical kernel causal states.

Our second simulated example is data from a high-dimensional molecular dynamics simulation of the *n*-butane molecule. We show that the empirical kernel causal states distinguish between the three well-known low-energy conformations of the molecule within the first two causal diffusion components, with finer differences in molecular positions revealed by considering higher dimensions in the causal diffusion embedding.

A. The pendulum

For a sufficiently simple system, the causal states and their kernel equivalents can be written down exactly and compared to the computational results of the empirical kernel causal state algorithm. We demonstrate this using a classical pendulum as sketched in Fig. 4 (left). The pendulum is purely deterministic and so the structure of the causal state set is especially straightforward. The purpose here is primarily to introduce notation and build intuition around working with causal states and kernel causal states for continuous-valued systems.
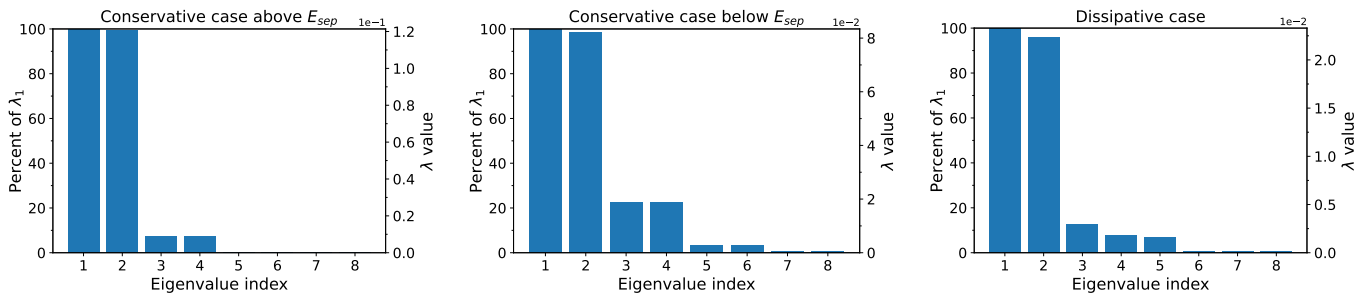
FIG. 5. Eigenvalues and spectral gaps for the simple pendulum empirical kernel causal states. The two conservative cases match those in Fig. 4 and the dissipative case Fig. 6.

The equation of motion of a pendulum is given by:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\sin\theta \ , \tag{12}$$

where L is the length of the (massless) pendulum rod, θ is the angle of the pendulum with the vertical, and g is strength of gravity. We obtain a set of coupled first-order equations by considering the angular velocity $\dot{\theta}$ as a separate variable:

$$\frac{d\theta}{dt} = \dot{\theta}$$
$$\frac{d\dot{\theta}}{dt} = -\frac{g}{L}\sin\theta \ .$$

The position of the pendulum at time t relative to an origin placed at the bottom of the swing is given by the generalized coordinates:

$$q_1(t) = L\sin\theta(t)$$
$$q_2(t) = L\left(1 - \cos\theta(t)\right) \ .$$

Despite its simplicity, the system of differential equations cannot be solved in terms of elementary functions without the small angle approximation, which reduces the dynamics of the pendulum to simple harmonic motion. However, the solutions can be computed numerically. A phase portrait of pendulum trajectories in $\left(\theta, \dot{\theta}\right)$ at different total energy levels is given in Fig. 4 (middle).

There are two modes of behavior: First, a pendulum without sufficient energy to rotate entirely around the pivot, in which case the phase portrait is a limit cycle that grows increasingly elliptical as energy is increased; second, a pendulum with enough energy to swing all the way around, in which case the phase portrait is a sinusoidal curve. These modes are separated by a boundary, called a *separatrix*, found at the energy level E_{sep} at which the pendulum can stand straight up in an unstable equilibrium at $\theta = 180°$.

Now, let's construct the causal state set of an idealized pendulum that has been swinging for infinite time at energy level $E \neq E_{\text{sep}}$. Since there is no stochasticity in the system, this is simple: each point on the phase space trajectory corresponds to exactly one causal state. Let us parameterize the phase space trajectory with a single coordinate denoted $\phi \in [0, 2\pi)$. When $E \leq E_{\text{sep}}$ we use the angle from the positive horizontal axis for this purpose; when $E > E_{\text{sep}}$ we use $\theta \mod 2\pi$. Either way, each causal state σ_ϕ is labeled by its value of ϕ.

The time indices for which the pendulum will be at exactly ϕ are given by $\tau_\phi = \{\tau + nT : n \in \mathbb{Z}\}$, where τ is the time it takes the pendulum to travel from its initial condition to ϕ and T is the period of the oscillation. For our idealized pendulum, each causal state σ_ϕ is associated with a set of infinitely many pasts $\{x_\tau : \tau \in \tau_\phi\}$. Since the pendulum is deterministic, each past in this set induces not only the same *distribution* over futures $\mu_{\epsilon[\phi]}$ as required by the equivalence relation but also the exact same *realization* of the future: the exactly solvable trajectory of the pendulum for $t > \tau$, $\tau \in \tau_\phi$.

So the causal state set $\boldsymbol{S}$ of the simple pendulum has infinite cardinality but is one dimensional, in the sense that we can uniquely label each casual state with a single variable. It should be noted that this analysis does not depend on the total energy of the pendulum.[3]

Now let's construct the kernel causal states. Recalling that the conditional distributions over futures μ_ϕ are delta distributions over a single future, we have:

$$\sigma_{k,\phi} = \int_{\mathcal{Z}^{\mathbb{N}}} k^Y(y, \cdot)\, d\mu_\phi(y) = k^Y(y_\phi, \cdot) \ .$$

Following Section III A, the inner product between the kernel causal states of the pendulum reduces to kernel evaluations between futures: $\langle \sigma_{k,\phi}, \sigma_{k,\phi'} \rangle = k^Y(y_\phi, y_{\phi'})$. If we assume that the kernel is constructed to agree with the Euclidean distance between positions of the pendulum in physical space, then kernel causal states corresponding to nearby points on the pendulum trajectory in phase space will be close to each other in the kernel Hilbert space. Furthermore, the kernel causal state set $\boldsymbol{S}_k$ will inherit

————

[3] Except in the case of the unstable equilibrium. For a pendulum that stands at $\theta = 180°$ indefinitely, the recurrent causal state set is a single point.

the topology of the trajectory in phase space, which is to say the circle S^1.

Let's compare this to the structure of the empirical kernel causal state set $\widehat{\mathcal{S}}$. Consider the empirical kernel causal states of the pendulum at two different energy levels, as plotted in Fig. 4 (middle). The first has $E < E_{\text{sep}}$ and is given in blue, the second is given in red and has $E > E_{\text{sep}}$. In both cases we use the generalized coordinates (q_1, q_2) as input to the empirical kernel causal state algorithm and take $L_f = L_p = T$ so that our past and future sequences are arrays of shape $T \times 2$:

$$x_t = \left(\begin{bmatrix} q_{1,t-T+1} \\ q_{2,t-T+1} \end{bmatrix}, \ldots, \begin{bmatrix} q_{1,t-1} \\ q_{2,t-1} \end{bmatrix}, \begin{bmatrix} q_{1,t} \\ q_{2,t1} \end{bmatrix} \right)$$

$$y_t = \left(\begin{bmatrix} q_{1,t+1} \\ q_{2,t+1} \end{bmatrix}, \begin{bmatrix} q_{1,t+2} \\ q_{2,t+2} \end{bmatrix}, \ldots, \begin{bmatrix} q_{1,t+T} \\ q_{2,t+T} \end{bmatrix} \right) .$$

We use a Gaussian kernel with variance 1 for comparing values in each sequence and a product of such kernels for both k^X and k^Y.

After the empirical kernel causal states are constructed, they are each embedded using the diffusion transform from Section III C. The two cases are plotted using their respective first two components in Fig. 4 (right) and color coded to match their respective trajectories in the phase space plot. Although the empirical kernel causal state sets are point sets, in the plot the $\widehat{\mathcal{S}}$ sets appear as smooth ellipsoids. This is due to the density of the sampling of the pendulum trajectories. If this sampling was reduced, we would begin to observe gaps between the empirical kernel causal states, as we do in the real data examples in Section V.

As is visually apparent, the empirical kernel causal states $\widehat{\sigma}_x$ and the diffusion eigenvectors are identical for both $E > E_{\text{sep}}$ and $E < E_{\text{sep}}$. This is just as expected—as noted above, in nearly every case the causal states are not dependent on the pendulum's energy level, and so the empirical kernel causal state do not distinguish between these two cases.

In both cases we require two components (ψ_1, ψ_2) to embed the empirical kernel causal states. This is because the diffusion mapping algorithm retains the topology of the kernel causal states when embedding them in Euclidean space and S^1 is minimally embedded in $\mathbb{R}^2$. We indeed note a spectral gap with the third eigenvalue in Fig. 5.

The difference between the two cases is explained by the difference in the nature of the trajectories in (q_1, q_2) space. Below the separatrix, the points near each extrema of the pendulum, with $q_1 < 0$ or $q_1 > 0$ but with the same q_2, are also the points where the pendulum has the lowest velocity. Hence the trajectories, sampled at constant rate, consist of more points near these top positions than at the bottom. The kernel bandwidth thus has to be reduced in order to correctly separate these points, which become closer and closer as we approach the separatrix. Above the separatrix, this is not so much of an issue, because the pendulum never swings back. So, past-future sequences

are similar only to those at the same location along the full loop.

