Nonequilibrium Statistical Mechanics and Optimal Prediction of Partially-Observed Complex Systems

Adam Rupe,[1, 2, *] Velimir V. Vesselinov,[2, †] and James P. Crutchfield[3, ‡]

[1] *Center For Nonlinear Studies, Theory Division, Los Alamos National Laboratory*

[2] *Computational Earth Science, Earth and Environmental Sciences Division, Los Alamos National Laboratory*

[3] *Complexity Sciences Center, Department of Physics and Astronomy, University of California Davis*

(Dated: September 8, 2022)

Only a subset of degrees of freedom are typically accessible or measurable in real-world systems. As a consequence, the proper setting for empirical modeling is that of partially-observed systems. Notably, data-driven models consistently outperform physics-based models for systems with few observable degrees of freedom; e.g., hydrological systems. Here, we provide an operator-theoretic explanation for this empirical success. To predict a partially-observed system's future behavior with physics-based models, the missing degrees of freedom must be explicitly accounted for using data assimilation and model parametrization. Data-driven models, in contrast, employ delay-coordinate embeddings and their evolution under the Koopman operator to implicitly model the effects of the missing degrees of freedom. We describe in detail the statistical physics of partial observations underlying data-driven models using novel Maximum Entropy and Maximum Caliber measures. The resulting nonequilibrium Wiener projections applied to the Mori-Zwanzig formalism reveal how data-driven models may converge to the true dynamics of the observable degrees of freedom. Additionally, this framework shows how data-driven models infer the effects of unobserved degrees of freedom implicitly, in much the same way that physics models infer the effects explicitly. This provides a unified implicit-explicit modeling framework for predicting partially-observed systems, with hybrid physics-informed machine learning methods combining implicit and explicit aspects.

I. INTRODUCTION

Most Earth Science investigations access only a subset of a high-dimensional dynamical system's degrees of freedom due to limited instrumentation. Predicting the future behavior of partially-observed systems is a central challenge for many areas of Earth Science, and one that dates back to the earliest uses of scientific computing [1, 2].

Traditional prediction employing *physics-based* (or *process-based*) *models* relies on *explicit representations*: systems are modeled via closed-form equations of motion that determine how a system evolves forward in time through interactions among all its degrees of freedom. Predictions are extracted from numerical approximations of solutions of the equations of motion. This requires knowing the full state of the system at each time, but limited instrument measurements of the true system provide only a partial view of the underlying state. *Data assimilation* is then used to generate a *data-image* through model inversion. The result is a coarse-grained approximation of the full system state that is most consistent with the instrument observations and assumptions of the underlying physics.

In contrast, *data-driven prediction* (typically) does not rely on explicit closed-form models and thus does not require interpolated data-images. For the prediction task that evolves only instrument measurements forward in time, data-driven models learn *implicit representations* for this evolution directly from the observations themselves.

The explicit nature of physics models hinges on our understanding of the underlying physics governing the system being encapsulated in closed-form differential equations-of-motion. This is what explicit representations attempt to approximate. The equivalent governing physics—the "ground truth"—for the evolution of the measurement observables is given by linear, infinite-dimensional Koopman operators. The implicit representations of data-driven models thus attempt to learn projections of the Koopman operators' action [3].

In fact, the governing equations of motion for the measurement observables are given by the Mori-Zwanzig equation [4, 5], derived from expanding the action of the Koopman operator in terms of projection operators onto the observable degrees of freedom [6, 7]. A key insight from the Mori-Zwanzig formalism is that predictive models of partially-observed systems require a history dependence—past observations of the observable degrees of freedom generally contain information relevant for future predictions.

Recently, the connection between the history dependence of predictive models and the intrinsic geometry of delay-coordinate embeddings [8, 9] has been explored [10–13]. Past values of partial observations, in the form of delay embeddings, implicitly stand in for the missing degrees of freedom. This parallels how, for physics-based models, data images act explicitly to fill in the gaps of the missing degrees of freedom when predicting partially-

[*] adamrupe@lanl.gov

[†] vvv@lanl.gov

[‡] chaos@ucdavis.edu

observed systems.

Most data-driven modeling and prediction relies on Hilbert space methods that learn a *target function* living in a Hilbert space of functions [14]. Optimal Hilbert space models take the form of a *conditional expectation* of future observations given past observations. This optimum is equivalent to a nonlinear projection of the action of the Koopman operator (that gives the future value of the measurement observables) onto the Hilbert subspace of functions of only the observable degrees of freedom. History-dependent target function models can be expressed as functions of delay-coordinate embeddings using Wiener projections [6]. The optimal model is then the nonlinear Wiener projection of the action of the Koopman operator onto functions of observed delay embeddings.

There is evidence that Wiener projection models may converge to the true dynamics of the measurement observables if sufficient past observations are taken into account [15]. Here, we provide a new perspective on the behavior of history-dependent data-driven models and their relation to the true underlying physics of partial observations. We do so using insights from the logical inference approach to statistical mechanics given by Jaynes' Maximum Entropy principle [16]. This further builds on the connections between nonequilibrium statistical mechanics and optimal prediction of partially-observed systems [17].

Optimal Hilbert space models are typically formulated in terms of an invariant "equilibrium" measure. However, we show there is a natural family of time-dependent "nonequilibrium" measures induced by partial observations using Maximum Entropy and its time-varying generalization Maximum Caliber [18, 19]. Constructively, these measures support more general nonasymptotic behaviors—behaviors that cannot be modeled with an invariant measure. Importantly, though, they provide unique insights into the convergence of optimal models to the true governing physics of partial observations. They do this by directly constructing *predictive distributions*—probabilities over future observations given past observations. In particular, we express the possible convergence of history-dependent models as a thermodynamic limit in which the variance of predictive distributions vanishes as the length of past observations increases. This again shows how the action of Koopman operators on delay embeddings implicitly account for the effects of unobserved degrees of freedom.

Formulating optimal data-driven models as expectations of predictive distributions suggests a more general stochastic framework for modeling partially-observed systems. Rather than returning the expectation of predictive distributions, optimal stochastic models simply return the predictive distributions themselves [20], which then may be sampled for ensemble forecasts. Our direct construction of predictive distributions using Maximum Caliber measures leads naturally to such optimal stochastic models for partially-observed systems. A sequel gives this stochastic formulation of optimal prediction of partially-observed systems.

A. Implicit versus explicit representations

The physical insights that emerge shed light on why data-driven models can outperform traditional physics models for predicting systems with relatively few observed degrees of freedom. Indeed, this has become increasingly common for hydrological systems [21–24]. In these cases, the implicit approach that uses delay-coordinate embeddings is more effective than the explicit approach that uses data assimilation. For example, while many details, e.g., subsurface morphology, are crucial for geophysical prediction, given limited available subsurface measurements, reconstructing informative data-images for them is exceedingly difficult. This leads to less effective physics-based methods that rely on the latter.

Perhaps unsurprisingly in this light, due to their empirical successes in scientific applications, data-driven predictive models are increasingly employed. That said, they are widely considered to be an entirely new paradigm—a paradigm with little to no relation with governing physics and physics-based models. We aim to show that they are in fact quite similar.

Our framework, together with numerical examples, shows that data-driven models do implicitly what physics-based models do explicitly to account for unobserved degrees of freedom. We also clarify how the action Koopman and Perron-Frobenius operators on delay-coordinate embeddings may converge to the true system dynamics on the full system state. Generating partitions on maps of the unit interval are discussed as a rigorous example displaying this behavior. Said another way, the physics underlying history-dependent data-driven models is the same as the physics underlying traditional physics-based models.

The resulting unified modeling framework shows that the distinction is not so much "data-driven versus physics-based", but rather the emphasis should be on where approaches land in the "implicit versus explicit" representation spectrum. The class of *physics-informed machine learning* models [25–27], now rapidly gaining popularity, are thus seen to lie between fully-explicit physics-based models and fully-implicit data-driven models. Such hybrid models explicitly enforce certain physical properties as *inductive biases* [28, 29], with any remaining properties learned implicitly from the data.

B. Synopsis

Our development unfolds as follows. Section II introduces Platonic models as the true dynamics of a given physical system. This is what models attempt to predict. Next, Section III formalizes partial observations and the resulting stochastic processes over the observable degrees of freedom, which we call dynamical processes. These are the main objects of study. To set the stage for the development of implicit data-driven models, Section IV first reviews the explicit physics-based modeling approach. Next, Section V gives the physics of partial observations expressed in terms of Koopman and Perron-Frobenius operators. This section also discusses connections to statistical mechanics and introduces Maximum Entropy measures.

Section VI overviews implicit data-driven models and their Hilbert space formulation for the case of instantaneous prediction. Section VII details the Mori-Zwanzig formalism, motivating history-dependent models. Section VIII discusses histories of past observations in the form of delay-coordinate embeddings. Section IX then expresses the Mori-Zwanzig formalism in terms of delay embeddings using Wiener projections. This provides the formulation of history-dependent Hilbert space models, using both the equilibrium invariant measure and nonequilibrium Maximum Caliber measures. In the nonequilibrium case, the Maximum Caliber measures allow for the direct construction of predictive distributions, providing insights into the convergence behavior of optimal history-dependent models. Section X provides examples demonstrating the ability of data-driven models to implicitly learn the effects of the unobserved degrees of freedom. Finally, Section XI uses the prior development to formally connect implicit data-driven models with explicit physics-based models. This shows the underlying similarity between the two approaches and offers a unified implicit-explicit modeling framework.

II. SYSTEMS AND PLATONIC MODELS

After centuries of intellectual inquiry, physical scientists collectively have come to believe in having a solid grasp of the basic physics governing measurable phenomena. For example, many Earth Science systems are governed by classical field theories. Atmospheric circulation, shown in Fig. 1, is governed by the laws of fluid mechanics and thermodynamics [30].

Saying that one "understands" these system's basic physics means, more specifically, that the governing principles are encapsulated in the form of explicit differential equations-of-motion [31]. Formally, the system state ω evolves according to:

$$\begin{aligned}\dot{\omega} &= \frac{d\omega}{dt} \\ &= \Phi(\omega) ~,\end{aligned}$$

where the governing equations Φ are a function of ω. For spatially-extended field theories, ω itself is a function of spatial coordinates, too. Φ then typically includes finitely-many spatial derivatives of ω, signifying the state dynamics are governed by local interactions.

The "unreasonable effectiveness of mathematics" in physics has been repeatedly noted since Ref. [32] highlighted the puzzle. Noting that governing equations Φ are almost always given in *closed form* the effectiveness is all the more intriguing. Our development further highlights that demanding physical systems always be expressed in closed form rather restricts the class of mathematical models used to describe the physical world.

Here, we represent a given physical system as a differential dynamical system (Ω, Φ) that, for a shorthand, we call the *Platonic model*. A system's true dynamics, given by the Platonic model, may be well approximated with closed-form equations of motion. The Navier-Stokes partial differential equations come to mind as an approximation to the Platonic model of fluid flow. However, we need not assume a particular functional form for Platonic models.

That said, there are three important properties we do assume for Platonic models. Note that we are primarily concerned here with phenomena that occur at classical energy scales, such as found in Earth Systems. The first property is that system states evolve continuously—they are continuous trajectories in the state space over time.

The next two properties define what the system *state* $\omega \in \Omega$ actually is. The second property assumes Platonic models are *Markovian*: Determining a later state $\omega_t = \Phi^t \omega_0$ only requires knowing the state at a single prior time ω_0. The third property assumes Platonic models are *deterministic*: The same initial condition ω_0 always produces the same later state $\omega_t = \Phi^t \omega_0$.

The latter two properties impose a *closure relationship* among the *degrees of freedom* constituting the system state ω. That is, ω is considered a vector with each component ω^i being a degree of freedom. The dynamic $\Phi(\omega)$ captures the physically-relevant interactions among the degrees of freedom by determining how they evolve forward in time. The system's governing physics is appropriately captured or modeled when, with sufficiently-many degrees of freedom comprising ω, there is a closure in their dynamics: For every ω^i, its time evolution is a deterministic and Markovian function of a subset of the other $\{\omega^i\}$, i.e., the system state ω.

As there are many parallels to statistical mechanics, note that there is an important property we are *not* assuming of (Ω, Φ)—that the system is Hamiltonian. In the partially-observed setting, introduced shortly, the Platonic model (Ω, Φ) is analogous to a "microsystem". Statistical mechanics would take it to be Hamiltonian. This is too restrictive for our purposes. Importantly, Hamiltonian systems are conservative and volume-preserving, via Liouville's theorem [4]. Volume-preserving dynamics admit a natural invariant probability distribution over Ω, known as the *microcanonical ensemble* in statistical mechanics [4]. While such invariant probability measures are convenient mathematically, many physical systems of interest display transient nonasymptotic behavior that cannot be captured by invariant measures. This is particularly notable for fluid flows.

To accommodate nonasymptotic behaviors within our formalism, we do not assume Platonic models are necessarily volume-preserving, although they may be. More generally, while it is standard to assume the dynamics is *measure-preserving* such that there is a probability measure over Ω which is invariant under Φ, our formalism does not require an invariant measure. Rather, one of our main contributions is introducing natural time-dependent measures for partially-observed systems that can support nonasymptotic behaviors. In the language of statistical mechanics, our approach is a nonequilibrium formalism that generalizes the equilibrium setting using asymptotic invariant measures. For more details on ergodicity, invariant measures, and dissipative systems, see Appendix A.

Additionally, in what follows, we assume a system's dynamic is reversible, so that:

$$(\Phi^t)^{-1} = \Phi^{-t} \ .$$

This, however, is an assumption for notional convenience and simplicity. It can be lifted without much difficulty. An added advantage of our time-dependent formulation is that we need not assume reversible dynamics. Note though that many systems of interest are reversible in this way, such as all finite-dimensional systems of ordinary differential equations.

III. PARTIAL OBSERVATIONS AND DYNAMICAL PROCESSES

The semigroup formalism of dynamical systems [33, Ch. 7] is particularly apt for our development. Consider a dynamical system $(\Omega, \Sigma_\Omega, \nu, \Phi)$. The state space Ω is a Euclidean space or manifold for finite-dimensional systems or a general Hilbert space for spatially-extended systems. Σ_Ω is the Borel σ-algebra and ν the Lebesgue reference measure that gives a "volume" to state space.

Φ is the dynamic—the infinitesimal generator of a continuous semigroup of measurable flow maps $\{\Phi^t : \Omega \to \Omega\}_{t\in\mathbb{R}}$, with:

$$\begin{aligned}\Phi(\omega) &= \lim_{\tau\to 0} \frac{1}{\tau}\big(\Phi^{t+\tau}(\omega) - \Phi^t(\omega)\big) \\ &= \frac{d}{dt}\Phi^t(\omega)|_{t=0} \ .\end{aligned}$$

Thus, the orbits $\{\omega_t = \Phi^t(\omega_0) : t \in \mathbb{R}_{(\geq 0)}\}$ are continuous functions of time t. When the dynamic is specified by a system of differential equations, Φ is the time derivative of the orbits:

$$\begin{aligned}\dot{\omega} &= \frac{d}{dt}\omega \\ &= \Phi(\omega) \ .\end{aligned}$$

For a given dynamical system under study, let $x \in \mathcal{X}$ be the subset of system variables that are observable, measurable, or generally accessible. Through experimental or observational measurements or numerical simulations, they may be collected in a *time series* $\{x_0, x_1, \ldots, x_{T-1}\}$—a time-ordered set of observations of x taken at uniform time intervals $\{t_0, t_1, \ldots, t_{T-1}\}$ with $t_i = (i-1)\Delta t$. The observations x are generated by the dynamical system under the continuous and measurable mapping $X : \Omega \to \mathcal{X}$ so that $x_t = X(\omega_t)$. In practice, the measurement observables are given as a vector of real numbers, so that $\mathcal{X} = \mathbb{R}^n$.

We are interested in the case of a *partially-observed* dynamical system for which the map X is many-to-one and not invertible. Due to this, an observation x_t is insufficient for determining the full state ω_t of the underlying dynamical system at any given time. That is, there are unobservable, unmeasurable, or inaccessible degrees of freedom in ω. And so, measurement data can only ever provide a limited view of the system's true state ω. An important example is weather prediction, shown in Fig. 1.

We refer to collections of arbitrarily-long time series of observables $\{\ldots, x_{-1}, x_0, x_1, \ldots\}$ as a *dynamical process*, signifying that it is a stochastic process derived from a deterministic dynamical system through partial observations. They are the objects we wish to model. If the underlying system is governed by noninvertible dynamics we consider the time index of a dynamical process to correspond to *observation time*. That is, x_0 is not an initial condition, but rather the present moment of observation. The leading dots then indicate that we allow measurements from arbitrarily far in the past.

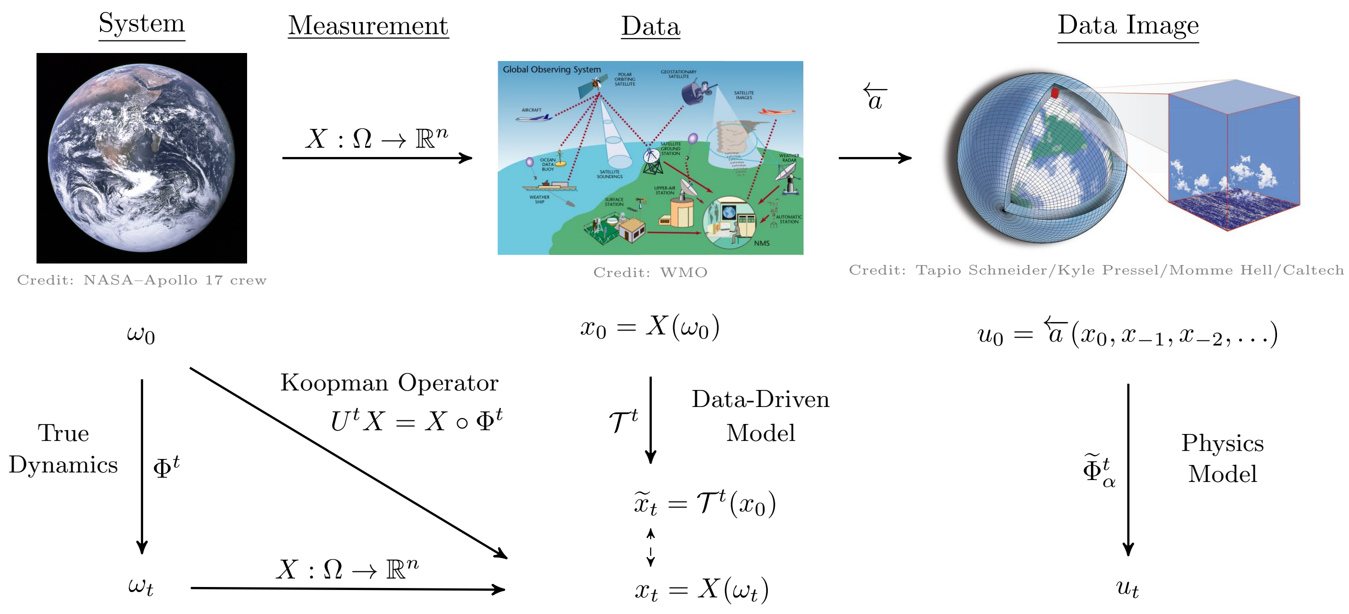
Various properties of dynamical processes will be given shortly, using the Koopman and Perron-Frobenius operators. First though, we detail the standard approach for modeling partially-observed systems using physics-based models.

IV. EXPLICIT PREDICTIVE MODELS

Given a physical system's Platonic model—its governing physics—and partial observations from instrument measurements, how do we predict the system's future behavior? Our main interest is to explain the effectiveness of *implicit* approaches learned by data-driven models. To set the stage, though, we first overview the more familiar *explicit* approach using physics-based models. Figure 1 shows the relation between data-driven and physics-based methods for modeling systems from partial observations. After formulating the physics underlying implicit data-driven models, a formal connection with explicit physics models is given in Section XI.

A. Physics-Based Models

In essence, physics-based models simply attempt to solve the governing equations that constitute or well-



Credit: NASA–Apollo 17 crew

Credit: WMO

Credit: Tapio Schneider/Kyle Pressel/Momme Hell/Caltech

FIG. 1. Predicting complex systems (left) from partial observations: Instantaneous data-driven modeling (middle) versus physics-based models (right). Instrument measurements provide a partial view x of the true system state ω through a noninvertible mapping X. From an initial measurement observation $x_0 = X(\omega_0)$, the value of the instruments at later time t is found by letting the actual system evolve, given by the dynamic Φ^t, and taking a measurement $x_t = X(\omega_t)$ at this time. Koopman operators U^t, provide an alternate point of view by providing a *future measurement* function $x_t = U^t X(\omega_0)$ that gives the instrument readings x_t at time t from the current system state ω_0. It provides the ground-truth that data-driven models $\mathcal{T}^t$ try to approximate as functions from current observation x_0 to future observation x_t. In contrast, physics-based models create a coarse-grained approximation u_0 of the full system state ω_0 most consistent with past observations using data assimilation and model inversion as $u_0 = \overleftarrow{a}(x_0, x_{-1}, x_{-2}, \ldots)$. Data images are then evolved through numerical approximation $u_t = \widetilde{\Phi}^t(u_0)$ of the system dynamics $\omega_t = \Phi^t(\omega_0)$.

approximate the Platonic model Φ. There are three main challenges when predicting a physical system using differential equation models: nonlinearity, high-dimensionality, and calibration.

First, most systems of interest are governed by nonlinear equations that cannot be solved analytically. Thus, numerical approximations to solutions are necessary. For complicated systems like those encountered in Earth Sciences, this introduces a second challenge.

Even with the arrival of massive high performance computing, today's largest machines still do not have the computational resources required to fully account for all known physical effects in a system at the necessary scales. This second challenge is certainly the case for numerical models of the atmosphere, as depicted in Fig. 1's right column. The effects that are not directly computed are accounted for using *parametrization* schemes to replace processes that are too small-scale or complicated to be directly computed in the model. This simplifying procedure produces a deterministic, Markovian closure.

While parametrization schemes are often heuristic choices, increasingly they are being informed by separate models specifically targeting the effects being parameterized. This includes, for example, using cloud-resolving models to inform cloud formation parametrizations in large-scale atmospheric models. Moreover, for spatially-extended field theories, continuous spatial coordinates must be discretized into a finite grid or mesh and then the effects of subgrid-scale processes must be parameterized using approximate closure models. However they are arrived at, the parametrizations represent a modeler's choices, and these choices necessarily induce *conceptual error* that affects the model's predictive capabilities. Poor choice of generic model may also lead to conceptual error in prediction.

Assume, for a given physical system, that we know effective differential equations $\Phi(\omega)$ that govern the system. A *generic model* of $\Phi(\omega)$ is an auxiliary set of equations $\widetilde{\Phi}^\tau_\alpha : u_t \mapsto u_{t+\tau}$ whose solutions $\{u_t\}$ can be solved numerically and that approximate the solutions of Φ. The generic model typically contains a set of parameters α that include those associated with parametrization schemes as well as physical parameters of the model, such as viscosity in the Navier-Stokes equations. The generic model $\widetilde{\Phi}^\tau_\alpha$ acts on *data images* u_t that are coarse-grained approximations of the state ω_t of the physical system. Neglecting numerical round-off error, numerical models are also Markov and deterministic, like the differential equation models they approximate.

