SIAM J. Applied Dynamical Systems
Vol. 20, No. 4, pp. 2558–2601

Data-Driven Learning for the Mori–Zwanzig Formalism: A Generalization of the Koopman Learning Framework*

Yen Ting Lin†, Yifeng Tian‡, Daniel Livescu‡, and Marian Anghel†

Abstract. A theoretical framework which unifies the conventional Mori–Zwanzig formalism and the approximate Koopman learning of deterministic dynamical systems from noiseless observation is presented. In this framework, the Mori–Zwanzig formalism, developed in statistical mechanics to tackle the hard problem of construction of reduced-order dynamics for high-dimensional dynamical systems, can be considered as a natural generalization of the Koopman description of the dynamical system. We next show that, similar to the approximate Koopman learning methods, data-driven methods can be developed for the Mori–Zwanzig formalism with Mori's linear projection operator. We have developed two algorithms to extract the key operators, the Markov and the memory kernel, using time series of a reduced set of observables in a dynamical system. We have adopted the Lorenz '96 system as a test problem and solved for the above operators. These operators exhibit complex behaviors, which are unlikely to be captured by traditional modeling approaches in Mori–Zwanzig analysis. The nontrivial generalized fluctuation-dissipation relationship, which relates the memory kernel with the two-time correlation statistics of the orthogonal dynamics, was numerically verified as a validation of the solved operators. We present numerical evidence that the generalized Langevin equation, a key construct in the Mori–Zwanzig formalism, is more advantageous in predicting the evolution of the reduced set of observables than the conventional approximate Koopman operators.

Key words. Mori–Zwanzig formalism, memory effects, generalized Langevin equations, generalized fluctuation-dissipation relationship, dynamic mode decomposition, extended dynamic mode decomposition, approximate Koopman learning, data-driven, reduced-order dynamical system

AMS subject classifications. 37M19, 37M99, 46N55, 65P99, 82C31

DOI. 10.1137/21M1401759

1. Introduction.

The Mori–Zwanzig formalism [29, 50, 51, 9] was first developed in statistical physics for the difficult task of constructing coarse-grained models from high-dimensional microscopic models. The goal of model coarse-graining is to construct equations describing the

*Received by the editors March 1, 2021; accepted for publication (in revised form) by G. Froyland August 24, 2021; published electronically December 16, 2021.
https://doi.org/10.1137/21M1401759
Funding: This project was supported by the Laboratory Directed Research and Development Program at Los Alamos National Laboratory (project XWX8: Machine Learning for Turbulence). This work has been authored by employees of Triad National Security, LLC, which operates Los Alamos National Laboratory (LANL) under contract 89233218CNA000001 with the U.S. Department of Energy/National Nuclear Security Administration. The work has been supported by the LDRD (Laboratory Directed Research and Development) program at LANL under project 20190059DR (Machine Learning for Turbulence). The first author was partially supported by project LDRD-20190034ER (Massively Parallel Acceleration of the Dynamics of Complex Systems: A Data-Driven Approach) for finalizing the manuscript.
†Information Sciences Group, Computer, Computational and Statistical Sciences Division (CCS-3), Los Alamos National Laboratory, Los Alamos, NM 87545 USA (yentingl@lanl.gov, manghel@lanl.gov).
‡Computational Physics and Methods Group, Computer, Computational and Statistical Sciences Division (CCS-2), Los Alamos National Laboratory, Los Alamos, NM 87545 USA (yifengtian@lanl.gov, livescu@lanl.gov).

evolution of a smaller set of variables which are measurable, or quantities of interest. These quantities are often referred to as the relevant or resolved variables. For example, a microscopic model can be an all-atom molecular dynamics simulation of a protein in a solvent, and a relevant variable can be the distance between two atoms of interest. It is desirable to construct a closed dynamical system which includes only the resolved variables without the information of other degrees of freedom. On the one hand, it is easier to perform analysis on a closed lower-dimensional system and to shed important insights on the interactions between the resolved variables. On the other hand, it is more efficient to simulate the reduced-dimensional system computationally.

The major challenge of coarse-graining modeling is that the resolved variables may be influenced by the unresolved variables. In the above example, the distance between the two specific atoms may be influenced by nearby water molecules that are not in the set of resolved variables. To solve this difficult closure problem, Mori [29] and Zwanzig [50] developed the projection-based methods to express the effect of the unresolved variables in terms of the resolved ones. The generalized Langevin equation, the main result of the Mori–Zwanzig formalism, decomposes the evolutionary equations of the resolved variables into three parts: a Markov term, which captures the interaction within the resolved variables, a memory term, which is history dependent and captures the interactions between the resolved and the unresolved variables, and a term representing the orthogonal dynamics, which captures the unknown initial condition of the unresolved variables. Although the generalized Langevin equation is formally exact, it is challenging to theoretically derive closed-form expressions of these terms in the generalized Langevin equation without approximations. Conventionally, the applications of the Mori–Zwanzig formalism rely on modeling self-consistent operators based on the mathematical structure of the generalized Langevin equation [38, 13, 6, 22, 21, 23, 32, 44, 14].

In a seemingly unrelated research area, approximate Koopman learning methods such as dynamic mode decomposition [35, 36] and extended dynamic mode decomposition (EDMD) [46], have been actively developed for data-driven modeling of dynamical systems. The general idea is that by collecting enough data of a dynamical system, possibly from a high-fidelity simulation of the microscopic system, one would be able to learn important features in the dynamics, e.g., spectral [27] and dynamic modes [35, 36, 46]. The theoretical foundation of these methods, the Koopman theory, is a formulation for general dynamical systems [18, 17]. Instead of the typical description of a possibly nonlinear system in the physical space, in Koopman theory, the dynamics are described as a linear dynamical system of the observables, which are functions of the physical-space variables, in an infinite-dimensional Hilbert space. Because the interactions in this framework are linear (but with a caveat that the space is infinite-dimensional), learning from data is a convex problem which is easier to solve, in contrast to the nonlinear regression in the conventional physical-space picture.

The major aim of this article is to bridge these two seemingly disconnected research areas in dynamical systems: the Mori–Zwanzig formalism and the approximate Koopman learning. We will establish that *the Mori–Zwanzig formalism with Mori's linear projector is functionally identical to the approximate Koopman learning methods in a shared Hilbert space*. This connection helps to bring the advantage of one research area to another. On the one hand, the approximate Koopman learning methods can be generalized for data-driven modeling of the operators in Mori–Zwanzig formalism. As will be seen in this article, the

operators describing the Markov and memory terms in the generalized Langevin equation can be numerically learned from the simulation of the microscopic system. Surprisingly, these operators, inferred from data, are highly nontrivial and are unlikely to be modeled accurately without in-depth knowledge of the system. On the other hand, the memory terms in the Mori–Zwanzig formalism can be considered as higher-order corrections of the Koopman learning methods. We will show that by including the memory kernel and the history of the resolved variables, the generalized Langevin equation predicts more accurately than the EDMD.

This article is organized as follows. In section 2, we provide a gentle introduction to the two descriptions of a dynamical system: the description of a finite-dimensional but possibly nonlinear dynamics in the physical space, and the Koopman description of infinite-dimensional but linear dynamics of the observables. For completeness, we include a self-contained introduction and review of the Mori–Zwanzig formalism in section 3. We establish that the Mori–Zwanzig formalism is a generalization, in the sense that it contains the higher-order memory effect, of the EDMD [46] in section 4. Two novel algorithms, motivated by the EDMD to extract the key operators in the Mori–Zwanzig formalism by simulation data, are presented in section 5. We perform numerical experiments on a Lorenz '96 model [25] and present the results in section 6. In section 6.2, we demonstrate the advantage of the Mori–Zwanzig formalism over the conventional EDMD in predicting future dynamics. Finally, we provide a discussion and future outlook in section 7.

2. Preliminaries.

There exist two equivalent formulations to describe a dynamical system. In the first formulation [1, 39], the system is characterized by a collection of *physical-space variables*, often called the state of the system. For example, a physical-space variable can be one component of the position of an atom in a many-particle system, or one component of the velocity field at a specific location in a fluid dynamical system. The aim of this first formulation is to describe the evolution of these physical-space variables. Suppose the state of the system is fully characterized by N physical-space variables ϕ_i, $i = 1, \ldots, N$, and we denote the state of the system at time t by $\mathbf{\Phi}(t) := [\phi_1(t), \ldots, \phi_N(t)]^T$, an $N \times 1$ column vector. Then the evolution of the variables in the physical space (assumed to be $\mathbb{R}^N$ for simplicity) is described by the deterministic evolutionary equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_i(t) = R_i(\mathbf{\Phi}(t), t) \quad \text{and} \quad \mathbf{\Phi}(0) = \mathbf{\Phi}_0, \quad i = 1, \ldots, N, \tag{2.1}$$

where the flow $R_i : \mathbb{R}^N \to \mathbb{R}$ is a function which maps the state $\mathbf{\Phi}$ to a real number that characterizes the velocity of the physical-space variable ϕ_i at time t, and $\mathbf{\Phi}_0$ is the given $N \times 1$ column vector specifying the initial condition of the system's state. In this article, we exclusively consider autonomous dynamical systems, where R_i does not explicitly depend on the time t. Thus, the evolutionary equation (2.1) can be written in a terse form $\dot{\mathbf{\Phi}} = \mathbf{R}(\mathbf{\Phi})$, where the flow $\mathbf{R}$ is defined as $[R_1(\mathbf{\Phi}(t)), \ldots, R_N(\mathbf{\Phi}(t))]^T$ and is implicitly time dependent. In general, the flow $\mathbf{R}$ can be nonlinear in $\mathbf{\Phi}$.

In the second formulation proposed by Koopman (see [18, 17]), the system is characterized by a collection of *observables* which are functions of the physical-space variables. For example, an observable can be a component of the total angular momentum of a subset of all atoms in a particle system, or the locally averaged density in a fluid dynamical system. The

Koopmanian formulation describes how observables evolve in an infinite-dimensional Hilbert space $\mathcal{H}$, which is composed of all the possible observables. The advantage of this formulation is that the evolution of the observables, which is a vector in the infinite-dimensional Hilbert space $\mathcal{H}$, is *always linear*, even for systems that are nonlinear in the physical-space picture. The disadvantage of this formulation is that the state space of the system, which consists of all possible observables, is infinite-dimensional.

To illustrate the difference of the formulations, we consider a one-dimensional nonlinear dynamical system in the physical-space formulation: $\dot{\phi}(t) = R(\phi(t))$ and $\phi(0) := \phi_0$, where $R(x) := -x^2$ is a nonlinear function and ϕ_0 is the initial condition. While the analytic solution exists for this simple problem ($\phi(t) = 1/(t+1/\phi_0)$), it is challenging to derive the closed-form solution for general multidimensional ($N > 1$) nonlinear dynamical systems. Note that $\phi(t)$ is a nonlinear function of the initial condition ϕ_0. In the Koopman formulation, the dynamics are characterized by observables of ϕ. It is sufficient for us to consider a set of observables which will serve as the basis functions. For this example, we consider $g_k(\phi) := \phi^k$, $k \in \mathbb{Z}_+$. Other observables can be expressed as a weighted linear superposition of these basis functions via Taylor series expansion. In contrast to the first formulation, Koopman's theory describes the dynamics of the basis functions g_k:

$$\frac{\mathrm{d}}{\mathrm{d}t} g_k(t) := \frac{\mathrm{d}}{\mathrm{d}t}\left[g_k \circ \phi(t)\right] = \frac{\mathrm{d}g_k}{\mathrm{d}\phi} \cdot \frac{\mathrm{d}\phi}{\mathrm{d}t} = k\phi^{k-1}(t) \cdot \left[-\phi^2(t)\right] = -k g_{k+1}(t). \tag{2.2}$$

Throughout this article, we will use the symbol $\circ$ to denote the composite functions. Here, the observable functions g_k are functions of the physical-space variable ϕ, which is a function of the physical time t. The dynamics of $g_k(t)$ is always linearly dependent on $g_{k+1}(t)$, but is not closed unless the system involves infinitely many k's. The evolution of lower-order nonlinearity involves higher-order nonlinearity, similar to the common phenomenon in Carleman linearization [5, 19] and moment expansion methods [2, 37]. Nevertheless, we can choose two simple functions, $g_0(\phi) := \phi^0$ and $g_{-1}(\phi) = \phi^{-1}$, and their dynamics are closed:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} g_0(t) \\ g_{-1}(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} g_0(t) \\ g_{-1}(t) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} g_0(0) \\ g_{-1}(0) \end{bmatrix} = \begin{bmatrix} 1 \\ \phi_0^{-1} \end{bmatrix}. \tag{2.3}$$

The above *linear* ordinary differential equations (ODEs) are solved to derive the analytical solution of $g_{-1}(\phi(t)) = t + \phi_0^{-1}$, which is then used to calculate $\phi(t)$. The choice of this invariant set of functions is not unique. We can even choose a smaller set which contains only one observable $g_e(\phi) := \exp(-1/\phi)$, which satisfies a one-dimensional linear ordinary equation $\dot{g_e}(t) = -g_e(t)$.

The above examples illuminate two key features of the Koopman theory. First, the dynamics of observables are always linearly dependent on other observables. Second, deriving closed-form solution in the Koopman theory is equivalent to identifying a set of observables whose dynamics are invariant in a subspace which is linearly spanned by the set of the observables. In general, it is challenging to identify the finite set of observables that closes the dynamics, and one has to resort to approximation methods to close the system. In the next section, we illustrate how the Mori–Zwanzig formalism leverages the projection operators to close the dynamics.

3. The Mori–Zwanzig formalism. Here, we provide a review to the Mori–Zwanzig formalism. For completeness, we provide two comprehensive derivations of the major result of the Mori–Zwanzig formalism, the generalized Langevin equation (GLE). We begin with the operator algebraic derivation [7, 8, 10] in section 3.1. To make the connection to the Koopman representation of the dynamics, we provide an alternative derivation of the GLE based on the the Koopman eigenfunctions in section 3.2. Although the first derivation is terse and elegant, it is not easy to build intuition to understand the action of operators in the GLE. The second derivation has two advantages: (1) it provides a more transparent representation of the Mori–Zwanzig operators, and (2) its terminology acts as a bridge to approximate Koopman analysis such as EDMD [46]. In fact, the second approach's geometric representation in the functional space is identical to Mori's original construct [29], and the derivation is very close to the variation of the constants method presented in Zwanzig's own derivation [51]. We will thus adopt the terminology of the second derivation throughout the rest of the paper. After the GLE is set up, we provide a geometric interpretation of the GLE in section 3.3 and a detailed discussion on the projection operator in section 3.4. In section 3.5 we discuss the consequence of the GLE on the evolutionary equations of the covariance matrices and the projected image. Section 3.6 is dedicated to the emergence of the self-consistent generalized fluctuation-dissipation relationship. We conclude the review of the Mori–Zwanzig formalism by discussing its applicability to discrete-time dynamics in section 3.7.

3.1. Operator algebraic derivation of the generalized Langevin equation. We begin with the evolutionary equation (2.1), where the flow field $\mathbb{R} : \mathbb{R}^N \to \mathbb{R}^N$ is assumed to be locally Lipschitz continuous such that a unique $\mathbf{\Phi}(t)$ exists $\forall t \geq 0$. More generally, the state space can be any compact Riemannian manifold endowed with the Borel σ-algebra and a measure; for brevity, we will consider the state space as $\mathbb{R}^N$ below. The solution $\mathbf{\Phi}(t; \mathbf{\Phi}_0)$ nonlinearly maps the initial condition $\mathbf{\Phi}_0$ to the phase-space configuration at physical time t. Next, one defines the Liouville operator $\mathcal{L} := \sum_{i=1}^{N} R_i(\mathbf{x}) \partial_{x_i}$, with a dummy variable $\mathbf{x} \in \mathbb{R}^N$, and considers the partial differential equation (PDE)

$$\frac{\partial \psi(t, \mathbf{x})}{\partial t} = \mathcal{L}\psi(t, \mathbf{x}), \tag{3.1a}$$

$$\psi(0, \mathbf{x}) = g(\mathbf{x}), \tag{3.1b}$$

where $g : \mathbb{R}^N \to \mathbb{R}$ is a real-valued function of the state of the system, $\mathbf{x} \in \mathbb{R}^N$. One can show that the solution to the above first-order PDE is $u(t, \mathbf{x}) \equiv g(\mathbf{\Phi}(t; \mathbf{x}))$ by the method of characteristics. Note that $\mathbf{x} \in \mathbb{R}^N$ can be any initial state. Thus, solving the above linear PDE fully solves the function g evaluated at the trajectory of the nonlinear system given any initial condition $\mathbf{\Phi}_0$: $g(\mathbf{\Phi}(t; \mathbf{\Phi}_0)) = \psi(t, \mathbf{\Phi}_0)$. We adopt the slightly abused notation in published literature [7, 8, 10] and denote the solution $\psi(t, \mathbf{x})$ with the initial condition $\psi(0, \mathbf{x}) = g(\mathbf{x})$ by $g(t, \mathbf{x})$. A special choice of g is $g(\mathbf{x}) := x_i$, where g extracts the ith component of the multivariate vector. In this case $\psi(\mathbf{\Phi}_0, t)$ is the solution of the ith component, $\psi(t, \mathbf{\Phi}_0) \equiv \phi_i(t; \mathbf{\Phi}_0)$. The semigroup notation tersely represents the solution of the above PDE (3.1) as $\psi(t, \mathbf{x}) = e^{t\mathcal{L}} g(\mathbf{x})$, with an evolutionary equation

$$\frac{\partial}{\partial t}\left[e^{t\mathcal{L}} g\right](\mathbf{x}) = \left(\mathcal{L} e^{t\mathcal{L}} g\right)(\mathbf{x}) = \left(e^{t\mathcal{L}} \mathcal{L} g\right)(\mathbf{x}). \tag{3.2}$$

The above equation applies to any $\mathbf{x} \in \mathbb{R}$, so "$(\mathbf{x})$" is often neglected in calculations.