We also wish to consider the case of the damped simple pendulum, in which total energy of the system changes over time. The equation of motion is adjusted to add a drag term:

$$\frac{d^2\theta}{dt^2} = -\frac{b}{m}\frac{d\theta}{dt} - \frac{g}{L}\sin\theta , \qquad (13)$$

where $m = 1$ is the mass of the pendulum bob and $b = 0.1$ is the drag coefficient. The trajectory of a pendulum bob originating at $\theta_0 = -\frac{\pi}{2}$ and zero initial velocity is shown in $(\theta, \dot{\theta})$ phase space in Fig. 6 (middle).

As in the undamped case, each point in the phase space trajectory corresponds to exactly one causal state. However, in this case, each point is only visited once and corresponds to exactly one time index—except for the final point at the bottom of the oscillation, which the pendulum reaches as $t \to \infty$ and then remains at indefinitely.[4] We call these *transient causal states*. However, the advantage of the empirical kernel causal states algorithm is that because it does not rely on clustering or frequentist estimation of probability measures, it is just as capable of constructing the transient causal states, as we show here.

As is visually clear, the damped pendulum is quasi-periodic, with the duration of each swing decreasing over time due to drag. We take the length of the past and future to be the average quasi-period $\overline{T}$ as inferred over the full simulation, so $\overline{T} = L_p = L_f$.

The right side of Fig. 6 shows the empirical kernel causal states of the pendulum plotted using their first two components (ψ_1, ψ_2). Once again, only two components are necessary to embed the pendulum's $\widehat{\mathcal{S}}$ (Fig. 5). In this case we more clearly observe a mild distortion in the empirical kernel causal states as compared to the trajectory in phase space. This is expected, as the causal states are independent of the frame of reference of the underlying system and so the diffusion map algorithm (or indeed, any dimension reduction algorithm) applied to the empirical kernel causal states will not return the original system's coordinate system in general.

Here we do not discuss the dynamics of the ϵ-machine or its kernel variant, intentionally, for now, limiting our scope, which focuses only on the structure of the kernel causal state set. However, in this case it is unsurprisingly rather straightforward: the causal states smoothly evolve along the causal state parameter ϕ. The same holds true for the kernel causal states. For the empirical kernel causal states, we could in principle infer the evolution of $\widehat{\sigma}_x$ in terms of the diffusion coordinates using a variety of algorithms [4, 10–13]. However, we leave this to the future.

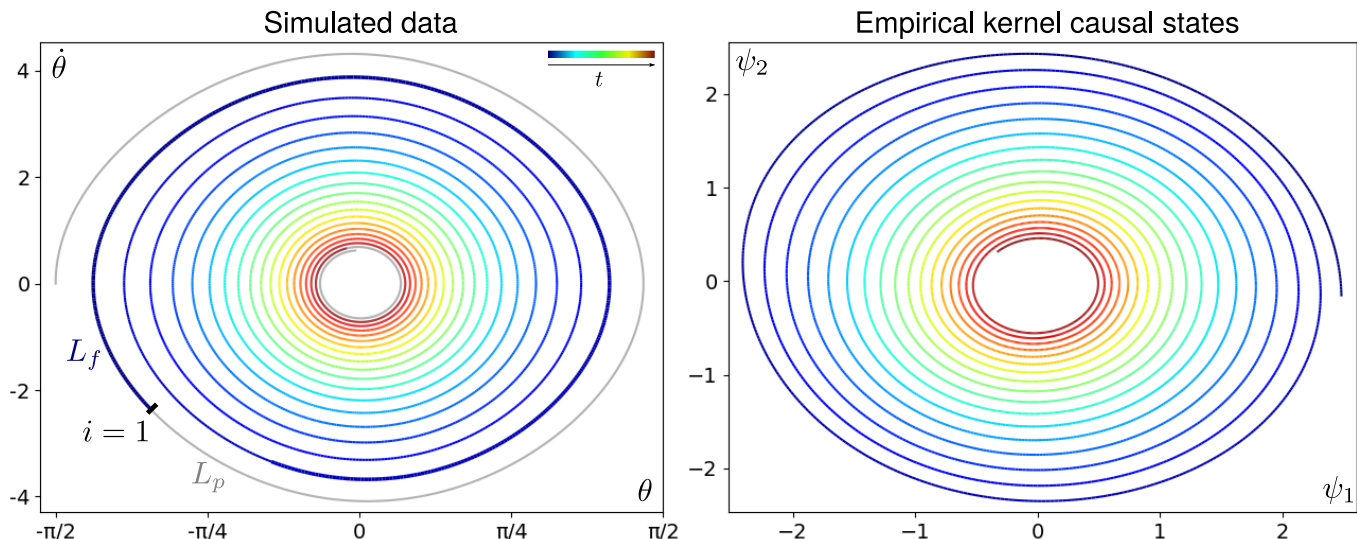[4] As in the E_{sep} case, the recurrent causal state set is a single point.

FIG. 6. Left: The pendulum trajectories in the angle and angular velocity $(\theta, \dot\theta)$ parameter plane colored by time. The gray sections at the beginning and end are the first L_p-length past and last L_f-length future, respectively, and so do not correspond to any embedded causal states. The bold segment denotes the L_f-length future for the first embedded history. Right: The empirical kernel causal states plotted using their first two diffusion coordinates.
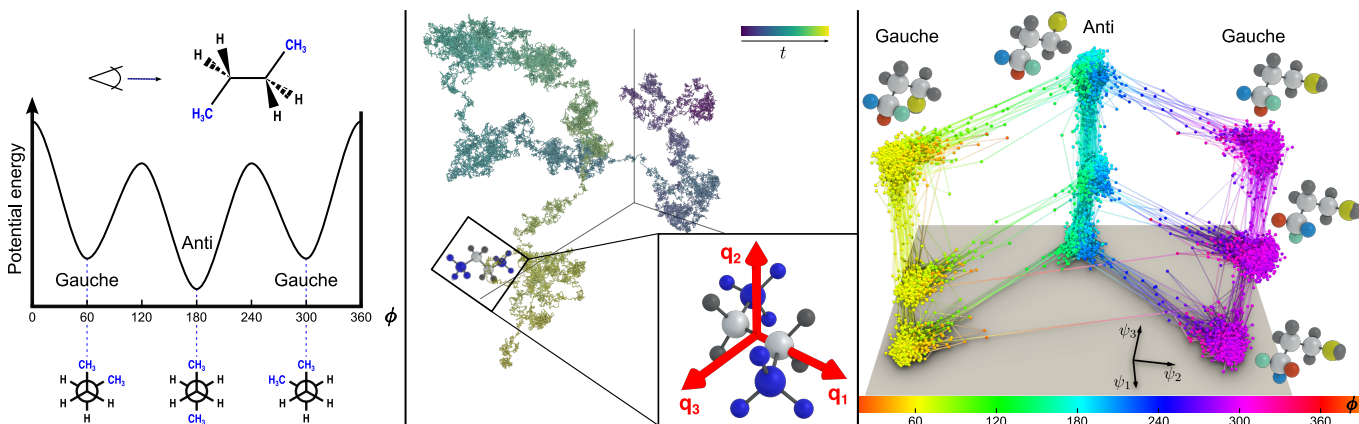

FIG. 7. Left: The potential energy of n-butane molecule plotted along the dihedral angle ϕ. The Newman projection for each of the three low energy conformations is shown below the plot. Middle: The global trajectory of the simulated n-butane molecule under Brownian motion. The inset shows the local frame of the molecule position used as input to the empirical kernel causal state algorithm. Only the carbon in the back is mobile in this local frame. Right: The empirical kernel causal state set of the n-butane molecule, embedded using the first three causal diffusion components. The frame shows the base vectors in $\mathbb{R}^3$, scaled down to fit in the figure. These give the directions along which the values $\psi_{j,i}$ are plotted, for $j = 1, 2, 3$ indicated on the figure and i each data index. Dots are plotted at the locations of each $\widehat{\sigma}_{x_i}$ and lines connect successive samples in time. Nine clusters are clearly apparent. The average position of each atom in a selection of these clusters is shown, with the free carbon (yellow) and hydrogens of the fixed methyl group (red, green, and blue) highlighted.

B.   n-Butane molecule

Now, let's move on to a discovery challenge that showcases how the empirical kernel causal state method applies to a high-dimensional, stochastic system. The n-butane molecule $H_3C - CH_2 - CH_2 - CH_3$ consists of four carbon atoms and ten hydrogens arranged as depicted in Fig. 7. The energy landscape and conformations of n-butane are well known and characterized by the dihedral angle ϕ between the two methyl groups. There are three low-

energy conformations: the *antiperiplanar* conformation, for which $\phi = 180°$ and the methyl groups are anti-aligned, and the two *gauche* conformations (from left to right, $\phi = 60°$ and $\phi = -60°$) for which the methyl groups are staggered. As we will show, the empirical kernel causal states find these conformations even without taking the dihedral angle as input to the algorithm.

Molecular dynamics simulators begin with the initial positions and velocities of a set of atoms and simulate their motion over time according to the influence of atomic

interaction forces and possibly external forces (an applied temperature gradient or magnetic field, for example). The data set we use is the full time series of the positions of the fourteen atoms of the n-butane molecule as calculated by the molecular dynamics simulator AmberTools15. The n-butane molecule was simulated for an interval of 10 ns with a step-size of 2 fs. The resulting data was then downsampled by a factor of 100, of which we retain a trajectory of 25 000 data points. See [39] for details, where this data set was first introduced.