B. Data Assimilation

The final challenge in using a generic physics model to predict the future behavior of a partially-observed physical system comes during *calibration*. The generic model must be made into a *specific model* that appropriately captures the particular circumstances of the physical system of interest. This includes specification of the parameters α and boundary conditions, as well as *initialization* of the model. (For simplicity, we include specification of the boundary conditions in α.) The Markov property allows for generating an orbit u_t of the specific model from a single initial state u_0. For this orbit to provide a prediction of the true system's orbit $\{\omega_t\}$ requires aligning the model's initial state u_0 as well as possible to the true physical system's initial state ω_0.

To emphasize the difficulty of initialization in particular, consider the commonly-encountered case of predicting a spatially-extended system using approximated solutions of a classical field theory—i.e., Φ is a set of partial differential equations. It is not possible to determine the system's configuration over a continuum of spatial coordinates. Rather, as depicted in the top of Fig. 1's middle column, measurements derive from a variety of instruments collecting data over a relatively small subset of the spatial domain. However, solving the model equations—say, using a finite element method—requires an initial condition on a grid over the spatial domain. Our instruments, though, do not necessarily provide full coverage over the grid. Thus, the calibration methods produce a data image u_0 (top right of Fig. 1) that represents inferred values over the full grid.

Model calibration, including parameter and boundary condition specification, as well as initialization of the data image u_0, are carried out using *model inversion* and *data assimilation* [34, 35]. Since these techniques require multiple past observations, calibration is sometimes also referred to as *history matching*. Given a history of past observations $\overleftarrow{x}_t^k := \{x_t, x_{t-1}, \ldots, x_{t-k}\}$, calibration attempts to find the initial data image $u_t = \overleftarrow{a}(\overleftarrow{x}_t^k)$ and parameter set α such that the model output $\{u_t, \widetilde{\Phi}_\alpha^{-1}(u_t), \ldots, \widetilde{\Phi}_\alpha^{-k}(u_t)\}$ is as consistent with the past observations $\overleftarrow{x}_t^k$ as possible. (Recall that we are assuming reversible dynamics for notational simplicity, but this is not strictly required.)

Due to the many-to-one nature of the partial observation map X, the calibration process is typically not unique. Multiple parameter sets and initial data images may produce orbits of data images that are equally consistent with the observations up to time t. Therefore, there will be multiple specific models that are equally consistent with past observations, but make different predictions for future behaviors. Therefore, a specific model used for prediction generally has *calibration error*. This combined with conceptual error leads to the model's overall prediction error. Prediction error can accumulate rapidly, particularly for deterministic chaotic systems whose inherent instabilities exponentially amplify small variations. This is one reason why weather is so hard to predict.

We stress here that model parametrizations and initialization $u_0 = a(x_0)$ are both means of *explicitly* accounting for unobserved or unrepresented degrees of freedom not in $x = X(\omega)$. For instance, in atmospheric circulation, imagine we do not have instruments on a remote island in the Pacific. As a consequence, atmospheric variables—temperature, pressure, wind speed, and the like—at that spatial location are not in $x = X(\omega)$. However, when a data image u is created these variables *are* approximated at that location.

Note also that the primary concerns are the predictive capability of physics models and how it relates to the Platonic model's true dynamics. In a sense, though, we are agnostic as to whether a specific model is *valid* or not [36, 37]. Loosely speaking, validity measures how well a physical system's specific model approximates its Platonic model. The specific model's predictive skill is, of course, related to how well it approximates the Platonic model. And, this is a question we care about here.

That said, there is a deeper concern about how well a specific model approximates the Platonic model. This involves the question of how much we can infer about the underlying physical and causal processes governing the true system, given a specific model of that system and its predictive capability. That is, how well can we explicitly formulate a Platonic model given a skillful specific model? It is in this deeper mechanistic sense that we are agnostic to the question of model validity. As the saying goes, "All models are wrong, but some are useful" [38].

V. PHYSICS OF PARTIAL OBSERVATIONS

Platonic equation-of-motion models are given in terms of the underlying system state—the full set of degrees of freedom. Due to their explicit nature, the connection between physics models and the system's true governing physics described by Platonic models is clear. Data-driven models of partially-observed systems do not generally attempt to explicitly infer the full Platonic system state, as physics-based models do. And so, it is less clear how they relate to the physics of Platonic models. To discuss Platonic models and the true governing physics in a meaningful way for partially-observed system requires the operator-theoretic formulation of dynamical systems [3, 33]. Both Koopman and Perron-Frobenius operators, defined shortly, provide alternative descriptions of a system's temporal evolution: Koopman operators give the evolution of observables, while Perron-Frobenius operators evolve state distributions. These *evolution operators* are the classical analogs of the Heisenberg and Schrödinger formulations of quantum mechanics, respectively.

A. Koopman Operators

A *Koopman operator* U^t acts on functions of the system state, known as *observables* $f : \Omega \to \mathbb{R}$, where f is an element of a function space $\mathcal{F}$. The action of $U^t : \mathcal{F} \to \mathcal{F}$ on observable f is given by composition with the dynamic Φ^t, also known as the *pullback* of f along Φ^t:

$$U^t f = f \circ \Phi^t, \tag{1}$$

$$\begin{aligned}[U^t f](\omega) &= f_t(\omega) \\ &:= f\big(\Phi^t(\omega)\big) \ .\end{aligned} \tag{2}$$

That is, U^t's action on observable $f \in \mathcal{F}$ gives the time-shifted observable $f_t = U^t f$ whose value at state ω is obtained by evaluating f at the future state $\omega_t = \Phi^t(\omega)$. Recall that the flow maps $\{\Phi^t\}$ form a semigroup in that $\Phi^{t+s} = \Phi^t \circ \Phi^s$. The set $\{U^t\}$ inherits this semigroup structure, so that $U^t \circ U^{\Delta t} = U^{t+\Delta t}$.

Each U^t is a linear infinite-dimensional operator when $\mathcal{F}$ is a vector space. As discussed more below in relation to Perron-Frobenius operators, it is most natural to take $\mathcal{F} = L^\infty(\Omega, \nu)$—the almost-everywhere bounded functions of ω—but the square-integrable functions $L^2(\Omega, \nu)$ are often used for mathematical convenience. The following uses Koopman operators on $L^2(\Omega, \nu)$ since it is a Hilbert space and the development requires orthogonal projections.

Recall that we are interested in observable functions X that are generally multidimensional. With this, an observable function f is a component of a vector-valued observable function; e.g., $f = X_i$. A Koopman operator that acts on an observable X is then the product over the component operators acting on X_i. To avoid excessive notation, we denote these product operators as U^t.

For a dynamical system with initial condition ω_0 at time t_0, the measurement observable at a later time $t > t_0$ is given by:

$$\begin{aligned}x_t &= X(\omega_t) \\ &= X\big(\Phi^t(\omega_0)\big) \\ &= X_t(\omega_0) \\ &= [U^t X](\omega_0) \ .\end{aligned}$$

The dynamical process, therefore, is a function of the underlying system's (unknown) initial condition:

$$\begin{aligned}&\{\ldots, x_{-1}, x_0, x_1, \ldots\} = \\ &\qquad \{\ldots, U^{-1}X(\omega_0), U^0 X(\omega_0), U^1 X(\omega_0), \ldots\} \ .\end{aligned} \tag{3}$$

This transparently relates the evolution of partial measurement observations of the dynamical process $\{\ldots, x_{-1}, x_0, x_1, \ldots\}$ to the physics of the Platonic model $\Phi^t(\omega_0)$ through the action of Koopman operators on the observable map X.

B. Perron-Frobenius Operators

Koopman operators connect dynamical processes to the Platonic model (Ω, Φ) via an unknown initial Platonic state ω_0. If we do not seek to directly infer ω_0, as done with physics models, it becomes useful to formulate the problem in terms of distributions over possible ω_0. The dynamics of these distributions is provided by *Perron-Frobenius operators.*

In appropriately defined spaces, Perron-Frobenius operators are dual to Koopman operators. It is most common to consider Perron-Frobenius operators acting on $L^1(\Omega, \nu)$ densities and, thus, their Koopman duals evolve observables in $L^\infty(\Omega, \nu)$. However, as often done, the following considers both operators acting on $L^2(\Omega, \nu)$ functions. In this case, Perron-Frobenius operators act on L^2 measures. If the L^2 measure ν_ρ is absolutely continuous with respect to the reference measure ν, ν_ρ is related to the density ρ through the reference measure:

$$\nu_\rho(B) = \int_B \rho d\nu \ ,$$

for density $\rho \in L^1(\Omega, \nu)$ and $B \in \Sigma_\Omega$.

For continuous-time dynamical systems there is a continuous semigroup $\{P^t\}$ of Perron-Frobenius operators that evolve measures μ through the *pushforward* of μ along Φ^t:

$$\begin{aligned}\mu_t &= P^t \mu \\ &:= \mu \circ \Phi^{-t} \ .\end{aligned} \tag{4}$$

The measure μ_t defines the probability space $(\Omega, \Sigma_\Omega, \mu_t)$ that quantifies uncertainty in system state ω_t at time t. In turn, this casts observables, given by the measurable map $X : \Omega \to \mathcal{X}$, as random variables X_t distributed according to the pushforward measure:

$$\mu^X_t(B_\mathcal{X}) = \mu_t\left(X^{-1}(B_\mathcal{X})\right) \ ,$$

for $B_\mathcal{X} \in \Sigma_\mathcal{X}$. Thus, we can write X_t's distribution in terms of the initial measure μ_0:

$$\begin{aligned}\Pr(X_t \in B_\mathcal{X}) &= \int_{B_\mathcal{X}} d\mu^X_t \\ &= \int_{X^{-1}(B_\mathcal{X})} d\mu_t \\ &= \int_{\Phi^{-t}\left(X^{-1}(B_\mathcal{X})\right)} d\mu_0 \ .\end{aligned} \tag{5}$$

This is analogous to writing, as done in Eq. (3), observations x_t in terms of the initial state ω_0 and the action of Koopman operators on the measurement observable X. Recall that the two operators are dual, so that these two perspectives are equivalent. If there is initial uncertainty over system states, then the observables become random variables. The Koopman operator then evolves observable random variables that are distributed according to the action of Perron-Frobenius operators on the initial distribution.

Thus, given an initial uncertainty measure over system states, a dynamical process is a *stochastic process* $\{\ldots, X_{-1}, X_0, X_1, \ldots\}$—a time series of random variables—with *realizations* $\{\ldots, x_{-1}, x_0, x_1, \ldots\}$. Note that the random variables in $\{\ldots, X_{-1}, X_0, X_1, \ldots\}$ are actually (measurable) functions of two variables: $X_t = X(t, \omega_0)$. Fixing ω_0 produces a realization, or *sample path*, $\ldots, x_{-1}, x_0, x_1, \ldots$ of the stochastic process. From our setup, the realization $\ldots, x_{-1}, x_0, x_1, \ldots$ for a given ω_0 is the result of applying the map X to each ω_t in the orbit generated by ω_0. We consider continuous maps X so that realizations $\{x_t = X(\omega_t) : \omega_t = \Phi^t(\omega_0)\}$ are also continuous curves in t.

We emphasize again that evolution operators are defined in Eqs. (2) and (4) in terms of the Platonic model Φ. As such, they are yet other ways of expressing the true governing physics of a given system. In particular, they provide the true physics of partial observations.

C. Nonequilibrium Statistical Mechanics

Equation (3) expresses the time series of measurement observations in terms of Koopman operators and an unknown initial Platonic state ω_0. In contrast, Eq. (5) expresses the measurement observables as a continuous stochastic process using Perron-Frobenius operators and an initial probability distribution μ_0 over the Platonic states. To compensate for not knowing the exact initial state ω_0, one can ask, is there a natural choice for an initial distribution μ_0 over Ω induced by observations? This key question leads directly to statistical mechanics.

The standard choice for μ_0 is the invariant measure μ_* given by $P^t \mu_* = \mu_*$. For the ergodic systems considered here μ_* is guaranteed to exist and to be reached asymptotically (see Appendix A). The following employs this commonly-invoked "equilibrium case" to review instantaneous data-driven models. Note that, by definition, taking the invariant measure μ_* as μ_0 leads to $\mu_t = \mu_*$ for all t. Due to this, the random variable observables in Eq. (5) have time-independent distributions. In this case, the stochastic process over measurement observables is a *stationary* stochastic process. Clearly though, assuming the invariant measure μ_* precludes nonasymptotic "nonequilibrium" behaviors that we ultimately wish to also capture.

The preceding defined dynamical processes as stochastic processes generated by deterministic dynamical systems. To set the stage for the nonequilibrium generalization with time-dependent measures used later for history-dependent models, recall that underlying system states ω_t can not be uniquely identified from an observation x_t due to the noninvertibility of the measurement observable function X. This setup admits a natural nonasymptotic measure induced by a single observation $x_t = X(\omega_t)$ that we now define.

Consider a dynamical system $(\Omega, \Sigma_\Omega, \nu, \Phi)$ and a single observation $x_t = X(\omega_t)$ at an arbitrary time t. Since the observation mapping X is not invertible there can be many $\omega_t \in \Omega$ yielding the observed value x_t under X. (This is directly related to the non-uniqueness of model inversion when assimilating physics-based models). Thus, for a given observation x_t define the set $B_t \in \Sigma_\Omega$ as:

$$B_t = X^{-1}(x_t) = \{\omega_t \in \Omega \mid X(\omega_t) = x_t\} \ . \qquad (6)$$

Note that B_t is ν-measurable.

Following Refs. [16, 19]'s minimal bias argument there is a natural measure $d\mu_t = \rho_t d\nu$ defined through the density ρ_t that is constant over B_t and zero elsewhere, so that $\Pr(\omega_t \in b \subseteq B_t) = \nu(b)/\nu(B_t)$. The *Maximum Entropy Principle* (MEP) says that the distribution which maximizes entropy subject to known constraints creates the minimally-biased prior distribution that is spread out as much as possible, up to given constraints. If the only constraint given is the support set, MEP reduces to the *Principle of Indifference* and assigns uniform probability over the set.

In what way is the noninvariant measure μ_t a nonequilibrium generalization of the equilibrium measure μ_*? The nonequilibrium behaviors allowed by ergodic systems with dynamics Φ which have no explicit time-dependence are those of relaxation processes [39]. According to the attractor-basin formalism described in Appendix A, these processes limit to equilibrium distributions given by the invariant measure μ_*. Theorem 4.5 in Ref. [39] establishes the correspondence between the invariant measure μ_* and thermodynamic equilibrium for these systems. Hence, any other measure μ is a nonequilibrium distribution that asymptotically limits to the equilibrium distribution. The measure μ_t is a nonequilibrium measure naturally induced through partial observations.

Note that there is a wide range of nonequilibrium phenomena beyond relaxation processes. For instance, an invariant measure may correspond to an equilibrium steady state or a nonequilibrium steady state [40] that absorbs and dissipates energy from its surroundings. These more general far-from-equilibrium processes that include thermal driving require explicit time-dependence in the dynamics [41]. Detailed thermodynamic analysis is not our primary concern as yet, but the formalism introduced here readily extends to such settings by including explicit time-dependence in the Koopman and Perron-Frobenius operators. See, for example, the dynamics governed by Ref. [42]'s time-dependent rate-matrices. In that language, the development here applies in the special case of a fixed time-independent protocol with relaxation to the associated invariant distribution.

As seen shortly, the two measures μ_* and μ_t represent different sets of assumptions used to motivate and interpret the behavior of data-driven models. Neither is typically known explicitly, but is rather inferred approximately from observations. Kernel methods are particularly useful for this in practice [3, 43]. The nonequilibrium measure μ_t is more closely aligned with the modeling approach of physics-based models. Its construction requires knowledge of the set B_t of Platonic states consistent with the observation x_t, much like the construction of the data image u_t that is the approximation of the Platonic state most consistent with x_t. Ultimately though, both μ_* and μ_t are insightful in their own way for understanding implicit data-driven models, and so both are discussed in detail in what follows.

VI. INSTANTANEOUS IMPLICIT MODELS

With the physics of dynamical processes laid out using the machinery of Koopman and Perron-Frobenius evolution operators, we return to the question of optimal prediction. Section IV outlined the challenges of using physics-based modeling for prediction. Circumventing explicit inference of unobserved degrees of freedom requires learning, directly from observation, evolution rules for the variables that are accessible through instrument measurements. Pushing this further, we explore learning *implicit models* that predict the evolution of the observables—models given in a more flexible, possibly more abstract, form than differential equations-of-motion.

A. Instantaneous Predictive Distributions

The most basic form of prediction for a dynamical process is instantaneous: Given a single observation $x_t = X(\omega_t)$, predict the observable at a single time in the future $x_{t+\tau} = X(\omega_{t+\tau})$. Before reviewing the functional Hilbert space approach for learning instantaneous implicit models, we first theoretically analyze the problem using evolution operators. We argue that the maximal instantaneous predictive information available is given in an instantaneous predictive distribution. We will later see that the optimal Hilbert space model for instantaneous prediction is the expectation value of the instantaneous predictive distribution.

Given a single observation x_t, Eq. (6) defined B_t as the set of all possible Platonic states ω_t consistent with the observation x_t such that $X(\omega_t) = x_t$. This then defines the set of all possible observables $x_{t+\tau}$ that may be seen at a later time by evolving each $\omega_t \in B_t$ under Φ^τ and applying the observable mapping X. Said another way, the set of all possible future observables $x_{t+\tau}$ is given through the action of the Koopman operator by applying the time-shifted observable $X_\tau = [U^\tau X](\omega_t)$ to all ω_t in B_t.

Furthermore, we use the MEP measure μ_t over B_t and the Perron-Frobenius operator to define the distribution over possible future observables, supported on the set $\{x_{t+\tau} = [U^\tau X](\omega_t), \quad \text{for all } \omega_t \in B_t\}$. This distribution—the *instantaneous predictive distribution*—is given as the pushforward of the time-evolved measure $\mu_{t+\tau} = P^\tau \mu_t$ along X, following Eq. (5):

$$\Pr(U^\tau X | X_t = x_t) \ \sim \ \mu^X_{t+\tau} \ . \tag{7}$$

To define the instantaneous predictive distribution, we need *some* initial measure μ_t. Without additional information on the system, the choice of a MEP measure is most natural. What matters for our purposes is that μ_t is supported on the set B_t and, thus, the instantaneous predictive distributions are supported on $\{x_{t+\tau} = [U^\tau X](\omega_t), \quad \text{for all } \omega_t \in B_t\}$. In practice, these measures are estimated empirically from data and the MEP is not typically invoked for μ_t's empirical construction.

Also note that the formalism for instantaneous models we now review is given in terms of the equilibrium measure μ_*, as is standard. However, instantaneous predictive distributions cannot be expressed directly in terms of μ_*. That said, the nonequilibrium construction of predictive distributions just given is instructive for understanding instantaneous models built using μ_*. The equilibrium and nonequilibrium formulations of data-driven models are more closely connected below for history-dependent models using Wiener projections. Our introduction of nonequilibrium measures is most impactful for history-dependent models, as they provide insights not available through use of the invariant measure.

B. Instantaneous Data-Driven Models

We now review instantaneous data-driven models and their Hilbert space formalism. Following established practice, we use the invariant measure μ_*.

Given the current observation data x_t, the goal is to construct a model $\mathcal{T}^\tau : \mathcal{X} \to \mathcal{X}$, called a *target function*, that predicts what the instruments will read at a later time $t + \tau$ [14]. This is depicted in Figure 1's middle column. On the one hand, recall that for physics-based models we assume the system's governing equations of motion Φ are known, and that one of the main challenges is to infer from partial observations $x_t = X(\omega_t)$ the underlying Platonic state ω_t that the equations evolve. On the other hand, the data-driven paradigm flips this around to work directly with the measurement observations x_t, without directly inferring ω_t. In point of fact, an appropriate set of governing equations for x_t is generally not know a priori. They may not even be desired. Instead, the goal is to *learn* a model $\mathcal{T}^\tau$ from the measurement data.

In some cases we can learn $\mathcal{T}^\tau$ as closed-form equations using Galerkin projections of Φ onto $\mathcal{X}$ [44]. In many cases, though, the evolution of x_t cannot be adequately described by a set of closed-form equations [45]. Thus, we seek more general forms for $\mathcal{T}^\tau$ that are measurable mappings from $\mathcal{X}$ into $\mathcal{X}$. For example, neural networks

[46] are universal function approximators [47] and so are able, in principle, to represent $\mathcal{T}^\tau$.

As the name suggests, a modeler cannot simply write down an implicit model. Rather, implicit models are implemented algorithmically and learned from data. From the discussion above, the Koopman operator provides the ground truth for prediction—the equivalent of the Platonic model. Following Ref. [14], the mean-squared error for model $\mathcal{T}^\tau$ is given as:

$$\left\|\mathcal{T}^\tau \circ X - U^\tau X\right\|^2_{L^2(\mu_*)} . \tag{8}$$

From the Koopman operator definition and Fig. 1's commutation relations, the measurement data $\{x_0, x_1, \ldots, x_T\}$, used to learn $\mathcal{T}^\tau$, contain samples of the Koopman operator's action. That is, for an observation $x_t = X(\omega_t)$, a later observation is:

$$\begin{aligned} x_{t+\tau} &= X(\omega_{t+\tau}) \\ &= X_\tau(\omega_t) \\ &= [U^\tau X](\omega_t) . \end{aligned}$$

Empirically, the Koopman operator's action is approximated through the action of the *shift operator* [3, 14]. And so, the ground truth for training $\mathcal{T}^\tau$ is found by simply looking up in the observed data $\{x_0, x_1, \ldots, x_T\}$ what happens after time τ. In this way, given $x_t \in \{x_0, x_1, \ldots, x_T\}$, $0 \leq t < (T-\tau)$, $\widetilde{x}_t = \mathcal{T}^\tau(x_t)$ is the prediction made by $\mathcal{T}^\tau$. With the ground truth given by $x_{t+\tau} \in \{x_0, x_1, \ldots, x_T\}$, a parametric model (e.g., neural network) $\mathcal{T}^\tau$ is trained by minimizing $\|x_{t+\tau} - \widetilde{x}_t\|^2$ over the training data, $0 \leq t < (T-\tau)$.

C. Analog Forecasting

Analog forecasting, dating back at least to Ref. [48], is one of the oldest methods for approximating $\mathcal{T}^\tau$ implicitly from data. The basic procedure is to predict a system's future by finding the value recorded in past observations (the analog) that is most similar to the present observation and then use the following value in the recorded history as the forecast.