Conventionally, the goal of the Mori–Zwanzig procedure is to construct the evolutionary equations for a set of components $\hat{\phi} := \{\phi_i\}_{i=1}^{M}$, $M < N$, referred to as the resolved components. These are the components which we can measure as the dynamics move forward in time. Because the state space $\mathbb{R}^N$ can be fully characterized by N coordinates ϕ_i, $i = 1, \ldots, N$, knowing $M < N$ resolved observables could not fully specify the system's state for constructing a closed dynamical system. Consequently, one would need to postulate another $N - M$ underresolved components, often denoted by $\tilde{\phi}$. The Mori–Zwanzig procedure proceeds with a postulated joint distribution $\mathrm{d}\mu \equiv \rho(\mathbf{x})\,\mathrm{d}^N\mathbf{x}$, where ρ is the probability density and $\mathrm{d}^N\mathbf{x}$ is the Borel measure in $\mathbb{R}^N$, for asserting the initial distribution of the underresolved $\tilde{\phi}$ conditioned on a given set $\hat{\phi}$. The choice of $\mathrm{d}\mu$ is model specific but is often the equilibrium (or nonequilibrium stationary) distribution of the system. Despite the conventional choice of using the components of the state (i.e., $g(\mathbf{x}) = x_i$) as the resolved and underresolved observables, g can be any function of the state (for example, g can be the center of mass of a molecule whose full configurations are specified by the position and momentum of all its atoms). A technical condition on g is that it has to be L^2-integrable with respect to the measure $\mathrm{d}\mu$ for constructing an inner product of a Hilbert space in which the Mori–Zwanzig formalism operates.

The Mori–Zwanzig procedure proceeds with a specified projection operator $\mathcal{P}$, which maps a function of the full-space configuration, $g : \mathbb{R}^N \to \mathbb{R}$, to a function of only the resolved observables $\mathcal{P}g : \mathbb{R}^M \to \mathbb{R}$, assumed to be L^2-integrable with respect to $\mathrm{d}\mu$. The complement of the projection operator is defined as $\mathcal{Q} := I - \mathcal{P}$. Applying the Dyson identity [9]

$$e^{t(A+B)} = e^{tB} + \int_0^t e^{(t-s)(A+B)} A e^{sB}\,\mathrm{d}s \tag{3.3}$$

to operators $A := \mathcal{P}\mathcal{L}$ and $B := \mathcal{Q}\mathcal{L}$, one obtains

$$e^{t\mathcal{L}} = e^{t\mathcal{Q}\mathcal{L}} + \int_0^t e^{(t-s)\mathcal{L}} \mathcal{P}\mathcal{L} e^{s\mathcal{Q}\mathcal{L}}\,\mathrm{d}s. \tag{3.4}$$

The operator is applied to (3.2), resulting in the following expression for any g with the fact that $\mathcal{P} + \mathcal{Q} = I$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left[e^{t\mathcal{L}} g\right] &= \left[e^{t\mathcal{L}} \mathcal{L} g\right] = \left[e^{t\mathcal{L}} \left(\mathcal{P} + \mathcal{Q}\right) \mathcal{L} g\right] \\ &= \left[e^{t\mathcal{L}} \mathcal{P}\mathcal{L} g\right] + \left[e^{t\mathcal{Q}\mathcal{L}} \mathcal{Q}\mathcal{L} g\right] + \int_0^t e^{(t-s)\mathcal{L}} \left[\mathcal{P}\mathcal{L} e^{s\mathcal{Q}\mathcal{L}} \mathcal{Q}\mathcal{L} g\right] \mathrm{d}s. \end{aligned} \tag{3.5}$$

Specifically for $g(\mathbf{x}) = x_i$, $i = 1, \ldots, M$, we define the Markov transition as $M_i(\hat{\mathbf{x}}) := [\mathcal{P}R_i](\mathbf{\Phi}(t, \mathbf{x}))$, the orthogonal dynamics as $F_i(t, \mathbf{x}) := \left[e^{t\mathcal{Q}\mathcal{L}} \mathcal{Q}\mathcal{L} g\right](\mathbf{x})$, and the memory function as $K_i(t, \hat{\mathbf{x}}) := -[\mathcal{P}\mathcal{L}F_i](t, \mathbf{x})$ to obtain the GLE describing the evolution of resolved components given an initial condition $\mathbf{\Phi}_0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\phi}_i(t, \mathbf{\Phi}_0) = M_i\left(\hat{\mathbf{\Phi}}(t, \mathbf{\Phi}_0)\right) - \int_0^t K_i\left(\hat{\mathbf{\Phi}}(t, \mathbf{\Phi}_0), t - s\right) \mathrm{d}s + F_i(t, \mathbf{\Phi}_0). \tag{3.6}$$

Note that we follow the original sign convention that Mori [29] and Zwanzig [50] adopted: the memory term has a negative sign, in contrast to later publications [7, 8, 10] in which the memory term was defined with a positive sign.

The difference between Mori and Zwanzig is their choice of the projection operator. With Mori's construction [29], one relies on an inner product defined as

$$\langle f, g\rangle = \int_{\mathbb{R}^N} f(\mathbf{x})\, g(\mathbf{x})\, \rho(\mathbf{x})\, \mathrm{d}^N\mathbf{x}, \quad f, g \in L^2(\mu), \tag{3.7}$$

to define a projection operator given a set of resolved observables $\hat{\phi} = \{\phi_i\}_{i=1}^M$:

$$[\mathcal{P}f]\left(\hat{\phi}\right) := \sum_{i,j=1}^{M} \langle f, \phi_i\rangle \left[\mathbf{C}_0^{-1}\right]_{i,j} \phi_j, \tag{3.8}$$

where $\mathbf{C}^{-1}(0)$ is the inverse of an $M \times M$ matrix $\mathbf{C}_0$ whose (i, j) entry is $\langle \phi_i, \phi_j\rangle$. A more geometric interpretation of Mori's projector will be presented in section 3.4. Note that $\mathcal{P}f$ is a linear function of the resolved observables, ϕ_j, $j = 1, \ldots, M$, and thus Mori's projector is often referred to as a linear projection. In contrast, with the same set of observables, Zwanzig [50] does not rely on the inner product but relies on the direct marginalization of the underresolved observables:

$$[\mathcal{P}f](\hat{\mathbf{x}}) := \frac{\int_{\mathbb{R}^{N-M}} f(\hat{\mathbf{x}}, \tilde{\mathbf{x}})\, \rho(\hat{\mathbf{x}}, \tilde{\mathbf{x}})\, \mathrm{d}\tilde{\mathbf{x}}}{\int_{\mathbb{R}^{N-M}} \rho(\hat{\mathbf{x}}, \tilde{\mathbf{x}})\, \mathrm{d}\tilde{\mathbf{x}}}. \tag{3.9}$$

Note that the resulting function $\mathcal{P}f$ is generally nonlinear in the resolved observables. Also termed the nonlinear projection [8] and infinite-rank projection [10], Zwanzig's projection operator can lead to a nonlinear Markov transition and nonlinear memory kernel in the GLE [7, 8, 14].

After applying Mori's projection operator to M, F, and K, one obtains a linear GLE [51],

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\phi}_i(t, \mathbf{\Phi}_0) = \sum_{j=1}^{M} [\mathbf{M}]_{i,j}\, \hat{\phi}_j(t, \mathbf{\Phi}_0) - \int_0^t [\mathbf{K}(t-s)]_{i,j}\, \hat{\phi}_j(t, \mathbf{\Phi}_0)\, \mathrm{d}s + F_i(t, \mathbf{\Phi}_0), \tag{3.10}$$

where $\mathbf{M}$ is an $M \times M$ constant matrix, as well as $\mathbf{K}(t)$, $t \geq 0$. We will illuminate the physical meaning of this linear GLE in section 3.3. In this manuscript, we will focus on Mori's projection and its connection to the approximate Koopman learning methods.

3.2. Generalized Langevin equation in Koopman representation. It is not easy to build intuition from the terse derivation presented in section 3.1. Zwanzig even commented, "*The derivation to be given here is based on abstract operator manipulations that were designed to get to the desired result as quickly as possible*" and provided a more lengthy motivating derivation based on a variation of the constants method prior to the formal operator algebraic derivation. In this section, we aim to provide a similar derivation using Koopman representation of the dynamics for better understanding the GLE. By introducing a Koopman representation of the dynamics, we also aim to make a natural and formal connection between Mori–Zwanzig and Koopman formulations.

As shown in the motivating example in section 2, in Koopman representation of the dynamics, one aims to describe the evolution of the observables, which are functions of the system's state $\mathbf{\Phi}$. In the space of all L^2-integrable observables, the evolution is always linear in other observables, but the dimensionality of the operating space can be infinite. Formally, given a measure $\mathrm{d}\mu$, we denote the space of all L^2-integrable real-valued observables (that can be generalized to complex-valued observables) of the state space by $\mathcal{F} = L^2\left(\mathbb{R}^N, \mu\right)$. Together with a defined inner product, such as (3.7), these functions form a Hilbert functional space $\mathcal{H}$, in which the functions evolve forward in time. The continuous-time Koopman operator $\mathcal{K}_t : \mathcal{F} \to \mathcal{F}$ is defined by

$$\left(\mathcal{K}_t g\right)\left(\mathbf{\Phi}_0\right) = g \circ \mathbf{\Phi}\left(t; \mathbf{\Phi}_0\right) \equiv g\left(\mathbf{\Phi}\left(t; \mathbf{\Phi}_0\right)\right) \quad \forall g \in \mathcal{F}, \forall \mathbf{\Phi}_0 \in \mathbb{R}^N. \tag{3.11}$$

In other words, the Koopman operator $\mathcal{K}_t$ transforms the function g into a function $\mathcal{K}_t g$ of any initial condition $\mathbf{\Phi}_0$. At any time $t \geq 0$, $\mathcal{K}_t g$ is equivalent to the observable g evaluated at the solution of the dynamics $\mathbf{\Phi}\left(t; \mathbf{\Phi}_0\right)$, which is also a function of $\mathbf{\Phi}_0$. The above equation shows the *dual representations* of the dynamics: The left-hand side of the equation is the Koopman picture (analogous to the Heisenberg picture in quantum mechanics) that the observable is evolving (transformed by $\mathcal{K}_t$) forward in time and is always evaluated at the fixed state $\mathbf{\Phi}_0$, while the right-hand side of the equation is the Perron–Frobenius picture (analogous to Schrödinger's picture in quantum mechanics) that the state is evolving forward in time $\left(\mathbf{\Phi}\left(t; \mathbf{\Phi}_0\right)\right)$ and evaluated by a fixed observable g.

The linear Koopman operator $\mathcal{K}_t$ can be characterized by its eigenvalues and eigenfunction. A function $\phi : \mathbb{R}^N \to \mathbb{R}$ (or $\mathbb{C}$) is defined as a Koopman eigenfunction if it satisfies $\left(\mathcal{K}_t \phi\right) = e^{\lambda t}\phi$. Here, we drop the "as a function of initial-condition" annotation "$(\mathbf{\Phi}_0)$" again. The space of the eigenfunctions is infinite-dimensional: Given two pairs (λ_1, ϕ_1) and (λ_2, ϕ_2), one can generate infinitely many eigenfunctions $(m\lambda_1 + n\lambda_2, \phi_1^m \phi_2^n)$, $m, n \in \mathbb{N}$. The infinitesimal generator of $\mathcal{K}_t$, $\lim_{t \downarrow 0}\left(\mathcal{K}_t - I\right)/t$ is the Liouville operator $\mathcal{L} := \sum_{i=1}^N R_i\left(\mathbf{x}\right)\partial_{x_i}$, which is the Lie derivative with respect to the flow field $\mathbf{R} : \mathbb{R}^N \to \mathbb{R}^N$ [26]. Note that $\mathcal{L}\phi = \lambda\phi$. As such, the Koopman eigenfunctions are the eigenfunctions of the Liouville operator and are special initial data following a coherent evolution $\psi(t, \mathbf{x}) = \phi(\mathbf{x})\exp\left(\lambda t\right)$ by (3.1).

We aim to construct the evolution of a set of linearly independent observables, $\mathcal{M} := \{g_i\}_{i=1}^M$. Given a time $t \geq 0$, we would like to know how $g_i(t) := \mathcal{K}_t g_i$, a function of the initial condition $\mathbf{\Phi}_0$ parametrized by time t, changes with respect to time t. We remark that such dynamical variable notation ("$g_i(t)$") was first introduced by Mori [29] and has been the mainstream notation in the physics literature [51]. Using the modern Koopman notation, $g(t)$ is expressed as $\mathcal{K}_t g_i$, and with the algebraic notation in section 3.1 as $g_i(t, \cdot)$. Note that $g_i(0) = \mathcal{K}_0 g_i = g_i$. Although the Koopman operator may contain a continuous spectrum [18, 43, 27] for chaotic dynamical systems, in this derivation, we consider systems with only point spectra for brevity. For these systems, the observables g_i can be expressed as a linear combination of the countably infinite eigenfunctions [34, 46]:

$$g_i = \sum_{j=1}^{\infty} v_{i,j}\phi_j, \quad i = 1, \ldots, M. \tag{3.12}$$

In contrast to the above equation which decomposes a function into Koopman eigenfunctions, the Mori–Zwanzig formalism utilizes the inner product in the Hilbert space to decompose the

space into the subspace linearly spanned by the set of observables, $\mathcal{H}_{\mathbf{g}} := \mathrm{Span}(\mathcal{M})$, and an orthogonal subspace $\mathcal{H}_{\bar{\mathbf{g}}} = \{\bar{g} \in \mathcal{F} : \langle \bar{g}, g_i \rangle = 0, g_i \in \mathcal{M}\}$. One proceeds with constructing a complete set of basis functions in $\mathcal{H}$, with the natural choice of using $\mathcal{M}$ as the set of basis functions in $\mathcal{H}_{\mathbf{g}}$. One can then use the Gram–Schmidt process to construct the basis functions in the orthogonal space from the Koopman eigenfunctions, $\{\phi_i\}_{i=1}^{\infty}$. We denote this infinite set of basis functions by $\bar{\mathcal{M}} := \{\bar{g}_i\}_{i=1}^{\infty}$. Similarly to (3.12), we can decompose $\bar{g}_i$'s in terms of the eigenfunctions: $\bar{g}_i = \sum_{j=1}^{\infty} \bar{v}_{i,j}\phi_j$, $i \in \mathbb{N}$. By construction, $\langle g_i, \bar{g}_j \rangle = 0$, $i \in \{1, \ldots, M\}$, $j \in \mathbb{N}$. Note that we did not require orthogonality between the basis functions in the same subspace, that is, $\langle g_i, g_j \rangle$ and $\langle \bar{g}_i, \bar{g}_j \rangle$ are not required to be 0 if $i \neq j$. However, we assume linear independence between any of the pairs of the basis functions, and the combined set $\mathcal{M} \cup \bar{\mathcal{M}}$ forms a complete set of basis functions in $\mathcal{H}$. Consequently, one can express any Koopman eigenfunction ϕ_i in terms of these new basis functions,

$$\phi_i = \sum_{j=1}^{M} \omega_{i,j} g_j + \sum_{j=1}^{\infty} \bar{\omega}_{i,j} \bar{g}_j. \tag{3.13}$$

Applying the Koopman operator $\mathcal{K}_t$ to the above equation, we obtain the relationship $\forall t \geq 0$, $\phi_i(t) = \sum_{j=1}^{M} \omega_{i,j} g_j(t) + \sum_{j=1}^{\infty} \bar{\omega}_{i,j} \bar{g}_j(t)$. Now, we can express the evolution of the basis functions g_i, $i = 1, \ldots, M$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} g_i(t) &= \lim_{s \downarrow 0} \frac{\mathcal{K}_s g_i(t) - g_i(t)}{s} = \sum_{j=1}^{\infty} v_{i,j} \lambda_j e^{\lambda_j t} \phi_j = \sum_{i=0}^{\infty} v_{i,j} \lambda_j \phi_j(t) \\ &= \sum_{\ell=1}^{M} \left(\sum_{j=1}^{\infty} v_{i,j} \lambda_j \omega_{j,\ell} \right) g_\ell(t) + \sum_{\ell=1}^{\infty} \left(\sum_{j=1}^{\infty} v_{i,j} \lambda_j \bar{\omega}_{j,\ell} \right) \bar{g}_\ell(t), \end{aligned} \tag{3.14}$$

and similarly for $\bar{g}_i$, $i \in \mathbb{N}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{g}_i(t) = \sum_{\ell=1}^{M} \left(\sum_{j=1}^{\infty} \bar{v}_{i,j} \lambda_j \omega_{j,\ell} \right) g_\ell(t) + \sum_{\ell=1}^{\infty} \left(\sum_{j=1}^{\infty} \bar{v}_{i,j} \lambda_j \bar{\omega}_{j,\ell} \right) \bar{g}_\ell(t). \tag{3.15}$$

We have established that at any time, the evolution of $g_i(t)$ and $\bar{g}_i(t)$ are linear functions of themselves, which is the major consequence of the Koopman representation. We now adopt a terse vector notation $\mathbf{g}_{\mathcal{M}}(t) = [g_1(t), \ldots, g_M(t)]^T$ and $\mathbf{g}_{\bar{\mathcal{M}}}(t) = [\bar{g}_1(t), \bar{g}_2(t) \ldots]^T$ and concisely express the full dynamics as

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t) \end{bmatrix} = \mathbf{L} \cdot \begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t) \end{bmatrix} := \begin{bmatrix} \mathbf{L}_{\mathcal{M}\mathcal{M}} & \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} \\ \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} & \mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t) \end{bmatrix}. \tag{3.16}$$

Here, the matrices $\mathbf{L}_{i,j}$, $i, j \in \{\mathcal{M}, \bar{\mathcal{M}}\}$, quantify the effect from set j to set i in the linear evolution. While these matrices can be found explicitly from (3.14) and (3.15) should one know the Koopman eigenfunctions, we derive them just for the purpose of establishing the linear evolutionary equation (3.16). We emphasize that although $g_i(0) = g_i \in \mathcal{H}_{\mathbf{g}}$ and $\bar{g}_i(0) = \bar{g}_i \in \mathcal{H}_{\bar{\mathbf{g}}}$ for $t > 0$, $g_i(t)$ and $\bar{g}_i(t)$ are not necessarily in $\mathcal{H}_{\mathbf{g}}$ and $\mathcal{H}_{\bar{\mathbf{g}}}$, respectively, if the interactions

between the spaces, $\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}}$ and $\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}$, are not zero. In other words, in general, both $g_i(t)$ and $\bar{g}_i(t)$ have nonzero components of the basis functions in $\mathcal{H}_{\mathbf{g}}$ and $\mathcal{H}_{\bar{\mathbf{g}}}$. Nevertheless, the above linear equation (3.16) always holds.