The molecule's trajectory in physical space is plotted in Fig. 7 (middle). The Brownian motion of a molecule at a constant nonzero temperature is typically modeled with the addition of a stochastic term—i.e., the Wiener process—rather than explicitly calculating the collisions with surrounding molecules. As such, this dataset is intrinsically stochastic, in contrast to the previous examples of purely deterministic systems. However, the presence of Brownian motion does pose one complication. If we analyze the movement of the molecule using the fixed simulation frame of reference, we will model not only the relative atomic positions—which determines the conformation—but also the global position of the molecule, which is irrelevant to the conformational dynamics.

To factor this this out, we introduce a local frame of reference internal to the molecule. The middle carbon bond is used to define one unit vector, see Fig. 7 (middle) for reference. The second basis vector is formed by the cross-product of the first with another carbon bond vector, and the third is computed as the cross-product of the first two vectors. All atom positions are expressed in that local basis, rendering our past and future sequences independent from the global location and orientation of the molecule in space. Thus, the inputs to the causal-state embedding algorithm are sequences consisting of 42 dimensional (three spatial coordinates times fourteen atoms) position data over one time step, for both the past and future sequences.

The empirical kernel causal states are shown in Fig. 7 (right) using their first three diffusion coordinates. They are colored according to the the dihedral angle of the molecule during the past associated with that empirical kernel causal state (since the history length is one time step, this is a unique value). The points in the figure are the empirical kernel causal state locations and the lines connect $\widehat{\sigma}_{x_t}$ to $\widehat{\sigma}_{x_{t+1}}$ for all t.

The immediately obvious structure of the $\widehat{\mathcal{S}}$ is nine clearly identifiable conformational clusters. Recall that in the limit of identical predictions over futures the kernel causal states map to the exact same point in Hilbert space and that the diffusion mapping translates the difference in the predictions over futures into distance in the diffusion coordinates space. This means that a tight cluster in the diffusion coordinates space represents a group of time indices in the observed trajectory where the predicted future was very similar. Jumping between these clusters then represents a significant change in the next predicted

molecular conformation. Typically, we can see that these transitions occur over one or two time steps.

The nine empirical kernel causal state clusters are arranged into three super-clusters along ψ_2 of three sub-clusters along ψ_3. We depict the position of each atom for five of these clusters, averaged over their matching time indices. (The other four clusters follow a similar pattern and can be reproduced with the provided code.) First consider only the position of the freely moving carbon as dictated by the local reference frame. This is the carbon highlighted in yellow in the ball-and-stick depictions. Comparing its position across the super-clusters shows ψ_1 separates the gauche and antiperiplanar conformations, as is also apparent using the coloring scheme. It does not distinguish between the two gauche conformations—recall that both gauche conformations have the same potential energy, as shown in Fig. 7 (left). If we take into account the second component ψ_2, the gauche conformations are separated. And, this gives us an interpretation of the superclusters: all three low-energy conformations are uniquely associated with a supercluster in the empirical kernel causal state set. Together, the first two components describe the position of the carbon chain.

Now, consider the positions of the hydrogens attached to the fixed carbon (highlighted in Fig. 7 (right) in red, green, and blue). These positions do not change when comparing across the top subcluster of each supercluster. Only by comparing the subclusters of a single supercluster does it become clear: the subdivisions in the conformational superclusters along ψ_1 represent changes in the orientation of the fixed methyl group. The first three components shown in Fig. 7 do not capture the second methyl group orientation. This is why the positions of the hydrogen atoms of that methyl group are not distinguished in the stick-and-ball plots. Further components (Fig. 8, left) add subclusters that do distinguish the orientation of the second methyl group. This preference for tracking the orientation the fixed carbon methyl group is purely algorithmic: the orientation of the methyl group attached to the freely moving carbon atom is washed out, in a sense, by the more relevant motion of the carbon. This comes down to the way one defines the local frame and does not represent a real asymmetry in the chemical importance of the two methyl groups.

We note that the frame in Fig. 7 (right) could be slightly rotated to better align ψ_2 with the dihedral angle. This would also give the gauche and antiperiplanar subclusters the same vertical coordinates. Recall, however, that the empirical kernel causal state algorithm has no notion of potential energy, dihedral angle, or methyl group.

In particular, ψ_3 separates out the rotation of fixed carbon methyl group not because there is a *physical* meaning to its orientation, but because its orientation probabilistically influences the future of the entire molecule. As is clear in Fig. 7 (right), it is possible for the molecule to transition between methyl group orientations, and it is possible for the molecule to transition between conformations, but it is highly improbable for these transitions to happen in
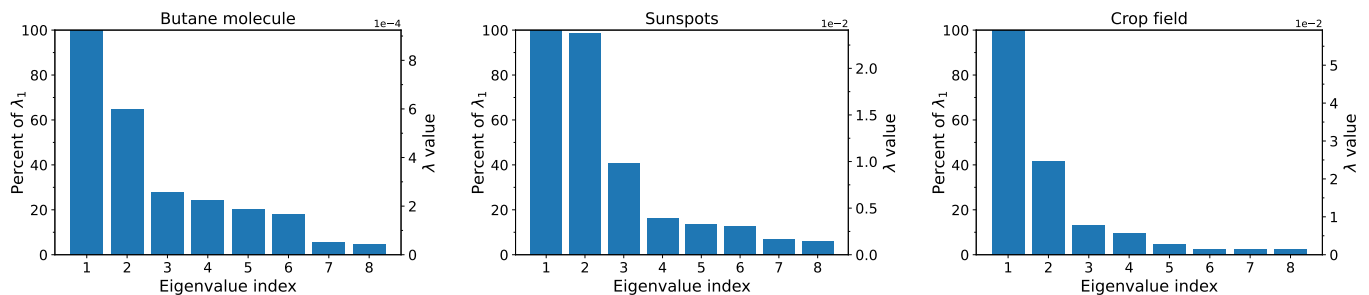
FIG. 8. Eigenvalues and spectral gaps for the empirical kernel causal states of the butane molecule (Fig. 7), the sunspots (Fig. 9) and the crop example (Fig. 10).

the same time step. The takeaway is that while structural analysis of the $\widehat{\mathcal{S}}$ can be very powerful, the interpretation of the uncovered structures and any possible physical meaning of the causal diffusion components must be done in conjunction with discipline knowledge.

Comparing this to the simple pendulum in Section IV A demonstrates the our method's ability to handle highly heterogeneous types of systems. The structure of the *n*-butane molecule empirical kernel causal state set is totally noncyclical and highly stochastic.

Indeed, the empirical kernel causal states naturally organize into something that looks very much like a finite-state machine. Modeling this process at a larger time scale, we could discretize each cluster as distinct "meta"-causal states, with their own probability distribution of dwell times and instantaneous transitions to the next cluster. This would give a continuous-time, discrete-state ϵ-machine of a renewal process; see [40, 41]. Though out of present scope, it will be considered in future explorations of nested causal structure, together with a comparison of the dwell times in each conformation with the theoretical distribution for residency in each energy well.

V. REAL DATA APPLICATIONS

This section presents two examples of the kernel causal-state embedding algorithm applied to real data. The first is a classic and well-known example in observational astronomy: the number of sunspots visible on the sun's surface each month. We show that the causal diffusion components discover several known empirical patterns in the sunspot sequence. The second example is data from a crop field. This example shows the use of multi-variate and heterogeneous measurements, some of which exhibit missing values and low data quality. We show that the empirical kernel causal states detects changes in crop species over a decade-long observation period.

A. Sunspots

Let's start with a deceptively simple and well-known example. One of the longest-running time series of direct observations without interruptions available is the number of sunspots. Sunspots have been observed on the sun's surface since the advent of the telescope. Sunspots are temporary dark spots on the Sun's surface caused by concentrations of magnetic flux. High-quality monthly counts are provided by the Sunspot Index and Long-term Solar Observations (SILSO) databank maintained by the Royal Observatory of Belgium [42] from 1749 to the present day, as shown in Fig. 9 (left). Regularizations are performed to account for differences in measurement technique over time.

Driven by the Sun's turbulent convection, the sunspot sequence is a well-known benchmark in nonlinear time series analysis and it is known to be very difficult to predict [43]. Fundamental aspects of how and why the number of sunspots vary over time—and how those changes are related to solar physics—are not well-understood to this day. Our goal is not to build a full predictive model, but rather to show how the geometry of the embedded kernel causal states aids in a structural analysis of the underlying dynamical system.

Sunspots are one experimental observation of the broader solar magnetic activity cycle. This cycle is also reflected in the quasi-periodic variation in other solar activity—such as solar flares and coronal loops and mass ejections. The physical driver of the magnetic activity cycle is the Sun's magnetic field reversing polarity—that is, the Sun's north and south magnetic poles swap places.