Formally, let $\{x_0, x_1, x_2, \ldots, x_T\}$ be the finite set of historical observations—the training data set. Then, given the current observation $X_t = x$, with $t > T$, identify x's analog x_a in the training set. This is typically implemented with Euclidean distance:

$$a = \underset{i \in \{0, \ldots, T-\tau\}}{\operatorname{argmin}} \; D(x, x_i) .$$

The forecast $x_{t+\tau}$ of x for τ time steps into the future is given by the analog forecast $x_{a+\tau} \in \{x_0, x_1, x_2, \ldots, x_T\}$. That is, the analog forecast simply looks up what happened in the training data set τ time steps after the analog:

$$\mathcal{T}^\tau_{\mathrm{AF}}(x) = x_{a+\tau} . \tag{9}$$

Analog forecasting is used for the data-driven prediction examples given below in Section X.

D. Optimal Hilbert Space Models

As data-driven models, target functions $\mathcal{T}^\tau$ map a single input to a single output. As such, target functions live in a function space. Since inner products and orthogonal projections play an important role in the development, we seek target functions as elements of a Hilbert space. The following reviews the Hilbert space formulation of instantaneous data-driven models [3, 14, 15].

Recall that partial observations of a dynamical system (Ω, Φ) induce a time-dependent probability measure μ_t over Ω. For simplicity when using instantaneous models, in the asymptotic limit we employ the invariant ergodic measure μ_*. This leads to the probability space $(\Omega, \Sigma_\Omega, \mu_*)$ and measurement observables as random variables given by the measurable map $X : \Omega \to \mathcal{X}$. The space $\mathcal{X}$ is often referred to as the *covariate space* and X the *covariate map.*

More generally, we may consider a *response space* $\mathcal{Y}$ and the (measurable) *response map* $Y : \Omega \to \mathcal{Y}$. The target function is then a measurable mapping from covariates to a response: $\mathcal{T}^\tau : \mathcal{X} \to \mathcal{Y}$. In our dynamical setting, the covariate and response spaces are the same—the observation instrument readings: $\mathcal{Y} = \mathcal{X} = \mathbb{R}^n$. The response map is the time-shifted measurement observable $Y = X_\tau = X \circ \Phi^\tau : \Omega \to \mathcal{X}$. Note that X and $Y = X_\tau$ are both random variables over the same probability space $(\Omega, \Sigma_\Omega, \mu_*)$ since $x_t = X(\omega_t)$ and $y_t = x_{t+\tau} = X_\tau(\omega_t)$. This is what makes $\mathcal{T}^\tau$ predictive.

Consider the Hilbert spaces:

$$\begin{aligned} H &:= \{f : \Omega \to \mathcal{X} : \int_\Omega f^2(\omega) d\mu_*(\omega) < \infty\} , \\ V &:= \{g : \mathcal{X} \to \mathcal{X} : g \circ X \in H\} , \text{ and} \\ H_X &:= \{f \in H : f = g \circ X \text{ for some } g \in V\} . \end{aligned} \tag{10}$$

Note that $H = L^2(\mu_*)$—the set of square-integrable functions *of the full system state* ω. And, H_X is the Hilbert subspace of $L^2(\mu_*)$ containing functions f_X that depend *only* on the observable degrees of freedom $x_t = X(\omega_t)$. Then $V = L^2(\mu_*^X)$, where μ_*^X is the pushforward of μ_* along X, is the set of functions over the measurement observables such that composition with the observable map X is square integrable. Since the observables $X \in H$ are what is accessible, the set V is what is at our disposal to build implicit models. However, since Koopman operators U^τ act on functions in H, we must compose elements of V with X for proper comparison with U^t's action. Specifically, the ground-truth for the future observation is given by $X_{t+\tau} = U^\tau X$ which lives in the space H, while the functions available for us to learn are in the subspace H_X.

Equation (8) identifies the unique minimizer—the optimal $\mathcal{T}^\tau$. Denote it Z^τ. In statistics, this estimator is known as the *regression function*, as well as the *conditional expectation function*. That is:

$$\begin{aligned}\mathbb{E}[U^\tau X|X] &= Z^\tau \\ &:= \underset{g\in V}{\operatorname{argmin}} \|g\circ X - U^\tau X\|^2_{L^2(\mu_*)} \ . \end{aligned} \qquad (11)$$

The conditional expectation $\mathbb{E}[\cdot|X]$ is the (nonlinear) orthogonal projection $P_X : H \to H_X$ from H onto H_X. Thus, $Z^\tau = P_X U^\tau X_t$. This is the best approximation of $X_{t+\tau} = U^\tau X \in H$ available using functions restricted to the subspace H_X.

For a given learned target function (data-driven model) $\mathcal{T}^\tau$, we decompose its error via the regression function Z^τ:

$$\mathcal{E}(\mathcal{T}^\tau) := \|\mathcal{T}^\tau \circ X - U^\tau X\|^2 \qquad (12)$$

$$= \Theta(\mathcal{T}^\tau) + \Xi^\tau_X \ . \qquad (13)$$

The *excess generalization error* $\Theta(\mathcal{T}^\tau) = \|\mathcal{T}^\tau - Z^\tau\|^2$ measures how far a given model is from the optimal solution, while Ξ^τ_X is the *intrinsic error* due to the partial observations X of a given system (Ω, Φ). For a given physical system with instrument measurements X, the regression function Z^τ represents the maximum predictive skill an instantaneous data-driven model can achieve. Ξ^τ_X is then the unavoidable error incurred from only being able to measure X. Increasing instrument coverage and expanding X can decrease Ξ^τ_X.

The conditional expectation $\mathbb{E}[U^\tau X|X]$ can be expressed as the expectation of a probability measure supported on the set $\{x_{t+\tau} = [U^\tau X](\omega_t), \text{ for all } \omega_t \in B_t\}$—the instantaneous predictive distribution. Although we directly formulated the instantaneous predictive distribution in Eq. (7) using the nonequilibrium measure μ_t, these predictive distributions cannot be directly formulated as an $L^2(\mu_*)$ measure [3]. That said, they provide insight into the intrinsic error of Z^τ, regardless of which measure is used for the nonlinear projection of Eq. (11). As Z^τ is the expectation of this distribution, the intrinsic error is then seen as the variance of the predictive distribution. Later, we will give explicit constructions of history-dependent generalizations of predictive distributions.

Empirical Hilbert Spaces While $L^2(\mu_*)$ is theoretically convenient, it is not a workable space for empirical models [49]. This is because functions in $L^2(\mu_*)$ cannot be distinguished with a finite set of samples, but the latter is what is empirically available. Data-driven algorithms thus typically employ reproducing kernel Hilbert spaces (RKHSs), which have well-defined point evaluations. For more on RKHS methods, as well as empirical sample measures and their convergence, see Refs. [3, 14, 50, 51].

VII. MORI-ZWANZIG FORMALISM OF DYNAMICAL PROCESSES

Although Eq. (11) defines the optimal instantaneous model, the optimal model still has an associated intrinsic error—the variance of the instantaneous predictive distribution. With the same Hilbert space and projection operator formalism used to define Z^τ, the Mori-Zwanzig formalism provides the full equations of motion for the observable degrees of freedom by projecting the system dynamics onto those degrees of freedom [5]. The composition of the Mori-Zwanzig equation reveals terms in addition to Z^τ that lead to the intrinsic error when not accounted for in instantaneous models. Crucially, the additional terms show that partial observation induces a memory dependence in dynamical processes. This then motivates the use of history-dependent models for increased predictive skill over instantaneous models.

The Mori-Zwanzig setting is a special case of the partially-observed dynamical systems considered so far. There is an underlying true system (Ω, Φ) and a noninvertible mapping $X : \Omega \to \mathcal{X}$. In the Mori-Zwanzig setting, the variables $x = X(\omega)$ are known as ω's *resolved degrees of freedom*. Denoting the remaining *unresolved degrees of freedom* $\widetilde{x}$, then $\omega = (x, \widetilde{x})$. The standard formulation of the Mori-Zwanzig equation in statistical mechanics assumes (Ω, Φ) to be Hamiltonian and considers projections of densities $\rho(\omega)$ and their time evolution by the Liouville operator [4]. Importantly, the equation can be derived in our more general setting of dissipative systems using the Koopman operator [6, 15].

The goal is to predict the future values of the resolved degrees of freedom using only information available from them. That is, the task is to express the evolution of the resolved variables—the dynamics governing the dynamical process—in terms of the resolved variables as much as possible. We do this by projecting the Koopman operator's action onto the resolved degrees of freedom. This is possible since the dynamics of dynamical processes is given in terms of Koopman operators, as shown above.

Referring Eq. (10)'s Hilbert spaces—i.e., $H = L^2(\mu_*)$—the discrete-time derivation expands the Koopman operator $U^{t+1} : H \to H$ via the Dyson formula:

$$U^{t+1} = \sum_{k=0}^{t} U^{t-k} PU(QU)^k + (QU)^{t+1} \ . \qquad (14)$$

In this, P is an orthogonal projection operator from H to a subspace $H_{\mathbf{\Psi}} \subseteq H_X \subset H$ spanned by basis functions $\mathbf{\Psi}(x)$ that depend *only* on the resolved variables $x = X(\omega)$. And, $Q = I - P$ is the orthogonal projection to the unresolved variables.

Recall that $X \in H$ is the observation function that returns data gathered from measurement recordings of an underlying physical system Ω. Forming new observable functions in the projected space $H_{\mathbf{\Psi}}$ uses observation measurements in X and functions $\psi(x)$ of them. This is in contrast to introducing new measurement instruments—instruments that would enlarge the resolved-variable space H_X. In finite subspace projection algorithms, P projects into the subspace $H_{\mathbf{\Psi}}$ spanned by the basis of dictionary functions $\mathbf{\Psi}$.

In discrete time, the dynamics of the resolved variables are generated via:

$$\begin{aligned} x_{t+1} &= X(\omega_{t+1}) \\ &= X_{t+1}(\omega_0) \\ &= [U^{t+1}X](\omega_0) \ . \end{aligned}$$

The discrete-time *Mori-Zwanzig equation* then follows by applying the expansion in Eq. (14) to the unit-shift observable $X_{t+1} = U^{t+1}X \in H$. Skipping algebra and notational simplifications [15], this yields:

$$x_{t+1} = M_0(x_t) + \sum_{k=1}^{t} M_k(x_{t-k}) + \xi_{t+1}(\omega_0) \ . \qquad (15)$$

The key is that this expression is exact. It gives the true evolution of the measurement observables, equivalent to the action of the full infinite-dimensional Koopman operator.

The first term $M_0(x_t)$ describes Markovian evolution. It gives the best Markov approximation of Φ^1 under projection P. That is, M_0 is the best approximation of the unit-step dynamics by a function of the current observable only. It is the optimal target function Z^1 defined above when the nonlinear projection given in Eq. (11) is used in Eq. (14). The last term $\xi_{t+1}(\omega_0)$ is the orthogonal term originating from the initial unresolved components. The second term captures longer-range statistical dependencies with a discrete convolution of a memory kernel that depends on the orthogonal terms: $M_k \circ X = P(\xi_k \circ \Phi) \circ X$. (Statistical mechanics refers to this orthogonal dependence of memory as a *fluctuation-dissipation relation.*) All terms depend on the particular projection operator P used. (For example, the terms M_0 and $\{M_k\}$ may be linear—i.e., matrices—for certain choices of P [7].)

Appendix B gives an alternative derivation of the Mori-Zwanzig equation, following the original Hamiltonian statistical mechanics treatment. For those unfamiliar with the Mori-Zwanzig formalism, this provides additional insight and physical intuition, starting from Hamilton's equations.

A comparison is in order between the Mori-Zwanzig perspective of Koopman operator projections in Eq. (15) versus the data-driven approaches for finite-dimensional Galerkin projections of the Koopman operator. The latter are presented in Appendix D; namely, Dynamic Mode Decomposition (DMD) and Extended Dynamic Mode Decomposition (EDMD). Both DMD and EDMD are instantaneous models and, as such, only approximate the Markovian term M_0. Both do so using linear finite subspace projections onto $H_{\mathbf{\Psi}} \subseteq H_X$. DMD uses the simple dictionary $\mathbf{\Psi} = \{f_X\}$ consisting of only the identity function f_X; while EDMD uses arbitrary dictionaries $\mathbf{\Psi}$ of basis functions. In contrast, while data-driven approaches to Mori-Zwanzig evolution operators also use finite subspace projections $H_{\mathbf{\Psi}} \subseteq H_X$ [5, 7, 52], they do so for the memory kernels as well as for the Markovian component.

Comparing further, EDMD seeks a Galerkin approximation of the Koopman operator itself, with a single matrix, while Mori-Zwanzig evolution approximates projections of the Koopman operator *action* specifically on functions of the resolved degrees of freedom. Paraphrasing Ref. [7]: "EDMD seeks a point $\mathbf{U}_X^\tau \mathbf{\Psi}(x_t)$ in $H_{\mathbf{\Psi}}$ that minimizes the error between the point and $\mathbf{\Psi}(x_{t+\tau})$, whereas Mori-Zwanzig simply projects $\mathbf{\Psi}(x_{t+\tau})$ onto $H_{\mathbf{\Psi}}$".

The crucial insight of the Mori-Zwanzig equation Eq. (15) is that partially observing dynamical systems induces a memory dependence in the observable degrees of freedom. Optimal instantaneous models are thus not fully optimal as data-driven Hilbert space models. History-dependent models will reduce intrinsic error and improve predictive skill. Moreover, the dependence of the memory kernels on the orthogonal unresolved degrees of freedom indicates that the memory dependence accounts for the effects of the unresolved variables on the dynamics of the resolved variables. Recall that much of the effort in physics-based models comes in explicitly inferring the unobserved degrees of freedom and their dynamical effects.

VIII. DELAY-COORDINATE EMBEDDINGS

A key step in bridging physics-based and data-driven approaches to prediction comes through the formulation of memory as reconstruction embeddings [53]. Their intrinsic geometry illuminates how memory of partial observations implicitly encodes effects of the unobserved degrees of freedom.

Starting with a scalar time series $\{x_t, t \in \mathbb{N}^+\}$, the task is to reconstruct an effective state space of *embedding dimension* m in which the effective states evolve as a deterministic dynamical system. In short, m is set large enough that the orbits in the reconstructed state space do not intersect. A *derivative-coordinate embedding* of a measurement observable x_t develops a reconstructed state space from $\overleftarrow{x}_t^m = \{\dot{x}_t, \ddot{x}_t, \ldots, d^m x_t/dt^m\}$ [8]. A *delay-coordinate embedding* uses $\overleftarrow{x}_t^{m,\delta} = \{x_t, x_{t-\delta}, x_{t-2\delta}, \ldots, x_{t-(m-1)\delta}\}$ with lag δ [9]. Due to its familiarity we discuss delay-coordinate embeddings, despite the extra required optimization over lag δ that may be required in practice. Unless otherwise stated, we take $\delta = 1$. For continuous-time systems the lag is given in units of the measurement sample rate Δt.

The original work on coordinate embeddings established that the geometry of the asymptotic attractor of (Ω, Φ) can be reconstructed, up to diffeomorphism, from embeddings $\overleftarrow{x}^m$ of partial observations $x = X(\omega)$ for sufficiently large m [8, 9]. The intuitive idea is that the additional values in the embedding $\overleftarrow{x}^m$ essentially act to fill in the degrees of freedom of ω missing from X. The reconstructed orbit $\overleftarrow{x}^m_t$ of the embedding traces out an attractor that is geometrically equivalent to that generated by the full system state ω_t.

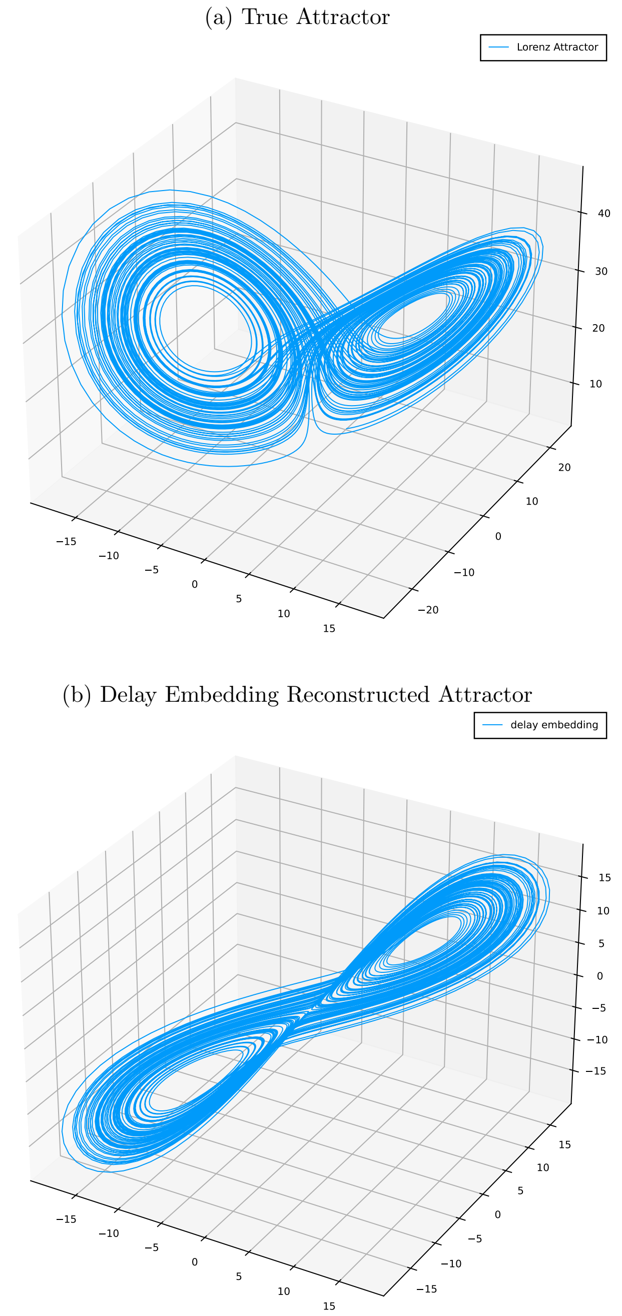
Geometrically, embeddings encode the unobserved degrees of freedom in the histories of the observed degrees of freedom. Moreover, the Koopman operator acting on a delay embedding observables implicitly encodes the unobserved degrees of freedom in a dynamically useful way [11–13]. In fact, the Koopman operator acting on delay-coordinate embeddings corresponds to the Laplace-Beltrami operator describing the attractor geometry [12]. In the asymptotic limit with evolution on the attractor, this correspondence allows employing geometric tools, such as heat kernels and diffusion maps, in data-driven modeling [3].

The details of how evolution operators acting on delay-coordinate embeddings dynamically encode the unobserved degrees of freedom will be examined thoroughly below. First though, we show the classical example of how delay embeddings geometrically encode unobserved degrees of freedom with the Lorenz 63 attractor. Later, we will return to this example to demonstrate the dynamical encoding of the unobserved degrees of freedom using analog forecasting.

Example Reconstruction The following gives an empirical demonstration that delay embeddings can geometrically "fill in the gaps" of missing degrees of freedom the three-dimensional Lorenz 63 system:

$$\begin{aligned}\dot{x} &= \sigma(y - x)\\ \dot{y} &= x(\rho - z) - y\\ \dot{z} &= xy - \beta z\ .\end{aligned}$$

Figure 2(a) shows the attractor revealed by their numerical solution with parameters $\sigma = 10.0$, $\rho = 28.0$, and $\beta = 8/3$. For comparison, Fig. 2(b) shows the attractor reconstructed using delay-coordinate embeddings of the x variable alone with $\delta = 4$ and $m = 7$. The delay-reconstructed attractor is a "squished" version of the original, but is (approximately) geometrically equivalent. The simulation and delay embedding reconstruction were performed using the DynamicalSystems.jl package in Julia [54].


(a) True Attractor

(b) Delay Embedding Reconstructed Attractor

FIG. 2. (a) Full 3D attractor from numerical solutions of Lorenz equations. (b) Delay-embedding reconstruction attractor using x variable only with embedding dimension $m = 7$ and lag $\delta = 4$. The first three dimensions $\{x_t, x_{t-4}, x_{t-8}\}$ are shown.

IX. HISTORY-DEPENDENT MODELS

Many history-dependent model classes, such as recurrent neural networks [46] , are more readily understood as mappings $\overleftarrow{\mathcal{T}}^\tau$ from pasts (delay embeddings) to future observations, rather than as fitting the paradigms of Markov and memory kernels from the Mori-Zwanzig equation. Generalizing the instantaneous case given in Eq. (8), the mean-squared error of history-dependent Hilbert space models is:

$$\|\overleftarrow{\mathcal{T}}^\tau \circ \overleftarrow{X}^k - U^\tau X\|^2_{L^2(\mu_*)}\ . \tag{16}$$

As in the instantaneous case, we identify the unique minimizer of Eq. (16) as the optimal history-dependent

Hilbert space model. This optimum is achieved by formally connecting the Mori-Zwanzig formalism with delay-coordinate embeddings using Wiener projections of the Koopman operator.

Before detailing Wiener projections, it is helpful to first review two standard orthogonal projections. Both use the $L^2(\mu)$ inner product. For now, we follow Ref. [6] and use the invariant measure μ_*:

$$\langle f, g\rangle_{L^2(\mu_*)} = \int_\Omega f\; g\; d\mu_* \ .$$

Equation (11) defined the optimal instantaneous model Z^τ as the conditional expectation function that minimizes the $L^2(\mu_*)$ norm between $g \circ X$ and $U^\tau X$. This is known as the nonlinear or *infinite-rank* projection, used by Ref. [55], of $U^\tau X$ from H into H_X.

In contrast, the linear projection used by Ref. [56], also known as a *finite-rank* or finite-subspace projection, is defined in terms of an orthogonal set $\mathbf{\Psi}$ of size N on the space V of functions of the observed variables. That is:

$$Pf = \sum_{i=1}^{N} \langle f, \phi_i\rangle_{L^2(\mu_*)}\; \psi_i \ , \tag{17}$$

where $\{\phi_i = \psi_i \circ X\}_{i=1}^N$.

In the infinite-rank limit $\mathbf{\Psi} \to V$, this linear projection converges to the nonlinear conditional expectation projection. As with EDMD, the challenge for data-driven methods that employ finite-subspace projections [7, 14, 15, 57] is to find an effective finite basis $\mathbf{\Psi}$.

A. Equilibrium Wiener Projections

Using these instantaneous projections, the standard Mori-Zwanzig formalism embodies history dependence of the observed variables in the collection of memory kernels. In contrast, *Wiener projections* incorporate history dependence directly into the projection operators via delay-coordinate embeddings. Specifically, the linear Wiener projections replace the single $L^2(\mu_*)$ inner product with:

$$\langle f, g\rangle_W := \lim_{k\to\infty} \frac{1}{k} \sum_{\tau=1}^{k} \int_\Omega [U^{-\tau} f][U^{-\tau} g] d\mu_* \ , \tag{18}$$

in the linear projection in Eq. (17). That is, $L^2(\mu_*)$ inner products of the reverse-time-shifted observables are taken at all times into the infinite past.