To obtain a closed-form evolution for our observables of interest in $\mathcal{M}$, we first solve for the observables in set $\bar{\mathcal{M}}$ implicitly. Treating $\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}\mathbf{g}_{\mathcal{M}}$ as an inhomogeneous driving term of the linear system, we solve the linear evolutionary equation for $\mathbf{g}_{\bar{\mathcal{M}}}$:

$$\mathbf{g}_{\bar{\mathcal{M}}}(t) = \int_0^t e^{(t-s)\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} \cdot \mathbf{g}_{\mathcal{M}}(s)\,\mathrm{d}s + e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{g}_{\bar{\mathcal{M}}}(0). \tag{3.17}$$

The implicit solution of $\mathbf{g}_{\bar{\mathcal{M}}}$ is in turn used to express closed evolutionary equations for the observables in the set $\mathcal{M}$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{g}_{\mathcal{M}}(t) = \mathbf{L}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(t) + \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}}\int_0^t e^{(t-s)\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} \cdot \mathbf{g}_{\mathcal{M}}(s)\,\mathrm{d}s \\ + \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}}e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{g}_{\bar{\mathcal{M}}}(0).\end{aligned} \tag{3.18}$$

Equation (3.18) is almost closed in our chosen set of variables except for the last term. The first term is the instantaneous configuration of the set of observables applied to the physical-space configuration at time t, and the second is a delayed impact of the set of observables applied to the physical-space variables at an earlier time $s < t$. Both terms depend only on the resolved observables $\mathbf{g}_{\mathcal{M}}$. However, the third term is induced by the initial setting of the underresolved observables, $\mathbf{g}_{\bar{\mathcal{M}}}(0)$, which cannot be generally resolved. Equation (3.18) is exact if one knows both $\mathbf{g}_{\mathcal{M}}(0)$ and $\mathbf{g}_{\bar{\mathcal{M}}}(0)$, in which case the system is fully resolved. Unfortunately, we do not have direct access to $\mathbf{g}_{\bar{\mathcal{M}}}(0)$ as they are underresolved observables, and one has to postulate their configurations in practice.

This simple analysis illustrates the essential intuition of the Mori–Zwanzig formalism. Because we only resolve a set of observables $\mathbf{g}_{\mathcal{M}}(t)$ of the full dynamics, the impact from other observables $\mathbf{g}_{\bar{\mathcal{M}}}(t)$ in (3.16) cannot be directly accessed. Instead, we indirectly estimate the effect $\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} \cdot \mathbf{g}(t)$ from (3.17), which contains two parts: the impact of resolved observables $\mathbf{g}_{\mathcal{M}}(s)$ at an earlier time s, and the initial conditions of the orthogonal observables $\mathbf{g}_{\bar{\mathcal{M}}}$. The former characterizes the "echo" of the set of our interested observables to itself: at an earlier time s, these observables made an impact to the underresolved observables $\mathbf{g}_{\bar{\mathcal{M}}}$ (via $\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}$), and such an impact propagates among the underresolved observables $\mathbf{g}_{\bar{\mathcal{M}}}$ for $t-s$ time via $e^{(t-s)\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}}$ before coming back to affect the resolved observables at time t via $\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}}$. The second part is a generic impact from the initial configuration of the orthogonal set of observables, $\mathbf{g}_{\bar{\mathcal{M}}}(0)$, which has propagated in the underresolved observables until the current time t and affects the resolved observables. In the end, (3.18) tells us that the accurate evolutionary equations of $\mathbf{g}_{\mathcal{M}}(t)$ always depend on (1) their instantaneous configuration, (2) their past history, and (3) an external "driving force" which depends on the initial configurations in the orthogonal space.

We make a remark that the second and third terms are zero if the dynamics is closed in $\mathcal{M}$, that is, $\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} = 0$, which corresponds to the scenario when we have a complete set of observables to describe the full dynamics. For example, this would be the $g_e := \exp(-1/x)$ for the dynamics $\dot{x} = -x^2$. The memory and the external driven force exist only because we have an incomplete observable set in $\mathcal{H}$.

We now drop the subscript $\mathcal{M}$ in $\mathbf{g}_{\mathcal{M}}$, as we only care about the dynamics of the resolved observables. By defining an $M \times M$ matrix $\mathbf{M} := \mathbf{L}_{\mathcal{M}\mathcal{M}}$, an $M \times M$ matrix parametrized by $\mathbf{K}(s) := -\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{s\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}$ parametrized by $s \in \mathbb{R}^+$, and an $M \times 1$ matrix $\mathbf{F}(t) := \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \cdot \mathbf{g}_{\bar{\mathcal{M}}}(0)$, we arrive at the GLE

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{g}(t) = \mathbf{M} \cdot \mathbf{g}(t) - \int_0^t \mathbf{K}(t-s) \cdot \mathbf{g}(s)\, \mathrm{d}s + \mathbf{F}(t). \tag{3.19}$$

In the rest of the paper, we refer to $\mathbf{M}$ as the *Markov transition matrix*, $\mathbf{K}(s)$ as the *memory kernel*, and $\mathbf{F}(t)$ as the *orthogonal dynamics*. Although $\mathbf{F}(t)$ is fully deterministic in the Mori–Zwanzig formalism for deterministic systems, it is often referred to as *noise* because of its resemblance to a Langevin noise in a Langevin equation. Note that the operators $\mathbf{M}$ quantify the interactions within the group of the relevant observables $\{g_i\}_{i=1}^M$ (that is, $\mathbf{L}_{\mathcal{M}\mathcal{M}}$). In contrast, the memory kernel combines the effects of the rest of the interactions ($\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}}$, $\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}$, and $\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}$)

Finally, we remark that despite their different terminologies, (3.19) and the operator algebraic formulation (3.10) are equivalent. When $g_i(\mathbf{x}) = x_i$, given an initial condition $\mathbf{\Phi}_0$, $g_i \circ \mathbf{\Phi}_0$ in (3.19) *is* $\phi_i(t, \mathbf{\Phi}_0)$ in (3.10).

3.3. Geometric interpretation of the GLE. The GLE describes the exact evolution of $\mathbf{g}(t)$ in $\mathcal{H}$. Figure 1 illustrates a schematic diagram of the dynamics of $\mathbf{g}(t)$ in the space $\mathcal{H}$. Because the GLE (3.19) is linear, the observables $\mathbf{g}(t)$ can be decomposed into two components: $\mathbf{g}(t) = \mathbf{g}_{\parallel}(t) + \mathbf{g}_{\perp}(t)$. We define the parallel component $\mathbf{g}_{\parallel}(t)$ as the general solution of the linear system satisfying

$$\dot{\mathbf{g}}_{\parallel}(t) = \mathbf{M} \cdot \mathbf{g}_{\parallel}(t) - \int_0^t \mathbf{K}(t-s) \cdot \mathbf{g}_{\parallel}(s)\, \mathrm{d}s, \tag{3.20a}$$

$$\mathbf{g}_{\parallel}(0) = \mathbf{g}(0). \tag{3.20b}$$

Because (3.20) is linear, it is clear that $\mathbf{g}_{\parallel}(t)$ is just a linear combination of the initial observables, i.e., we can always express $\mathbf{g}_{\parallel}(t) = \sum_{i=1}^M \alpha_i(t)\mathbf{g}_i(0)$ with some time-dependent coefficients $\alpha_i(t)$. Then, for any $t \geq 0$, $\left(\mathbf{g}_{\parallel}(t)\right)_i \in \mathcal{H}_{\mathbf{g}}$, $i = 1, \ldots, M$. The orthogonal component $\mathbf{g}_{\perp}(t)$ is the particular solution of the linear system with the driven force $\mathbf{F}(t)$ and satisfies

$$\dot{\mathbf{g}}_{\perp}(t) = \mathbf{M} \cdot \mathbf{g}_{\perp}(t) - \int_0^t \mathbf{K}(t-s) \cdot \mathbf{g}_{\perp}(s)\, \mathrm{d}s + \mathbf{F}(t), \tag{3.21a}$$

$$\mathbf{g}_{\perp}(0) = 0. \tag{3.21b}$$

Recall that the orthogonal dynamics $(\mathbf{F}(t))_i \in \mathcal{H}_{\bar{\mathbf{g}}}$, $i = 1, \ldots, M$. Then, again due to the linearity of (3.21), the orthogonal component $(\mathbf{g}_{\perp}(t))_i \in \mathcal{H}_{\bar{\mathbf{g}}}$ for any time $t \geq 0$.

The geometric representation in Figure 1 illustrates the closure problem in the Hilbert space: to evolve the exact $\mathbf{g}(t)$, one needs to provide the orthogonal dynamics $\mathbf{F}(t)$. Solving $\mathbf{F}(t)$ is as difficult as solving the full dynamics in the physical-space picture—we would need all the information of the orthogonal components, and thus the Mori–Zwanzig formalism has little to no advantage over the physical-space picture if the task is to solve for the exact

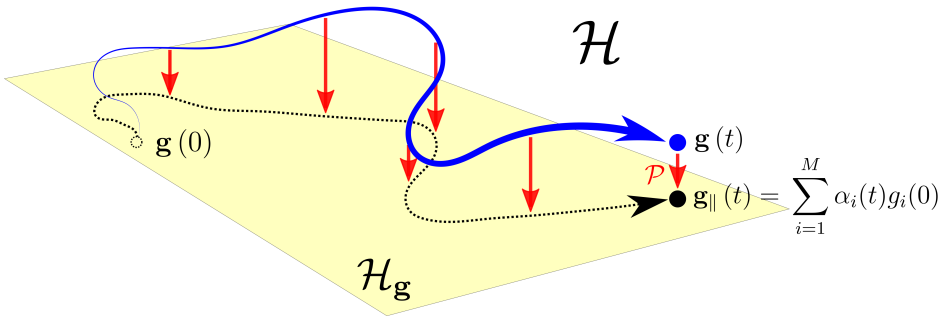

Figure 1. *Schematic diagram of the Mori–Zwanzig formalism. The Hilbert space $\mathcal{H}$ contains all possible observables which are functions of the initial conditions of the physical-space variables $\mathbf{\Phi}_0$. The subspace $\mathcal{H}_\mathbf{g}$ is linearly spanned by the set of selected observables Span$(\{g_i\}_{i=1}^M)$. We define the functions $g_i(t)$ parametrized by time t to be $g_i(t) \circ \mathbf{\Phi}_0 := g_i \circ \mathbf{\Phi}_t$, noting that $\mathbf{\Phi}(t)$ are functions of the initial condition $\mathbf{\Phi}_0$. At time $t = 0$, the vector $\mathbf{g}(0) = [g_1(0), \ldots, g_M(0)]^T$ is exactly the vector of the basis $\{g_i\}_{i=1}^M$, and thus $\mathbf{g}(0)$ is in $\mathcal{H}_\mathbf{g}$. As the nonlinear dynamics evolves, $\mathbf{g}(t)$ does not necessarily stays invariantly in $\mathcal{H}_\mathbf{g}$; in other words, $\mathbf{g}(t)$ is not necessarily a linear combination of $\{g_i\}_{i=1}^M$. Mori's projection operator $\mathcal{P}$ projects $\mathbf{g}(t)$ into $\mathcal{H}_\mathbf{g}$. The image is termed $\mathbf{g}_\parallel(t)$, which is a linear combination of $\{g_i\}_{i=1}^M$ and $\mathbf{g}_\parallel(t)$ satisfies (3.20).*

$\mathbf{g}(t)$. In fact, there is no free lunch by changing the Perron–Frobenius picture to the dual Koopman picture. To move forward, traditional analysis aims to select a set of slow-evolving observables—often referred to as the coarse-grained variables—and a noise model to replace the orthogonal dynamics $\mathbf{F}(t)$. The rationale of this approach is that if the selected observables are complete to describe the slow dynamics, the orthogonal dynamics shall live at a much faster timescale, and a proper noise model should suffice to *model* the dynamics of $\mathbf{F}(t)$. When a set of observables is given, and when the timescales of the resolved and underresolved observables are well-separated, Gottwald, Crommelin, and Franzke [12] provide a systematic slow-fast asymptotic analysis to homogenize the effect of the orthogonal noise. The challenge is that it is not a priori known what these observables are and what the corresponding noise model is, and identification of the observables and noise model is often made from educated guesses supported by domain-specific knowledge.

3.4. Mori's linear projection operator. Both the Koopman [18, 17, 27] and Mori–Zwanzig formulations operate in a Hilbert space, whose inner product is commonly defined as the inner product of two functions as the expected value of the product of the two with respect to a chosen measure $\mathrm{d}\mu$ (3.7). Despite the freedom to choose this measure, $\mathrm{d}\mu$ is conventionally set as a natural measure that is specific to the dynamics. For example, it is natural to adopt the canonical equilibrium distribution (Gibbs' measure) for equilibrium Hamiltonian systems [29]. For nonequilibrium systems, we can adopt the stationary measure as $\mathrm{d}\mu$ [9], as shown below. For stochastic and transient systems, it is natural to choose the induced time-dependent distribution [47].

With a defined inner product, Mori used an operator $\mathcal{P}$ that projects any function (of the initial condition $\mathbf{\Phi}_0$) $f \in \mathcal{H}$ onto the subspace $\mathcal{H}_\mathbf{g} := \mathrm{span}(\mathcal{M}) = \mathrm{span}\big(\{g_i\}_{i=1}^M\big)$. As such, the projected image can be expressed by $\mathcal{P}f := \sum_{i=1}^M \alpha_i g_i$ with coefficients α_i. Now, we decompose the function f by $f = \mathcal{P}f + f_\perp$, where $f_\perp \in \mathcal{H}_{\bar{\mathbf{g}}}$. Because $\langle f_\perp, g_j \rangle = 0$,

$j = 1, \ldots, M$, the inner product $\langle f, g_j \rangle$ is

$$\langle f, g_j \rangle = \langle \mathcal{P} f, g_j \rangle = \sum_{i=1}^{M} \alpha_i \langle g_i, g_j \rangle . \tag{3.22}$$

In vector notation, $\mathbf{g} := [g_1 \ldots g_M]^T$ and $\alpha := [\alpha_1 \ldots \alpha_M]$, the above equation can be expressed concisely by $\langle f, \mathbf{g}^T \rangle = \alpha \cdot \langle \mathbf{g}, \mathbf{g}^T \rangle$. Because the basis functions g_i's are linearly independent, $\langle \mathbf{g}, \mathbf{g}^T \rangle$ is a full-rank and invertible matrix. Then $\alpha = \langle f, \mathbf{g}^T \rangle \cdot \langle \mathbf{g}, \mathbf{g}^T \rangle^{-1}$, and we obtain the final expression for the projection operator

$$\mathcal{P} f = \langle f, \mathbf{g}^T \rangle \cdot \langle \mathbf{g}, \mathbf{g}^T \rangle^{-1} \cdot \mathbf{g}. \tag{3.23}$$

Note that if and only if the preselected set of functions $\{g_i\}_{i=1}^{M}$ is orthonormal with respect to the defined inner product, $\langle \mathbf{g}, \mathbf{g}^T \rangle^{-1}$ is an identity matrix and the expression can be simplified by $\mathcal{P} f = \langle f, \mathbf{g}^T \rangle \cdot \mathbf{g}$. Finally, we emphasize that the projection operator crucially depends on the choice of the defined inner product. In the rest of this article, we consider an inner product defined by averaging the observables over a long trajectory of the dynamical system:

$$\langle f, g \rangle := \lim_{T \to \infty} \frac{1}{T} \int_0^T (f \circ \mathbf{\Phi}(s)) \, (g \circ \mathbf{\Phi}(s)) \, \mathrm{d}s, \tag{3.24}$$

where $\mathbf{\Phi}(t)$ is the solution of (2.1) from any initial condition, assuming the choice of the initial condition does not change the long-time statistics.