The beginning of the cycle is called the *solar minimum*, and is marked by observation of very few sunspots. Over time, observed solar activity increases until it peaks at the *solar maximum* midway through the cycle. Conventionally, solar cycles are numbered from solar cycle 1, which began in February 1755 and ended in June 1766. At the time of writing we are in solar cycle 25, which began in December 2019. A full solar cycle from solar minimum to solar minimum takes on average eleven years, meaning it takes roughly twenty two years for the magnetic poles to return to their original positions. This magnetic field
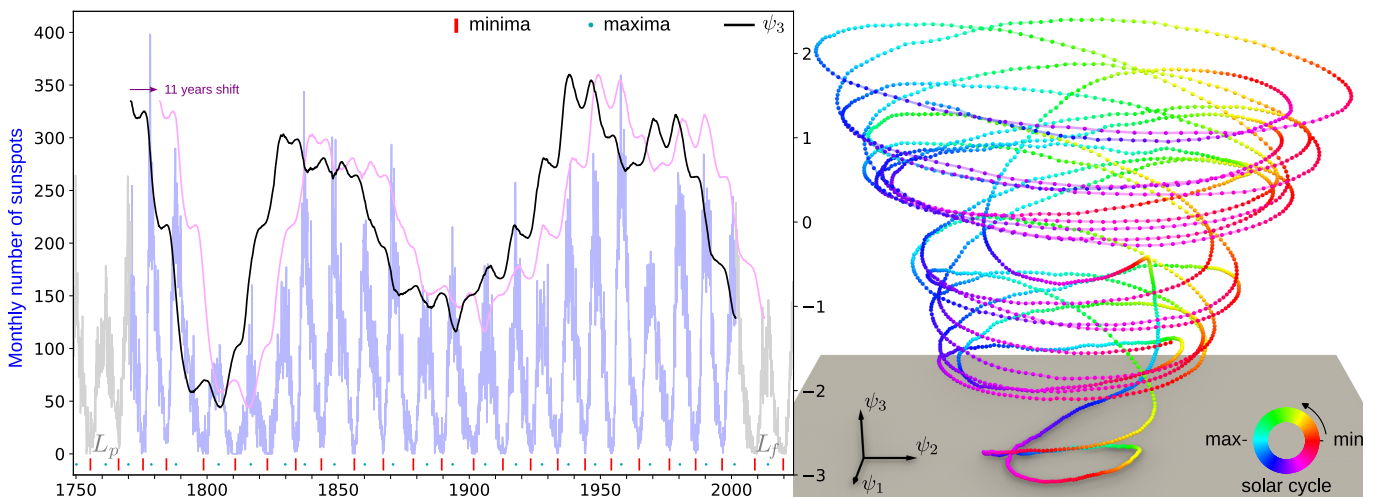
FIG. 9. Left: Monthly number of sunspots since 1749. Dates of solar minima and maxima are indicated below the main plot. The third diffusion component is overlaid on the data in black. Shifting this component by eleven years clarifies that ψ_3 approximately predicts the amplitude of the next solar cycle. Right: The empirical kernel causal states of the sunspot sequence, embedded using the first causal diffusion components. The indicated frame, dots and lines in the figure have the same interpretation as in Fig. 7. The color map indicates the solar cycle phase at the time t associated with each $\widehat{\sigma}_{x_t}$. This phase was computed by linearly interpolating the time between each solar minima and maxima.

periodicity is called the *Hale cycle.*

(The magnetic fields of celestial bodies, such as stars and planets, are theorized to be produced by the dynamo mechanism, which describes the motion of plasma or molten metals moving within the body's core. The Earth's magnetic field also undergoes polarity reversal. Though, unlike the Sun's solar cycle, statistical analysis of the Earth's reversals have not revealed any obvious periodicity [44, 45].)

The solar cycle is the most obvious quasi-periodic cycle in the sunspot sequence. However, there are other proposed periodic or quasi-periodic cycles hypothesized to influence solar activity. The most relevant here is the *Gleissberg cycle.* This cycle's period is said to be on the order of 70 to 100 years and modulates the amplitude of the solar cycle. There are other noncyclical observed patterns in the data. The Waldmeier effect states that the length of an individual solar cycle is inversely proportional to its amplitude maximum. The puzzling Gnevyshev-Ohl rule notes that sums of sunspot numbers over odd cycles are highly correlated with sums over preceding even cycles. The correlation is weaker between sums over even cycles and their preceding odd cycles. This perhaps suggests a preferential labeling of the Hale cycle. However, the physical mechanisms behind these patterns are still unknown.

Let's examine the results of the empirical kernel causal state embedding. We set the past and future length $L_p = L_f$ to be twenty two years—the length of the Hale cycle. The kernel bandwidth is set as the average amplitude of a solar cycle. Figure 9 (right) shows the empirical kernel causal state embedded using the first three components and colored according to the solar cycle phase. This phase

is estimated from data: after estimating the solar minima and maxima with a low-pass filter, the approximate phase of the solar cycle for each data point was computed by linear interpolation. Each empirical kernel causal state $\widehat{\sigma}_{x_t}$ was then colored according to the phase at time t.

The empirical kernel causal state set $\widehat{\mathcal{S}}$ traces out a cone-like surface in ψ space or, more fancifully, evokes a tornado. An immediate observation when comparing to the n-butane example in Fig. 7 is how smooth the time ordering is in causal state space. At this scale, we see no obvious time scale separation, indicating that the distributions over futures change smoothly over time, rather than exhibit sudden changes.

As is visually obvious from the color mapping, the first two components ψ_1 and ψ_2 capture the eleven-year solar cycle. Each eleven-year cycle traces out a roughly circular path in this plane and the phase coloration approximately aligns cycle to cycle. This alignment is not perfect because the solar cycle is not perfectly periodic. We can visually identify phase slipping, as we expect for a system hypothesized to be modulated by multiple, competing, periodic and quasi-periodic cycles. Yet, Fig. 8 (center) shows eigenvalues λ_1 and λ_2 are almost identical, indicating no particular preference for the phase of that cycle, unlike the crop example next.

The cycles are further organized along the third dimension ψ_3, here aligned with the vertical. Overlaying ψ_3 on the raw data, as in Fig. 9 (left), shows that the third component captures the amplitude modulation of the solar cycles with a slight offset—in fact, a lag of approximately eleven years. When ψ_3 is plotted with an eleven year shift, it tracks the amplitude modulation of the solar cycle. This implies that the third component at any given

time is predicting the amplitude of the *next* solar cycle, capturing the patterns described the proposed Gleissberg cycle. Additionally, an interesting section of the sunspot sequence is the *Dalton minimum*, which lasted from about 1790 to 1830 and covers solar cycles 4 through 7. During this minimum the solar activity was unusually low. In the empirical kernel causal state set the Dalton minimum corresponds to the segment of empirical kernel causal states that traverse through the center of the "cone", indicating a significant change in solar behavior as compared to other points in the sunspot series.

B.   Grignon crop field

Our final example showcases how the empirical kernel causal state algorithm handles heterogeneous data with a system of six different measurements from a crop field over a period of eleven years. This example also highlights how to work with multivariate data with periods of missing or low-quality measurements. These data are provided by the Integrated Carbon Observation System (ICOS), a research collaboration dedicated to producing standardized, high-quality climate observations to better understand the carbon cycle and greenhouse gas balance in Europe and adjacent regions. The collaboration maintains over forty atmospheric monitoring sites and over one hundred ecosystem stations in over a dozen European countries. These stations are strongly standardized and all observations are made publicly available by the ICOS network [46].

The Grignon ecosystem station is located in a nineteen hectare crop field about forty kilometers west of Paris. The site is near a cattle farm and exposed to heavy pollution from the Paris metropolitan area, as well as clean wind from the southwest. The growing practices at the site are representative of standard arable crop farming in France, with a rotation of crops including wheat, maize, and oilseed rape.

The actual measurement device is an eddy covariance flux tower. Eddy covariance is the standard method used by ecosystem scientists to monitor gas exchange between the land (soil and vegetation) and the atmosphere. The flux tower measures the vertical movement of greenhouse gases carried by eddies, which over time indicates whether the ecosystem is acting as a carbon sink or source. The advantage of the eddy covariance method is that it measures fluxes directly without disturbing the ecosystem. Flux towers also can continuously operate for decades, providing a long-term picture of how an ecosystem is changing over time.

The tower at the Grignon site makes measurements of carbon fluxes and meteorological conditions every half hour. We obtained measurements of air temperature (°C), solar influx (incoming shortwave radiation W/m^2), vapor pressure deficit (hPa), precipitation (mm), evapotranspiration (more precisely its proxy latent heat flux W/m^2),

and the CO_2 flux over the field (net ecosystem exchange µmol/m^{2}s) for a period covering 2004 to 2015. This six-dimensional data set was measured daily over a period eleven years, totaling 4018 data points. These measurements are plotted in Fig. 10 (left), overlaid with the crop type the field was growing at the time: either mustard (yellow), maize (blue), cereals (green), Phacelia—a nitrogen holder and weed suppressant (pink), and rapeseed (red). During the study period the Grignon field grew cereals (wheat, barley, triticale) the most often.

Each measurement source is accompanied by a quality control flag indicating the validity of the data at that time. In regions marked as low quality the data may be irrelevant or missing entirely. This results in gaps in the time series data in the latent heat flux and the CO_2 flux series, highlighted in orange in Fig. 10 (left). We filled the missing values with the technique outlined in Appendix C. In brief, this method involves training a linear model in diffusion coordinates space, constrained by the valid measurement series, and using the predictions of that model to fill the gaps. While simple, this method was effective for our purposes and could be replaced by more sophisticated gap-filling methods in the future.

Constructing kernels for heterogeneous data is discussed at length in Appendix B and so we will not go into depth here. However, we do note that since all measurements are real valued the underlying metric (Euclidean) was consistent across each measurement source. The past and future history lengths of $L_f = L_p = 91$ days (on average, three months) were also held constant across measurement sources. The Gaussian kernels bandwidth were set individually, as the standard deviation of each source. The kernels were aggregated using the product rule with no temporal decay.