Although the original formulation of Wiener projections, Eq. (18), is given in terms of the equilibrium invariant measure, the physical intuition is similar to a move common in nonequilibrium statistical mechanics. Statistics of a dynamical variable can equivalently be thought of in two ways. They can be either the stochastic time-evolution of a "state variable", for example Langevin dynamics, or the idea of "state variable" can be replaced by dynamical trajectories, with statistics over trajectories considered. The standard formulation of the Mori-Zwanzig in Eq. (15) is akin to the first perspective. In fact, the Mori-Zwanzig equation is often known as the generalized Langevin equation. The use of Wiener projections is akin to the second, where projections of time-dependent "state variables" are replaced with projections of trajectories (delay coordinate embeddings).

Applying a discrete-time Wiener projection P_W to the discrete-time Koopman operator, as first introduced in Ref. [6], results in:

$$U^{t+1} = U^t P_W U + (Q_W U)^{t+1} \ . \tag{19}$$

As expected, there is no longer a temporal convolution over memory kernels; only a Markov term and an orthogonal term. In the setting of quantum statistical mechanics, Ref. [58] gives a similar expression using time-dependent projection operators.

Furthermore, Ref. [15] argues that, if the conditions of the delay embedding theorem [9] are met, the orthogonal term vanishes. Thus, *in the ideal case*, Mori-Zwanzig evolution with delay-coordinate embeddings reduces to only a single Markov term that corresponds to the nonlinear Wiener projection of $U^{t+1}X$:

$$\begin{aligned} x_{t+1} &= \mathbb{E}[U^{t+1} X | \overleftarrow{X}] \\ &= \overleftarrow{M}_0(\overleftarrow{x}_t) \ , \end{aligned} \tag{20}$$

for embedding dimension m sufficiently large to satisfy the delay-embedding theorem.

In the nonideal case, particularly with finite k, the orthogonal term does not vanish and there may be memory effects at Markov order larger than that spanned by finite pasts $\overleftarrow{x}^k_t$. Therefore, as with standard (instantaneous) Mori-Zwanzig Markov approximations, the stochastic evolution of finite pasts is modeled by the finite Markov operator $\overleftarrow{M}^k_0(\overleftarrow{x}^k)$ plus an effective "noise" term:

$$\Pr(X_{t+1} | \overleftarrow{X}^k = \overleftarrow{x}^k) = \overleftarrow{M}^k_0(\overleftarrow{x}^k) + \text{noise} \ . \tag{21}$$

A finite model of this form is found in the HAVOK method [10, 13], based on Hankel DMD [11]. This finds the best-fit linear approximation for $\overleftarrow{M}^k_0$ with the leading components of the singular value decomposition of the Hankel matrix, whose columns are time-ordered delay embeddings. The last few components are then fit to the noise.

We emphasize that the Wiener projection approach to Mori-Zwanzig evolution is useful as it provides a direct connection to delay-coordinate embeddings and their intrinsic geometry. Theoretically, though, it merely rearranges memory dependence in the observable degrees of freedom. Delineating the practical advantages or disadvantages of Wiener projections over the standard Mori-Zwanzig formalism requires further investigation. Note, though, that the algorithms given by Ref. [7] for reconstructing the Markov and memory kernels of the latter employ two-time correlation functions of the observed variables. These are closely related to the instantaneous predictive distributions described above.

B. Nonequilibrium Wiener Projections

For another perspective on how the orthogonal term in Eq. (19) may vanish, it is instructive to formulate Wiener projections using nonequilibrium time-dependent measures, rather than the equilibrium invariant measure. In statistical mechanics, in fact, the invariant measure is taken to be exactly that of the equilibrium distribution, with the L^2 inner products being equilibrium correlations and the Mori-Zwanzig equation's validity holding only near equilibrium [41].

First, we introduce the history-dependent generalization of the time-dependent Maximum Entropy measures introduced in Section V. The history-dependent generalization of the Maximum Entropy Principle is known as Maximum Caliber [18, 19]. In short, given a time series of constraints up to the present moment, Maximum Caliber constructs the least biased distribution at the current time by maximizing the entropy while accommodating all time-evolving constraints. If the constraints are given in the form of expectation values, as is typical in statistical mechanics, this results in generally intractable spacetime path integrals.

As in the instantaneous case though, constraints for partially-observed systems are simply support sets of possible ω consistent with observations $x = X(\omega)$. Now, however, there are multiple time-evolving observations $\{x_t, x_{t-1}, \ldots, x_{t-k}\}$ in the form of delay embeddings to constrain the support sets over Ω.

Equation (6) defined the instantaneous set $B_t \in \Sigma_\Omega$ as the set $X^{-1}(x_t)$ of all ω_t consistent with the observation x_t such that $X(\omega_t) = x_t$. Rather than a single instantaneous observation, consider now two sequential observations x_{t-1} and x_t. Define $\overleftarrow{B}_t^2 \in \Sigma_\Omega$ as the set of ω_t consistent with both observations such that $X(\omega_t) = x_t$ *and* $[U^{-1}X](\omega_t) = X(\Phi^{-1}(\omega_t)) = x_{t-1}$. Note that $\overleftarrow{B}_t^2 \subseteq B_t$. If the two sets are not equal, we say that $\overleftarrow{B}_t^2 = B_t \cap X^{-1}(x_{t-1})$ *refines* B_t.

For a depth-k past $\overleftarrow{x}_t^k = \{x_t, x_{t-1}, \ldots, x_{t-k}\}$—a $k+1$-dimensional delay embedding—we define the set $\overleftarrow{B}_t^k$ as:

$$\begin{aligned}\overleftarrow{B}_t^k := \{\omega_t \mid X(\omega_t) = x_t, [U^{-1}X](\omega_t) = x_{t-1}, \ldots, \\ [U^{-k}X](\omega_t) = x_{t-k}\} \ . \qquad (22)\end{aligned}$$

Given the set $\overleftarrow{B}_t^k$, the Maximum Caliber distribution is uniform over $\overleftarrow{B}_t^k$ and zero elsewhere, as with the instantaneous Maximum Entropy case. For finite k, $\overleftarrow{B}_t^k$ is ν-measurable and the Maximum Caliber measure $\overleftarrow{\mu}_t^k$ is defined through the density $\overleftarrow{\rho}_t^k$ that is constant over $\overleftarrow{B}_t^k$ and zero elsewhere. If $\lim_{k\to\infty} \overleftarrow{B}_t^k$ is a discrete set, $\overleftarrow{\mu}_t^\infty$ is given as a sum of equally-weighted delta distributions.

With the Maximum Caliber measures in hand, we can define nonequilibrium Wiener projections using the nonequilibrium inner products:

$$\langle f_t, g_t \rangle_{\overleftarrow{x}_t^k} := \frac{1}{k} \sum_{\tau=1}^{k} \int_\Omega [U^{-\tau} f_t][U^{-\tau} g_t] d\overleftarrow{\mu}_t^k \ , \qquad (23)$$

with $f_t = f(\omega_t)$ and $g_t = g(\omega_t)$ to emphasize that the integrals are all carried out over values of ω_t for all terms in the sum. Note that this inner product is identical to its equilibrium counterpart in Eq. (18) except for the change of measure. The conditional expectation in Eq. (20) can be similarly defined using nonequilibrium Wiener projections defined by Eq. (23)'s inner product.

Unlike Eq. (18)'s equilibrium case, Eq. (23)'s inner product is contingent on the observation $\overleftarrow{x}_t^k$ that then defines the measure $\overleftarrow{\mu}_t^k$. In particular, we can analyze the sequential refinement behavior of $\overleftarrow{\mu}_t^k$ and their support sets $\overleftarrow{B}_t^k$ with increasing depth k of the observed past $\overleftarrow{x}_t^k$. This is shown in Fig. 3. Moreover, we can directly construct the predictive distributions whose expectation gives Eq. (20)'s conditional expectation using the (nonlinear) nonequilibrium Wiener projections.

C. Predictive Distributions

In the instantaneous case, recall that $B_t = X^{-1}(x_t)$ is the set of all ω_t consistent with observation x_t such that $X(x_t) = \omega_t$. The set of possible observables $x_{t+\tau}$ that may be seen at a later time τ are given by the Koopman operator as $\{x_{t+\tau} = [U^\tau X](\omega_t) \text{ for all } \omega_t \in B_t\}$. This set is the support of the instantaneous predictive distribution $\mu_{t+\tau}^X$—the pushforward of $P^\tau \mu_t$ along X.

For two sequential observations, we may expect that there are some, if not many, ω_t in B_t that are not in $\overleftarrow{B}_t^2$. That is, there may be ω_t such that $X(\omega_t) = x_t$ but $X\big(\Phi^{-1}(\omega_t)\big) \neq x_{t-1}$. As more observations are recorded, there may be increasingly fewer ω_t whose reverse orbit is consistent with the observed values $\overleftarrow{x}_t^k$. Therefore, $\overleftarrow{B}_t^k \subseteq B_t$ and, in some cases, $\overleftarrow{B}_t^k$ is a proper subset of B_t with $\nu(\overleftarrow{B}_t^k) < \nu(B_t)$. That is, the state space volume $\nu(\overleftarrow{B}_t^k)$ is monotonically nonincreasing as k grows and it may decrease for increasing k. Figure 3 illustrates the dynamical refinement of Maximum Caliber support sets.

We are now ready to examine the consequences of refinement on history-dependent predictive distributions. The latter are constructed as for instantaneous predictive distributions, using uniform initial measures over $\overleftarrow{B}_t^k$ rather than B_t. The *predictive distribution* $\Pr(U^{t+\tau}X | \overleftarrow{X}_t^k = \overleftarrow{x}_t^k)$ is supported on the set $\{x_{t+\tau} = [U^\tau X](\omega_t) \text{ for all } \ \omega_t \in \overleftarrow{B}_t^k\}$ and is distributed according to the pushforward of $P^\tau \overleftarrow{\mu}_t^k$ along X, as shown in Fig. 4 (b).

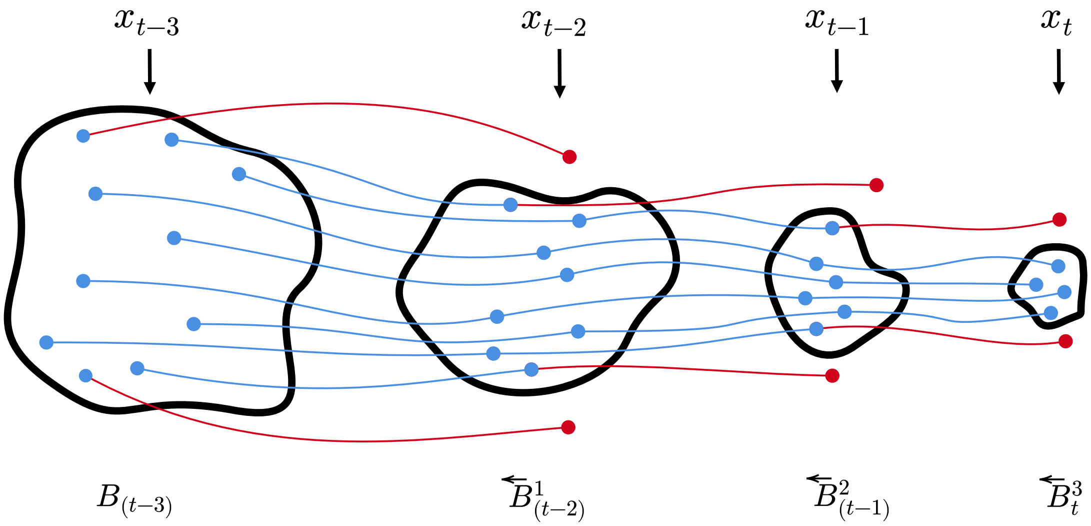


FIG. 3. Sequential support sets of Maximum Caliber measures for past lengths $k = \{0, 1, 2, 3\}$: The earliest observation x_{t-3} alone produces the $k = 0$ support set $B_{(t-3)} = X^{-1}(x_{t-3})$. The next observation x_{t-2}, together with x_{t-3} yields $\overleftarrow{B}^1_{(t-2)} = B_{(t-3)} \cap X^{-1}(x_{t-2})$. Similarly, x_{t-1} gives $\overleftarrow{B}^2_{(t-1)} = \overleftarrow{B}^1_{(t-2)} \cap X^{-1}(x_{t-1})$ and, finally, x_t gives $\overleftarrow{B}^3_t = \overleftarrow{B}^2_{(t-1)} \cap X^{-1}(x_{t-1})$. Unit-length orbits of $\omega_\tau \in \Omega$ consistent with the subsequent observation $x_{\tau+1}$ are shown in blue, while those inconsistent with the subsequent observation are red. The depiction shows strict refinement at each time, so that $\overleftarrow{B}^3_t \subset \overleftarrow{B}^2_{(t-1)} \subset \overleftarrow{B}^1_{(t-2)} \subset B_{(t-3)}$.

If $\nu(\overleftarrow{B}^k_t) < \nu(B_t)$, then the initial measure $\overleftarrow{\mu}^k_t$ is more constrained than μ_t. And so, $P^\tau \overleftarrow{\mu}^k_t$ is also be more constrained than $P^\tau \mu_t$. In this case, the predictive distribution $\Pr(U^{t+\tau} X | \overleftarrow{X}^k_t = \overleftarrow{x}^k_t)$ has lower entropy than the instantaneous $\Pr(U^{t+\tau} X | X_t = x_t)$. Here, "entropy" refers the size of a distribution's support. The volume of $\{x_{t+\tau} = [U^\tau X](\omega_t) \text{ for all } \omega_t \in \overleftarrow{B}^k_t\}$ is no larger than that of its instantaneous counterpart $\{x_{t+\tau} = [U^\tau X](\omega_t) \text{ for all } \omega_t \in B_t\}$ since $\overleftarrow{B}^k_t \subseteq B_t$ and $U^\tau X$ is measurable. Note that for distributions with a density ρ, the size of the effective support set—the *typical set*—is given by $2^{h(\rho)}$, where $h(\rho)$ is the differential entropy of ρ [59].

Taking $\tau = 1$, consider the optimal history-dependent target function $\overleftarrow{Z}^k = \mathbb{E}[U^{t+1} X | \overleftarrow{X}^k_t = \overleftarrow{x}^k_t]$ and the instantaneous optimal $Z = \mathbb{E}[U^{t+1} X | X_t = x_t]$. From the arguments above, $\nu(\overleftarrow{B}^k_t) < \nu(B_t)$ implies that $\overleftarrow{Z}^k$ is a more accurate estimator of $x_{t+1} = [U^{t+1} X](\omega_t)$ than the instantaneous optimal Z. This follows since there is less variance in $\Pr(U^{t+1} X | \overleftarrow{X}^k_t = \overleftarrow{x}^k_t)$—it is more tightly concentrated about its mean than $\Pr(U^{t+1} X | X_t = x_t)$. Such a conclusion is in line with the intuition that the intrinsic error of instantaneous models derives from the unobserved degrees of freedom and that including past observations in the form of delay-coordinate embeddings accounts for the missing degrees of freedom.

It is natural to ask what happens in the $k \to \infty$ limit of infinitely-many past observations. Reference [15] concludes that, for sufficiently large k, the estimator $\mathbb{E}[U^{t+1} X | \overleftarrow{X}^t] = X_{t+1}$ is the identity map that yields the true evolution of x_{t+1}. In the nonequilibrium case, this implies that, as $k \to \infty$, the size $\nu(\overleftarrow{B}^\infty_t)$ vanishes, with only a single ω_t consistent with the infinite set of observations in $\overleftarrow{x}^\infty_t$. As a consequence, there is a unique value $x_{t+1} = [U^1 X](\omega_t)$ for $\omega_t \in \overleftarrow{B}^\infty_t = \{\omega_t\}$. Similarly, $\overleftarrow{\mu}^\infty_t$ is a δ-distribution at ω_t in this case, with $\overleftarrow{\mu}^\infty_{t+1} = P^1 \overleftarrow{\mu}^\infty_t$ also a δ-distribution at $\omega_{t+1} = \Phi^1(\omega_t)$. The conditional distribution $\Pr(U^{t+1} X | \overleftarrow{X}^\infty_t \overleftarrow{x}^\infty_t)$ is then a δ-distribution with support on the single $x_{t+1} = X\big(\Phi^1(\omega_t)\big)$. See Fig. 4(c).

This clearly shows how coordinate embeddings recover the unobserved degrees of freedom and effectively act as an equivalent to the full underlying Platonic state ω. In the ideal case, as the length of an embedding increases to infinity, the corresponding size of the set $\overleftarrow{B}^k_t$ of possible initial conditions goes to zero, as does the variance of the resulting predictive distribution. Thus, there is an a.e. one-to-one correspondence between infinite-length delay embeddings and Platonic system states ω. The associated predictive distribution converges, in a thermodynamic limit, to the true value of the next observable, given by the Wiener projection of $U^{t+1} X$.

Consider, for example, the simple harmonic oscillator. The underlying state ω is two-dimensional: $\omega = (r, p)$. Assume access only to position: $X(\omega) = r$. The evolution of position is described by sine waves $r(t) = \sin(t)$. At a given time instant, we cannot determine the momentum from the instantaneous position alone and, therefore, cannot determine the full underlying state ω. How-

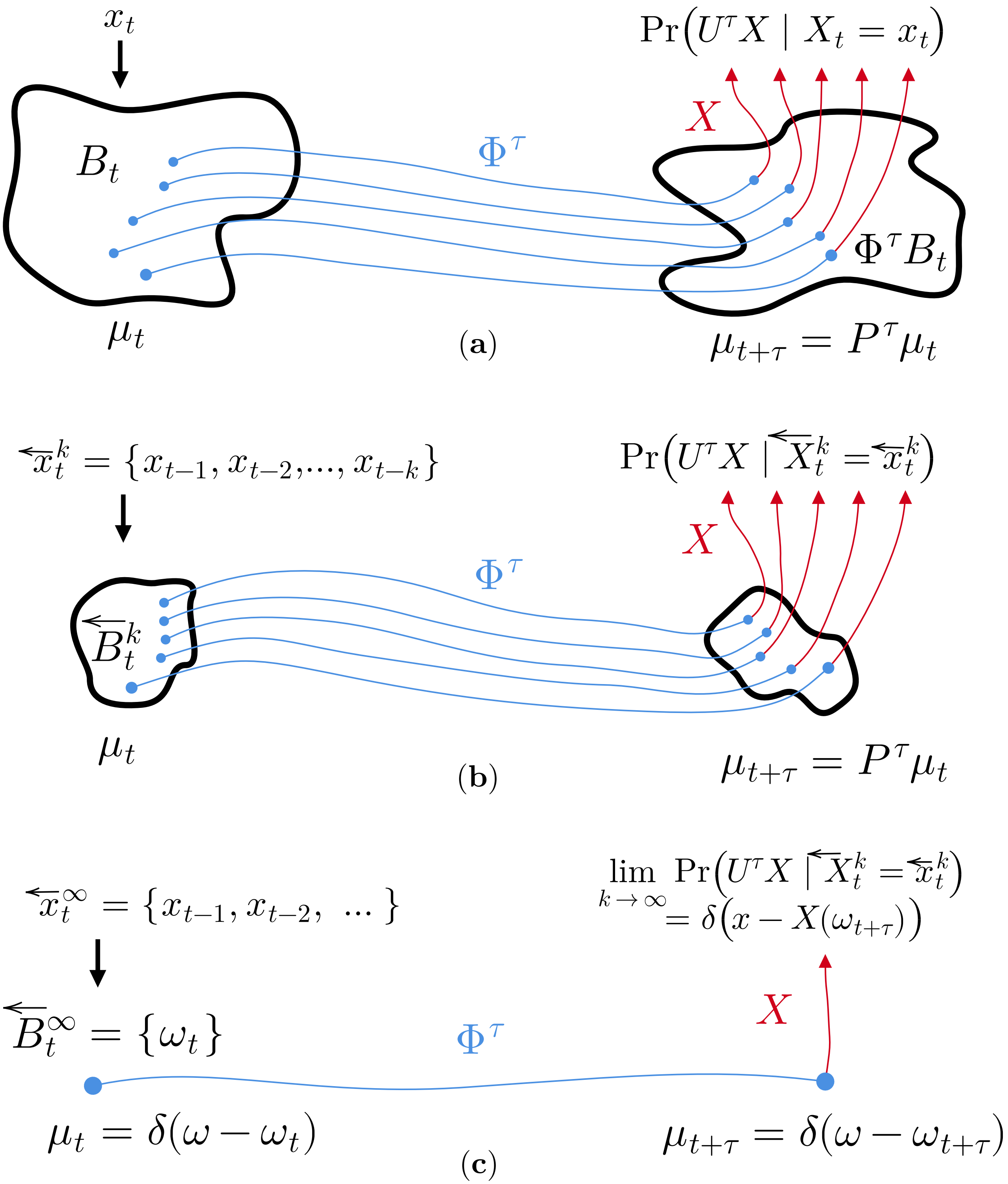

x_t
$\Pr\big(U^\tau X \mid X_t = x_t\big)$
X
B_t
Φ^τ
$\Phi^\tau B_t$
μ_t
(a)
$\mu_{t+\tau} = P^\tau \mu_t$
$\overleftarrow{x}_t^k = \{x_{t-1}, x_{t-2}, \ldots, x_{t-k}\}$
$\Pr\big(U^\tau X \mid \overleftarrow{X}_t^k = \overleftarrow{x}_t^k\big)$
X
Φ^τ
$\overleftarrow{B}_t^k$
μ_t
(b)
$\mu_{t+\tau} = P^\tau \mu_t$
$\lim_{k\to\infty} \Pr\big(U^\tau X \mid \overleftarrow{X}_t^k = \overleftarrow{x}_t^k\big)$
$= \delta\big(x - X(\omega_{t+\tau})\big)$
$\overleftarrow{x}_t^\infty = \{x_{t-1}, x_{t-2}, \ldots\}$
X
$\overleftarrow{B}_t^\infty = \{\omega_t\}$
Φ^τ
$\mu_t = \delta(\omega - \omega_t)$
(c)
$\mu_{t+\tau} = \delta(\omega - \omega_{t+\tau})$

FIG. 4. Operator-theoretic construction of predictive distributions: (a) Instantaneous case, with depth-$k = 0$ past consisting of a single observation x_t. (b) Intermediate case, with depth-k past $\overleftarrow{x}_t^k$. (c) Limiting case, with $\lim k \to \infty$ depth past $\overleftarrow{x}_t^\infty$. In all cases, the initial measure μ_t is uniform on the set of all ω_t consistent with observation $\overleftarrow{x}_t^k$. The initial measure is then propagated forward in time with the Perron-Frobenius operator P^τ and, finally, the predictive distribution is given according to the pushforward of $\mu_{t+\tau} = P^\tau \mu_t$ along the observable mapping X. (c) depicts the case when there is one and only one initial Platonic state ω_t consistent with the ∞-length observation $\overleftarrow{x}_t^\infty$. Then, the prediction converges to the true evolution $x_{t+\tau} = X\big(\Phi^\tau(\omega_t)\big) = [U^\tau X](\omega_t)$.

ever, there are only two momenta associated with each instantaneous position—call them *positive* and *negative*. Following Fig. 4's procedure for constructing predictive distributions, there are two corresponding system states and so two corresponding future position values at time $t + dt$. Call these *up* (for positive momentum) and *down* (for negative momentum). If we include a single infinitesimal past value from time $t - dt$, this is a.e. sufficient to distinguish if the momentum term is positive or negative at that time. Thus, except for the turning points (that are measure zero), including that single past value produces a δ-distribution for the inferred μ_0 and, thus, gives the exact future prediction using Eq. (20).