3.5. The evolution of the projected image and time correlation matrix. The GLE (3.19) is not a closed dynamical system for the exact $\mathbf{g}(t)$ because it contains the generalized Langevin noise, $\mathbf{F}$, which is not known and is hard to obtain. However, because $\mathbf{g}(t) = \mathbf{g}_{\|} + \mathbf{g}_{\perp}$, the projected image $\mathcal{P}\mathbf{g}(t) = \mathcal{P}\mathbf{g}_{\|}(t) + \mathcal{P}\mathbf{g}_{\perp}(t) = \mathbf{g}_{\|}$ follows the closed evolutionary equation (3.20) and with an initial condition $\mathcal{P}\mathbf{g}(0) = \mathbf{g}(0)$. The L^2-norm of the residual error of the projected image $\|\mathbf{g} - \mathcal{P}\mathbf{g}\|_2^2 := \langle \mathbf{g} - \mathcal{P}\mathbf{g}, \mathbf{g} - \mathcal{P}\mathbf{g} \rangle$ is minimal among all the schemes that decompose $\mathbf{g}(t)$ into a *parallel* component and a *nonparallel* component. Thus, the projected image $\mathcal{P}\mathbf{g}(t)$ is the best approximation to the evolution of $\mathbf{g}(t)$ in the parallel space, and in this sense one can conceive it as the optimal predictor of the exact evolution $\mathbf{g}(t)$ into the future ($t > 0$). Somewhat interestingly, this shows that the optimal prediction into the future defined as above only depends on the choice of the inner product (and consequently the projection operator) and does not depend on the orthogonal dynamics $\mathbf{F}(t)$.

A related way to close the dynamics is to apply $\langle \cdot, \mathbf{g}^T \rangle$ to the GLE (3.19), resulting in the evolutionary equation for the two-time correlation function $\mathbf{C}(t) := \langle \mathbf{g}(t), \mathbf{g}^T \rangle$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{C}(t) = \mathbf{M} \cdot \mathbf{C}(t) - \int_0^t \mathbf{K}(t-s) \, \mathbf{C}(s) \, \mathrm{d}s, \quad \mathbf{C}(0) = \langle \mathbf{g}, \mathbf{g}^T \rangle . \tag{3.25}$$

Here, $\mathbf{C}(t)$ is the expected two-time correlation of the observables with respect to an initial condition $\mathbf{\Phi}_0$ distributed according to a long-time statistics $\mathrm{d}\mu$ (cf. (3.24)). Multiplying $\mathbf{C}^{-1}(0) \cdot \mathbf{g}(0)$ to (3.25) from the right, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{C}(t) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{g}(0) = \mathbf{M} \cdot \mathbf{C}(t) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{g}(0) - \int_0^t \mathbf{K}(t-s) \, \mathbf{C}(s) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{g}(0) \, \mathrm{d}s. \tag{3.26}$$

Comparing (3.26) to the evolutionary equation (3.20), we immediately identify the solution of the projected image $\mathbf{g}_{\|}(t)$:

$$\mathbf{g}_{\|}(t) = \mathbf{C}(t) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{g}(0). \tag{3.27}$$

Thus, the temporal correlation matrix $\mathbf{C}(t)$ encodes the information for optimally predicting the dynamics into the future. In addition, it is straightforward to solve the orthogonal component $\mathbf{g}_{\perp}(t)$ as a superposition of the orthogonal driven force $\mathbf{F}(t)$:

$$\mathbf{g}_{\perp}(t) = \int_0^t \mathbf{C}(t-s) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}s. \tag{3.28}$$

We illustrate the analysis in Appendix A.

As will be seen in section 5, (3.25) plays a pivotal role for data-driven learning of the Markov transition $\mathbf{M}$ and memory kernel $\mathbf{K}(s)$, $s \geq 0$. We will also see that the solution of the optimal prediction equation (3.27) establishes the equivalence between the Mori–Zwanzig formalism and the approximate Koopman learning algorithm in section 4.

3.6. Generalized fluctuation-dissipation relationship. With a suitable choice of the inner product, there exists a subtle relationship—often referred to as the generalized fluctuation-dissipation relationship—between the memory kernel $\mathbf{K}$ and the orthogonal dynamics $\mathbf{F}$:

$$\mathbf{K}(s) = \left\langle \mathbf{F}(s), \mathbf{F}^T(0) \right\rangle \mathbf{C}^{-1}(0). \tag{3.29}$$

Using the notation defined in section 3.2, we illustrate how this subtle relationship emerges. Explicitly, from (3.18), we identify the orthogonal dynamics

$$\mathbf{F}(t) = \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{g}_{\bar{\mathcal{M}}}, \tag{3.30}$$

because $\mathbf{g}_{\bar{\mathcal{M}}}(0) \equiv \mathbf{g}_{\bar{\mathcal{M}}}$. Next, with the choice of the temporal averaging inner product (3.24), we can obtain

$$\left\langle \mathbf{F}(t), \mathbf{F}^T(0) \right\rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{g}_{\bar{\mathcal{M}}} \circ \mathbf{\Phi}(s) \cdot [\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} \mathbf{g}_{\bar{\mathcal{M}}} \circ \mathbf{\Phi}(s)]^T \, \mathrm{d}s. \tag{3.31}$$

Induced by the dynamics equation (2.1), $\mathbf{g}_{\bar{\mathcal{M}}}(s) \equiv \mathbf{g}_{\bar{\mathcal{M}}} \circ \mathbf{\Phi}(s)$ and $\mathbf{g}_{\mathcal{M}}(s) \equiv \mathbf{g}_{\mathcal{M}} \circ \mathbf{\Phi}(s)$ satisfy (3.16), so we use the identity

$$\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} \mathbf{g}_{\bar{\mathcal{M}}}(s) = \frac{\mathrm{d}}{\mathrm{d}s} \mathbf{g}_{\mathcal{M}}(s) - \mathbf{L}_{\mathcal{M}\mathcal{M}} \mathbf{g}_{\mathcal{M}}(s) \tag{3.32}$$

to replace $\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} \mathbf{g}_{\bar{\mathcal{M}}}(s)$ in (3.31):

$$\begin{aligned}
\left\langle \mathbf{F}(t), \mathbf{F}^T(0) \right\rangle &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{g}_{\bar{\mathcal{M}}}(s) \cdot \left[\frac{\mathrm{d}}{\mathrm{d}s} \mathbf{g}_{\mathcal{M}}(s) - \mathbf{L}_{\mathcal{M}\mathcal{M}} \mathbf{g}_{\mathcal{M}}(s) \right]^T \mathrm{d}s \\
&= \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{g}_{\bar{\mathcal{M}}}(s) \cdot \frac{\mathrm{d}}{\mathrm{d}s} \mathbf{g}_{\mathcal{M}}^T(s)\, \mathrm{d}s \\
&\quad - \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \left\langle \mathbf{g}_{\bar{\mathcal{M}}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle \mathbf{L}_{\mathcal{M}\mathcal{M}}^T \\
&= \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{g}_{\bar{\mathcal{M}}}(s) \cdot \frac{\mathrm{d}}{\mathrm{d}s} \mathbf{g}_{\mathcal{M}}^T(s)\, \mathrm{d}s.
\end{aligned} \tag{3.33}$$

In the last line, we have used the fact that $\mathbf{g}_{\mathcal{M}}$ and $\mathbf{g}_{\bar{\mathcal{M}}}$ are orthogonal with respect to the inner product by construction. Next, we perform an integration by parts, assume that the boundary terms are bounded and thus converge to 0 after dividing by T in the limit $T \to \infty$, and use the full dynamics equation (3.16) again to obtain

$$\begin{aligned}
\left\langle \mathbf{F}(t), \mathbf{F}^T(0) \right\rangle &= -\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \frac{\mathrm{d}}{\mathrm{d}s} \mathbf{g}_{\bar{\mathcal{M}}}(s) \cdot \mathbf{g}_{\mathcal{M}}^T(s)\, \mathrm{d}s \\
&= -\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \left[\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} \mathbf{g}_{\mathcal{M}}(s) + \mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}} \mathbf{g}_{\bar{\mathcal{M}}}(s)\right] \cdot \mathbf{g}_{\mathcal{M}}^T(s)\, \mathrm{d}s \\
&= -\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \left[\mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} \left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle + \mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}} \left\langle \mathbf{g}_{\bar{\mathcal{M}}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle\right] \\
&= -\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}} \left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle. \qquad (3.34)
\end{aligned}$$

Again, we used the orthogonality $\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}_{\bar{\mathcal{M}}} \rangle = 0$. Finally, the memory kernel $\mathbf{K}(t) := -\mathbf{L}_{\mathcal{M}\bar{\mathcal{M}}} e^{t\mathbf{L}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}} \mathbf{L}_{\bar{\mathcal{M}}\mathcal{M}}$ can be expressed as the two-time correlation statistics of the orthogonal dynamics $\mathbf{F}(t)$ and the autocorrelation of the observables $\left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle$:

$$\mathbf{K}(t) = \left\langle \mathbf{F}(t), \mathbf{F}^T(0) \right\rangle \cdot \left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}_{\mathcal{M}}^T \right\rangle^{-1}. \qquad (3.35)$$

The above relationship between the two-time statistics of the orthogonal dynamics $\mathbf{F}(t)$ and the memory kernel $\mathbf{K}(t)$ is referred to as the generalized fluctuation-dissipation (GFD) relationship. In the operator algebraic derivation (cf. section 3.1), GFD holds when the Liouville operator $\mathcal{L} = \mathcal{L}(\Phi)$ is anti-self-adjoint with respect to the chosen inner product, i.e., for any test functions f and h of the physical-space variable $\mathbf{\Phi}$,

$$\langle f, \mathcal{L}h \rangle = -\langle \mathcal{L}f, h \rangle. \qquad (3.36)$$

For Hamiltonian systems, the anti-self-adjointness is guaranteed directly from the volume-preserving property of the dynamics [18]. For nonequilibrium systems, using long-time averaging as the inner product also has the anti-self-adjoint property:

$$\begin{aligned}
\langle f, \mathcal{L}h \rangle &= \left\langle f, \frac{\mathrm{d}}{\mathrm{d}t} h \right\rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T f \circ \mathbf{\Phi}(t) \frac{\mathrm{d}}{\mathrm{d}t} \left[g \circ \mathbf{\Phi}(t)\right] \mathrm{d}t \\
&= -\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \left[f \circ \mathbf{\Phi}(t)\right] g \circ \mathbf{\Phi}(t)\, \mathrm{d}t = -\left\langle \frac{\mathrm{d}}{\mathrm{d}t} f, h \right\rangle = -\langle \mathcal{L}f, h \rangle. \qquad (3.37)
\end{aligned}$$

All we need are the minor conditions that the integration by parts is valid, and negligible boundary terms which can be guaranteed for bounded systems.

Furthermore, if the dynamical system is ergodic, the temporal average (3.24) converges to (3.7), in which $\mathrm{d}\mu = \rho_{\mathrm{stat}}(\mathbf{\Phi}_0)\, \mathrm{d}^N\mathbf{x}$, where ρ_{stat} is the stationary density function and satisfies $e^{t\mathcal{L}^*} \rho_{\mathrm{stat}} = \rho_{\mathrm{stat}}$; here, the adjoint $\mathcal{L}^*$ defines the Perron–Frobenius operator. In the literature, it is generally presented that the anti-self-adjointness is valid for Hamiltonian systems in a heat bath, e.g., where there is an induced Gibbs measure [8]. Our analysis shows that GFD is generally valid, even for nonequilibrium (not necessarily Hamiltonian) systems, so long as we choose time averaging as the inner product and the observables along a long trajectory are bounded.

Finally, GFD can be considered as a self-consistent condition of the memory kernel $\mathbf{K}$ and the orthogonal dynamics $\mathbf{F}$. A real stochastic Langevin system with white noise (no time correlation) would result in $\mathbf{K}(t) = \mathbf{K}_0\,\delta(t)$, where $\mathbf{K}_0$ is a constant matrix and δ is the Dirac δ-distribution. The Mori–Zwanzig formalism shows that when a dynamical system is not fully resolved, in general, there exists a nonzero memory kernel $\mathbf{K}$, and thus the self-consistent orthogonal dynamics $\mathbf{F}$ must be a color noise. In section 5 we provide data-driven algorithms to extract the memory kernel and the noise directly from the measured observables along a long trajectory. The GFD can serve as a very stringent self-consistency check for the algorithms.

3.7. A discrete-time Mori–Zwanzig formalism. Even though we are interested in a continuous-time model, it is common that the observations—either empirical measurements of the physical system or the output of computer simulations—are discrete in time. It is also desirable to store the trajectories of a large set of observables sparsely sampled in time to mitigate storage limitations. In this case, the continuous-time Mori–Zwanzig formalism is not adequate. In this section, we provide the result of a discretized Mori–Zwanzig formulation that is more suitable for discrete-time data. For brevity, we only present the result in this section and leave the tedious yet straightforward derivation to Appendix B.

We consider observing the continuous-time system at discretization of times, $t = k\Delta$, $k \in \mathbb{Z}_{\geq 0}$ and with Δ not necessarily small. After integrating the GLE (3.19) and the evolutionary equations for the correlation matrix $\mathbf{C}(t)$ (equation (3.25)) and the optimal prediction $\mathcal{P}\mathbf{g}(t)$ (equation (3.27)) at the discrete times, the snapshots of the observables of interest satisfy very similar mathematical structures of the continuous-time formulations. Specifically, in Appendix B.1, we establish the discrete-time GLE (cf. (3.19))

$$\mathbf{g}((k+1)\Delta) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \cdot \mathbf{g}((k-\ell)\Delta) + \mathbf{W}_k, \tag{3.38}$$

where the $\mathbf{\Omega}_{\Delta}^{(\ell)}$'s are $M \times M$, Δ-dependent matrices which can be defined in terms of the continuous-time Markov transition matrix $\mathbf{M}$ and memory kernel $\mathbf{K}$, and $\mathbf{W}_k$ is the discrete-time orthogonal dynamics which can be decomposed into linear functions of the continuous-time orthogonal dynamics $\mathbf{F}$. Similarly, the snapshots of the correlation matrix $\mathbf{C}$ satisfy (cf. (3.25))

$$\mathbf{C}((k+1)\Delta) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \cdot \mathbf{C}((k-\ell)\Delta), \tag{3.39}$$

and the snapshots of the projected image $\mathcal{P}\mathbf{g}(t)$ satisfy (cf. (3.27))

$$\mathcal{P}\mathbf{g}((k+1)\Delta) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \cdot \mathcal{P}\mathbf{g}((k-\ell)\Delta). \tag{3.40}$$

It is tempting to associate the operator $\mathbf{\Omega}_{\Delta}^{(0)}$ to the continuous-time Markov matrix $\mathbf{\Omega}_{\Delta}^{(0)} \approx \mathbf{I} + \mathbf{M}\Delta$, where $\mathbf{I}$ is the $M \times M$ identity matrix, and to associate the operator $\mathbf{\Omega}_{\Delta}^{(\ell)}$, $\ell > 1$, to the

continuous-time memory kernel by $\mathbf{\Omega}_{\Delta}^{(\ell)} \approx \Delta^2 \mathbf{K}(\ell\Delta)$. We remark that these representations involving the infinitesimal $\Delta \ll 1$ are mathematically valid, but in general for finite Δ the expressions of the operators $\mathbf{\Omega}_{\Delta}^{(\ell)}$ in terms of the continuous-time objects ($\mathbf{M}$ and $\mathbf{K}(s)$) are not simple; see Lemma B.1 in Appendix B.1. Nevertheless, the above equations (3.38), (3.39), and (3.40) are always valid regardless of the choice of Δ. We also establish the GFD relationship in Appendix B.4.

For a generic discrete-time dynamical system, it is also possible to directly derive its Mori–Zwanzig formula (see Appendix B.2 and [24]). We remark that while the mathematical expression of the results of this approach look identical to (3.38)–(3.40), there exists a subtlety between discretizing the correlation matrix $\mathbf{C}(t)$ of a continuous-time system (Appendix B.1) and the generic discrete-time correlation matrix (Appendix B.2). We discuss such a subtlety in Appendix B.3. Furthermore, in Appendix B.4.2, we show that the GFD relationship is valid when the discrete-time operator $\mathcal{L}_d$, which forward propagates the observables (i.e., $\mathbf{g}((k+1)\Delta) = \mathcal{L}_d \mathbf{g}(k\Delta)$), is anti-self-adjoint with respect to the invariant measure of the discrete-time dynamics (cf. section 3.6).

4. Relation to the approximate Koopman learning methods. The discretized formulations presented in section 3.7 provide the mathematical structure for making a comparison to the approximate Koopman learning framework [35, 36, 46]. In this section, we establish that the Mori–Zwanzig formalism with Mori's projection operator is not only compatible with, but also a generalization of the Koopman learning framework.