Figure 10 (right) presents two views of the empirical kernel causal states in diffusion components space. The left subfigure uses ψ_3 for the vertical axis, while the right uses ψ_4. Both three-dimensional representations use ψ_1 and ψ_2, which are also plotted in the plane below. This projection is thus the same for both subfigures and it is colored according to the day of year at time t for each $\widehat{\sigma}_{x_t}$. In both of the three-dimensional plots, the empirical kernel causal states are colored according to the crop being grown at their matching time. That said, it is important to keep in mind that the $\widehat{\sigma}_{x_t}$ are defined using the preceding three months and predict the next three months of field dynamics (Eqs. (5) and (6)).

We can visually identify that the first two components capture the seasonal cycle. If we compare the phase alignment in this example to the phase alignment in the sunspot $\widehat{\mathcal{S}}$ we can see that the alignment is a bit more consistent, as would be expected for the regularity of the seasonal cycle compared to the sunspots irregular 11-years cycle. However, the eigenvalues in Fig. 8 (right) reveal that ψ_1 is more informative than ψ_2 with respect to predicting the crop dynamics. One possible interpretation is that fluctuations in the data sources (temperature, rain,
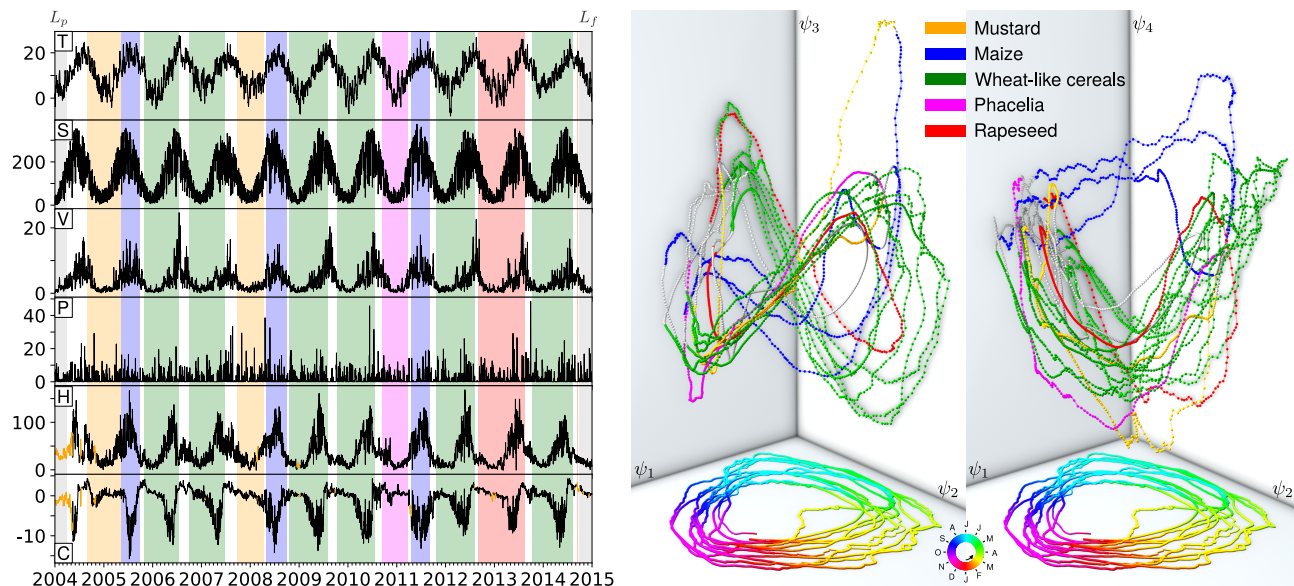
FIG. 10. Left: The six measurement sources over a period of 2004 to 2015 taken at the Grignon field site. From top to bottom: T) Temperature (°C); S) Solar influx (W/m^2); V) Vapor pressure deficit (hPa); P) Precipitation (mm); H) Heat flux (W/m^2); C) CO_2(μmol/m^{2}s). Missing or low-quality values that were filled with the gap-filling technique in Appendix C are in orange. Periods between seedling and cutting/harvesting are also indicated for each culture type. Right: The empirical kernel causal states embedding on the left subplot is shown using the first three components (ψ_1, ψ_2, ψ_3), and with ψ_4 instead of ψ_3 on the right. Trajectories are colored by culture type. Below, their projection on the (ψ_1, ψ_2) plane is shown for both subplots, colored by the day of the year.

CO_2, etc) in May or November (the alignment of ψ_1) are more impactful on the future of the system than fluctuations in February or August (the alignment of ψ_2).

Both views on the empirical kernel causal states reveal a saddle-like geometry in the three-dimensional space. In the left subplot using ψ_3, the low points of the saddle are roughly aligned with the summer (green) and the winter (pink), while the high points occur in fall (turquoise) and spring (yellow). In the ψ_4 subplot the saddle is inverted— the low points of the saddle are aligned with fall and spring and the high points are aligned with summer and winter. In late fall and early winter, the field will be bare for the most of the prediction window (three months) and as such we see that the empirical kernel causal states are mapped to a relatively small region in ψ space.

In the other seasons, the predictions made by the empirical kernel causal states begin to diverge. This divergence begins to appear as early as February. At least some of this divergence appears to be related to what pattern of crops is being grown that year. Most years, the field grows only a single cereal, but in three years (2005, 2008, and 2011) the field grows a cover crop in the spring (mustard in 2005 and 2008, Phacelia in 2011) followed by maize in the summer and fall. The empirical kernel causal state associated with these years are mapped away from the body of the saddle by both ψ_3 and ψ_4, although it is easier to see with ψ_4 in the right subplot. One possible explanation for these diverging trajectories is that, given the same environmental conditions, different plants produce different responses. This is reflected in

particular in the CO_2 and latent heat series.

VI. CONCLUSION

The development here covered a lot of ground—including, beyond the example applications, reviewing the theory of computational mechanics (Section II) and its extension to kernel ϵ-machine introduced in [18]. We then detailed the empirical kernel causal states algorithm (Section III B), including the diffusion map embedding it uses (Section III C) to produce *causal diffusion components*. Finally, we gave four examples of empirical kernel causal states computed on both simulated and real data, for both deterministic and highly stochastic systems. We performed preliminary structural analysis of these results, showing how the causal diffusion components uncover "hidden" structure. Three systems were arbitrarily high-dimensional, in the sense that the number of degrees of freedom of the system was unknown a priori.

A major advantage of computational mechanics-based methods is the interpretability the framework provides. By first transforming the sequences into the kernel causal states space (Section III B) and then applying manifold-learning techniques (Section III C), the empirical kernel causal states algorithm provides a meaningful separation between the discovery of causal structure and geometric structure. To the best of our knowledge the ability to analyze the global geometry of the causal state set in this way is novel.

In a sense, for data analysis and modeling purposes, our causal diffusion components plays a similar role as applying PCA to time-delayed embeddings. Both exploit the time dependencies of the signal and perform dimension reduction. However, our causal diffusion components are designed to maximize predictability, not variance, and they do so in a principled way, with sound theoretical foundations [17, 27].

The causal diffusion components can be intuitively seen as capturing the most impactful factors driving the system dynamics, as can be seen from the various examples recounted here: recovery of the phase space for the simple pendulum (Section IV A), of the main energy wells and the structural conformation of the butane molecule (Section IV B), of the major components of the solar cycle (phase, amplitude, Section V A), of the seasonal cycle and the culture types in a cultivated crop (Section V B).

Finally, we emphasize the potential power of the full kernel ϵ-machine: it is not only able to perform the kind of structural analysis done here, but also, given sufficient data for its estimation, the kernel ϵ-machine is theoretically an optimally predictive model of the system. These aspects require further development and will be the topic of a future article in this series.

AUTHORS' CONTRIBUTIONS

Author Jurgens clarified and extended the theory of RKHS causal states and wrote the main part of the article. Author Brodu wrote the code, designed and ran the examples and wrote the first draft of the article. Both authors contributed equally to the analysis and the interpretation of the results.

ACKNOWLEDGMENTS

[1] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Let.*, vol. 45, p. 712, 1980. 1

[2] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical Physics*, vol. 65, pp. 579–616, 1991. 1

[3] E. Tan, S. Algar, D. Corrêa, M. Small, T. Stemler, and D. Walker, "Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 33, no. 3, 2023. 2

[4] S. Mangiarotti and M. Huc, "Can the original equations of a dynamical system be retrieved from observational time series?," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 2, 2019. 2, 10

[5] M. Budišić, R. Mohr, and I. Mezić, "Applied koopmanism," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 4, 2012. 2

[6] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz, "Modern koopman theory for dynamical systems," *arXiv preprint arXiv:2102.12086*, 2021. 2

[7] S. Klus, I. Schuster, and K. Muandet, "Eigendecompositions of transfer operators in reproducing kernel hilbert spaces," *J. Nonlin. Sci.*, vol. 30, no. 1, pp. 283–315, 2020. 2

[8] P. J. Schmid, "Dynamic mode decomposition and its variants," *Annual Review of Fluid Mechanics*, vol. 54, pp. 225–254, 2022. 2

[9] D. Giannakis, "Data-driven spectral decomposition and forecasting of ergodic dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 47, no. 2, pp. 338–396, 2019. 2

[10] J. P. Crutchfield and B. McNamara, "Equations of motion from a data series," *Complex systems*, vol. 1, pp. 417–452, 1987. 2, 10

[11] P. J. Baddoo, B. Herrmann, B. J. McKeon, and S. L. Brunton, "Kernel learning for robust dynamic mode decomposition: linear and nonlinear disambiguation optimization," *Proceedings of the Royal Society A*, vol. 478, no. 2260, p. 20210830, 2022. 2