This convergence, however, is not guaranteed. The size of $\overleftarrow{B}_t^k$ *may* decrease with increasing k, but it does not necessarily always do so. Interestingly, while chaotic instabilities make future predictions challenging, they actually aid in this convergence. Trajectory divergence in forward time [31] means convergence in reverse time. Generating partitions $X_{\mathbb{G}}$ of symbolic processes, detailed in Appendix C, are a rigorous case where this is known to hold [60]. In that setting, $\overleftarrow{B}_t^k$ is the element of the dynamical refinement of the generating partition corresponding to the observed symbol sequence $\overleftarrow{x}_t^k$. In the limit of infinitely-many observations the size $\nu(\overleftarrow{B}_t^k)$ of the refined partition elements vanishes and almost-every infinite-length symbol sequence corresponds to a unique system state $\omega \in [0, 1]$. Generating partitions on chaotic maps of the unit interval are a rigorous case where the a.e. convergence of $\overleftarrow{B}_t^\infty$ to a single $\{\omega_t\}$ is achieved.

X. SUPPORTING EXAMPLES

We now provide examples to demonstrate the ability of delay-coordinate embeddings and Wiener projections of Koopman operators to dynamically encode unobserved degrees of freedom in practice. The data-driven models for these demonstrations mostly employ analog forecasting, Eq. (9), due to its simplicity and flexibility. Various classes of analog forecasting target functions are formed based on what inputs are given, with appropriate distances computed to find the analog. Fully-observed instantaneous, partially-observed instantaneous, and partially-observed history-dependent target functions are all considered.

We emphasize that our theoretical framework applies to *all* history-dependent data driven models. Analog forecasting, as one such model, is ideal for our demonstrations due to how transparently different types of input—e.g., delay embeddings—are used to make predictions and how predictions made from different inputs can be compared in a straightforward manner. In practice, analog forecasting may not be the most efficient or effective method for history-dependent data-driven forecasting [61].

A. Lorenz 63

First, we demonstrate that delay coordinate embeddings can act to fill-in missing degrees of freedom in a dynamically meaningful way; Fig. 5 provides a dynamical complement to Fig. 2's geometric demonstration of encoding unobserved degrees of freedom in the Lorenz 63 system. That is, analog forecast target functions for both cases shown to be geometrically equivalent in Fig. 2 have the same predictive skill.

As a baseline, consider the fully-observed case with $X(\omega_t) = \omega_t = (x_t, y_t, z_t)$. Figure 5(a) shows an instantaneous analog forecast that employs the full state variable as $\mathcal{T}(x_t, y_t, z_t)$. This is plotted alongside a numerical integration of the equations of motion in units of Lyapunov time.

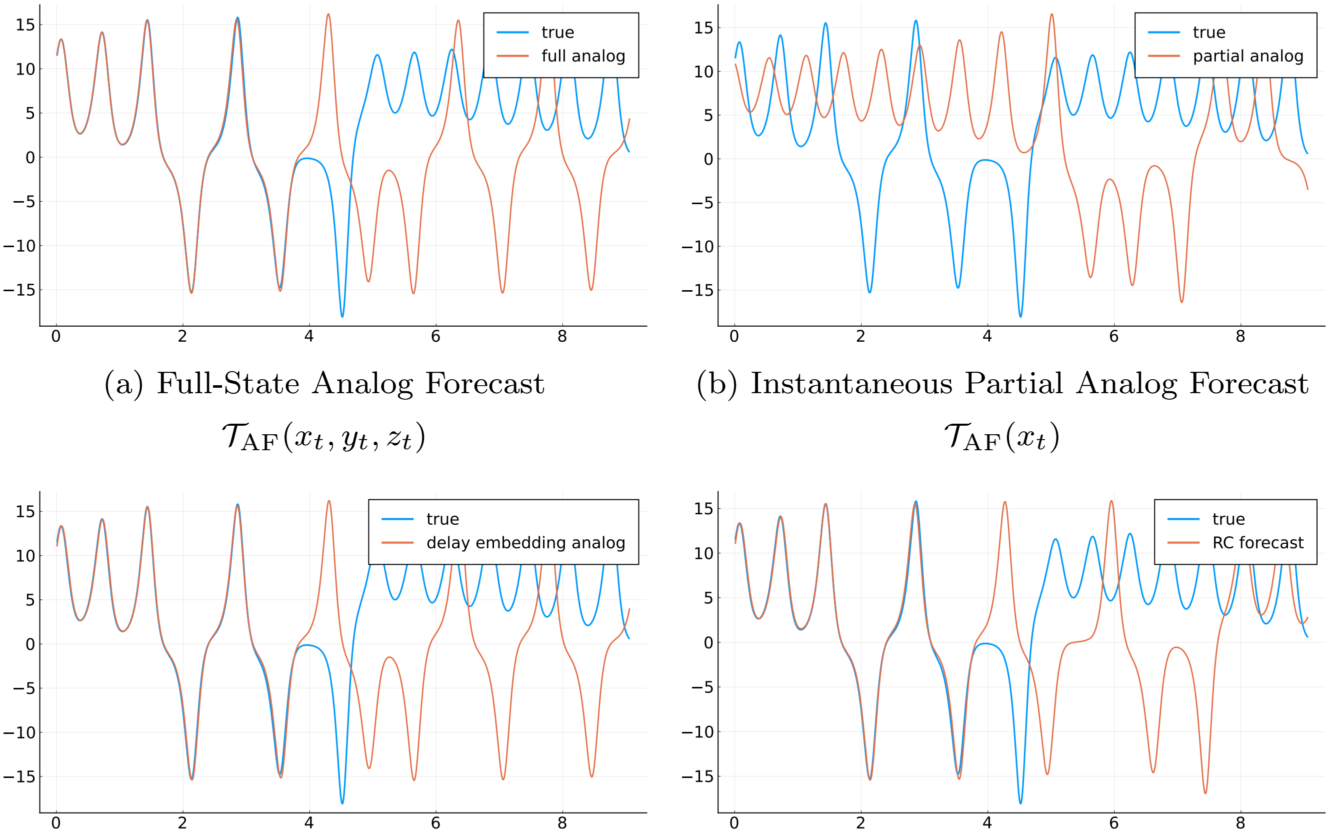
For comparison, Fig. 5(b) shows an instantaneous analog forecast via $\mathcal{T}(x_t)$ using only the first coordinate x. While the analog forecast using the full state variable (x, y, z) tracks the numerical integration for almost four Lyapunov times, the analog forecast using only the instantaneous x variable diverges immediately. Despite this quantitative divergence, the analog forecast using only x produces a qualitatively consistent forecast, capturing behavior expected of the Lorenz system with oscillations within, and jumps between, the two attractor lobes.

Fig. 5(c) shows an analog forecast using delay-coordinate embeddings $\overleftarrow{x}_t^m$ of the x variable with $\mathcal{T}(x_t, x_{t-\delta}, \ldots, x_{t-(m-1)\delta})$. This is, in fact, the same delay embeddings used above in Fig. 2(b) with $\delta = 4$ and $m = 7$. As with the forecast shown in Fig. 5(b), the forecast in (c) only has access to information in the x variable. However, as can be seen, including information from past observations of x, in the form of delay embeddings, results in a model with essentially the same predictive skill as the fully-observed case shown in (a), with divergence again occurring at about four Lyapunov times.

This shows that, at least in simple low-dimensional cases, delay embeddings fill in the gaps and so become effective proxies for the missing degrees of freedom in implicit predictive models. As far as we are aware, the equivalence of predictive skill between an instantaneous full-state model and a history-dependent partially-observed model has not been previously demonstrated.

To emphasize the generality of our theoretical results, Fig. 5 (d) shows a reservoir computing forecast of the x variable alone using an echo-state neural network [62]. As a form of recurrent neural network, echo-state network reservoir computers are history dependent, $\mathcal{T}^{\tau}_{\mathrm{RC}}(x_t, x_{t-1}, \ldots, x_{t-k})$, with a "fading memory" (or "echo state") property that means they do not depend on values arbitrarily-far in the past [63]. Although inputting past x values is not exactly the same as the delay embedding used in Fig. 5 (c), we see a similar forecast.

The connection between delay-coordinate embed-



(a) Full-State Analog Forecast $\mathcal{T}_{\mathrm{AF}}(x_t, y_t, z_t)$

(b) Instantaneous Partial Analog Forecast $\mathcal{T}_{\mathrm{AF}}(x_t)$

(c) Delay Embedding Analog Forecast $\mathcal{T}_{\mathrm{AF}}(x_t, x_{t-\delta}, \ldots, x_{t-(m-1)\delta})$

(d) Reservoir Computing Forecast $\mathcal{T}_{\mathrm{RC}}(x_t, x_{t-1}, \ldots, x_{t-k})$

FIG. 5. Analog forecasting the Lorenz 63 system using different prediction variables. (a) All three Lorenz variables (x, y, z) are used to compute the analog, with only the trajectory of the x variable shown. (b) The x variable alone is used to compute the analog. (c) Delay embeddings of the x variables alone, using the same embedding parameters $\delta = 4$ and $m = 7$ used in Fig. 2(b). The forecast using delay-embeddings in (c) is essentially identical to the forecast using the full system state in (a). (d) A forecast of the x variable alone using an echo-state network reservoir computer is given for comparison. The forecasts of the reservoir computer are made as functions of past observations of x, although it is not the same delay embedding used in the analog forecast in (c).

dings and reservoir computing was explored recently in Ref. [64]. It was found that reservoir computers are, in fact, equivalent to nonlinear regression on delay embeddings. As Ref. [65] pointed out, a linear regression on delay embeddings is essentially equivalent to predictions made using Hankel DMD [11]. The linear feature vector of delay embeddings used for regression is a flattened Hankel matrix. Thus, the nonlinear feature vectors used in Ref. [64] show that reservoir computer network predictions are akin to an extended Hankel DMD method (EHDMD). Note that a connection between EDMD and reservoir computers was utilized in Ref. [66] for finite-dimensional Koopman approximation.

As detailed in Appendix D, the extended dynamical mode decomposition (EDMD) reconstructs a finite-dimensional approximation of the Koopman operator using a *dictionary* of basis functions for finite subspace projection. The approximated EDMD operator is a generalized Galerkin projection of the Koopman operator in the finite subspace spanned by the dictionary functions. The original dynamic mode decomposition (DMD) is a projection into a finite subspace spanned by linear monomials, with a dictionary of just the identity function. We can thus understand Hankel DMD as an approximation using linear finite subspace Wiener projections. An extended Hankel DMD thus would use Wiener projections with a larger finite subspace spanned by nonlinear functions on delay coordinate embeddings. Recall that optimal history-dependent data-driven models in Eq. (20) are given by the full nonlinear Wiener projections that finite subspace projections limit to with increasingly large dictionaries. Thus the nonlinear regression of delay embeddings in Ref. [64] will limit to optimal prediction with increasingly-many nonlinear feature vectors. This regression on delay embeddings is in fact very similar to the NARMAX method in Ref. [6] that motivated the introduction of Wiener projections.

B. Lorenz 96

The Lorenz 63 model is useful to connect the geometric encoding of the unobserved degrees of freedom by delay embeddings with the dynamical encoding of the unobserved degrees of freedom by history-dependent models. However, it is a low-dimensional system, so that even when just a single degree of freedom is accessible, there are only two that are inaccessible.

In this example we demonstrate the effects in increasing-length input pasts with a higher-dimensional system using a 50-dimensional Lorenz 96 model. Each degree of freedom x^i in the model evolves according to local interactions as

$$\frac{dx^i}{dt} = (x^{i+1} - x^{i-2})x^{i-1} - x^i + F \ ,$$

with periodic boundary conditions. We set $F = 4.6$ and numerical integration is performed with a time step $\Delta t = 0.01$. Again we use analog forecasting as the data-driven model, which has access to only a single degree of freedom $x_t = X(\omega_t) = x^1_t$.

Figure 6 shows four analog forecast predictions of x^1_t in the Lorenz 96 system using history-dependent target functions of the form $\mathcal{T}(x_t, x_{t-1}, \ldots, x_{t-k})$. Each prediction is made with a different value of past depth k. As delay-coordinate embeddings, the embedding dimension is k and we use a unit lag $\delta = 1$. Like those in Fig. 5, predictions are made iteratively, with the target functions outputting a single prediction at the next time step.

The lowest-memory prediction is made with $k = 3$ and is shown in Fig. 6(a). The forecast follows the numerical integration for over 200 integration time steps before diverging. The mean-squared error of the prediction over the 1000 time-step window is 3.35. Predictions made with $k = 20$ are shown in Fig. 6(b). While it diverges from the numerical integration sooner than the $k = 3$ model, the quasi-periodicity of the Lorenz 96 system allows the forecast to closely track the numerical integration again at later times. The mean-squared error for the $k = 20$ model is thus slightly lower, at 3.18. Increasing the past depth to $k = 80$, shown in Fig. 6(c), provides a more noticeable improvement in predictive skill. This is apparent visually, and is reflected in the mean-squared error value of 1.72. Finally, increasing to $k = 120$, roughly the ideal $2N + 1$ embedding dimension, provides a dramatic improvement. The data-driven model closely follows the numerical integration for most of the 1000 time-step prediction window, giving a mean-squared error of 0.15.

Similar results are shown in Fig. 4 of Ref. [15] for a 5-dimensional Lorenz 96 model with $F = 8.0$. The authors use the more sophisticated kernel analog forecasting algorithm, first introduced in Ref. [57]. Related results can also be found in Ref. [67].

We emphasize again that convergence of history-dependent data-driven models to the true dynamics is not guaranteed. The results shown in these experiments should not be expected as generic behavior for data-driven models. However, they serve as clear demonstrations of the ability for data-driven models to implicitly encode the effects of unobserved degrees of freedom. This ability provides a physical basis for the efficacy of data-driven predictive models, and provides a bridge connecting them to physics-based models.

C. Transient Dynamics of the Circle Map Lattice

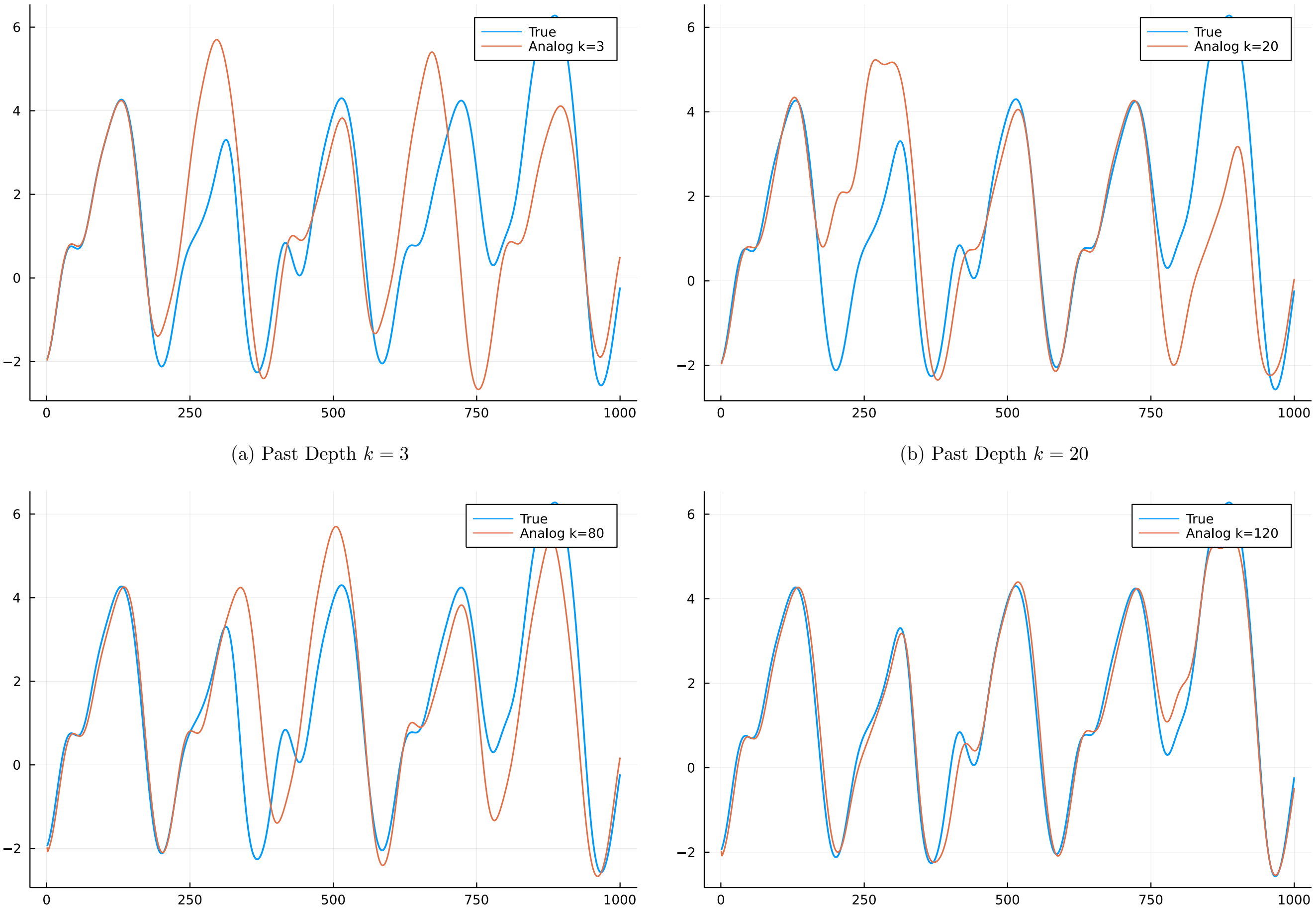
The Lorenz-63 and Lorenz-96 models are the standard proving grounds for data-driven forecasting of dynamical systems. As we have seen, they are useful for demonstrating the convergence of delay embedding models to the true dynamics of the full system. That said, the above examples involve forecasting asymptotic dynamics on an attractor. As we have emphasized throughout, however, our formalism does not require an asymptotic invariant measure that corresponds to equilibrium dynamics on an attractor. Our MaxCal measures also apply to transient nonequilibrium dynamics that include relaxation to an asymptotic attractor.

Nonequilibrium transient dynamics are particularly relevant to spatially-extended dynamical systems for which the relaxation time to an attractor may grow faster than exponential in system size [68]. Many real-world forecasting problems, particularly in Earth Sciences, involve forecasting components of large spatial systems. In such cases, it is not a given that there is an asymptotic invariant measure. Nonetheless, data-driven models have been empirically successful in forecasting real-world spatial systems [21–24]. Crucially then, our formalism applies to large spatial systems that are not guaranteed to be evolving on an invariant attractor.

Here, we demonstrate nonequilibrium transient dynamics of spatial systems using a coupled map lattice [68]. A one-dimensional *map lattice* is a spatially-extended dynamical system that evolves configurations on a discrete spatial lattice $\mathbb{Z}$ in discrete time steps according to the local dynamics:

$$\begin{aligned}\omega(\mathbf{r}, t+1) = (1-\alpha) f\big(\omega(\mathbf{r}, t)\big) \\ + \frac{\alpha}{2}\big[f\big(\omega(\mathbf{r}+1, t)\big) + f\big(\omega(\mathbf{r}-1, t)\big)\big],\end{aligned} \tag{24}$$

where $\mathbf{r}$ is the spatial index, t is the time index, α is the coupling strength, and f is an iterated map of the unit interval: $\omega(t+1) = f\big(\omega(t)\big)$, $\omega \in [0, 1]$. We use the circle



(a) Past Depth $k = 3$

(b) Past Depth $k = 20$

(c) Past Depth $k = 80$

(d) Past Depth $k = 120$

FIG. 6. Analog forecasting the Lorenz 96 system using input pasts of increasing depth. All predictions are made iteratively with analog forecast target functions of the form $\mathcal{T}(x_t, x_{t-1}, \ldots, x_{t-k})$. Model predictions as a function of integration time-steps are given for four values of past depth k.

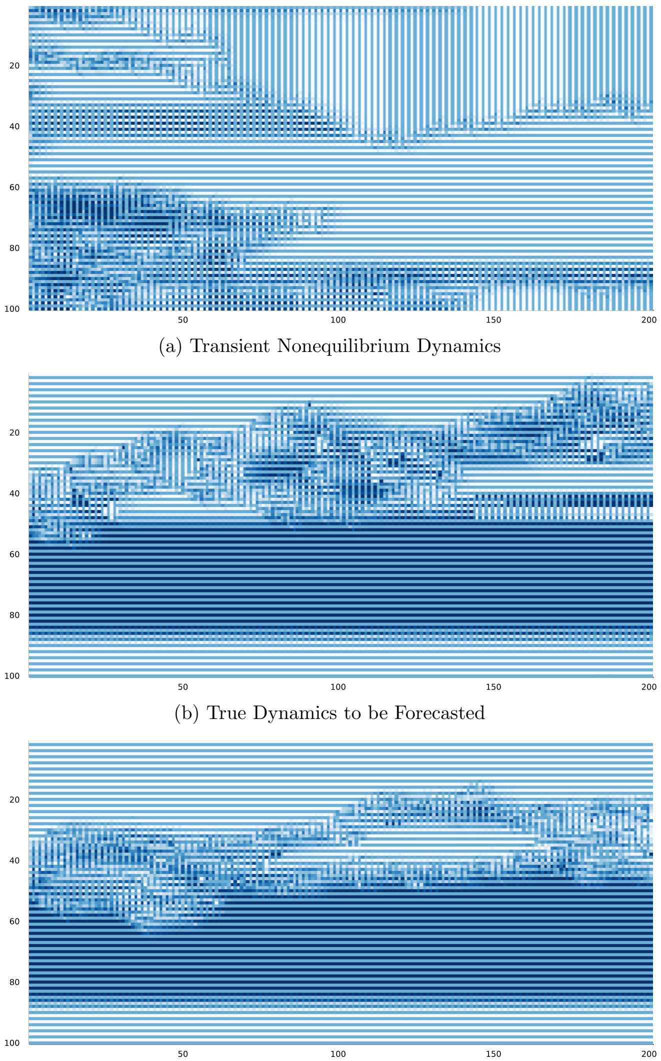
map [69]:

$$f(\omega) = \omega + \gamma - \frac{K}{2\pi}\sin(2\pi\omega) \mod 1 \ ,$$

where γ is a phase shift and K is the strength of the nonlinearity.