We begin with a short summary of the approximate Koopman learning methods, specifically, the extended dynamic mode decomposition (EDMD) [46]. EDMD takes a long trajectory of a set of M observables, $\{g_i\}_{i=1}^{M}$, of a nonlinear dynamical system as an input. The snapshots of these descriptors on a uniformly separated temporal grid were recorded as $g_i(j\delta)$, $j = 0, \ldots, N-1$. Generally, δ is conceived as a small time separation. The goal of the EDMD is to identify the *approximate Koopman operator* $\mathbf{K}_{\delta}^{\mathrm{Koop}}$—note the unfortunate convention of $\mathbf{K}$ for memory kernel in the context of Mori–Zwanzig—which linearly maps the previous snapshots to the following consecutive ones with a minimal squared residual error. Operationally, EDMD stacks up the snapshots into an array of "dependent variables" $\mathbf{Y}$ and "independent variables" $\mathbf{X}$,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{g}_1(1\delta) & \cdots & \mathbf{g}_1((N-1)\delta) \\ \mathbf{g}_2(1\delta) & \cdots & \mathbf{g}_2((N-1)\delta) \\ \vdots & \ddots & \vdots \\ \mathbf{g}_M(1\delta) & \cdots & \mathbf{g}_M((N-1)\delta) \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{g}_1(0\delta) & \cdots & \mathbf{g}_1((N-2)\delta) \\ \mathbf{g}_2(0\delta) & \cdots & \mathbf{g}_2((N-2)\delta) \\ \vdots & \ddots & \vdots \\ \mathbf{g}_M(0\delta) & \cdots & \mathbf{g}_M((N-2)\delta) \end{bmatrix}, \tag{4.1}$$

and the $M \times M$ matrix $\hat{\mathbf{K}}_{\delta}^{\mathrm{Koop}}$ is the linear operator which minimizes the mean squared error

$$\varepsilon^2 := \frac{1}{N-1} \sum_{j=1}^{N-1} \sum_{i=1}^{M} \left(\mathbf{y} - \mathbf{K}_{\delta}^{\mathrm{Koop}} \cdot \mathbf{x} \right)_{i,j}^2. \tag{4.2}$$

When the observables form a complete set of basis functions, the dynamics is closed in the linear spanned space $\mathcal{H}_{\mathbf{g}}$, and ε^2 can be minimized to zero. Nevertheless, for general problems,

it is challenging to identify a complete set of basis functions. Consequently, there exist nonzero residuals. The learning problem is fundamentally a linear regression problem, which is concave. Formally, the unique minimizer $\hat{\mathbf{K}}_\delta^{\text{Koop}}$ conditioned on a pair of $\mathbf{Y}$ and $\mathbf{X}$ is

$$\mathbf{K}_\delta^{\text{Koop}} = \left(\mathbf{Y}\cdot\mathbf{X}^T\right)\cdot\left(\mathbf{X}\cdot\mathbf{X}^T\right)^{-1}, \tag{4.3}$$

when the set of the observables is linearly independent. When the set of the observables is not linearly independent, the problem is underdetermined and there exists a family of minimizers [46]; in such a case, the inverse matrix $\left(\mathbf{X}\cdot\mathbf{X}^T\right)^{-1}$ can be operationally carried out by the Moore–Penrose pseudoinverse [46] to identify one of the minimizers. In the analysis below, we assume that the set of the basis functions is carefully chosen so that they are linearly independent.

Equation (4.3) exhibits the exact mathematical expressions of the projected image $\mathbf{g}_\parallel(t)$ in the continuous-time Mori–Zwanzig formalism (equation (3.27)). In the limit of infinitely long snapshots separated by δ, the matrices $\mathbf{Y}\cdot\mathbf{X}^T$ and $\mathbf{X}\cdot\mathbf{X}^T$ are exactly $\mathbf{C}(\delta)$ and $\mathbf{C}(0)$ where the inner product is defined to be with respect to the distribution induced by the long trajectories:

$$\mathbf{C}(0) = \left\langle \mathbf{g}, \mathbf{g}^T\right\rangle = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{g}\circ\mathbf{\Phi}\,(s)\cdot\mathbf{g}^T\circ\mathbf{\Phi}\,(s)\,\mathrm{d}s \approx \mathbf{X}\cdot\mathbf{X}^T, \tag{4.4a}$$

$$\mathbf{C}(\delta) = \left\langle e^{\delta\mathcal{L}}\mathbf{g}, \mathbf{g}^T\right\rangle = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{g}\circ\mathbf{\Phi}\,(s+\delta)\cdot\mathbf{g}^T\circ\mathbf{\Phi}\,(s)\,\mathrm{d}s \approx \mathbf{Y}\cdot\mathbf{X}^T. \tag{4.4b}$$

Thus, both formulations predict the exact propagator forward δ-time $\mathbf{C}(\delta)\cdot\mathbf{C}^{-1}(0)$.

It is intriguing that the Mori–Zwanzig formulation relies on the projection operator $\mathcal{P}$, which requires the equipped inner product of the Hilbert space, but the approximate Koopman learning framework only relies on the mean L^2-norm of the error, which relies on a less strict norm space. Nevertheless, operationally, we can use the geometric interpretation provided in section 3.3 to illustrate the intuition of the two formulations. The Koopman learning framework seeks a point in $\mathcal{H}_\mathbf{g}$ that minimizes the L^2-norm between the point and $\mathbf{g}(\delta)$, which is generally outside $\mathcal{H}_\mathbf{g}$ (see Figure 1). In contrast, the Mori–Zwanzig formalism simply projects $\mathbf{g}(\delta)$ onto $\mathcal{H}_\mathbf{g}$. These two formulations are formally identical because the projected image $\mathcal{P}\mathbf{g}(\delta)$ would be the unique point on $\mathcal{H}_\mathbf{g}$ such that the L^2-error of the projected image to $\mathbf{g}(\delta)$, $\|\mathbf{g}(\delta)-\mathcal{P}\mathbf{g}(\delta)\|_2^2 = \langle\mathbf{g}(\delta)-\mathcal{P}\mathbf{g}(\delta), \mathbf{g}(\delta)-\mathcal{P}\mathbf{g}(\delta)\rangle$, is minimized. We note the subtle difference between the two frameworks. In the Koopman framework, orthogonality between the residual and the $\mathcal{H}_\mathbf{g}$ is not established—there is no notion of an inner product. In the Mori–Zwanzig formalism with the equipped inner product in $\mathcal{H}$, the unique operator $\mathbf{C}(k\delta)\cdot\mathbf{C}^{-1}(0)$, $k\in\mathbb{N}$, propagates $\mathbf{g}(0)$ to $\mathbf{C}(k\delta)\cdot\mathbf{C}^{-1}(0)\cdot\mathbf{g}(0)$, which is *always* the projected image of $\mathbf{g}(k\delta)$, and the residual $\mathbf{g}(k\delta)-\mathcal{P}\mathbf{g}(k\delta)$ is *always* orthogonal to $\mathcal{H}_\mathbf{g}$, with respect to the induced inner product. We remark that $\mathbf{C}(\delta)\cdot\mathbf{C}^{-1}(0)$ can also be derived from the Rayleigh–Ritz variational principle of the leading eigenvalues of either the Perron–Frobenius or Koopman operators using related algorithms such as time-lagged independent component analysis and Algorithm for Multiple Unknown Signals Extraction, commonly used and historically established in the Perron–Frobenius picture by the molecular dynamics community [42, 28, 33, 30, 31, 48, 16, 47].

We have established the equivalence of the Mori–Zwanzig formalism and the approximate Koopman learning framework. Mori–Zwanzig is more general than the existing Koopman learning methods. First, (3.40) prescribes the optimal prediction $\mathcal{P}\mathbf{g}(\Delta) = \mathbf{\Omega}_\Delta^{(0)}\mathbf{g}(0)$ if $\mathbf{g}(0)$ was sampled from the measure $\mathrm{d}\mu$ which was used to define the inner product, and the horizon Δ does not have to be small. Our analysis in Appendix B.1 shows that it is always possible to identify the unique (but Δ-dependent) operator $\mathbf{\Omega}_\Delta^{(0)}$, regardless of how large Δ is. Interestingly, in our derivation provided in Appendix B.1, we can see that $\mathbf{\Omega}_\Delta^{(0)}$ depends not only on the Markov transition $\mathbf{M}$ but also on the memory kernel $\mathbf{K}(s)$, $s \in [0, \Delta)$. This formally states that when the system is not fully resolved, the forward operator cannot be simply approximated by $e^{t\mathbf{M}}$. Instead, the linear operator $\mathbf{\Omega}_\Delta^{(0)}$, which can be estimated by our proposed algorithm in section 5, has an implicit memory-kernel ($\mathbf{K}(s)$, $s \in [0, \Delta)$) dependence, and thus we cannot simply exponentiate it for predicting the system further than Δ into the future. Second, the discretized GLE (3.38) states that the prediction shall be made with the past history, when it is available and when the discretized memory kernel $\mathbf{\Omega}_\Delta^{(\ell)}$, $\ell \geq 1$, is not zero. By taking into the account the past history, we will be simultaneously using the information in the parallel component $\mathbf{g}_\parallel(j\Delta)$ and the perpendicular component $\mathbf{g}_\perp(j\Delta)$ to forward propagate the system and thus reduce the prediction error. In contrast, in approximate Koopman learning, we only use the current time to predict the next time step and neglect the orthogonal contribution which would be estimated by the past trajectory in the context of the Mori–Zwanzig formalism. The advantage of the Mori–Zwanzig's history-dependent prediction will be numerically illustrated in section 6.2.

5. Learning the operators in the Mori–Zwanzig formalism from data. Conventionally, the aim of Mori–Zwanzig analysis is to select a set of coarse-grained variables to construct a parallel space $\mathcal{H}_\mathbf{g}$ in which the projected image $\mathcal{P}\mathbf{g}(t)$ captures the slow modes of the dynamics, mostly described by the Markov transition matrix $\mathbf{M}$. If the parallel space fully captures the slow modes, the orthogonal dynamics would predominantly be the fast modes of the dynamics. In such construction by timescale separation, the orthogonal dynamics $\mathbf{F}(t)$ is usually modeled by a generic stochastic process, and the self-consistent memory kernel can be constructed to establish the GLE (3.19) describing the slow modes of the dynamics [38, 13, 6, 22, 21, 23, 32, 44, 14].

The conventional approach is challenging because the selection of the observables requires sophisticated understanding of how to separate the modes with different timescales in a complex dynamical system. Nevertheless, the GLE (3.19) is always mathematically correct, and thus if we have collected a large enough set of data from a dynamical system, it is possible to numerically estimate the Markov transition matrix $\mathbf{M}$, memory kernel $\mathbf{K}$, and the orthogonal dynamics $\mathbf{F}$ directly from the collected data. Here, we provide two algorithms, one for continuous-time models (Algorithm 5.1) and the other for discrete-time models (Algorithm 5.2), for estimating the key quantities in the Mori–Zwanzig formalism directly from long trajectories of the observables. We remark that the scope of this study is restricted to noiseless observations of deterministic systems. For systems with some form of randomness, such as noisy observation on a deterministic system or a generic stochastic system, the estimation of stochastic Koopman operators [27, 46, 14, 48] requires additional mathematical construction (a probability space) that is beyond the scope of our theoretical analysis in the previous

sections. As such, we will defer these cases to future studies.

The procedure begins with calculating the two-time correlation functions $\mathbf{C}(t)$ of the observables from the recorded long trajectory. Then we exploit (3.25) for continuous-time systems and (3.39) for discrete-time systems to estimate continuous-time $\mathbf{M}$ and $\mathbf{K}$ and various orders of discrete-time $\mathbf{\Omega}^{(\ell)}$. After solving these operators, we will use (3.19) and (3.38) to solve for the orthogonal dynamics $\mathbf{F}$ (continuous time) and $\mathbf{W}_k$ (discrete time). Specifically, for the continuous-time system, we first set $t = 0$ in (3.25), leading to

$$\dot{\mathbf{C}}(0) = \mathbf{M} \cdot \mathbf{C}(0). \tag{5.1}$$

Thus, we can solve for the Markov matrix $\mathbf{M} = \dot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0)$. The estimation of $\dot{\mathbf{C}}(0)$ can be made by the finite-difference method or, alternatively, directly computed from the right-hand side of the dynamical equation (2.1) in the microscopic simulator. Next, we differentiate (3.25) with respect to time t once and obtain

$$\ddot{\mathbf{C}}(0) = \mathbf{M} \cdot \dot{\mathbf{C}}(0) - \mathbf{K}(0) \cdot \mathbf{C}(0). \tag{5.2}$$

The memory kernel evaluated at $t = 0$ is $\mathbf{K}(0) = \mathbf{M} \cdot \dot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0) - \ddot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0)$. Again, $\ddot{\mathbf{C}}(0)$ can either be accessed directly in the simulator or estimated by finite-difference data. Then we set $t = \delta \ll 1$ to (3.25) and approximate the memory integration by trapezoidal rule

$$\int_0^\delta \mathbf{K}(\delta - s) \cdot \mathbf{C}(s)\, \mathrm{d}s \approx \frac{\delta}{2} \left(\mathbf{K}(0) \cdot \mathbf{C}(\delta) + \mathbf{K}(\delta) \cdot \mathbf{C}(0)\right) \tag{5.3}$$

to solve for $\mathbf{K}(\delta) \approx \delta^{-1}\left[2\mathbf{M} \cdot \mathbf{C}(\delta) - 2\dot{C}(\delta) - \delta \mathbf{K}(0) \cdot \mathbf{C}(\delta)\right] \cdot \mathbf{C}^{-1}(0)$. Similarly, we can recursively solve for $\mathbf{K}((k+1)\,\delta)$, $k \in \mathbb{N}$, as functions of previously obtained $\mathbf{K}((\ell)\,\delta)$, $\ell \le k$, and the measured correlation matrices $\mathbf{C}$:

$$\begin{aligned} &\mathbf{K}((k+1)\delta) \\ &\quad \approx -2\left[\frac{\dot{\mathbf{C}}(k\delta) - \mathbf{M} \cdot \mathbf{C}(k\delta)}{\delta} + \sum_{\ell=1}^{k} \mathbf{K}(\ell\delta) \cdot \mathbf{C}((k-\ell)\,\delta) + \frac{\mathbf{K}(0) \cdot \mathbf{C}(k\delta)}{2}\right] \cdot \mathbf{C}^{-1}(0). \end{aligned} \tag{5.4}$$

Once $\mathbf{M}$ and $\mathbf{K}$ are solved, we use (3.19) and the measured trajectory to solve for the orthogonal dynamics $\mathbf{F}(t)$. A detailed description of our proposed procedure is presented as Algorithm 5.1.

As for the discrete-time dynamics, we use (3.39). Setting $k = 0$ in (3.39), $\mathbf{C}(\Delta) = \mathbf{\Omega}_\Delta^{(0)} \mathbf{C}(0)$ indicates that $\mathbf{\Omega}_\Delta^{(0)} = \mathbf{C}(\Delta) \cdot \mathbf{C}^{-1}(0)$, exactly the approximate Koopman operator that one would obtain by carrying out EDMD analysis (cf. section 4). Then, recursively, we can solve $\mathbf{\Omega}_\Delta^{(k)}$, $k \in \mathbb{N}$, in terms of the correlation matrices $\mathbf{C}$ and previously solved lower-order $\Omega_\Delta^{(\ell)}$, $\ell < k$, using (3.39):

$$\mathbf{\Omega}_\Delta^{(k)} = \left(\mathbf{C}((k+1)\,\Delta) - \sum_{\ell=0}^{k-1} \mathbf{\Omega}_\Delta^{(\ell)} \cdot \mathbf{C}((k-\ell)\,\Delta)\right) \cdot \mathbf{C}^{-1}(0). \tag{5.5}$$

Once the operators $\mathbf{\Omega}_\Delta^{(k)}$'s are solved, we use (3.38) and the measured trajectory to solve for the discrete-time noise $\mathbf{W}$. A detailed description of the procedure is presented as Algorithm 5.2.

Algorithm 5.1. Data-driven learning of the continuous-time Mori–Zwanzig operators. This algorithm uses a long trajectory of the reduced-order dynamics to estimate the continuous-time Markov transition matrix $\mathbf{M}$, memory kernel $\mathbf{K}(s)$, and orthogonal dynamics $\mathbf{F}(t)$ in the GLE (3.19). The algorithm requires the trajectories of M a priori selected observables $\{g_i\}_{i=1}^{M}$, measured at finely and evenly discretized times $t = k\delta$, $\delta \ll 1$, $k = 0, \ldots, N-1$, along the long ($N \gg 1$) trajectory. A number $0 \leq h < N$ is required to specify the longest estimated horizon of the memory kernel $\mathbf{K}$. The algorithm will deliver estimates of the (1) Markov transition matrix $\mathbf{M}$, (2) memory kernel at the discretized time points $\mathbf{K}(k\delta)$, and (3) the orthogonal dynamics $\mathbf{F}(k\delta|i)$, evaluated at discrete times $k = 0, \ldots, h-1$, conditioned on the system started at the ith snapshot (i.e., the system's initial condition was set at $\mathbf{g}(i\delta)$).

for k in $\{0, \ldots, h+1\}$ **do**
  $C_{ij}(k\delta) \leftarrow \frac{1}{N-k}\left[\sum_{\ell=0}^{N-k-1} g_i((k+\ell)\,\delta) \times g_j(\ell\delta)\right]$
end for
for k in $\{1, 2\}$ **do**
  $C_{ij}(-k\delta) \leftarrow C_{ji}(k\delta)$
end for
for k in $\{-1, 0, \ldots, h\}$ **do**
  $\dot{C}_{ij}(k\delta) \leftarrow \frac{1}{2\delta}\left[C_{ij}((k+1)\delta) - C_{ij}((k-1)\delta)\right]$
end for
for k in $\{0, \ldots, h\}$ **do**
  $\ddot{C}_{ij}(0) \leftarrow \frac{1}{2\delta}\left[\dot{C}_{ij}(\delta) - \dot{C}_{ij}(-\delta)\right]$
end for
$\mathbf{M} \leftarrow \dot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0)$
$\mathbf{K}(0) \leftarrow \mathbf{M} \cdot \dot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0) - \ddot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0)$
for k in $\{1, \ldots, h\}$ **do**
  $\mathbf{K}(k\delta) \leftarrow -2\left[\frac{\dot{\mathbf{C}}(0) - \mathbf{M}\cdot\mathbf{C}(k\delta)}{\delta} + \sum_{\ell=1}^{k-1} \mathbf{K}(\ell\delta) \cdot \mathbf{C}((k-\ell)\,\delta) + \frac{\mathbf{K}(0)\cdot\mathbf{C}(k\delta)}{2}\right] \cdot \mathbf{C}^{-1}(0)$
end for
for i in $\{0, \ldots, N-h-1\}$ **do**
  for k in $\{0, \ldots, h\}$ **do**
    $m \leftarrow \frac{\delta}{2}\left[\mathbf{K}(0) \cdot \mathbf{g}(k\delta) + 2\sum_{\ell=1}^{k-1} \mathbf{K}(\ell\delta) \cdot \mathbf{C}((k-\ell)\,\delta) + \mathbf{K}(k\delta) \cdot \mathbf{g}(0)\right]$
    $\mathbf{F}(k\delta|i) \leftarrow \frac{1}{\delta}\left[\mathbf{g}((k+1)\,\delta) - \mathbf{g}(k\delta)\right] - \mathbf{M} \cdot \mathbf{g}(k\delta) + m$
  end for
end for
return $\mathbf{M}$, $\mathbf{K}(k\delta)$, $\mathbf{F}(\ell\delta|i)$

6. Numerical experiment.

6.1. Test model.

In this section, we present the application of our proposed algorithms. Our aim is to illustrate the usage and the capability of the algorithms to extract the Mori–Zwanzig operators from a long trajectory of simulation.