[12] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[13] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. Barbosa, "Next generation reservoir computing," *Nature communications*, vol. 12, no. 1, p. 5564, 2021. 2, 10

[14] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, "Approaching complexity by stochastic methods: From biological systems to turbulence," *Phys. Rep.*, vol. 506,

no. 5, pp. 87–162, 2008. 2

[15] Y. Wang, H. Fang, J. Jin, G. Ma, X. He, X. Dai, Z. Yue, C. Cheng, H.-T. Zhang, D. Pu, *et al.*, "Data-driven discovery of stochastic differential equations," *Engineering*, vol. 17, pp. 244–252, 2022. 2

[16] J. P. Crutchfield, "Between order and chaos," *Nature Physics*, vol. 8, no. January, pp. 17–24, 2012. 2, 3

[17] S. P. Loomis and J. P. Crutchfield, "Topology, convergence, and reconstruction of predictive states," *Physica D: Nonlinear Phenomena*, vol. 445, p. 133621, 2023. 2, 3, 4, 5, 17, 20

[18] N. Brodu and J. P. Crutchfield, "Discovering causal structure with reproducing-kernel hilbert space ε-machines," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 32, no. 2, 2022. 2, 3, 4, 5, 6, 7, 16, 20

[19] J. C. Sprott, *Chaos and Time-Series Analysis*. Oxford, United Kingdom: Oxford University Press, second ed., 2003. 2

[20] M. Cross and H. Greenside, *Pattern Formation and Dynamics in Nonequilibrium Systems*. Cambridge, United Kingdom: Cambridge University Press, 2009. 2

[21] M. C. Cross and P. C. Hohenberg, "Pattern formation outside of equilibrium," *Rev. Mod. Phys.*, vol. 65, no. 3, pp. 851–1112, 1993.

[22] A. Rupe and J. P. Crutchfield, "On principles of emergent organization," *Physics Reports*, vol. 1071, pp. 1–47, 2024. 2

[23] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *Journal of statistical physics*, vol. 104, pp. 817–879, 2001. 2, 3

[24] A. Jurgens and J. P. Crutchfield, "Shannon entropy rate of hidden Markov processes," *J. Statistical Physics*, vol. 183, 2020. arXiv.org:2008.12886. 3, 7

[25] A. Jurgens and J. P. Crutchfield, "Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes," *arxiv.org:2102.10487*, 2021. 3, 4

[26] S. Marzen and J. P. Crutchfield, "Structure and randomness of continuous-time discrete-event processes," *J. Stat. Physics*, vol. 169, no. 2, pp. 303–315, 2017. 3

[27] D. R. Upper, *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan. 4, 17

[28] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *J. Mach. Learn. Res.*, vol. 2, p. 67–93, mar 2002. 4

[29] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, aug 2010. 4, 19

[30] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert Space Embedding for Distributions," in *Algorithmic Learning Theory: 18th International Conference*, vol. 31, pp. 13–31, 2007. 5

[31] L. Song, K. Fukumizu, and A. Gretton, "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 98–111, 2013. 6

[32] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017. 5, 6

[33] S. Marzen and J. P. Crutchfield, "Statistical signatures of structural organization: The case of long memory in renewal processes," *Phys. Lett. A*, vol. 380, no. 17, pp. 1517–1525, 2016. 6

[34] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006. 6, 7

[35] T. Berry, D. Giannakis, and J. Harlim, "Nonparametric forecasting of low-dimensional dynamical systems," *Phys. Rev. E*, vol. 91, no. 3, p. 032915, 2015. 7

[36] S. P. Loomis and J. P. Crutchfield, "Exploring predictive states via Cantor embeddings and Wasserstein distance," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 32, p. 123115, 12 2022. 7

[37] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7432–7437, 2005. 8

[38] T. Berry and D. Giannakis, "Spectral exterior calculus," *Communications on Pure and Applied Mathematics*, vol. 73, no. 4, pp. 689–770, 2020. 8

[39] S. Klus, P. Koltai, and C. Schütte, "On the numerical approximation of the perron-frobenius and koopman operator," *Journal of Computational Dynamics*, vol. 3, no. 1, pp. 51–79, 2016. 12

[40] S. E. Marzen and J. P. Crutchfield, "Informational and causal architecture of discrete-time renewal processes," *Entropy*, vol. 17, no. 7, pp. 4891–4917, 2015. 13

[41] S. E. Marzen and J. P. Crutchfield, "Structure and randomness of continuous-time, discrete-event processes," *Journal of Statistical Physics*, vol. 169, no. 2, pp. 303–315, 2017. 13

[42] SILSO World Data Center, "The international sunspot number," *International Sunspot Number Monthly Bulletin and online catalogue*, 1749-2023. 13

[43] Y. Dang, Z. Chen, H. Li, and H. Shu, "A comparative study of non-deep learning, deep learning, and ensemble learning methods for sunspot number prediction," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2074129, 2022. 13

[44] T. M. Lutz, "The magnetic reversal record is not periodic," *Nature*, vol. 317, no. 6036, pp. 404–407, 1985. 14

[45] G. A. Glatzmaiers and P. H. Roberts, "A three-dimensional self-consistent computer simulation of a geomagnetic field reversal," *Nature*, vol. 377, no. 6546, pp. 203–209, 1995. 14

[46] "ICOS data for the Grignon site." `https://meta.icos-cp.eu/resources/stations/ES_FR-Gri`. Accessed: 2024-05-13. 15

[47] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet, "On the relation between universality, characteristic kernels and rkhs embedding of measures," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 773–780, JMLR Workshop and Conference Proceedings, 2010. 19

[48] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006. 20

[49] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schlimbach, M. Patwary, S. Maidanov, V. Lee, M. Prabhat, *et al.*, "Disco: Physics-based unsupervised discovery of coherent structures in spatiotempo-

ral systems," in *2019 IEEE/ACM Workshop on Machine* *Learning in High Performance Computing Environments (MLHPC)*, pp. 75–87, IEEE, 2019. 20

Appendix A: Code and data availability

The code for reproducing these examples is freely available under the MIT license from our project page: https://team.inria.fr/comcausa/continuous-causal-states/. Data for the examples is also provided together with the source code.

Appendix B: Metaparameters and kernel construction for sequences

There are three categories of metaparameters:

- The specifics of the site-wise comparison, including kernel choices for each data source;
- The method of kernel aggregation over time, including the temporal decay profile; and
- The method of kernel aggregation over data sources, including potential weighting of the influence of each data source.

We review these below.

1. Extended notations

Section III B introduced notations for a single sequence of observations and history lengths that do not depend on the data source. We recall and extend these notations to more general cases, matching what is actually handled by our algorithm:

- We assume that Z is constructed from measurements of D data sources: $z_t = \{m_t^1, m_t^2, \ldots, m_t^D\}$. Each of these has its own data type: vector of real values, string, graph, and the like.

- For each data source, indexed by $d = 1 \ldots D$, the pasts are sequences of length L_p^d and the futures are sequences of length L_f^d. Specifically, they are the finite length sequences:

$$x_t^d = \left(m_{t-L_p^d+1}^d, \ldots, m_{t-1}^d, m_t^d\right) \quad \text{and}$$
$$y_t^d = \left(m_{t+1}^d, m_{t+2}^d, \ldots, m_{t+L_f^d}^d\right)$$

- At each time, the pasts and futures are thus defined as lists of the above:

$$x_t = \left(x_t^1, \ldots, x_t^D\right) \quad \text{and}$$
$$y_t = \left(y_t^1, \ldots, y_t^D\right)$$

- The data is organized in K temporal blocks of consecutive measurements. In Section III B we considered only $K = 1$ block for simplicity, but the method can exploit measurements from multiple realizations of the same physical process. Each temporal block consists of T_k consecutive measurements $(z_t, \ldots, z_{t+T_k-1})$.

- With this data organization, have a library of $N = \sum_k \left(T_k - \max_D L_p^d - \max_D L_f^d + 1\right)$ pairs (x_i, y_i) of pasts $\boldsymbol{\mathcal{X}} = \{x_1, \ldots, x_N\}$ and futures $\boldsymbol{\mathcal{Y}} = \{y_1, \ldots, y_N\}$. Each of the $i = 1 \ldots N$ pairs corresponds to a time index in one of the temporal blocks.

- Some data may be missing or be tagged as low quality. These are replaced by not-a-number pseudo-values and possibly gap-filled with the technique described in Appendix C. Otherwise, each of these missing data reduces N by at most $2 \times \left(L_p^d + L_f^d\right) - 1$ values (consecutive missing data reduce by less than this amount).

The next subsections describe how to construct the two symmetric and positive definite reproducing kernels $k^X(x, \cdot) : \boldsymbol{\mathcal{X}} \to \mathcal{H}^X$ and $k^Y(y, \cdot) : \boldsymbol{\mathcal{Y}} \to \mathcal{H}^Y$, for pasts and futures respectively. These kernels generate the reproducing kernel Hilbert spaces $\mathcal{H}^X$ and $\mathcal{H}^Y$.

2. Sitewise comparison

As already noted, we do not constrain our measurements m^i to be anything other than drawn from a compact set. Typically, this is either a finite discrete set or a closed interval in the reals. This is because a *universal* reproducing kernel over a compact set has several desirable properties, namely that the measure embedding into the reproducing Hilbert space generated by the kernel given by Eq. (2) is injective and the norm convergence under the kernel inner product as defined in Eq. (3) is equivalent to convergence in distribution over measures [29].