An example of the transient dynamics of the circle map lattice is shown in Fig. 7 (a), with phase shift $\gamma = 0.5$, nonlinearity $K = 1.0$, and coupling strength $\alpha = 1.0$. The spatial lattice has $N = 100$ sites and has periodic boundary conditions. The vertical axis is the spatial dimension and the horizontal axis is time, evolving forward from left to right. The dynamics can be qualitatively described in terms of stable *domain* regions with domain wall *dislocations* between them [70]. In this case, the domains correspond to regions that are period-2 in time, space, or both. As can be seen around $\mathbf{r} = 70$ and $t = 100$ in Fig. 7 (a), wandering dislocations may undergo pairwise annihilation, resolving into a now-stable domain region. Thus, the dynamics asymptotically limit to either a single domain or multiple domains with stable dislocation interfaces; potentially with a single unstable dislocation meandering through space. Empirically, a single spatially-periodic domain is the most common attractor when evolved from IID random initial conditions.

With a spatial lattice of $N = 100$ sites, much of the transient behavior resolves into domains after just 100 time steps using nonlinearity $K = 1.0$ and coupling $\alpha = 1.0$. To train a data-driven model to predict transient dislocation behaviors, the model could be trained on an ensemble of many short-time runs with different initial conditions. Alternatively, we used slightly perturbed parameters with $K = 0.96$ and $\alpha = 1.002$ that produce qualitatively similar behavior, but with markedly longer-lived transients. The conditions that allow for dislocations to merge and resolve into domains are stricter, although after sufficiently many time steps ($\sim 10^6$) most initial conditions resolve to the spatially-periodic domain attractor.



(a) Transient Nonequilibrium Dynamics

(b) True Dynamics to be Forecasted

(c) Data-Driven Forecast

FIG. 7. Circle-map coupled-map lattice: (a) Transient nonequilibrium dynamics. (b) True dynamics of a particular instance of the map lattice that is forecasted in (c) using a composition of local analog forecasts.

Since the transient dynamics clearly manifest when viewing the system's full spatial evolution, we performed a composite data-driven prediction of the full system using partial observations in the following way. With $N = 100$ spatial sites, the systems can be considered as 100 coupled time series. Forecasts are made for each lattice site $n \in \{1, 2, \ldots, 100\}$ and the predicted time series are then shown together as an evolving spatial field in Fig. 7 (c). Each individual forecast, however, is made using an analog forecast of partial observation delay embeddings. In particular, for each site n the past values (delay embeddings) of the site plus the past values of all its 10 left nearest neighbors $\{n-1, n-2, \ldots, n-10\}$ and all its 10 right nearest neighbors $\{n+1, n+2, \ldots, n+10\}$ are used as inputs to $\mathcal{T}^{\tau}_{\mathrm{AF}}$. This amounts to combining 21 delay-embedding vectors. Therefore, principle component analysis (PCA) is used for dimensionality reduction for more efficient calculation of the analogs [71]. The global evolution of all 100 sites for 10^4 time steps is used for training. The forecast was performed using the Julia package TimeSeriesPrediction.jl, with `cubic_shell_embeddings` creating the local nearest-neighbor delay embeddings and their PCA reductions.

Comparing the composite forecast shown in Fig. 7 (c) with the true map lattice evolution in Fig. 7 (b), we see that while the details of the dislocation dynamics do not match, the over qualitative behavior is captured. In particular, the stable spatially-periodic domains are perfectly captured, with an imperfectly captured dislocation running between.

The difficulty in predicting the detailed behavior of dislocations is not surprising, particularly for analog methods. Using results from ergodic theory, particularly Kac's lemma, Ref. [61] shows the data requirements for a good analog forecast grow exponentially with the number of degrees of freedom for asymptotic dynamics on an attractor.

In light of this, the physically-reasonable analog forecasts shown Fig. 7 (c) are made possible by two factors. First, the local interactions imply the effective degrees of freedom for making local predictions is smaller than the number of sites on the lattice. In this case, the dynamics of an individual site depends only on the site's value and those of its immediate nearest neighbors. Second, the nonlinearity and coupling strength parameters we have chosen extend the transient dynamics so that there is a local *quasistationary* measure [68]. Although the global dynamic is out of equilibrium with transient dislocation behavior, the local dynamics that produce this behavior do not change over time. The system is conditionally stationary.

We emphasize that analog forecasting is used for this demonstration due to its transparency, particularly with the clear dependence on delay coordinate embedding inputs. Again, the generality of our theoretical framework applies to any history-dependent data-driven model of spatially-extended dynamical systems, like the popular Kuramoto-Sivashinsky model [71–73] or the growing body of work on turbulent fluid flows [74].

XI. A UNIFIED FRAMEWORK

Taken all together, our development provides a unified framework for modeling complex systems from partial observations. Now, with it laid out, we can connect physics-based and data-driven prediction. We find that the two, seemingly disparate, paradigms fall at two extremes of the same spectrum: physics-based models are fully *explicit* and data-driven models are fully *implicit*.

Recall that physics models reconstruct a coarse-grained data image $u_t = \overleftarrow{a}(\overleftarrow{x}^{k}_{t})$ that explicitly fills in missing degrees of freedom using data assimilation over past observations. The model dynamic $\widetilde{\Phi}^\tau_\alpha$ evolves data images u_t by explicitly computing the interactions among its degrees of freedom. In this way, the orbits of data images are generated by:

$$\begin{aligned} u_{t+\tau} &= \widetilde{\Phi}^\tau_\alpha(u_t) \\ &= \widetilde{\Phi}^\tau_\alpha\big(\overleftarrow{a}(\overleftarrow{x}^k_t)\big) \ . \end{aligned}$$

This is reminiscent of Hilbert space models $\overleftarrow{\mathcal{T}}^\tau(\overleftarrow{x}^k)$, although the above technically evolves data images. However, due to the explicit nature of data images we can define *simulated measurements* $\widetilde{x}_t = \widetilde{X}(u_t)$ that produce instrumental readings for the data images. In this way, we predict instrument readings using physics models via:

$$\begin{aligned} x_{t+\tau} &= \widetilde{M}^\tau_0(\overleftarrow{x}^k_t) \\ &= [\widetilde{X} \circ \widetilde{\Phi}^\tau_\alpha \circ \overleftarrow{a}](\overleftarrow{x}^k_t) \ , \end{aligned} \tag{25}$$

Again, this is all implemented *explicitly* in terms of interactions among the observed and inferred unobserved degrees of freedom.

Equation (25) now clearly parallels the optimal history-dependent Hilbert space model:

$$\begin{aligned} x_{t+\tau} &= \overleftarrow{M}^\tau_0(\overleftarrow{x}^k_t) \\ &= [P_W U^\tau X](\overleftarrow{x}^k_t) \ . \end{aligned} \tag{26}$$

Rather than explicitly fill-in the missing degrees of freedom with assimilated data images, data-driven models use the intrinsic geometry of coordinate embeddings to implicitly fill-in the missing variables. Whereas physics models attempt to directly approximate Platonic differential equations-of-motion, data-driven models attempt to approximate the action of the Platonic Koopman operator on embedding coordinates via Wiener projections. And, as we demonstrated, they may converge to the Platonic model in the limit of infinite-length embeddings.

Markovian closure in their dynamics motivated our introducing Platonic models. Model parametrizations, though, are used for physics models if the data images cannot provide adequate closure. Analogously, if a finite-dimensional embedding does not provide adequate closure, there is still a nonzero orthogonal component in the Mori-Zwanzig equation resulting from the Wiener projection on the action of the Koopman operator. In this case, a noise term can be added as a *stochastic parametrization* to alleviate the lack of closure [52].

Between fully-explicit and fully-implicit models lies a spectrum including history-dependent models that combine implicit and explicit modeling. In particular, the spectrum encompasses the recent trend in *physics-informed machine learning* (PIML) [25–27]. There, known physical constraints are explicitly incorporated into the model, usually in the form of conservation laws. The model is then trained from data to implicitly learn the dynamics while maintaining the explicitly enforced constraints.

The unified framework allows clearly evaluating the advantages and disadvantages of various modeling paradigms. Having explicit access to degrees of freedom in physics-based models allows for their direct manipulation in the model. This greatly facilitates, for example, making projections of future climate outcomes under various anthropogenic forcing scenarios. Such uses of physics-based models are becoming increasingly common under the heading of *digital twins* [75, 76].

That said, on the one hand, difficulties arise with physics models. This is particularly the case for prediction and especially when confronted with the poor coverage provided by observed degrees of freedom. Said simply, it is challenging to construct good data images that approximate the system state well. Similarly, if there are many important interactions to track, such as in the climate system, it is impossible to explicitly account for all interactions and so parametrizations are required. Compounding these problems, generally, it is not clear how to construct appropriate or effective parametrizations. Due to all these challenges, implicit approaches are increasingly being added to physics models, particularly to provide data-driven parametrizations [77–80].

On the other hand, though famously difficult to interpret, data-driven models often excel at straightforward prediction and forecasting tasks. This is no longer surprising. The unified framework provided a physical explanation for this success. While data-driven models can converge to the Platonic model in the limit, however, in practice, they must be learned from finite resources. Deep learning models, in particular, can be prohibitively computationally expensive to train. Adding known physical constraints, when applicable, can help such models converge more quickly. The lesson is that models should not implicitly learn already-known features; the latter should be incorporated explicitly.

XII. CONCLUSION

Many modeling applications attempt to predict a physical system's future behavior but can access only a small subset of the system degrees of freedom. Historically, predictions with partial observations have relied on physics-based models that explicitly fill-in missing degrees of freedom using data assimilation and parametrization. In this, the physics models provide approximate solutions to the governing equations of motion—the system's true dynamics or Platonic model. Data-driven approaches, in contrast, learn the dynamics of the observed variables using implicit representations of all the degrees of freedom through delay-coordinate embeddings.

We demonstrated how the maximal predictive information available to a data-driven model—information from past observations of the accessible variables—is given by the predictive distributions. We gave an explicit construction using Koopman and Perron-Frobenius operators. Most data-driven models are Hilbert space models, in the form of a target function, that map past observations forward in time. Maximum Caliber measures were used to develop a nonequilibrium version of Wiener projections for the Mori-Zwanzig formalism. Using this, we showed that optimal Hilbert space models correspond to expectation values of the predictive distributions. Building on the intuition from generating partitions of symbolic processes, this insight illuminated how optimal Hilbert space models converge to the true evolution of the accessible variables, in the limit of infinite-length coordinate embeddings. We also showed how Wiener projections provide a clear theoretical connection between data-driven and physics-based models.

At first blush, the empirical success of data-driven models is counterintuitive. Indeed, by definition, they know nothing of the underlying physics governing the full system. And yet, they still learn, and from only partial observations, to predict the true evolution; i.e., they converge to the Platonic model. Upon reflection, however, we recognize that our understanding and mathematical formulation of physical laws did not spontaneously manifest. They formed and evolved over generations precisely through our observations and interactions with the natural world. The development of science has been data-driven and successful at that. And so, it is not surprising that data-driven models "learn physics" from observations alone.

What is perhaps discomforting, though, is the implicit and often uninterpretable manner in which most data-driven methods learn to approximate the governing physics. We hope that the detailed investigations of implicit models given here alleviates at least some of this puzzle. It must also be remembered that initially there was a great deal of discomfort with explicit numerical physics models. After all, complicated numerical models are very much "black box" in ways similar to data-driven models, particularly deep learning models. The behavior that emerges in complicated physics models often cannot be deduced directly from the inputs given to that model. If a numerical model produces unphysical or otherwise pathological behaviors, it is often not immediately clear how to diagnose and address the concern [36, 37].

Returning to our motivating question, Is there a *best* way to predict a given physical system from partial observations? At present, it does not seem that there is a universally "best" approach. When working with finite data and finite computational resources, all methods have their advantages and disadvantages. We sought to convey the commonality among seemingly disparate approaches to predicting complex systems from partial observations. Our goal was to illuminate the theoretical underpinnings of implicit and explicit models. Finding commonality in a unified predictive framework should help build confidence in the models currently employed. And, hopefully, this will pave the way forward to models with ever more predictive skill and structural interpretability. At which point, the science of complex systems will have moved closer to automated theory building [81].

ACKNOWLEDGMENTS

We thank Nicolas Brodu, Derek DeSantis, and Jordan Snyder for insightful discussions, as well as Stefan Klus, Yen Ting Lin, Balu Nadiga, and Dimiter Vassilev for helpful conversations. The authors also thank the Telluride Science Research Center for hospitality during visits and the participants of the Information Engines Workshops there. JPC acknowledges the kind hospitality of the Santa Fe Institute, Institute for Advanced Study at the University of Amsterdam, and California Institute of Technology for their hospitality during visits. Part of this research was performed while AR was visiting the Institute for Pure and Applied Mathematics, which is supported by the National Science Foundation grant DMS-1440415. AR acknowledges the support of the U.S. Department of Energy through the LANL/LDRD Program and the Center for Nonlinear Studies. This material is also based upon work supported by, or in part by, APRA E Program: Design Intelligence Fostering Formidable Energy Reduction and Enabling Novel Totally Impactful Advanced Technology Enhancements (DIFFERENTIATE) award number DE-AR0001202, Templeton World Charity Foundation grant TWCF0570, Foundational Questions Institute and Fetzer Franklin Fund grant FQXI-RFP-CPW-2007, U.S. Army Research Laboratory and the U.S. Army Research Office grants W911NF-21-1-0048 and W911NF-18-1-0028, and U.S. Department of Energy grant DE-SC0017324.

Appendix A: Ergodicity and Invariant Measures

In contrast to conservative Hamiltonian systems—the default assumption for statistical mechanics—many physical, chemical, and biological systems display dissipative and nonasymptotic behaviors that demand attention for a full understanding. We now define these behaviors in detail. This, in turn, highlights the application breadth of the unified framework.

A system's *phase space* consists of all of its allowed configurations. A primary goal in dynamical systems theory is to identify the key state-space structures that guide and constrain a system's complex behaviors [31]. We wish to capture them explicitly in our development. This requires a slightly more general presentation than is usually given for ergodic theory.

Invariant sets are subsets of a system's states that map onto themselves under a system's dynamic. When perturbations from them return, they are stable invariant sets—called *attractors*. That set of states which tend asymptotically to a given attractor is the attractor's *basin of attraction*. A given dynamical system can be decomposed into its invariant sets including attractors and their basins and the *basin boundaries*. Specifying these objects delineates a system's *attractor-basin portrait*—its comprehensive dynamically-relevant architecture.

Multistable systems are those with multiple attractors. Transient, nonasymptotic behaviors reflect relaxation to an attractor from states starting in its basin. The asymptotic stability of attractors meanwhile allows for the standard long-time analysis of ergodic systems. That is, the standard setup for the measure-preserving and ergodic dynamical systems of interest to us describes the evolution on an attractor.

We now formally define these concepts.

Consider the measure space $(\Omega, \Sigma_\Omega, \nu)$ and a dynamic Φ, where Ω is the state space, Σ_Ω its Borel algebra, and ν the Lebesgue measure. A set $A \subset \Sigma_\Omega$ is a Φ^t*-invariant set* if $\left(\Phi^t\right)^{-1}(A) = A$ for all $t > 0$, where $\left(\Phi^t\right)^{-1}(A)$ is the pre-image of A under Φ^t. In contrast, A is a *forward-invariant* set of Φ^t if for every $\omega \in A$, $\Phi^t(\omega) \in A$ for all $t > 0$. Note that all Φ^t-invariant sets are necessarily also forward-invariant, but not all forward-invariant sets are Φ^t invariant.

An attractor of $(\Omega, \Sigma_\Omega, \nu, \Phi)$ is a set $A \subset \Sigma_\Omega$ with the following properties:

- A is a forward-invariant set of Ω under Φ^t,
- There exists an open set $\mathcal{B} \supset A$, called the basin of attraction of A such that for every $\omega \in \mathcal{B}$, $\lim\limits_{t\to\infty} \Phi^t(\omega) \in A$, and
- There is no proper subset of A with the first two properties.

By definition, an attractor is a forward-invariant set. However, due to the existence of its basin of attraction, an attractor is not a Φ^t-invariant set. There are points $\omega \in \mathcal{B} \setminus A$ that are in the pre-image $\left(\Phi^t\right)^{-1}(A)$ but not in A. However, the full basin of attraction $\mathcal{B}$ for a given attractor A *is* Φ^t-invariant. (The attractor itself is in its basin $A \subset \mathcal{B}$.) Every state in $\mathcal{B}$ limits to its attractor A. And so, if there are states in the pre-image of $\mathcal{B}$ that are not in $\mathcal{B}$ they, by definition, do not limit to A. Therefore, any state not in the pre-image of $\mathcal{B}$ is not in $\mathcal{B}$. In fact, an alternative definition of the basin $\mathcal{B}$ of attractor A is as the limit of pre-images of A: $\mathcal{B} = \lim\limits_{t\to\infty} \left(\Phi^t\right)^{-1} A$.

Attractors and their basins of attraction decompose a dynamical system into its dynamically-independent components—the system's attractor-basin portrait. For a multistable system with multiple attractors, the basins of attraction partition the state space Ω into equivalence classes of states based on the attractor to which they limit since orbits never cross basin boundaries. Without loss of generality, the development considers dynamical systems with a single attractor and Ω its basin of attraction, unless explicitly stated otherwise. For multistable systems, each attractor and its basin may be analyzed separately as if it were its own separate system. The full attractor-basin portrait becomes relevant, though, when one executes independent experimental trials that select a wide range of initial states. Moreover, real-world systems are never fully isolated and this typically introduces fluctuations that can drive a system between otherwise noncommunicating basins.

Decomposing a dynamical system into independent components raises the issues of ergodicity and ergodic measures [33]. A dynamical system $(\Omega, \Sigma_\Omega, \mu, \Phi)$ is *ergodic* and μ is an *ergodic measure*, if every Φ^t-invariant set B is such that $\mu(B) = 1$ or $\mu(B) = 0$. For an ergodic system, all Φ^t-invariant sets are trivial subsets of Ω. From the definition of basins of attraction, a dynamical system with a single basin of attraction or a multi-stable system restricted to a single basin is ergodic.

Ergodic theory often considers a dynamical system $(\Omega, \Sigma_\Omega, \mu, \Phi)$ with measure μ to also be *measure-preserving*: $\mu\big((\Phi^t)^{-1}(B)\big) = \mu(B)$ for $B \subset \Sigma_\Omega$. An equivalent statement is that the measure μ is *invariant* under the dynamics Φ.

This is mathematically convenient for casting the behavior of dynamical processes as stationary stochastic processes. However, it is too restrictive for our purposes, as it does not capture relaxation to an attractor. Transient behavior during relaxation to an attractor A is *dissipative* if it involves measurable subsets of $\mathcal{B}$ not in A, known as *wandering sets.* In essence, measure is "carried away" by wandering sets, and so the support of an invariant measure cannot include wandering sets. This can also be seen from the definitions of invariant measures and the Perron-Frobenius operator above in Eq. (4): a measure is invariant if and only if it is a fixed point of the Perron-Frobenius operator [33, Thm 4.1.1].

It is often of particular concern whether or not the Lebesgue reference measure ν is invariant under the dynamics. Since ν provides a measure of state space volume, dynamics that preserve ν are said to be *volume preserving.* Wandering sets, by definition, preclude volume preservation. Hamiltonian systems, on the other hand, are volume preserving due to Liouville's theorem [4]. Since we consider only ergodic systems, there will always be a physical invariant probability measure that may be used, whether the system preserves volume or not. If the Lebesgue measure is invariant (and so volume is preserved), the microcanonical distribution gives the equilibrium invariant probability distribution. If Lebesgue measure is not invariant, there will be still be a unique asymptotic invariant measure. (More on this shortly.) Our formalism works in all cases, but is particularly useful for generalizing to nonasymptotic behaviors of systems that do not preserve phase space volume.

Note that when considering probability measures, the terminology of *measure-preserving* dynamics should not be confused with what we might call *conservation of measure* (or *conservation of probability*). As standard, we assume the dynamics Φ to be *nonsingular* such that $\mu(\Phi^{-t}(B)) = 0$ for all sets B with $\mu(B) = 0$. This ensures the evolution of probability measures by Perron-Frobenius operators are still probability measures.

To include transient behavior (relaxation to an attractor), the following does not assume an invariant measure. However, by restricting to ergodic dynamics (considering single basins of attraction at a time), this guarantees the existence of a unique, nonsingular (with respect to the Lebesgue volume measure) *asymptotic invariant measure*:

$$\mu_*(B) = \int_B \rho_* d\nu \ .$$

This follows since the L^1 Perron-Frobenius operators have a unique invariant density ρ_* for ergodic dynamics: $P^t \rho_* = \rho_*$ [39, Thm 4.5]. (These measures play a role roughly analogous to equilibrium macrostates in thermodynamics.) That is, this measure is preserved by dynamics on the attractor, to which the system is restricted in the limit. Therefore, in the asymptotic limit the ergodic theorem applies and time averages equal state space averages for observables f:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f\big(\Phi^k(\omega)\big) = \frac{1}{\mu_*(\Omega)} \int_\Omega f(\omega) d\mu_* \ .$$

In the asymptotic limit, the system trajectories settle on the attractor and the resulting dynamical process is distributed according to the asymptotic invariant measure $\mu_*(B)$. Thus, by definition, the process is stationary only in the limit:

$$\begin{aligned} \Pr(X_t \in B_\mathcal{X}) &= \int_{X^{-1}(B_\mathcal{X})} d(\mu_*)_t \\ &= \int_{X^{-1}(B_\mathcal{X})} d(\mu_*)_{t+\tau} \\ &= \Pr(X_{t+\tau} \in B_\mathcal{X}) \ . \end{aligned}$$

Generally, though, $\mu_t \neq \mu_{t+\tau}$.

Note that singular measures will not necessarily limit to the nonsingular invariant measure. In particular, in the above we have made use of singular measures given as Dirac delta distributions and noted that the action of the Perron-Frobenius operator on such measures is equivalent to the dynamics acting on the center of the Dirac distribution. For the deterministic dynamical systems considered here, the action of the Perron-Frobenius operator does not "spread out" singular Dirac measures into nonsingular measures.

Appendix B: Statistical Mechanics Treatment of Mori-Zwanzig

This appendix briefly outlines the original statistical-mechanical derivation of the Mori-Zwanzig equation. This also gives context for the Koopman operator in Hamiltonian systems that are perhaps more familiar in statistical mechanics. This largely follows Secs. 10.3 and 10.4 in Ref. [4].

Recall that in Hamiltonian mechanics there are $2N$ degrees of freedom corresponding to canonical positions q_k and momenta p_k that evolve according to Hamilton's equations:

$$\dot{q}_k = \frac{\partial \mathcal{H}}{\partial p_k} \quad \text{and} \quad \dot{p}_k = -\frac{\partial \mathcal{H}}{\partial q_k} \ ,$$

$k = 1, \ldots, N$. As above, now consider the evolution of functions (observables) $A(q,p,t)$ and densities $\rho(q,p,t)$.