Throughout this section, we adopt the Lorenz '96 system [25] as the test problem. The

Algorithm 5.2. Data-driven learning of the discrete-time Mori–Zwanzig operators. This algorithm estimates the discrete-time operators $\mathbf{\Omega}_{\Delta}^{(\ell)}$ and the orthogonal dynamics $\mathbf{W}$ in the discrete-time GLE (3.38) from a long trajectory of the dynamical simulations of the dynamical system. The algorithm requires the snapshots of a set of M a priori selected observables $\{g_i\}_{i=1}^{M}$ measured at evenly distributed times $t = k\Delta$, $k = 0, \ldots, N-1$, along a long ($N \gg 1$) trajectory. In contrast to Algorithm 5.1, the time separation of the snapshots Δ does not necessarily have to be small. A number $0 \leq h < N$ is required to specify the highest order of the discrete-time operators ($\mathbf{\Omega}^{(h)}$). The algorithm delivers estimates of (1) the discrete-time operators $\mathbf{\Omega}(k\Delta)$ and (2) the orthogonal dynamics $\mathbf{W}(k\Delta|i\Delta)$, $k = 0, 1, \ldots, h$, conditioned on the system started at the ith snapshot (i.e., the system's initial condition was set as $\mathbf{g}(i\Delta)$).

for k in $\{0, \ldots h+1\}$ **do**
  $C_{ij}(k\Delta) \leftarrow \frac{1}{N-k}\left[\sum_{\ell=0}^{N-k-1} g_i((k+\ell)\Delta) \times g_j(\ell\Delta)\right]$
end for
$\mathbf{\Omega}_{\Delta}^{(0)} \leftarrow \mathbf{C}(\Delta) \cdot \mathbf{C}^{-1}(0)$
for k in $\{1, \ldots h\}$ **do**
  $\mathbf{\Omega}_{\Delta}^{(k)} \leftarrow \left[\mathbf{C}((k+1)\Delta) - \sum_{\ell=0}^{k-1} \mathbf{\Omega}^{(\ell)} \cdot \mathbf{C}((k-\ell)\Delta)\right] \cdot \mathbf{C}^{-1}(0)$
end for
for i in $\{0, \ldots N-h-1\}$ **do**
  for k in $\{0, \ldots h\}$ **do**
    $\mathbf{W}(k\Delta|i\Delta) \leftarrow \mathbf{g}(i+k+1) - \sum_{\ell=0}^{k} \mathbf{\Omega}^{(\ell)} \cdot \mathbf{g}((i+k-\ell)\Delta)$
  end for
end for
return $\mathbf{\Omega}_{\Delta}^{(k)}$, $\mathbf{W}(k\Delta|i\Delta)$

system consists of N physical-space variables which evolve nonlinearly by

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_i(t) = (\phi_{i+1} - \phi_{i-2})\,\phi_{i-1} - \phi_i + F, \qquad i = 1, \ldots, N, \tag{6.1}$$

with the periodic boundary condition $\phi_{-1} = \phi_{N-1}$, $\phi_0 = \phi_N$, and $\phi_{N+1} = \phi_1$. We fixed the model parameter $N = 20$ and $F = 8$, with which the system has a chaotic behavior.

We define our observables to be $g_1(t) := \phi_1(t)$, $g_2(t) := \phi_4(t)$, $g_3(t) := \phi_8(t)$, and $g_4(t) = \phi_{14}(t)$. We also include a constant function $g_0 = 1$ because the long-time average of the observables is not zero-meaned. We remark that the results below are conditioned on the choice of the observables we make. We use the general-purpose integrator LSODA (by `scipy.integrate.solve_ivp`) to solve the evolutionary equation (6.1). LSODA uses adaptive time steps for controlling the error of the numerical integration and can handle stiff ODE systems. We chose a randomized initial condition, integrated the trajectory until $t = 10^5$, and recorded the snapshot of the observables every $\delta = 0.01$. The total length $t = 10^5$ was deemed sufficient from the convergence of the computed correlation matrix $\mathbf{C}$ between these chosen observables. We checked that the choice of the initial condition does not affect the obtained two-time correlations. Numerically, the trajectory is long enough such that the contribution of the initial transient behavior of the trajectory converging to the attractor is negligible.

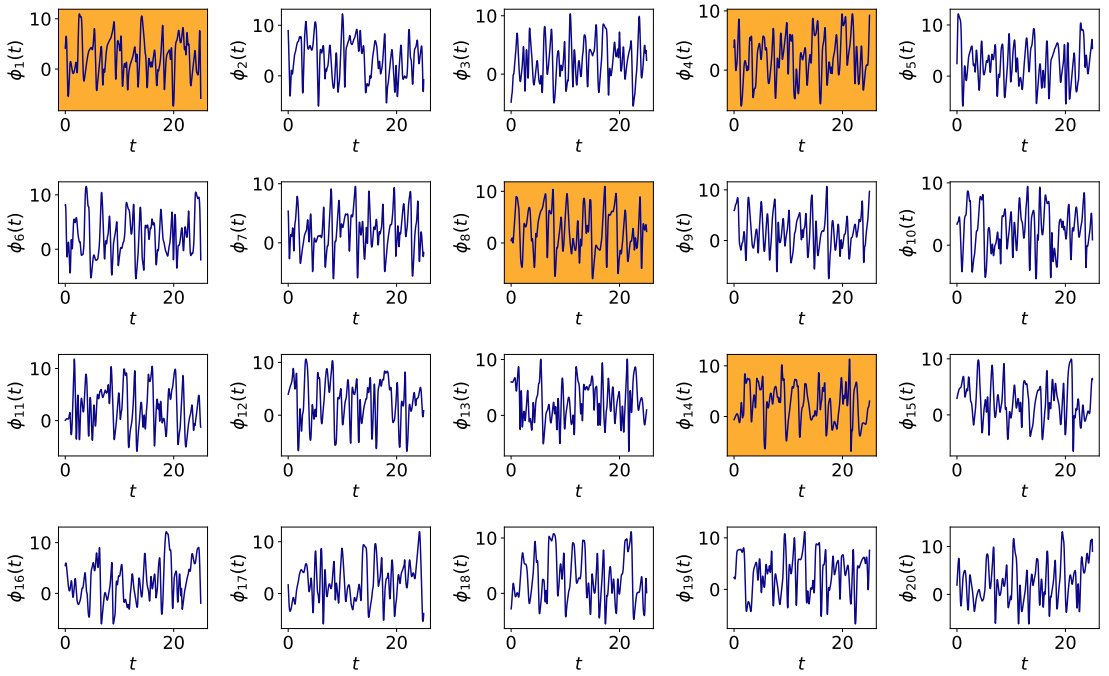

Figure 2. *The dynamics of the physical-space variables* $\{\phi_i\}_{i=1}^{20}$ *of the Lorenz '96 system (6.1). The highlighted panels are those chosen variables which serve as the observables* $g_1 := \phi_1$, $g_2 := \phi_4$, $g_3 := \phi_8$, *and* $g_4 := \phi_{14}$. *Because the observables are not zero-meaned, we also include a constant function* $g_0 := 1$. *For the reference of the reader, the estimated Lyapunov time* ≈ 0.622.

A short-time behavior of the dynamical system is shown in Figure 2, where the observables are highlighted. We numerically computed the Lyapunov exponents of the system and determined that the largest Lyapunov exponent ≈ 1.608, which corresponds to a Lyapunov time $\approx 1/1.608 \approx 0.622$. The estimated Kaplan–Yorke dimension [11] of the system ≈ 13.37.

We use the collected snapshots of the observables to compute two-time correlation function up to a lag $t = 10$, as shown in Figure 3. Then we apply Algorithm 5.1 to the numerically computed $\mathbf{C}(t)$ to extract the continuous-time Markov matrix

$$\mathbf{M} = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & -0.052 & -0.416 & 0.215 & -0.175 \\ 0.652 & 0.372 & 0.099 & -0.795 & 0.047 \\ -2.333 & -0.066 & 0.787 & -0.068 & 0.342 \\ 0.338 & 0.171 & 0.043 & -0.378 & 0.021 \end{bmatrix}, \tag{6.2}$$

and the memory kernel $\mathbf{K}(t)$, illustrated in Figure 4. With the calculated kernel $\mathbf{K}$, Algorithm 5.1 also quantifies the orthogonal dynamics, $\mathbf{F}(t, \mathbf{\Phi}(s))$ for $t \geq s$, which allows us to calculate the right-hand side of the GFD relationship (3.29). The GFD, which servers as a stringent self-consistent condition, is numerically verified and presented in Figure 4.

We also applied Algorithm 5.2 to extract the discrete-time operators $\mathbf{\Omega}_\Delta^{(\ell)}$. We first fix $\Delta = 30\delta$ and identify the lowest order $\mathbf{\Omega}_\Delta^{(0)}$ which serves like the Markov transition matrix in

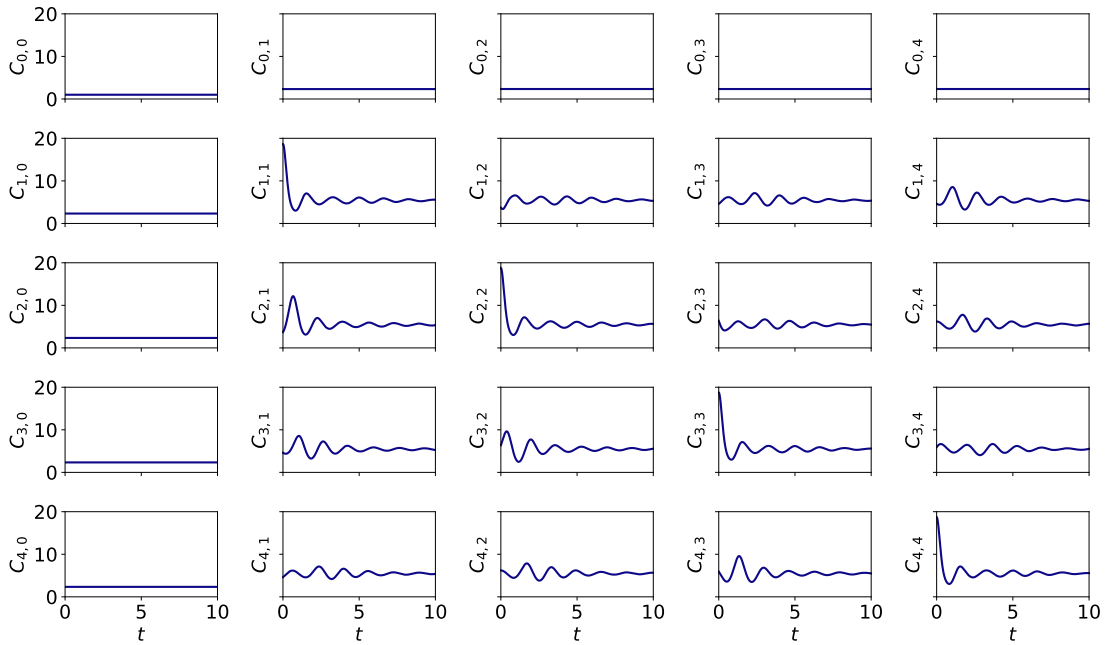

Figure 3. *The two-time correlation function* $\mathbf{C}(t)$ *computed from snapshots along a long* (10^5) *trajectory* $C_{ij}(k\delta) = 10^{-5} \times \sum_{\ell=1}^{10^5-k} g_i((k+\ell)\,\delta)\, g_j(k\delta)$. *The estimated Lyapunov time* ≈ 0.622.

the discrete-time formulation:

$$\mathbf{\Omega}_{\Delta}^{(0)} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.827 & 0.294 & -0.062 & 0.036 & -0.053 \\ 1.378 & 0.173 & 0.332 & -0.108 & 0.016 \\ 0.925 & -0.015 & 0.27 & 0.284 & 0.064 \\ 1.771 & 0.026 & 0.012 & -0.105 & 0.311 \end{bmatrix}, \tag{6.3}$$

We present the higher-order operators, $\mathbf{\Omega}_{\Delta}^{(\ell)}$, in Figure 5. Algorithm 5.2 also quantifies the discrete-time orthogonal dynamics $\mathbf{W}_k$, which we used to evaluate the right-hand side of the discrete-time GFD relationship (B.28). Again, the stringent self-consistent GFD illustrates that our numerical analysis quantifies the operators accurately.

For a comparison between the continuous-time and discrete-time operators, we compute the propagator $\exp(\Delta \times \mathbf{M})$ using the continuous-time Markov operator $\mathbf{M}$ computed in (6.2):

$$\exp\left(\Delta \times \mathbf{M}\right) = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.256 & 0.976 & -0.117 & 0.08 & -0.049 \\ 0.296 & 0.114 & 0.995 & -0.234 & -0.001 \\ -0.658 & -0.003 & 0.236 & 0.946 & 0.103 \\ 0.149 & 0.052 & -0.004 & -0.111 & 0.999 \end{bmatrix}. \tag{6.4}$$

The difference between the lowest-order Markov operators, (6.3) and (6.4), illustrates the difference between the continuous-time and the discrete-time formulations. The discrete Markov

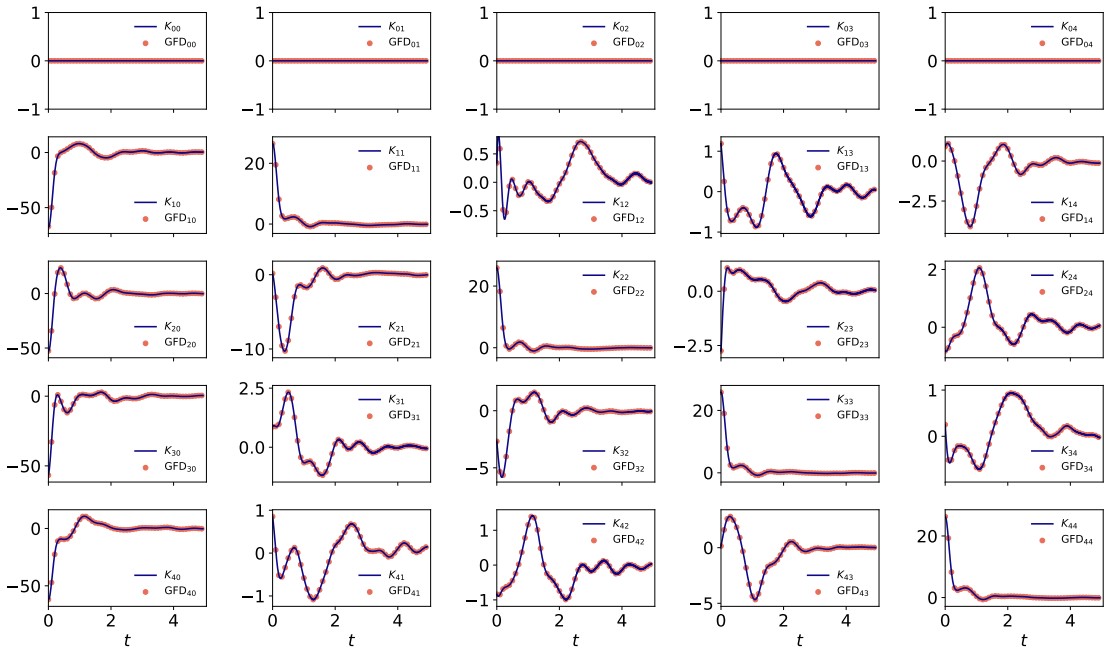

Figure 4. *The memory kernel* $\mathbf{K}(t)$ *is illustrated as the solid line. To verify the GFD relationship, the right-hand side of* (3.29)*,* $GFD_{ij} := \left[\langle \mathbf{F}(t) \cdot \mathbf{F}^T(0) \rangle \cdot \mathbf{C}^{-1}(0)\right]_{ij}$*, is calculated and plotted as the discrete dots, which perfectly align with the memory kernel* $\mathbf{K}$*. The timescale of nontrivial memory kernel exceeds the estimated Lyapunov time* ≈ 0.622*.*

operator $\mathbf{\Omega}_\Delta^{(0)}$ is identical to the approximate Koopman operator if one would carry out the EDMD [46] with a time separation $\Delta = 30\delta = 0.3$. Nevertheless, as pointed out in section 4, although the operator $\mathbf{\Omega}_\Delta^{(0)}$ can always be estimated by data (using Algorithm 5.2) and is optimal to predict one-step Δ into the future, it cannot be approximated by exponentiating the instantaneous Markov operator of the continuous-time formulation, $\exp(\Delta \times \mathbf{M})$. Our analysis in Appendix B.1 shows that $\mathbf{\Omega}_\Delta^{(0)}$ contains not only the effect of continuous-time Markov operator $\mathbf{M}$ which is the exact Koopman operator of the dynamics, but also the effect of the continuous-time memory kernel $\mathbf{K}(s)$, $0 \le s \le \Delta$.