A very large class of universal kernels are those which are both compactly supported and translation invariant in $\mathbb{R}^n$ [47]. This includes popular kernels such as the Gaussian and the Laplacian, that both depend only on the distance between their arguments. Typically, the distance is also scaled by a kernel width or radius that sets the characteristic scale. For instance, the Gaussian kernel is written:

$$k_G(m, m') = \exp\left(-\frac{1}{2}\left(\|m - m'\|/\xi\right)^2\right).$$

So, it is possible to define a kernel for a particular data source $k^d\left(m^d, m'^d\right)$ simply by defining a metric on that data source d and picking an appropriate bandwidth ξ, expressed in data units.

The choice of metric is fundamental to the empirical kernel causal states construction because it sets the underlying geometry that the kernel uses to embed the causal states, so it must be done consistently with the nature of the underlying system. However, we have considerable freedom in picking $\|m - m'\|$: it only needs to be a metric. Since metrics have been defined on all kinds of mathematical structures—strings, graphs, probability distributions, game theoretic strategies—our ability to build kernels is extraordinarily broad. In practice, a Euclidean metric is typically used when the underlying data is drawn from $\mathbb{R}^n$ and a discrete metric is used for symbolic data. In this paper examples, all the data sources are real numbers and we use the Euclidean metric together with the Gaussian kernel.

With these choices fixed, the most important parameter becomes the kernel bandwidth ξ, as it sets the scale for comparing two data values. For example, exploring the effect of temperature variations on the cultures at the scale of $0.1°C$ would make no sense for the crop example in Section V B. When a natural, characteristic scale exists from domain knowledge, it should be used to analyze the physical process operating at that scale. In a data exploration phase, it is also interesting to sweep through a range of bandwidths. This could highlight structure at multiple scales, possibly reflecting separate physical processes.

3. Kernel aggregation over time

Once kernels have been set to compare values for each data source, the next step is to combine them across time to get kernels over sequences.

Fortunately, kernels can be combined easily. Let $k^A\left(a, a'\right) : A \times A \to \mathbb{R}$ and $k^B\left(b, b'\right) : B \times B \to \mathbb{R}$ be reproducing kernels. Then the following operations result in a reproducing kernel k^C:

- Scalar multiplication by $\alpha > 0$: $k^C\left(a, a'\right) = \alpha k^A\left(a, a'\right)$, where $C = A$.

- Exponentiation by $\alpha > 0$: $k^C\left(a, a'\right) = \left(k^A\left(a, a'\right)\right)^{\alpha}$, where $C = A$.

- Kernel multiplication: $k^C\left(c, c'\right) = k^A\left(a, a'\right) k^B\left(b, b'\right)$, where $C = A \times B$.

- Kernel addition: $k^C\left(c, c'\right) = k^A\left(a, a'\right) + k^B\left(b, b'\right)$, where $C = A \times B$.

The above relations are not exhaustive (see [48, Section 6.2]) but they are enough to cover the most usual cases. They also preserve the translation-invariance property,

hence guarantee the universality of the combined kernels (with our compactness requirement).

Kernels over sequences can be then easily defined by combining the kernels across time, using schemes such as:

- Geometric weighting scheme:

$$k^{Y^d}\left(y_t^d, \cdot\right) = \prod_{\tau=1}^{L_f^d} k^d\left(m_{t+\tau}^d, \cdot\right)^{\omega(\tau)}$$

- Arithmetic weighting scheme:

$$k^{Y^d}\left(y_t^d, \cdot\right) = \sum_{\tau=1}^{L_f^d} \omega\left(\tau\right) k^d\left(m_{t+\tau}^d, \cdot\right)$$

where k^{Y^d} is the kernel over the future sequences Y^d of data source d. The kernel can be further normalized so that $k^{Y^d}\left(y_t^d, y_t^d\right) = 1$. Similar kernel definitions hold for the past sequences X^d of each data source.

In these schemes, we call $\omega\left(\tau\right)$ the decay profile. Typically, we desire a weighting scheme designed such that observations further away in time have less influence than the more immediate past/future observations. In practice, when $\|m^d - m'^d\|$ is the Euclidean metric, and when k^d is the Gaussian kernel, then using the product scheme becomes a natural choice:

$$k^{Y^d}\left(y_t^d, y_t'^d\right) = \prod_{\tau=1}^{L_f^d} \exp\left(-\frac{1}{2\xi^2}\|m_{t+\tau}^d - m_{t+\tau}'^d\|^2\right)^{\omega(\tau)}$$

$$= \exp\left(-\frac{1}{2\xi^2} \sum_{\tau=1}^{L_f^d} \omega\left(\tau\right) \|m_{t+\tau}^d - m_{t+\tau}'^d\|^2\right)$$

Then, the weighting scheme acts on the distance between measurements. It can be interpreted as scaling the bandwidth ξ by $1/\sqrt{\omega\left(\tau\right)}$ for each τ, hence reducing the sensitivity of the comparisons as ω decays.

Our previous work [18] introduced a power law decay. Exponential decay was successfully used in a related work [49], where it was noted to give superior performances compared to uniform weighting. An appealing avenue is to tie the weights to the autocorrelation of each data source. An information-theoretic based method was also proposed in [17]. Since the only constraint is that the aggregation of kernels be bounded, there is considerable flexibility in how one wishes to define $\omega\left(\tau\right)$. Mathematically, we must introduce some kind weighting scheme on the sums of kernels acting on infinite sequences that ensures their convergence [17]. A temporal decay fills this role naturally. Practically, however, we work with finite sequences and so a decay profile is not strictly mathematically necessary.

A common misconception is that without a decay the aggregated kernel is invariant under permutations of the sequence, and that would make it inappropriate for use

on time series. While the resulting kernel would indeed be invariant, this is the expected behavior and not a problem, even when working with time series. Choosing not to use a decay profile is, in effect, making the assumption that the system itself does not experience causality decay (or that the causality decay is not appreciable within the analyzed time window), in which case it is perfectly acceptable for the distance between x_1 and x_2 to be the same as the distance between their permuted versions x_1' and x_2'. This is not, in itself, a reason to select a causal decay profile which enhances preferentially certain permutations over others.

In the examples here, we only use the product scheme with uniform weighting—that is, without temporal decay. This choice was made to reduce the number of metaparameters to set, also to ease comparison between examples.

4. Kernel aggregation over data sources

Once the kernels k^{Y^d} are constructed for each data source, the procedure for combining them is similar to the aggregation over time. For example, using the arithmetic scheme:

$$k^Y\left(y_t, \cdot\right) = \sum_{d=1}^{D} w\left(d\right) k^{Y^d}\left(y_t^d, \cdot\right)$$

and similarly for the geometric scheme. The weighting profile $w\left(d\right)$ now sets the influence of each data source in the final, combined kernel. This may be useful for expressing domain knowledge, but otherwise uniform weighting should be considered. In principle, it would be possible to use a different weighting scheme for the kernel over futures k^Y and for the kernel over pasts k^X, which is built with the same procedure. Again, there is no motivation to do so without prior knowledge on the data.

Appendix C: Gap-filling missing data

When working with real measurements, there are sometimes sections with missing or invalid data. For our method, these missing measurements can be quite impactful: because of the way we construct libraries of past and future sequences, even a single missing (or invalid) measurement m_t^d in data source d at time t induces a gap length of $G = \times \left(L_p^d + L_f^d\right) - 1$. This because before the missing measurement, the last future sequence that can be computed is $y_{t-L_f^d-1}^d$ and after the gap, the first past history that can be computed is $x_{t+L_p^d}^d$. Consecutive missing values, however, only increase G by 1 for each additional missing value.

1. Linear interpolation of causal diffusion coordinates

Filling these gaps amounts to predicting the missing values. As a proof of concept, we train a linear transition operator T acting on the causal diffusion coordinates:

$$[\psi_{1,t+1}, \dots, \psi_{M,t+1}]^\intercal \triangleq T\left[\psi_{1,t}, \dots, \psi_{M,t}\right]^\intercal \qquad \text{(C1)}$$

T is typically fit by minimizing the least squared error for Eq. (C1) using all valid time transitions $\widehat{\sigma}_{x_t} \to \widehat{\sigma}_{x_{t+1}}$. For each data gap of size G, consider the time t for the last valid empirical kernel causal state $\widehat{\sigma}_t$ before this gap. The next available state is then $\widehat{\sigma}_{t+G+1}$. With a slight abuse of notation, remembering that T acts on the diffusion coordinates, we fill the gaps by applying T both forward:

$$\widehat{\sigma}_{t+g}^f = T^g\, \widehat{\sigma}_t$$

and backward:

$$\widehat{\sigma}_{t+g}^b = T^{g-G-1}\, \widehat{\sigma}_{t+G+1}$$

then combining both:

$$\widehat{\sigma}_{t+g}^p = \left(1 - \frac{g}{G+1}\right)\widehat{\sigma}_{t+g}^f + \frac{g}{G+1}\widehat{\sigma}_{t+g}^b\ . \qquad \text{(C2)}$$

This way, we get predicted states $\widehat{\sigma}_{t+g}^p$ that recover the valid states before ($g=0$) and after ($g=G+1$) the gap and interpolate between these within the gap ($1 \leq g \leq G$).

Once intermediate states are predicted, we still need to convert them into predictions in data space, for each data source. For this gap-filling proof of concept, we also fit these observation functions E^d as linear maps, using all valid measurements:

$$m_t^d \triangleq E^d\left[\psi_{1,t}, \dots, \psi_{M,t}\right]^\intercal\ .$$

In other words, each E^d attempts to recover the corresponding data source value from the diffusion coordinates of the empirical kernel causal state matching that time. We make predictions for all missing values by applying these E^d maps to the intermediate states $\widehat{\sigma}_{t+g}^p$ that were interpolated above.