Unlike the flow map semigroups used above, Hamiltonian dynamics are specified by a set of differential equations. Application of a Koopman operator to Hamiltonian systems thus requires the infinitesimal generator $\mathcal{L}$ of the Koopman semigroup, defined as:

$$\mathcal{L}A := \lim_{t \to 0} \ \frac{1}{t} \ \left[U^t A - A\right] \ .$$

The Koopman generator $\mathcal{L}$ is the Lie derivative of an observable A along the vector field $\Phi(\omega)$ when ω evolves according to the differential equation [82]:

$$\begin{aligned} \dot{\omega} &= \frac{d}{dt}\omega \\ &= \Phi(\omega) \ . \end{aligned}$$

The time derivative of observables, given by the Koopman generator, can be computed simply using the chain rule:

$$\begin{aligned} \frac{d}{dt}A &= \mathcal{L}A \\ &= \nabla A \cdot \Phi \ . \end{aligned}$$

For Hamiltonian systems then:

$$\begin{aligned} \frac{d}{dt}A(q,p,t) &= \mathcal{L}A \\ &= \sum_k \left[\frac{\partial A}{\partial p_k}\dot{p}_k + \frac{\partial A}{\partial q_k}\dot{q}_k\right] \\ &= \sum_k \left[\frac{\partial A}{\partial q_k}\frac{\partial \mathcal{H}}{\partial p_k} - \frac{\partial A}{\partial p_k}\frac{\partial \mathcal{H}}{\partial q_k}\right] \\ &= \{A, \mathcal{H}\} \\ &= i\widehat{L}A \ , \end{aligned}$$

where $\{\cdot,\cdot\}$ is the familiar Poisson bracket. As expected, we find the Koopman generator $\mathcal{L}$ is dual to the Liouville operator $\widehat{L} = -i\{\cdot, \mathcal{H}\}$ that evolves densities according to the well-known Liouville equation:

$$\frac{d}{dt}\rho(q,p,t) = -i\widehat{L}\rho \ .$$

This is the infinitesimal generator version of Koopman operators being dual to Perron-Frobenius operators—the Liouville operator is the generator of the Perron-Frobenius semigroup for Hamiltonian systems. Note that the Koopman generator $\mathcal{L}$ is sometimes conflated with the Liouville operator in statistical mechanics.

To further underscore the dual perspectives of observables and densities, consider the classical Hilbert space induced by the expectation-value inner product:

$$\begin{aligned}\langle A\rangle &= \int A(p,q)\rho(p,q)\ \ dqdp \\ &= \langle A|\rho\rangle\ ,\end{aligned}$$

with bra-ket notation used for simplicity.

Starting at an initial time $t = 0$, consider the time dependence residing in either the observable A or in the density ρ. (Again, these options are the analogs of the Heisenberg and Schrödinger pictures of quantum mechanics, respectively.) That is:

$$\begin{aligned}\langle A\rangle_t &= \langle A(t)|\rho(0)\rangle \\ &= \langle A(0)|\rho(t)\rangle\ .\end{aligned}$$

And, we can similarly express the time derivative as:

$$\begin{aligned}\frac{d}{dt}\langle A\rangle_t &= \langle i\widehat{L}A(t)|\rho(0)\rangle \\ &= \langle A(0)| - i\widehat{L}\rho(t)\rangle\ .\end{aligned}$$

Having set up the machinery of Hilbert space dynamics for Hamiltonian systems, we now examine the time evolution of densities and observables projected into a subspace of partial observations. As usual, consider the noninvertible observable mapping $X(p,q)$ that represents (functions of) some, but not all, of the canonical positions and momenta—the *accessible* degrees of freedom. The projection operator $\widehat{P} = |X\rangle\langle X|$ can act on both densities ρ and other observables A, yielding the component of each in the subspace of the accessible degrees of freedom. As before, there is an associated orthogonal projection operator $\widehat{Q}$ such that the sum of the two operators is the identity: $\widehat{P} + \widehat{Q} = \widehat{I}$.

First, consider the evolution of densities in the accessible subspace. Inserting the above identity into the Liouville equation yields:

$$\begin{aligned}\frac{d}{dt}|\rho(t)\rangle &= -i\widehat{L}|\rho(t)\rangle \\ &= -i\widehat{L}(\widehat{P}+\widehat{Q})|\rho(t)\rangle \\ &= -i\widehat{L}\big[|\rho_X(t)\rangle + |\rho_O(t)\rangle\big]\ ,\end{aligned}$$

where $|\rho_X(t)\rangle = \widehat{P}|\rho(t)\rangle$ is the accessible component of the density and $|\rho_O(t)\rangle = \widehat{Q}|\rho(t)\rangle$ is the orthogonal inaccessible component.

Next, apply the projection operator $\widehat{P}$ to both sides of the expanded Liouville equation to project the dynamics of $|\rho(t)\rangle$ onto the accessible degrees of freedom. This gives the time derivative of $|\rho_X(t)\rangle$:

$$\begin{aligned}\widehat{P}\frac{d}{dt}|\rho(t)\rangle &= \frac{d}{dt}|\rho_X(t)\rangle \\ &= -i\widehat{P}\widehat{L}\big[|\rho_X(t)\rangle + |\rho_O(t)\rangle\big]\ .\end{aligned} \tag{B1}$$

The first equality follows since the projection operator does not have explicit time dependence and so commutes with the time derivative.

Applying the orthogonal projection $\widehat{Q}$ gives the derivative of the orthogonal component $|\rho_O(t)\rangle$:

$$\frac{d}{dt}|\rho_O(t)\rangle = -i\widehat{Q}\widehat{L}\big[|\rho_X(t)\rangle + |\rho_O(t)\rangle\big]\ .$$

Similar to the Dyson expansion, Laplace transforms can be used to solve for $|\rho_O(t)\rangle$, yielding:

$$|\rho_O(t)\rangle = e^{-i\widehat{Q}\widehat{L}t}|\rho_O(0)\rangle - i\int_0^t e^{-i\widehat{Q}\widehat{L}\tau}\widehat{Q}\widehat{L}|\rho_X(t-\tau)\rangle\ d\tau\ .$$

Substituting into the expression for the time derivative of $|\rho_X(t)\rangle$ in Eq. (B1) gives the desired kinetic equation [83]:

$$\begin{aligned}\frac{d}{dt}|\rho_X(t)\rangle =& -i\widehat{P}\widehat{L}|\rho_X(t)\rangle \\ &- \int_0^t \widehat{P}\widehat{L}e^{-i\widehat{Q}\widehat{L}\tau}\widehat{Q}\widehat{L}|\rho_X(t-\tau)\rangle\ d\tau \\ &- i\widehat{P}\widehat{L}e^{-i\widehat{Q}\widehat{L}t}|\rho_O(0)\rangle\ .\end{aligned} \tag{B2}$$

This has the familiar form of, in order, a Markov term that depends only on the current value of the accessible component, a memory convolution of the accessible component, and a residual orthogonal component from the initial conditions. Note that Eq. (B2) is exact; no approximations were used.

Obtaining the Mori-Zwanzig equation applies the time derivative of $|\rho_X(t)\rangle$ in Eq. (B2) to compute the equation of motion for the expectation value $\langle X\rangle(t)$ of the accessible degrees of freedom. Plugging Eq. (B2) into the expression:

$$\frac{d}{dt}\langle X\rangle(t) = \Big\langle X\frac{d}{dt}|\rho_X(t)\Big\rangle \tag{B3}$$

yields three terms: Markovian, memory, and orthogonal.

The Markov term is:

$$\begin{aligned}\langle X| - i\widehat{P}\widehat{L}|\rho_X(t)\rangle &= \langle X| - i\widehat{P}\widehat{L}\widehat{P}|\rho(t)\rangle \\ &= \langle X|(-i\widehat{L})|X\rangle\langle X|\rho(t)\rangle \\ &= \langle\dot{X}|X\rangle\langle X\rangle(t) \\ &:= \Gamma\langle X\rangle(t)\ .\end{aligned} \tag{B4}$$

The orthogonal term is:

$$\begin{aligned}\langle X|\widehat{P}(-i\widehat{L})e^{-i\widehat{Q}\widehat{L}}\widehat{Q}|\rho(0)\rangle &= \langle i\widehat{L}X|e^{-i\widehat{Q}\widehat{L}t}\widehat{Q}|\rho(0)\rangle \\ &= \langle e^{i\widehat{Q}\widehat{L}t}\widehat{Q}i\widehat{L}X|\rho(0)\rangle \\ &= \langle F(t)|\rho(0)\rangle \\ &= \langle F\rangle(t) \ ,\end{aligned} \tag{B5}$$

where we have used $|F(t)\rangle := |e^{-i\widehat{Q}\widehat{L}t}\widehat{Q}(-i\widehat{L})X\rangle$.

Finally, the integrand of the memory term is given as:

$$\begin{aligned}&\langle X|\widehat{P}(i\widehat{L})e^{-i\widehat{Q}\widehat{L}\tau}\widehat{Q}(-i\widehat{L})\widehat{P}|\rho(t-\tau)\rangle \\ &\quad= \langle X|(i\widehat{L})e^{-i\widehat{Q}\widehat{L}\tau}\widehat{Q}(i\widehat{L}|X\rangle\langle X\rangle(t-\tau) \\ &\quad= \langle i\widehat{L}X|F(\tau)\rangle\langle X\rangle(t-\tau) \\ &\quad= \langle \widehat{Q}i\widehat{L}X|F(\tau)\rangle\langle X\rangle(t-\tau) \\ &\quad= \langle F(0)|F(\tau)\rangle\langle X\rangle(t-\tau) \\ &\quad:= K(\tau)\langle X\rangle(t-\tau) \ ,\end{aligned} \tag{B6}$$

noting that $\widehat{Q}|F(t)\rangle = |F(t)\rangle$, since $|F(t)\rangle$ is in the inaccessible subspace orthogonal to $|X\rangle$.

Combining the three terms in Eqs. (B4), (B6), and (B5), and using a delta weight function in the expectation value inner product, we arrive at the continuous time Mori-Zwanzig equation:

$$\frac{d}{dt}X(t) = \Gamma X(t) - \int_0^t K(\tau)X(t-\tau)d\tau + F(t) \ . \tag{B7}$$

The detailed expressions, as in Eq. (B2), allow for additional insight into the statistical mechanics of the Mori-Zwanzig formalism. In particular, we see the orthogonal subspace propagator $e^{-i\widehat{Q}\widehat{L}t}$ appear in both the orthogonal term and the memory term. For the orthogonal term, this represents degrees of freedom that were initially in the inaccessible orthogonal subspace and remained inaccessible for the duration up to time t. Whereas, its presence in the memory term represents degrees of freedom that were accessible at time $t-\tau$, then dissipated into and propagated through the inaccessible subspace and, finally, returned to the accessible subspace by time t.

Appendix C: Optimal Finite-Precision Instruments for Continuous Observables

The following temporarily leaves behind the fully-continuous dynamical processes setting. Instead, it considers discrete-time, discrete-valued *symbolic processes* [60] and how they relate to discrete measurements of continuous dynamical systems. In this, the mapping X corresponds to a coarse-grain partition $\mathbb{P}$ of the state space Ω. As with dynamical processes, X is many-to-one and noninvertible, yielding fully-discrete stochastic processes of observations. Rather than interpreting X as accessing only a subset of accessible degrees of freedom in ω, for symbolic processes X is interpreted as a collection of measurement instruments, each with access to all relevant degrees of freedom, but only report the result of finite-precision observations [84]. To avoid confusion, we denote the observation function for symbolic processes as $X_{\mathbb{P}}$.

1. Symbolic Processes from Generating Partitions

A symbolic measurement function $X_{\mathbb{P}} : \Omega \rightarrow \mathcal{A}$ generates a finite partition $\mathbb{P}$ of state space Ω, with every $\omega \in \Omega$ mapping to a partition element $\mathbb{P}_i$ such that $\mathbb{P}_i \cap \mathbb{P}_j = \emptyset$ for all $\mathbb{P}_i, \mathbb{P}_j \in \mathbb{P}$ and $\bigcup_i^K \mathbb{P}_i = \Omega$. Each partition element $\mathbb{P}_i$ carries a label, or *symbol* $a_i \in \mathcal{A}$. Without loss of generality, we will take label$(\mathbb{P}_i) = i$, with $\mathcal{A} = \{0, 1, \ldots, K-1\}$ for K partition elements in $\mathbb{P}$. Using this, we can explicitly write the piecewise constant symbolic measurement function $X_{\mathbb{P}}$ in terms of the partition elements as:

$$X_{\mathbb{P}}(\omega) = \sum_{i=0}^{K-1} i \mathbb{1}_{\mathbb{P}_i}(\omega) \ , \tag{C1}$$

where:

$$\mathbb{1}_{\mathbb{P}_i}(\omega) = \begin{cases} 1 & \omega \in \mathbb{P}_i \\ 0 & \omega \notin \mathbb{P}_i \end{cases}$$

is the indicator function for partition element $\mathbb{P}_i$.

Paralleling our development of dynamical processes, we now consider symbol sequences generated by measuring orbits of the underlying system (Ω, Φ). For an initial value ω_0 there is an initial symbol $a_0 = X_{\mathbb{P}}(\omega_0)$—an element of partition $\mathbb{P}$. Similarly, $X_{\mathbb{P}} \circ \Phi$ induces a partition over Ω, denoted $\Phi^{-1}\mathbb{P}$, such that each element $(\Phi^{-1}\mathbb{P})_i$ is the set of all ω for which $X_{\mathbb{P}}\big(\Phi(\omega)\big) = \mathbb{P}_i$. That is, $\Phi^{-1}\mathbb{P}$ is a partition over Ω at the initial time t_0 where every ω_0 in the same element of $\Phi^{-1}\mathbb{P}$ emits the same symbol $a_1 = X_{\mathbb{P}}(\omega_1) = X_{\mathbb{P}}\big(\Phi(\omega_0)\big)$ at the next time t_1. Each time step t_n generates a new partition $\Phi^n\mathbb{P}$ whose elements are all the points $\omega_0 \in \Omega$ such that $X_{\mathbb{P}}\big(\Phi^n(\omega_0)\big) \in \mathbb{P}_i$.

Importantly, an iterated partition *refines* the previous partition. For two partitions $\mathbb{P}$ and $\mathbb{Q}$, the *refinement* $\mathbb{P} \vee \mathbb{Q} = \{\mathbb{P}_i \cap \mathbb{Q}_j$, for all $\mathbb{P}_i \in \mathbb{P}$ and $\mathbb{Q}_j \in \mathbb{Q}\}$ is also a partition. The first refinement of $\mathbb{P}$ under Φ is $\mathbb{P} \vee \Phi^{-1}\mathbb{P}$. Its elements are all the points $\omega_0 \in \Omega$ that emit the same symbol $X_{\mathbb{P}}(\omega_0)$ for time t_0 and that emit the same symbol $X_{\mathbb{P}}\big(\Phi(\omega_0)\big)$ at the next time t_1. Therefore, the refinement $\mathbb{P} \vee \Phi^{-1}\mathbb{P}$ maps from Ω to two-symbol sequences $a_0 a_1$ in $\mathcal{A} \times \mathcal{A}$. In the limit, the full dynamical refinement $\mathbb{P} \vee \Phi^{-1}\mathbb{P} \vee \Phi^{-2}\mathbb{P} \vee \cdots$ maps points in Ω to infinite-length symbol sequences in $\mathcal{A} \times \mathcal{A} \times \mathcal{A} \times \cdots$.

A partition $\mathbb{P}$ is *generating* if there is a one-to-one correspondence, almost everywhere, between an initial condition $\omega_0 \in \Omega$ and the infinite sequence of symbols $\{X_\mathbb{P}(\omega_0), X_\mathbb{P}\big(\Phi(\omega_0)\big), X_\mathbb{P}\big(\Phi^2(\omega_0)\big), \ldots\}$ generated by ω_0. Thus, while the initial measurement symbol $a_0 = X_\mathbb{P}(\omega_0)$ is far from sufficient to fully determine ω_0, the full infinite sequence of subsequent symbols *does* (almost-everywhere) fully determine ω_0 if $\mathbb{P}$ is a generating partition. This occurs since the size of the dynamical refinement partition elements goes to zero in the infinite-time limit. And, in turn, this requires the system to be chaotic; exponential spreading of orbits in forward time corresponds to exponential convergence in reverse time.

Due to all this, generating partitions provide a rigorous notion of a "good" measurement device for which information lost by a coarse single-time measurement is recovered through an infinite-time limit of measurement observations. The one-to-one correspondence property of generating partitions emerges above when discussing the potential convergence of data-driven models of partially-observed systems. The set $\overleftarrow{B}^k_t$ is the element of the dynamical refinement of the generating partition corresponding to the observed symbol sequence $\overleftarrow{x}^k_t$. In the limit of infinitely-many observations the size $\nu(\overleftarrow{B}^k_t)$ of the refined partition elements vanishes and almost-every infinite-length symbol sequence corresponds to a unique system state $\omega \in [0, 1]$.

An important bridge between symbolic and dynamical processes arises from the fact that the evolution of partitions $\Phi^{-n}\mathbb{P}$ (not the dynamical refinements) is governed by discrete-time Koopman operators. The partition $\Phi^{-n}\mathbb{P}$ is generated by the time-shifted symbolic measurement function $X_\mathbb{P} \circ \Phi^n = U^n X_\mathbb{P}$. Therefore, again paralleling dynamical processes, the symbol sequences are given by $\{X_\mathbb{P}(\omega_0), [U X_\mathbb{P}](\omega_0), [U^2 X_\mathbb{P}](\omega_0), \ldots\}$.

2. Generating Partition of the Logistic Map

A common arena for investigating symbolic processes of chaotic dynamical systems considers continuous maps on the unit interval $\Omega = [0, 1]$ [85, 86]. Here, we examine the logistic map:

$$\begin{aligned}\omega_{n+1} &= \Phi(\omega_n) \\ &= r\omega_n(1 - \omega_n) \ ,\end{aligned}$$

We set $r = 4$.

The binary partition $\mathbb{G}$, shown in Fig. 8, with $\mathbb{G}_0 = [0, \frac{1}{2}]$ and $\mathbb{G}_1 = [\frac{1}{2}, 1]$, is a generating partition of the logistic map [86]. The corresponding symbolic measurement function is the step function:

$$X_\mathbb{G}(\omega) = \begin{cases} 0 & 0 \leq \omega \leq 0.5 \\ 1 & 0.5 \leq \omega \leq 1 \end{cases} .$$

Note this function is in the general form of Eq. (C1).

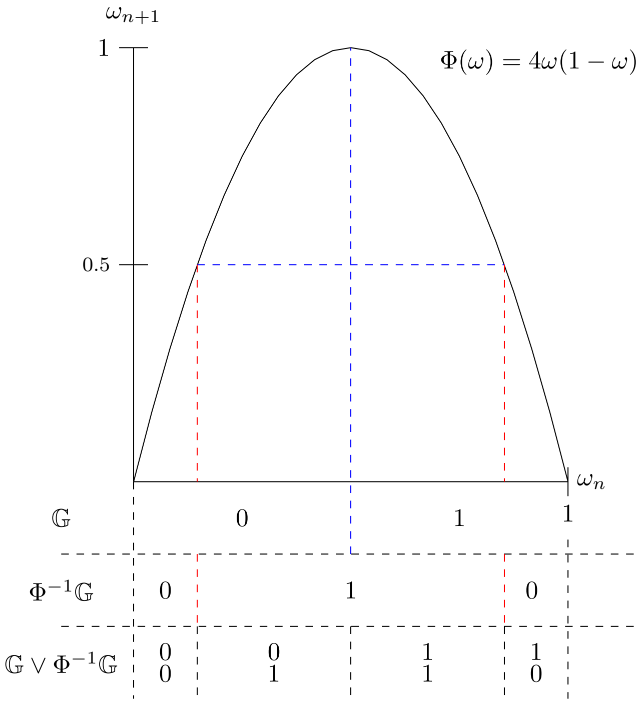


FIG. 8. Logistic map of the unit interval at $r = 4$: Shown with generating partition $\mathbb{G}$, the single-time evolved partition $\Phi^{-1}\mathbb{G}$, and the first dynamical refinement partition $\mathbb{G} \vee \Phi^{-1}\mathbb{G}$.

The single-time evolved partition $\Phi^{-1}\mathbb{G}$, also shown in Fig. 8, is given by the single-time shift symbolic measurement function:

$$X_\mathbb{G}\big(\Phi(\omega)\big) = [U X_\mathbb{G}](\omega) \tag{C2}$$

$$= \begin{cases} 1 & \frac{1-\sqrt{\frac{1}{2}}}{2} \leq \omega \leq \frac{1+\sqrt{\frac{1}{2}}}{2} \\ 0 & \text{otherwise} \end{cases} . \tag{C3}$$

The new boundary points $\left(1 - \sqrt{\frac{1}{2}}\right)/2$ and $\left(1 + \sqrt{\frac{1}{2}}\right)/2$ of $\Phi^{-1}\mathbb{G}$ are the pre-images $\{\Phi^{-1}(\omega)\}$ of the original boundary point $\omega = 1/2$ of $\mathbb{G}$.

Finally, the first dynamical refinement $\mathbb{G} \vee \Phi^{-1}\mathbb{G}$, mapping Ω to two-symbol sequences, is also shown in Fig. 8. From Fig. 8 we see that the dynamical refinement adds the boundary points of $\Phi^{-1}\mathbb{G}$ to the original boundary point of $\mathbb{G}$.

Beyond rigorously formulating good measurement devices—generating partitions—symbolic processes were historically important for introducing concepts and methods from discrete information and computation theories into dynamical systems and ergodic theory, as noted above. In particular, the Shannon entropy rate of a symbolic process has a (possibly nonunique) supremum over all possible partitions for a given iterated map. This is the *Kolmogorov-Sinai entropy*. That is, the supremum is achieved for generating partitions [87, 88]. Moreover, the Kolmogorov-Sinai entropy is bounded by the positive Lyapunov exponents of the underlying system [89].

This provides a rigorous link between the geometric instabilities of deterministic chaos and observed randomness. And, this explains, in part, why the weather is hard to predict [90, 91].

Appendix D: Data-Driven Koopman Approximation

Given that U^τ provides the ground-truth for $\mathcal{T}^\tau$, why not use a data-driven approximation of U^τ for $\mathcal{T}^\tau$? Finite-dimensional approximations of U^t are useful for global spectral analysis of nonlinear systems, but they are typically not optimal for predictive modeling, as we now show.