The continuous-time and discrete-time memory kernels shown in Figures 4 and 5 show similar behavior. Note that in the GLE (3.19), it is conventional that the memory kernel $\mathbf{K}$ comes with an overall negative sign in front, but in the discrete formulation (equation (3.38)), it is more natural to refrain from putting the -1 in front of the $\ell \ge 1$ terms. Thus, to make a comparison between Figures 4 and 5, one of them shall be flipped upside-down. We point out a few important observations in these numerical estimations of the memory kernels. First, it shows that in both formulations, the memory kernels decay to zero at a finite timescale. The finite timescale of the memory kernel indicates that operationally we do not necessarily need the full history of the system from $t = 0$ until the current time to make a prediction—the trajectory with the finite timescale is sufficient. Second, the analysis shows that the memory kernels could behave very nontrivially, and the kernels are not likely captured by

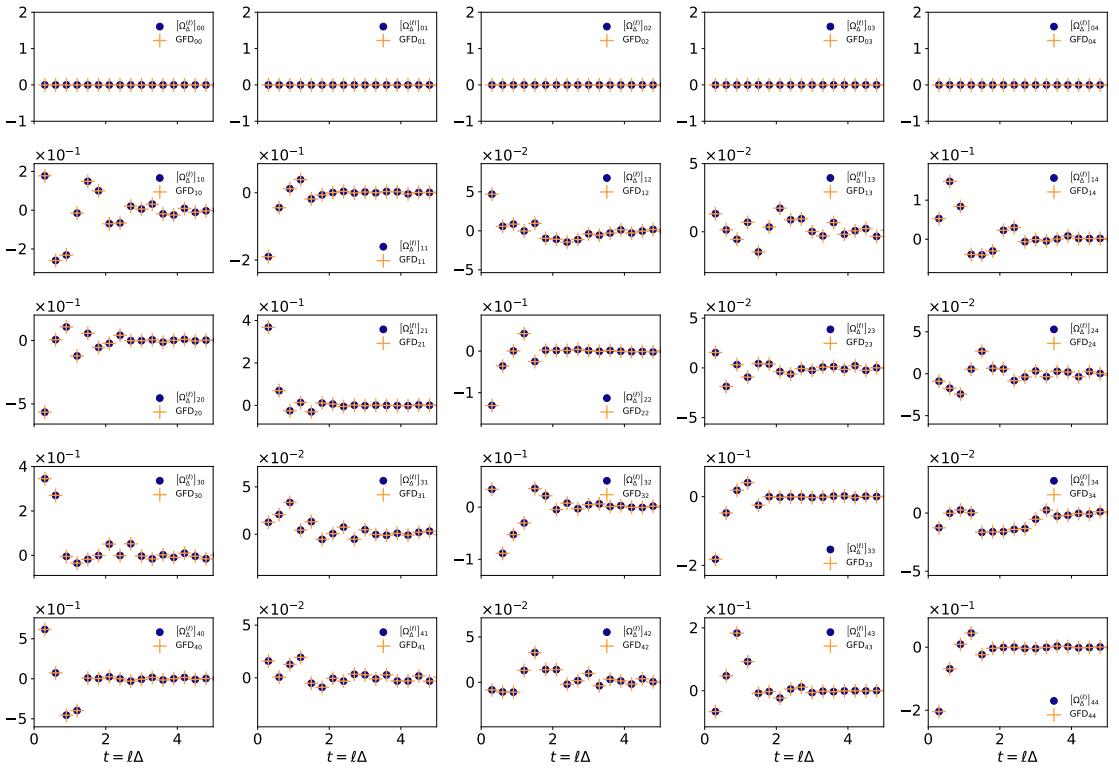

Figure 5. *We present the extracted higher-order discrete-time operators* $\mathbf{\Omega}_\Delta^{(\ell)}$, $\ell \geq 1$, *in discrete dots. To verify the discrete-time GFD relationship, the right-hand side of* (B.28), $GFD_{ij} := \left[\langle \mathbf{W}_\ell \cdot \mathbf{W}_0^T \rangle \cdot \mathbf{C}^{-1}\,(-\Delta)\right]_{ij}$, *is calculated and plotted as the discrete pluses, which align with the memory kernel* $\mathbf{\Omega}_\Delta^{(\ell)}$.

simple models. Third, the discrete-time memory kernel seems to be a coarse-grained and smoothed-out picture of the continuous-time kernel.

To investigate further into the Δ-dependence smoothing of the discrete memory operators, we calculate $\mathbf{\Omega}_\Delta^{(\ell)}$, $\ell \geq 1$, with $\Delta = 0.01$, 0.1, and 0.2. Note that the smallest $\Delta = 0.01$ corresponds to the smallest time resolution $\delta = 0.01$, which we used to compute the continuous-time operators. To properly scale and compare their behavior, we need to scale the discrete operators by Δ^{-2}. The first scaling Δ comes from the fact that for a fixed physical memory timescale, a larger Δ comes with fewer snapshots in the discrete sum of the past history. Another way to understand this scaling is that $\mathbf{\Omega}$ in (3.38) is analogous to $-\mathbf{K}(s)\,\mathrm{d}s$ in (3.19), and the scaling comes from $\Delta \sim \mathrm{d}s$. The second scaling comes from the fact that the discrete operators map the system forward to Δ. We present the scaled operators, $\Delta^{-2}\mathbf{\Omega}_\Delta^{(\ell)}$, in the same Figure 6, which shows that when $\Delta = \delta = 0.01$, the calculated discrete-time operator by Algorithm 5.2 converges to the continuous-time kernel calculated using continuous-time Algorithm 5.1. The smoothing as we increase Δ can also be observed. With this comparison, we recommend the discrete-time formulation as it requires fewer inputs but achieves comparable results of the continuous-time formulation which needs estimation of $\dot{\mathbf{C}}(0)$ and $\ddot{\mathbf{C}}(0)$, when the time separation Δ is set equal to the fine discretization δ of the continuous-time model.

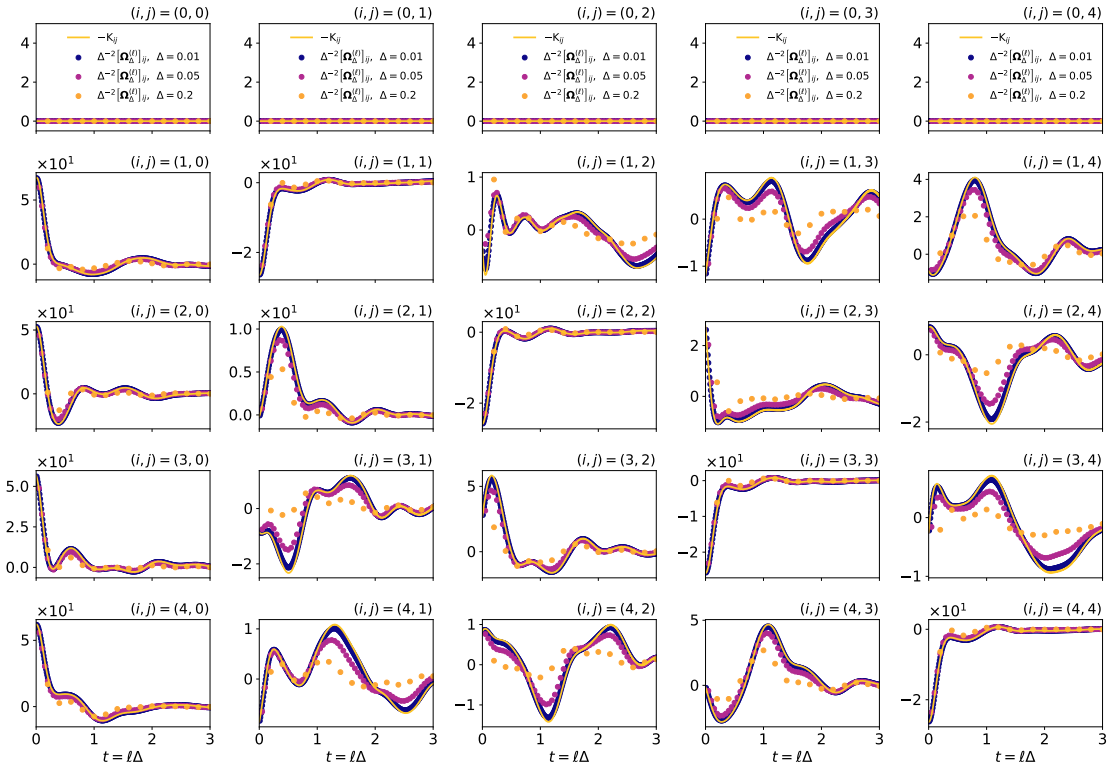

Figure 6. *Comparisons between the discrete-time memory operators* $\mathbf{\Omega}_\Delta^{(\ell)}$, $\ell \geq 1$, *with various* Δ *and the continuous-time memory kernel* $-\mathbf{K}(t)$. *Note that the continuous-time memory kernel in the generalized Langevin equation* (3.19) *has a sign different from the discrete-time formulation* (3.38). *To make a comparable visualization, we compare the discrete-time operators to the negative of the continuous-time memory kernel,* $-\mathbf{K}$.

6.2. Advantage of the Mori–Zwanzig formalism in prediction. In this section, we compare different numerical procedures for making a prediction to illustrate the advantage of the Mori–Zwanzig formalism over plain Koopman analysis when the set of basis functions is not complete. Specifically, we consider the following problem setup. Suppose we obtain snapshots of a set of observables along a long trajectory. These snapshots, separated by a fine time resolution δ, were used to compute either the approximate Koopman operator (cf. section 4) or the Mori–Zwanzig operators (cf. Algorithm 5.2). Next, suppose we were given the snapshots of the observables along a segment of trajectory of length $\zeta = m\delta$, $m \in \mathbb{N}$. We denote the snapshots by $\{\mathbf{g}(k\delta)\}_{k=1}^m$, noting that $\mathbf{g}$ is an $M \times 1$ column vector. We are interested in using the given snapshots to predict the observables $\eta = n\delta$, $n \in \mathbb{N}$, in the future. Specifically, we are interested in comparing the prediction errors of Koopman and discrete-time Mori–Zwanzig formulations:

1. The recursive Koopman computation. With this approach, we use the EDMD [46] to compute the approximate Koopman operator $\mathbf{K}_\Delta^{\text{Koop}}$ with a time separation $\Delta = k\delta$. Because the aim is to predict $n\delta$ into the future, k is restricted to the set of divisors of n. When $k = 1$,

the learned approximate Koopman operator would be closest to the true Markov operator in the Mori–Zwanzig formalism. When $k = n$, the time separation is chosen to be exactly the predictive horizon. We then apply n/k times the learned Koopman operator to advance the last known observables to make a prediction, i.e., $\mathbf{g}\left((m+n)\,\delta\right) \approx \left(\mathbf{K}_{\Delta}^{\text{Koop}}\right)^{n/k} \cdot \mathbf{g}\left(m\delta\right)$.

2. The recursive discrete-time Mori–Zwanzig computation. With this approach, we use Algorithm 5.2 to compute the Mori–Zwanzig operators $\mathbf{\Omega}_{\Delta}^{(\ell)}$. We are interested in different magnitudes of Δ, noting a restriction that they must be multiples of the finest time resolution δ. We chose $\Delta = 0.01$, 0.02, 0.05, 0.1, and 0.2. We are also interested in the predictive error as a function of the memory length, which is restricted as multiples of Δ. We will integrate (3.38) to advance $\mathbf{g}(t)$ to $\mathbf{g}(t+\Delta)$, noting that it is not possible to estimate the exact noise $\mathbf{W}_k$. Instead of injecting artificially generated samples from a noise model, we set the $\mathbf{W}_k$'s to zero. Note that the number of steps needed for such integration is also Δ-dependent. For example, when Δ matches the predictive horizon ($\Delta = n\delta$), we only need to integrate (3.38) once; when Δ is the finest time resolution ($\Delta = \delta$), we need to iteratively integrate (3.38) n times and accumulate the prediction $\mathbf{g}\left((m+k)\,\delta\right)$, $1 \le k < n$, as past history until the horizon is met.

We remark that when the memory length is set to zero in the recursive discrete-time Mori–Zwanzig approach, the method converges to the recursive Koopman approach because $\mathbf{K}_{\Delta}^{\text{Koop}} \equiv \mathbf{\Omega}_{\Delta}^{(0)}$. Thus, the second family of numerical procedures (with different settings of Δ and memory length) is a superset of the first one.

We adopt the L^2-norm as the measure to compare errors between different methods. That is, suppose a method made a prediction $\mathbf{g}^{\text{pred}}$ and suppose the ground truth is $\mathbf{g}^{\text{GT}}$; the error is computed by

$$\varepsilon^2 := \left\| \mathbf{g}^{\text{pred}} - \mathbf{g}^{\text{GT}} \right\|_2^2 \equiv \sum_{i=1}^{M} (g_i^{\text{pred}} - g_i^{\text{GT}})^2. \tag{6.5}$$

To collect the statistics of the prediction error, we generate 2×10^4 samples of segments, each of which contains $m = 500$ snapshots (separated by $\delta = 0.01$), sampled from a long ($t = 10^5$) test trajectory. Because we are interested in out-of-sample prediction, the long test trajectory used to generate test samples is different from the one we used to compute the correlation matrix (and the operators). Both trajectories were generated by integrating the same evolutionary equation (6.1) with two randomized initial conditions; the procedure ensures a consistent inner-product space which is defined by the long-time statistics.

We present the error statistics from the Mori–Zwanzig formulation with different memory and prediction horizon in Figure 7. In Figures 7(a) and (b), we present the average error over 2×10^4 samples when the prediction horizon is chosen as $n\delta = 0.2$ and 0.8 (20 and 80 finest time steps, $\delta = 0.01$), respectively. The recursive Koopman computation corresponds to those points with zero memory length. A few important observations can be made. First, for approximate Koopman computation, the most accurate way is to match the time separation Δ to the prediction horizon $n\delta$. This observation is consistent with our reasoning in section 3.5, i.e., that the optimal prediction $n\delta$ into the future is $\mathbf{C}(n\delta) \cdot \mathbf{C}^{-1}(0)$, which is the approximate Koopman operator $\mathbf{K}_{n\delta}^{\text{Koop}}$ (cf. section 4). Second, as the length of the past history is increased, the error decreases but levels off at a finite timescale which depends on the

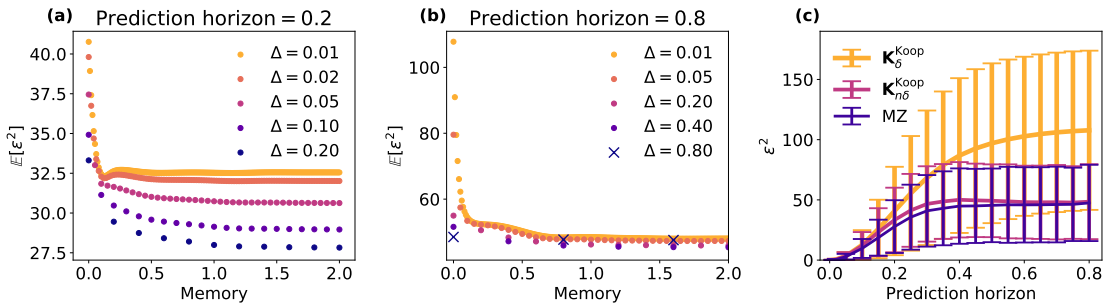

Figure 7. *Comparison of the prediction error of different numerical procedures.* (a) *The mean error of the Mori–Zwanzig formulation for predicting the observable* $n\delta = 0.2$ *into the future.* (b) *The mean error of the Mori–Zwanzig formulation for predicting the observable* $n\delta = 0.8$ *into the future.* (c) *We compare the statistics of the prediction error as the function of the prediction horizon* $n\delta$ *from* (1) *the approximate Koopman operator* $\mathbf{K}_\delta$, (2) *the approximate Koopman operator* $\mathbf{K}_{n\delta}$, *and* (3) *recursive Mori–Zwanzig computation. For the reader's reference, the estimated Lyapunov time* ≈ 0.622.

choice of Δ. The improvement comes from an important fact that the given segment of trajectory $\mathbf{g}(k\delta)$, $k = 1, \ldots, m$, contains information in both the parallel and orthogonal spaces; see the decomposition equations (3.20) and (3.21). By integrating GLE forward in time with the segment of trajectory, we can evolve both the parallel and orthogonal dynamics beyond $t = m\delta$, despite the fact that we cannot access the unknown orthogonal dynamics $\mathbf{F}(t \geq m\delta)$, which is thus set to be zero operationally for making predictions. In contrast, the Koopman approach always predicts from the last snapshot at $t = m\delta$ and can only optimally predict in the parallel space by (3.20) and omits the orthogonal contributions (3.21). Interestingly, the optimal choice of Δ, the separation between snapshots for learning, depends nontrivially on the memory length and the prediction horizon in the Mori–Zwanzig formulation. For a prediction horizon 0.2 (Figure 7(a)), it is optimal to match $\Delta = 0.2$ regardless of the memory length; however, for a prediction horizon 0.8 (Figure 7(b)), it is more advantageous to adapt a mismatched $\Delta = 0.4$ in the long-memory kernel regime. Regardless, in all cases, we found that including the memory contribution improves the prediction, compared to the memoryless Koopman prediction. Finally, we observed that the memory length does not need to be long before the accuracy levels off.