The linear maps E^d and operator T do not reflect the correct ϵ-machine dynamic. A full treatment would consider the possible distribution of intermediate states, given the observed values before and after the gap, and produce an ensemble of intermediate trajectories, from which data samples could be drawn. Still, a linear interpolation and prediction model is sufficient our purposes here.

2. Using all valid measurements

It is common that only a subset of the data sources have missing values, such as the latent heat and the CO_2 flux

in the example in Section V B. The other data sources present valid measurements at the same time indices, which shall be exploited to refine the gap filling.

Consider the forward computations $\widehat{\sigma}_{t+1}^f = T\,\widehat{\sigma}_t$. After each application of T, it is possible to apply the observation functions E^d on the predicted state $\widehat{\sigma}_{t+1}^f$, yielding a vector $\left(\widehat{m}_{t+1}^1, \ldots, \widehat{m}_{t+1}^D\right)$ of D measurement estimates. For the valid data sources, the measurement estimates and the actual data should ideally match. Noting $V \subset \{1 \ldots D\}$ the set of valid data source indices at time $t+1$, we implement the following optimization scheme:

$$\tilde{\sigma}_{t+1}^f = \text{argmin}_s \sum_{d \in V} w\left(d\right)\left(E^d s - m_{t+1}^d\right)^2$$
$$\text{s.t. } \left\|s - \widehat{\sigma}_{t+1}^f\right\| < \epsilon \left\|\widehat{\sigma}_{t+1}^f\right\| , \tag{C3}$$

using the CVXPY iterative constrained convex optimizer with starting point σ_{t+1}^f. In other words, we seek a causal state estimate $\tilde{\sigma}_{t+1}^f$ maximally consistent with the valid data $\left\{m_{t+1}^d\right\}_{d \in V}$, but not too far from the forward state estimate $\widehat{\sigma}_{t+1}^f$ computed with the transition operator T.

This procedure is repeated iteratively G times in the forward direction: $\widehat{\sigma}_{t+g+1}^f = T\,\tilde{\sigma}_{t+g}^f$ is used as a starting point for optimizing $\tilde{\sigma}_{t+g+1}^f$ in each of these steps.

The same procedure is applied backward and the final predicted states $\tilde{\sigma}_{t+g}^p$, for all indices g in the gap, are combined using Eq. (C2). Finally, predicted values $\widehat{m}_{t+g}^d = E^d \tilde{\sigma}_{t+g}^p$ are set for each missing data source d at each time index $t+g$ in the gap.

3. Adjusting the causal diffusion components

By construction the predicted states for the gaps respect the constraint $\left\|s - \widehat{\sigma}_{t+g}^f\right\| < \epsilon \left\|\widehat{\sigma}_{t+g}^f\right\|$ of Eq. (C3) at each step of the iterative process, and similarly for the backward pass. Errors may accumulate but, at most, only proportionally to a fixed number of iterations G. The predicted state $\tilde{\sigma}_{t+g}^p$ thus remains arbitrarily close to the

$\widehat{\sigma}_{t+g}^p$ of Eq. (C2). These were computed with the linear transition operator T, itself fit by a linear regression. Hence, the rank of the original $N \times N$ similarity matrix $G_{ij}^{\mathcal{S}_k} = \left\langle \widehat{\sigma}_{x_i}, \widehat{\sigma}_{x_j} \right\rangle$ should not be changed by the addition of the predicted states. The updated $\widetilde{G^{\mathcal{S}_k}}$ is now a similarity matrix of size $(N+G) \times (N+G)$ between all states, original and produced to fill the gap.

We take the conservative approach to first use only the gap-filled states for the $\max_D L_p^d + L_f^d - 1$ states before and after the missing values. Using only the backward or the forward pass, we also fill the sections with partial histories at the beginning and the end of each of the K temporal blocks. In other words, the optimization in Eq. (C3) was done only for time indices that match actual data, with no missing value, but for which an empirical kernel causal state could not be estimated due to lack of full histories. After this first step, the gaps consist only of the time indices with actually missing data. We recompute the diffusion transform on the updated $\widetilde{G^{\mathcal{S}_k}}$. The whole gap-filling procedure is then repeated in this updated causal diffusion components space, for filling these much smaller gaps. As a result, we get state estimates for every of the original data time indices, using both the original $\widehat{\sigma}$ embeddings and the predicted $\tilde{\sigma}^p$ for partial histories and missing data.

Figure 11 shows the data for the CO_2 flux over the whole series, together with zooms on gap-filled sections. The sections labeled "invalid data" (in red) actually correspond to a quality control flag of 0, meaning these data are not reliable at all and should be discarded. Hence, the invalid data in Fig. 11 should not be seen as a target for the gap-filled data (in green). The valid (blue) data could also be of low quality at each gap boundary. What is relevant is that gap-filled data in Fig. 11 look plausible and could have been measured instead of the invalid red values. The gap-filled data indeed look like they could have been measured, even for the first few months of 2004 for which the invalid data is clearly nonsensical.

In this proof of concept for gap-filling, the transition operator T and the observation functions E^d are linear, but they could be replaced by full-fledged estimators of the kernel ϵ-machine dynamics in future works. Compared to alternative gap-filling techniques, this method presents the advantage of exploiting all available information coming from all data sources at all time indices, and consistently so with the overall system dynamics.
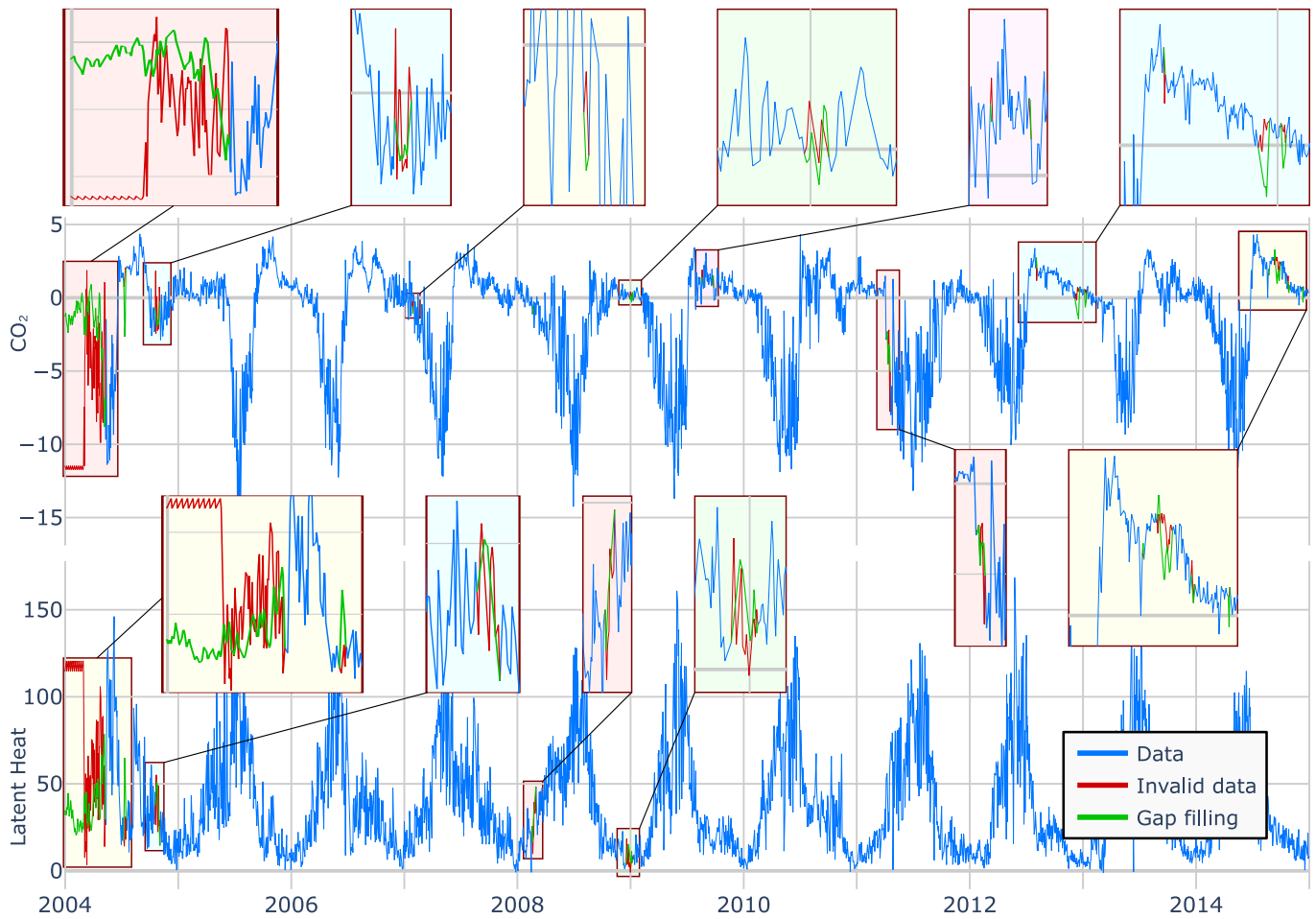
FIG. 11. CO_2 and Latent Heat flux data from the crop example in Section V B. Data in red is marked as invalid and should not be seen as a target for the gap-filled data (in green). This is especially apparent for the first few months of 2004. Data in blue could also be of low quality, especially at each gap boundary. The gap-filled sections look plausible and are consistent with the measurements of all other variables.