Data-driven finite-dimensional—i.e., matrix—approximations of U^τ are most generally understood through the *Extended Dynamic Mode Decomposition* (EDMD) algorithm [92, 93] shown in Fig. 9. Consider a dictionary $\mathbf{\Psi} = [\psi^1(x), \ldots, \psi^k(x)]^{\mathrm{T}}$ of basis functions in V. For simplicity, assume this is an orthonormal set so that $\mathbf{\Psi}$ defines the closed Hilbert subspace $H_{\mathbf{\Psi}} \subseteq H_X \subset H$ spanned by $\mathbf{\Psi} \circ X$. Given a set of training data $\{x_0, x_1, \ldots, x_T\}$, EDMD finds a (least squares) best-fit matrix $\mathbf{U}^\tau_X$ such that:

$$\mathbf{\Psi}(x_{t+\tau}) = \mathbf{U}^\tau_X \mathbf{\Psi}(x_t) \ . \tag{D1}$$

This is typically an overdetermined optimization, and so it is common to pick a solution by applying the pseudoinverse $\mathbf{\Psi}^+$, giving:

$$[\mathbf{U}^\tau_X]^{\mathrm{T}} = \mathbf{\Psi}(x_{t+\tau}) \mathbf{\Psi}^+(x_t) \ . \tag{D2}$$

In the infinite data limit, $\mathbf{U}^\tau_X$ converges to a Galerkin projection of U^τ onto $H_{\mathbf{\Psi}}$, so that:

$$\langle \psi_j, U^\tau \psi_i \rangle = \langle \psi_j, \mathbf{U}^\tau_X \psi_i \rangle \ , \tag{D3}$$

for all $i, j = 1, \ldots, k$.

In the fully-observed case, where X is the identity ($X(\omega) = x = \omega$), the Galerkin projection $\mathbf{U}^\tau$ converges to the true Koopman operator U^τ in the limit of an infinitely-large dictionary $\mathbf{\Psi}$, where $H_{\mathbf{\Psi}} \to H$ [94]. However, in the partially-observed case, the dictionary is restricted to functions of partial observations x only. Thus, in the limit of an infinitely-large dictionary $\mathbf{\Psi} \to V$, we only have that $H_{\mathbf{\Psi}} \to H_X$. Therefore, $\mathbf{U}^\tau_X$ cannot converge to the full U^τ. We include the subscript X in $\mathbf{U}^\tau_X$ to signify this fundamental restriction.

Several difficulties arise in using $\mathbf{U}^\tau_X$ as a predictive model. First and foremost, the identity function $f_X(x) = x$ must be included in $\mathbf{\Psi}$. (f_X is sometimes called the *full-state observable* in the Koopman literature, but we do not as it is confusing in the setting of partially-observed systems.) A prediction is then given as:

$$\begin{aligned} x_{t+\tau} &= \mathcal{T}^\tau_{\mathrm{EDMD}}(x_t) \\ &= \mathbf{U}^\tau_X [f_X \circ X](\omega_t) \ . \end{aligned}$$

That is, the forecast is determined by the action of $\mathbf{U}^\tau_X$ on the identity observable $f_X \circ X$.

To be clear, $\mathbf{U}^\tau_X$ is an operator on the Hilbert subspace $H_{\mathbf{\Psi}} \subseteq H_X \subset H$ of observables of the full underlying system Ω. However, due to the partial-observation constraint it must always act on observables composed with X. (And so, it can be thought of as acting on functions of x.) However, the identity observable f_X need not be included in constructing the dictionary $\mathbf{\Psi}$. The constraint of requiring $f_X \in \mathbf{\Psi}$ can be avoided through the use of autoencoder neural networks to construct $\mathbf{U}^\tau_X$ [73, 95, 96]. The decoder of the network learns a nonlinear map from $H_{\mathbf{\Psi}} \to \mathcal{X}$ that recovers f_X as a nonlinear combination of the elements of $\mathbf{\Psi}$.

In practice, the distinction between discrete-time and continuous-time systems can be important. For continuous time, the gEDMD algorithm [97] should be employed to approximate the Koopman generator. This is done by using finite differences or automatic differentiation of the observation time series.

A more serious difficulty in using EDMD for prediction comes from its its lack of closure—leakage out of the subspace $H_{\mathbf{\Psi}}$. If $H_{\mathbf{\Psi}}$ is not a finite Koopman-invariant subspace [98], then after several iterations $U^\tau[f_X \circ X]$ eventually no longer lies within $H_{\mathbf{\Psi}}$. Due to this, $\mathbf{U}^\tau_X$'s action differs from the true evolution given by U^τ's action. Note that if H_X is not a Koopman invariant subspace, then *all* instantaneous models $\mathcal{T}^\tau$ accrue a similar prediction error. This is the intrinsic error Ξ^τ_X discussed above, which is incurred for having only partial observations X of Ω.

In the infinite dictionary limit, $\mathbf{\Psi} \to V$ and so $\mathbf{U}^\tau_X[f_X \circ X] = g \circ X$ is always in $H_{\mathbf{\Psi}} = H_X$, for some $g \in V$. Thus, EDMD converges in the limit to optimal target function—regression function—Z^τ in Eq. (11). Given that the Koopman operator provides the ground-truth for data-driven models, it is not surprising that the EDMD approximation method for U^τ recovers the optimal instantaneous target function.

The difficulty is that EDMD never reaches the $\mathbf{\Psi} \to V$ limit. Therefore, generally the leakage of $\mathbf{U}^\tau_X[f_X \circ X]$ out of $H_{\mathbf{\Psi}}$ may still lie within H_X. Unlike the $\mathbf{\Psi} \to V$ limit, with leakage out of H_X, this leakage is avoidable, given a better choice of or larger dictionary $\mathbf{\Psi}$. Moreover, the prediction error from the leakage compounds over time. The choice of $\mathbf{\Psi}$ thus substantially impacts EDMD's predictive skill. In general, a finite invariant subspace cannot be determined a priori. And, for that matter, may not exist for a given physical system Ω with a given X—the set of measurements that can be made on Ω. Recent deep learning approaches [66, 73, 95, 96] attempt to learn an optimal $\mathbf{\Psi}$ from data. Similarly, kernel methods [99–101] are used to create a very large, implicitly-defined, dictionary.

Note that employing the trivial dictionary $\mathbf{\Psi} = \{f_X\}$, which includes only the identity, yields the exact Dynamic Mode Decomposition (DMD) algorithm [102]. For prediction DMD finds the optimal matrix (i.e. linear) so-

$$\begin{tikzcd}
\underline{\text{System}} & \underline{\text{Observations}} & \underline{\text{Galerkin Approximation}} \\
\omega_0 \in \Omega \arrow[r, "X : \Omega \to \mathcal{X}"] \arrow[ddd, "\Phi^t"] \arrow[dddr, "X_t = U^t X"] & x_0 = X(\omega_0) \arrow[r, "\boldsymbol{\Psi}"] & {[\psi^1(x_0), \dots, \psi^k(x_0)]^{\mathrm{T}}} \arrow[d, "\mathbf{U}_X^t"'] \\
 & & {[\mathbf{U}_X^t \psi^1(x_0), \dots, \mathbf{U}_X^t \psi^k(x_0)]^{\mathrm{T}}} \arrow[dd, dashed, leftrightarrow, "\text{minimize}"] \\
 & & \\
\omega_t = \Phi^t(\omega_0) \arrow[r, "X : \Omega \to \mathcal{X}"] & x_t = X(\omega_t) \arrow[r, "\boldsymbol{\Psi}"] & {[\psi^1(x_t), \dots, \psi^k(x_t)]^{\mathrm{T}}}
\end{tikzcd}$$

FIG. 9. EDMD algorithm's commuting diagram for finite-dimensional Galerkin approximation $\mathbf{U}_X^\tau$ of U^t onto $H_{\boldsymbol{\Psi}} \subseteq H_X \subset H$.

lution for $\mathcal{T}^\tau$ which minimizes the instantaneous target function error in Eq. (8).

For complex, nonlinear systems, using a linear model for prediction may seem like a bad idea. Interestingly, though, "linear plus noise" models, such as Linear Inverse Modeling (LIM) [102], can be reasonably effective and are frequently used in climate science [103]. We are not aware of attempts to generalize this to an Extended Linear Inverse Model that implements EDMD plus noise. The efficacy of LIM models suggests the tolerance induced by noise may help alleviate the effects subspace leakage.

Equations of Motion From Data A popular approach for data-driven modeling learns an explicit closed-form equation model for $\mathcal{T}^\tau$. This is referred to as *equation discovery.* The most common approach performs a dictionary regression; sometimes also called *symbolic regression* [45, 104]. Like EDMD, a dictionary $\boldsymbol{\Psi} = [\psi^i, \dots, \psi^k]$ of functions is chosen and a (typically sparse) regression is performed to find the best-fit coefficients a_i that minimize $\|\dot{x}_t - \sum_i a_i \psi^i(x_t)\|^2$. In fact, the dictionary regression approach to equation discovery is a special case of gEDMD [97].

Whatever form of equation discovery is used, the ultimate goal is to approximate $\dot{x} = \Phi_X(x)$ with a closed-form expression for Φ_X. For partially-observed dynamics, though, it is not guaranteed that Φ_X will be well-represented by closed-form equations of motion, even if Φ is [45]. The insight that dictionary regression is a special case of the gEDMD algorithm for approximating the Koopman generator illustrates that forcing $\mathcal{T}^\tau$ to be closed-form is an unnecessary restriction. There are certainly many advantages to having closed-form models, including interpretability and extracting adjustable physical parameters. In contrast, for prediction our unified framework demonstrates that it is often advantageous to use *implicit* models for $\mathcal{T}^\tau$.

[1] P. N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (The MIT Press, 2010).

[2] G. Dyson, *Turing's cathedral: the origins of the digital universe* (Pantheon, 2012).

[3] T. Berry, D. Giannakis, and J. Harlim, Bridging data science and dynamical systems theory, Notices of the American Mathematical Society **67**, 1336 (2020).

[4] R. Wilde and S. Singh, *Statistical Mechanics: Fundamentals and Modern Applications*, 1st ed. (Wiley & Sons, New York, 1998).

[5] A. J. Chorin, O. H. Hald, and R. Kupferman, Optimal prediction with memory, Physica D: Nonlinear Phenomena **166**, 239 (2002).

[6] K. K. Lin and F. Lu, Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism, Journal of Computational Physics **424**, 109864 (2021).

[7] Y. T. Lin, Y. Tian, D. Livescu, and M. Anghel, Data-driven learning for the mori–zwanzig formalism: A generalization of the koopman learning framework, SIAM Journal on Applied Dynamical Systems **20**, 2558 (2021).

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

[9] F. Takens, Detecting strange attractors in fluid turbu-

lence, in *Symposium on Dynamical Systems and Turbulence*, Vol. 898, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 1981) p. 366.

[10] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, Chaos as an intermittently forced linear system, Nature Comm. **8**, 1 (2017).

[11] H. Arbabi and I. Mezic, Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator, SIAM Journal on Applied Dynamical Systems **16**, 2096 (2017).

[12] D. Giannakis, Data-driven spectral decomposition and forecasting of ergodic dynamical systems, Applied and Computational Harmonic Analysis **47**, 338 (2019).

[13] M. Kamb, E. Kaiser, S. L. Brunton, and J. N. Kutz, Time-delay observables for koopman: Theory and applications, SIAM J. Appl. Dynamical Systems **19**, 886 (2020).

[14] R. Alexander and D. Giannakis, Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques, Physica D: Nonlinear Phenomena **409**, 132520 (2020).

[15] F. Gilani, D. Giannakis, and J. Harlim, Kernel-based prediction of non-Markovian time series, Physica D: Nonlinear Phenomena **418**, 132829 (2021).

[16] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. II, **106**, 620 (1957).

[17] A. J. Chorin, O. H. Hald, and R. Kupferman, Optimal prediction and the mori–zwanzig representation of irreversible processes, Proc. Natl. Acad. Sci. USA **97**, 2968 (2000).

[18] E. T. Jaynes, Macroscopic prediction, in *Complex Systems—Operational Approaches in Neurobiology, Physics, and Computers* (Springer, 1985) pp. 254–269.

[19] W. Grandy, *Entropy and The Time Evolution of Macroscopic Systems*, Vol. 10 (Oxford University Press, 2008).

[20] C. R. Shalizi and J. P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, J. Stat. Phys. **104**, 817 (2001).

[21] F. Kratzert, D. Klotz, M. Herrnegger, and S. Hochreiter, A glimpse into the unobserved: Runoff simulation for ungauged catchments with lstms, in *Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)* (2018).

[22] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences **23**, 5089 (2019).

[23] J. S. Read, X. Jia, J. Willard, A. P. Appling, J. A. Zwart, S. K. Oliver, A. Karpatne, G. J. Hansen, P. C. Hanson, W. Watkins, *et al.*, Process-guided deep learning predictions of lake water temperature, Water Resources Research **55**, 9173 (2019).

[24] X. Jia, J. Zwart, J. Sadler, A. Appling, S. Oliver, S. Markstrom, J. Willard, S. Xu, M. Steinbach, J. Read, *et al.*, Physics-guided recurrent graph model for predicting flow and temperature in river networks, in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* (SIAM, 2021) pp. 612–620.

[25] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, Integrating physics-based modeling with machine learning: A survey, arXiv preprint arXiv:2003.04919 (2020).

[26] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, Physics-informed machine learning, Nature Reviews Physics **3**, 422 (2021).

[27] K. Kashinath, M. Mustafa, A. Albert, J. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, *et al.*, Physics-informed machine learning: case studies for weather and climate modelling, Phil. Trans. Roy. Soc. A **379**, 20200093 (2021).

[28] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261 (2018).

[29] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, arXiv preprint arXiv:2104.13478 (2021).

[30] M. Ghil and V. Lucarini, The physics of climate variability and climate change, Rev. Mod. Physics **92**, 035002 (2020).

[31] J. D. Meiss, *Differential Dynamical Systems* (SIAM, 2007).

[32] E. P. Wigner, The unreasonable effectiveness of mathematics in the natural sciences, Comm. Pure Applied Math. **13**, 1 (1960).

[33] A. Lasota and M. C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*, Vol. 97 (Springer Science, 1994).

[34] F. Bouttier and P. Courtier, Data assimilation concepts and methods, Meteorological training course lecture series. ECMWF **718**, 59 (1999).

[35] D. Sanz-Alonso, A. M. Stuart, and A. Taeb, Inverse problems and data assimilation, arXiv preprint arXiv:1810.06191 (2018).

[36] L. F. Konikow and J. D. Bredehoeft, Ground-water models cannot be validated, Advances in water resources **15**, 75 (1992).

[37] J. Carrera, S. F. Mousavi, E. J. Usunoff, X. Sánchez-Vila, and G. Galarza, A discussion on validation of hydrogeological models, Reliability Engineering & System Safety **42**, 201 (1993).

[38] G. E. Box, Science and statistics, J. Am. Stat. Assoc. **71**, 791 (1976).

[39] M. C. Mackey, ed., *Time's Arrow: The Origins of Thermodynamic Behavior* (Springer-Verlag, New York, 1992).

[40] Y. Oono and M. Paniconi, Steady state thermodynamics, Progress of Theoretical Physics Supplement **130**, 29 (1998).

[41] M. te Vrugt and R. Wittkowski, Mori-zwanzig projection operator formalism for far-from-equilibrium systems with time-dependent hamiltonians, Phys. Rev. E **99**, 062118 (2019).

[42] M. T. Semaan and J. P. Crutchfield, Homeostatic and adaptive energetics: Nonequilibrium fluctuations beyond detailed balance in voltage-gated ion channels, arXiv preprint arXiv:2202.13038 (2022).

[43] N. Brodu and J. P. Crutchfield, Discovering causal structure with reproducing-kernel Hilbert space ϵ-machines, Chaos, A Journal of Nonlinear Science , in press (2022), arXiv:2011.14821.

[44] C. W. Rowley, T. Colonius, and R. M. Murray, Model reduction for compressible flows using POD and Galerkin projection, Physica D: Nonlinear Phenomena **189**, 115 (2004).

[45] J. P. Crutchfield and B. S. McNamara, Equations of motion from a data series, Complex Systems **1**, 417 (1987).

[46] A. Chattopadhyay, P. Hassanzadeh, and D. Subramanian, Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, Nonlinear Processes in Geophysics **27**, 373 (2020).

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

[48] E. N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, Journal of Atmospheric Sciences **26**, 636 (1969).

[49] E. Gonzalez, M. Abudia, M. Jury, R. Kamalapurkar, and J. A. Rosenfeld, Anti-koopmanism, arXiv:2106.00106 (2021).

[50] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, Kernel mean embedding of distributions: A review and beyond, Foundations and Trends in Machine Learning **10**, 1 (2017).

[51] S. P. Loomis and J. P. Crutchfield, Topology, convergence, and reconstruction of predictive states, arXiv preprint arXiv:2109.09203 (2021).

[52] A. J. Chorin and F. Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, Proceedings of the National Academy of Sciences **112**, 9804 (2015).

[53] T. Sauer, J. A. Yorke, and M. Casdagli, Embedology, J. Stat. Phys. **65**, 579 (1991).

[54] G. Datseris, Dynamicalsystems.jl: A julia software library for chaos and nonlinear dynamics, Journal of Open Source Software **3**, 598 (2018).

[55] R. Zwanzig, *Nonequilibrium statistical mechanics* (Oxford university press, 2001).

[56] H. Mori, Transport, collective motion, and brownian motion, Progress of theoretical physics **33**, 423 (1965).

[57] Z. Zhao and D. Giannakis, Analog forecasting with dynamics-adapted kernels, Nonlinearity **29**, 2888 (2016).

[58] T. Koide, Derivation of transport equations using the time-dependent projection operator method, Progress of theoretical physics **107**, 525 (2002).

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

[60] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, New York, 1995).

[61] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani, Predicting the future from the past: An old problem from a modern perspective, American Journal of Physics **80**, 1001 (2012).

[62] H. Jaeger and H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, science **304**, 78 (2004).

[63] L. Grigoryeva and J.-P. Ortega, Echo state networks are universal, Neural Networks **108**, 495 (2018).

[64] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. Barbosa, Next generation reservoir computing, Nature communications **12**, 1 (2021).

[65] E. Bollt, On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to VAR and DMD, Chaos: An Interdisciplinary Journal of Nonlinear Science **31**, 013108 (2021).

[66] M. Gulina and A. Mauroy, Two methods to approximate the koopman operator with a reservoir computer, Chaos: An Interdisciplinary Journal of Nonlinear Science **31**, 023116 (2021).

[67] A. C. Costa, T. Ahamed, D. Jordan, and G. Stephens, Maximally predictive ensemble dynamics from data, arXiv preprint arXiv:2105.12811 (2021).

[68] J. P. Crutchfield and K. Kaneko, Are attractors relevant to turbulence?, Phys. Rev. Let. **60**, 2715 (1988).

[69] J. Gravner and K. Johnson, Coupled map lattices as musical instruments, Computer Music J. **42**, 22 (2018).

[70] J. E. Hanson and J. P. Crutchfield, Computational mechanics of cellular automata: An example, Physica D **103**, 169 (1997).

[71] J. Isensee, G. Datseris, and U. Parlitz, Predicting spatio-temporal time series using dimension reduced local states, Journal of Nonlinear Science **30**, 713 (2020).

[72] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Physical review letters **120**, 024102 (2018).

[73] S. E. Otto and C. W. Rowley, Linearly recurrent autoencoder networks for learning dynamics, SIAM Journal on Applied Dynamical Systems **18**, 558 (2019).

[74] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Machine learning for fluid mechanics, Ann. Rev. Fluid Mech. **52**, 477 (2020).

[75] F. Tao, H. Zhang, A. Liu, and A. Y. Nee, Digital twin in industry: State-of-the-art, IEEE Transactions on Industrial Informatics **15**, 2405 (2018).

[76] S. Boschert and R. Rosen, Digital twin—the simulation aspect, in *Mechatronic futures* (Springer, 2016) pp. 59–74.

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

[78] K. Duraisamy, G. Iaccarino, and H. Xiao, Turbulence modeling in the age of data, Annual Review of Fluid Mechanics **51**, 357 (2019).

[79] A. P. Guillaumin and L. Zanna, Stochastic-deep learning parameterization of ocean momentum forcing, Journal of Advances in Modeling Earth Systems **13**, e2021MS002534 (2021).

[80] L. Zanna and T. Bolton, Deep learning of unresolved turbulent ocean processes in climate models, Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences , 298 (2021).

[81] J. P. Crutchfield and K. Young, Inferring statistical complexity, Phys. Rev. Let. **63**, 105 (1989).

[82] B. O. Koopman, Hamiltonian systems and transformation in Hilbert space, Proceedings of the National Academy of Sciences **17**, 315 (1931).

[83] R. Zwanzig, Ensemble method in the theory of irreversibility, The Journal of Chemical Physics **33**, 1338 (1960).

[84] M. Casdagli and S. Eubank, eds., *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity (Addison-Wesley, Reading, Massachusetts, 1992).

[85] J. Milnor and W. Thurston, On iterated maps of the

interval, Springer Lecture Notes **1342**, 465 (1988).

[86] P. Collet and J.-P. Eckmann, *Maps of the Unit Interval as Dynamical Systems* (Birkhauser, Berlin, 1980).

[87] A. N. Kolmogorov, Entropy per unit time as a metric invariant of automorphisms, Doklady of Russian Academy of Sciences **124**, 754 (1959).

[88] Y. G. Sinai, On the notion of entropy of a dynamical system, Doklady of Russian Academy of Sciences **124**, 768 (1959).

[89] Y. B. Pesin, Characteristic lyapunov exponents and smooth ergodic theory, Uspekhi Matematicheskikh Nauk **32**, 55 (1977).

[90] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).

[91] E. N. Lorenz, The problem of deducing the climate from the governing equations, Tellus **XVI**, 1 (1964).

[92] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition, Journal of Nonlinear Science **25**, 1307 (2015).

[93] S. Klus, P. Koltai, and C. Schütte, On the numerical approximation of the perron-frobenius and koopman operator, Journal of Computational Dynamics **3**, 51 (2016).

[94] M. Korda and I. Mezić, On convergence of extended dynamic mode decomposition to the koopman operator, Journal of Nonlinear Science **28**, 687 (2018).

[95] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator, Chaos: An Interdisciplinary Journal of Nonlinear Science **27**, 103111 (2017).

[96] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nature Comm. **9**, 1 (2018).

[97] S. Klus, F. Nüske, S. Peitz, J.-H. Niemann, C. Clementi, and C. Schütte, Data-driven approximation of the Koopman generator: Model reduction, system identification, and control, Physica D: Nonlinear Phenomena **406**, 132416 (2020).

[98] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control, PLoS one **11**, e0150171 (2016).

[99] M. O. Williams, C. W. Rowley, and I. G. Kevrekidis, A kernel-based method for data-driven koopman spectral analysis, Journal of Computational Dynamics **2**, 247 (2015).

[100] S. Klus, I. Schuster, and K. Muandet, Eigendecompositions of transfer operators in reproducing kernel hilbert spaces, Journal of Nonlinear Science **30**, 283 (2020).

[101] S. Das and D. Giannakis, Koopman spectra in reproducing kernel Hilbert spaces, Applied and Computational Harmonic Analysis **49**, 573 (2020).

[102] J. H. Tu, C. W. Rowley, S. L. Luchtenburg, D. M .and Brunton, and J. N. Kutz, On dynamic mode decomposition: Theory and applications, J. Computational Dynamics **1**, 391 (2014).

[103] M. A. Alexander, L. Matrosova, C. Penland, J. D. Scott, and P. Chang, Forecasting pacific ssts: Linear inverse model predictions of the pdo, Journal of Climate **21**, 385 (2008).

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