Consistent with these results, it is not expensive to improve the accuracy using the discrete-time Mori–Zwanzig formulation; we only need to store a finite length of snapshots to carry out the discrete sum in (3.38). In Figure 7(c), we present the error as a function of the prediction horizon $n\delta$. In addition to the average error, we also visualize the standard deviation of the ε^2 in 2×10^4 sample segments. We compare three methods: Koopman with the smallest time separation $\Delta = \delta$, Koopman with the matched time separation $\Delta = n\delta$, and discrete Mori–Zwanzig with $\Delta = n\delta$ with a fixed memory length 0.24. We conclude that the Mori–Zwanzig formulation is consistently more accurate in making predictions into the future, and if one must carry out the approximate Koopman analysis, it is best to match the time separation Δ to that of the prediction horizon—learning an approximate Koopman operator with a short time separation Δ and recursively applying $\mathbf{K}^{\text{Koop}}_{\Delta < n\delta}$ invokes more error.

Finally, it is worth pointing out that, similar to the Koopman analysis, the quality of the prediction crucially depends on the selection of the observables. In our test problems above, we only considered five simple observables $\{1, \phi_1, \phi_4, \phi_8, \phi_{14}\}$ and performed apples-to-apples comparison between the Koopman and Mori–Zwanzig predictions. We emphasize the result that the Mori–Zwanzig, with the same functional basis, improves the prediction of the plain Koopman analysis, despite the fact that the absolute error of the Mori–Zwanzig prediction is still large (see Figures 7(a) and (b)). To further improve the prediction, one then needs to optimize a set of observables, which is an important procedure but not the focus of this article.

7. Discussion. In this article, we showed that the Mori–Zwanzig formalism with Mori's projector [29] is not only consistent with but also a generalization of the existing Koopman learning procedures, such as the extended dynamic mode decomposition (EDMD [46]). The propagator of the projected image $\mathbf{C}(\Delta) \cdot \mathbf{C}^{-1}(0)$ for any time separation Δ is identified as the discrete-time approximate Koopman operator $\mathbf{K}_{\Delta}^{\text{Koop}}$. We identify that the propagator contains not only the effect of the instantaneous Markov transition matrix $\mathbf{M} \equiv \dot{\mathbf{C}}(0) \cdot \mathbf{C}^{-1}(0)$, but also the memory effect from the trajectory between $t = 0$ to $t = \Delta$ (Lemma B.1). Such a history dependence emerged because the dynamics is not fully resolved—that is, the dynamics does not evolves invariantly in functional subspace $\mathcal{H}_{\mathbf{g}}$ spanned by the selected observables. Although the Markov transition matrix $\mathbf{M}$ is the instantaneous Koopman operator, the finite-time propagator is $\exp(\Delta \mathbf{M})$ only when the dynamical system is fully resolved. In those partially observed cases, there always emerge a history-dependent term and the orthogonal dynamics from the unresolved degrees of freedom of the dynamics.

Motivated by the data-driven Koopman learning methods, we constructed two numerical algorithms which extract key operators in the Mori–Zwanzig formalism, specifically, with Mori's projection operator. One of our algorithms extracts the continuous-time operators, and the other the discrete-time ones. To the lowest order, the algorithms are operationally identical to the EDMD, and they compute the continuous-time Markov matrix $\mathbf{M}$ and the discrete-time approximate Koopman operator $\mathbf{\Omega}_{\Delta}^{(0)}$. The novelty of our algorithms is that they go beyond the lowest order and proceed with a recursive procedure which uses further two-time correlations to extract the memory kernels (continuous-time $\mathbf{K}(s)$ and discrete-time $\mathbf{\Omega}_{\Delta}^{(\ell)}$, $\ell \geq 1$). The orthogonal dynamics can be computed after the Markov transition matrix and the memory kernel are obtained.

Our proposed algorithms provide an alternative data-driven way of applying the Mori–Zwanzig formalism to study dynamical systems. Our approach bypasses the conventional need of modeling the memory kernel and the orthogonal dynamics. The numerical analysis on the test problem shows the complex behavior of the memory kernel which is not likely to be modeled by simple mathematical models. Importantly, we numerically verified that the extracted memory kernel and orthogonal noise satisfy the self-consistent generalized fluctuation-dissipation relationship. To our knowledge, it is the first time that the GFD is numerically verified on a nontrivial model whose analytic solution is not known. We are confident with the validity of the extracted memory kernel because of the rather stringent GFD relationship.

We showed that the Mori–Zwanzig formalism with the numerically extracted memory kernel and past history can significantly improve the accuracy of the prediction of the EDMD

[46]. The reason is that the past history contains partial information of the orthogonal dynamics, $\mathbf{g}_\perp(t)$, and the prediction can be improved by incorporating the information. We remark that such memory-dependent learning is fundamentally different from the recently proposed time-embedded Koopman learning methods [13, 3, 4, 20, 15, 45, 24], which include the past history in the set of the observables. These methods, motivated by the famous Takens embedding theorem [40], require one to expand the number in the set of variables by h times, where h is the number of the past snapshots and h has to be specified before the computation. With such a construction, the subspace that functions are projected to is $\mathrm{span}(\{\mathbf{g}\left((h-1)\,\Delta\right)\ldots\mathbf{g}(\Delta),\mathbf{g}(0)\})$, which is larger than $\mathcal{H}_\mathbf{g}$. Although the one-step projection using past history seems to be mapped by a memory kernel, a more proper way to interpret the operation is to regard the augmented configuration mapped by the Markov transition matrix to that at the next discrete time, i.e., $[\mathbf{g}\left(h\Delta\right)\ldots\mathbf{g}(2\Delta),\mathbf{g}(\Delta)]^T = \mathbf{M}\cdot[\mathbf{g}\left(h\Delta\right)\ldots\mathbf{g}(2\Delta),\mathbf{g}(\Delta)]^T+\mathbf{W}$, where $\mathbf{W}_0$ is the orthogonal noise. Note that with a finite ($h<\infty$) past history, the Mori–Zwanzig memory kernel would emerge if a history longer than h steps is given and if the past h-step observables do not linearly span an invariant manifold. Time-embedded analysis requires a more expensive inversion of the larger $\mathbf{C}(0)\in\mathbb{R}^{Mh\times Mh}$ matrix. In comparison, the memory kernel of the Mori–Zwanzig construction requires a single inversion of the $\mathbf{C}(0)\in\mathbb{R}^{M\times M}$, and the memory kernel is constructed by other two-time correlations $\mathbf{C}(k\Delta)$.

Our approach here is close to but different from the method presented by Lin and Lu [24], and it merits a more careful and detailed comparison. At the very high level, we share the same formal construction of the discrete-time GLE. Our approaches diverged as soon as the projection operator was chosen: we chose Mori's projection (referred to as the finite-rank projection in [24]), and Lin and Lu chose the Wiener projection. Formally, choosing the Wiener projection is identical to the time-embedding technique; as pointed out in [24], the subspace to which the Wiener projection projects is spanned by the past trajectory, and there is no Mori–Zwanzig memory kernel for a single prediction. Lin and Lu had to impose a "decaying memory condition," by which they meant the Markov transition weight of the distant past configuration must decay. In our case, the Mori–Zwanzig memory kernel naturally decays and no such a constraint is imposed. Computationally, our operations are always linear operations with explicit solution of the linear optimizer, and nonlinear optimization is needed in [24] because they invoke the rational approximations after the z-transformation. In this paper, we did not propose practical ways to model the orthogonal dynamics $\mathbf{W}_k$, but Lin and Lu provided a practical way forward using a multivariate Gaussian process model to match the power spectrum. As time-embedded analyses are much more expensive to fit but have the benefit of having smaller prediction errors (data not shown), it remains an interesting future research direction to objectively evaluate these two methods: Under the same computational budget, which one has a higher accuracy and with what metric (e.g., error in prediction, error in estimated Koopman spectrum, etc.)? Which one converges faster with the same finite data set?

Although the mathematical construction of the Mori–Zwanzig formalism and Koopman theory is general for any set of observables, the accuracy of the prediction crucially depends on the selection of the observables. In practice, it is preferable to adopt a set of observables which invoke smaller orthogonal dynamics $\mathbf{F}(t)$. Our algorithms, which are the first such algorithms to our knowledge, extract the exact orthogonal dynamics from the data and thus provide

an opportunity for us to study the statistics from the extracted orthogonal dynamics. One possibility in the future is to treat the orthogonal dynamics, $\mathbf{F}(t)$, as another dynamical system and recursively perform Mori–Zwanzig learning to telescope into the residual dynamics. We propose another possibility, analogous to the Gram–Schmidt process, to use the numerically extracted $\mathbf{F}(t)$ for identifying those predominant observables orthogonal to the existing set of observables. We expect by including these orthogonal observables, the predictive accuracy of the model can be improved. We remark that it is much cheaper to extract the memory kernel $\mathbf{K}(s)$ than the orthogonal dynamics $\mathbf{F}(t)$. Since the memory kernel is related to the two-time correlation function of the orthogonal dynamics by the GFD relationship, it will be an interesting research direction to identify those properties of the memory kernel that can be directly used for optimizing the observables, potentially incorporating the recent development of using deep neural networks as approximate functions by Yeung, Kundu, and Hodas [49]. The extracted operators provide multiple angles for such an optimization. For example, should the objective be to minimize the magnitude of the memory kernel (which $\propto \left\langle \mathbf{F}(t), \mathbf{F}^T(0) \right\rangle$ by (3.29)) or the timescale of the memory kernel, as one would like to achieve in the Markov state models?

We are currently working on generalizing the proposed methods to partially observed Markov stochastic systems with a focus on spectral analysis on the Mori–Zwanzig operators. In terms of applications, we are applying the proposed data-driven learning algorithm to study the modeling of turbulent flows [41] and molecular dynamical systems, both of which are extremely challenging engineering applications.

Appendix A. Solution of the orthogonal component. It is straightforward to check that (3.28),

$$\mathbf{g}_\perp(t) = \int_0^t \mathbf{C}(t-s) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}s, \tag{A.1}$$

is the solution to (3.21),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{g}_\perp(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^t \mathbf{C}(t-s) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}s\right] \\ &= \mathbf{F}(t) + \int_0^t \frac{\mathrm{d}\mathbf{C}(t-s)}{\mathrm{d}t} \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}s, \end{aligned} \tag{A.2}$$

and by (3.25),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{g}_\perp(t) = \mathbf{F}(t) &+ \int_0^t \mathbf{M} \cdot \mathbf{C}(t-s) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}s \\ &- \int_0^t \int_0^{t-s} \mathbf{K}(w) \cdot \mathbf{C}(t-s-w) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{F}(s)\, \mathrm{d}w\, \mathrm{d}s. \end{aligned} \tag{A.3}$$

We assume that the integrand satisfies the conditions which allow for a change in the order

of integrations:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{g}_\perp(t) &= \mathbf{F}(t) + \mathbf{M}\cdot\int_0^t \mathbf{C}\,(t-s)\cdot\mathbf{C}^{-1}\,(0)\cdot\mathbf{F}(s)\,\mathrm{d}s \\ &\quad - \int_0^t \mathbf{K}\,(w)\cdot\left[\int_0^{t-w}\mathbf{C}\,(t-w-s)\cdot\mathbf{C}^{-1}\,(0)\cdot\mathbf{F}(s)\,\mathrm{d}s\right]\,\mathrm{d}w \\ &= \mathbf{F}(t) + \mathbf{M}\cdot\mathbf{g}_\perp\,(t) - \int_0^t \mathbf{K}\,(w)\,\mathbf{g}_\perp\,(t-w)\,\mathrm{d}w,\end{aligned} \tag{A.4}$$

which is (3.28). In addition, $\mathbf{g}_\perp(t) = 0$ satisfies the initial condition in (3.21).

Appendix B. Derivation of the discrete-time Mori–Zwanzig formalism from continuous-time. We present two independent derivations of the discrete-time Mori–Zwanzig formalism, assuming the underlying process is a continuous-time and deterministic dynamics. In the first approach presented in section B.1, we begin with the continuous-time equations (3.19), (3.25), and (3.27), solving for the quantities evaluated on a temporally evenly sampled time grid $t = 0, \Delta, 2\Delta\ldots$, and we identify the recursive relationship between the solutions. In the second approach, presented in section B.2, we consider discretizing the dynamics first by transforming the continuous-time operator to a abstract discrete-time map, and then rederiving the Mori–Zwanzig equations (analogous to (3.19), (3.25), and (3.27)) to the discrete-time map. Importantly, in both approaches, we do not impose an infinitesimal constraint on Δ, and it can be any finite number.

Interestingly, the two constructs deliver the same recursive relationship that resembles the continuous-time Mori–Zwanzig equations. We shall refer to these relationships as the discrete-time Mori–Zwanzig equations. However, there exists a subtle difference: the same discrete-time Mori–Zwanzig equations link different mathematical objects in the two approaches. We discuss this difference between the two approaches, and a sufficient condition that the two descriptions agree, in section B.3. The GFD relationship in the discrete-time formulation is discussed in section B.4.

B.1. Discretization of the continuous-time Mori–Zwanzig equations. The GLE (3.19), the evolutionary equations for the correlation matrix $\mathbf{C}(t)$ (equation (3.25)), and the projected image $\mathcal{P}\mathbf{g}(t) = \mathbf{g}_\parallel(t)$ (equation (3.27)) exhibit a similar evolutionary operator which involves the Markov matrix $\mathbf{M}$ and the memory kernel $\mathbf{K}(t)$. Thus, here we consider the evolutionary operator applied on a test matrix $\mathbf{T}(t)$, which is $M \times P$, where M is the number of the observables serving as our basis functions spanning $\mathcal{H}_\mathbf{g}$. In the case of (3.19) and (3.25), $P = 1$, and in the case of (3.25), $P = M$. Suppose the evolutionary equation of the test matrix $\mathbf{T}(t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{T}(t) = \mathbf{M}\cdot\mathbf{T}(t) - \int_0^t \mathbf{K}\,(t-s)\cdot\mathbf{T}\,(s)\,\mathrm{d}s, \tag{B.1}$$

and suppose we only measure the snapshots of the observables at times on an evenly spaced grid $t = [0, \Delta, 2\Delta\ldots]$ with any finite temporal separation $\Delta > 0$. We also assume that we know the initial value $\mathbf{T}(0)$. The central aim of this section is to prove that the solution

$T(k\Delta)$, $k \in \mathbb{Z}_+$, can be written as

$$\mathbf{T}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}^{(k)} \cdot \mathbf{T}\left((k-\ell)\,\Delta\right), \tag{B.2}$$

where $\mathbf{\Omega}^{(k)}$ are prescribed $M \times M$ matrices. We break the proof down into a few separate lemmas.

Lemma B.1. *Given the evolutionary equation* (B.1) *with an initial condition* $\mathbf{T}(0)$ *and the continuous-time kernel* $\mathbf{K}(s)$*, for any* $t \geq 0$*, the solution* $\mathbf{T}(t)$ *can be expressed as a linear operator* $\mathbf{\Omega}_t$ *parametrized by the continuous-time* t *operated on the initial condition* $\mathbf{T}(0)$*:*

$$\mathbf{T}\left(t\right) = \mathbf{\Omega}_t^{(0)} \cdot \mathbf{T}\left(0\right). \tag{B.3}$$

Proof. We observe that the correlation matrix $\mathbf{C}(t)$ satisfies (3.25), which has a form similar to that of (B.1). It is easy to check that $\mathbf{\Omega}_t^0 := \mathbf{C}(t) \cdot \mathbf{C}^{-1}(0)$ is the solution:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{T}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\mathbf{C}(t) \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{T}(0)\right] \\
&= \left[\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{C}(t)\right] \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{T}(0) \\
&= \left[\mathbf{M} \cdot \mathbf{C}(t) - \int_0^t \mathbf{K}\left(t-s\right) \cdot \mathbf{C}\left(s\right)\,\mathrm{d}s\right] \cdot \mathbf{C}^{-1}(0) \cdot \mathbf{T}(0) \\
&= \mathbf{M} \cdot \mathbf{T}(t) - \int_0^t \mathbf{K}\left(t-s\right) \cdot \mathbf{T}\left(s\right)\,\mathrm{d}s.
\end{aligned} \tag{B.4}$$

∎

Theorem B.2. *Given* (1) *the evolutionary equation* (B.1)*,* (2) *the continuous-time memory kernel* $\mathbf{K}(s)$*,* $s \geq 0$*,* (3) *a nonnegative integer* $k \in \mathbb{Z}_+$*, and* (4) *snapshots of past history at discrete times* $\mathbf{T}(j\Delta)$*,* $j = 0, 1, \ldots, \leq k$*, we can express* $\mathbf{T}$ *at a future time* $k\Delta+\tau$*,* $0 < \tau \leq \Delta$*, in terms of linear superpositions of the past snapshots:*

$$\mathbf{T}\left(\tau + k\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_\tau^{(\ell)} \cdot \mathbf{T}\left((k-\ell)\,\Delta\right), \tag{B.5}$$

where the higher-order operators $\mathbf{\Omega}_\tau^{(k)}$*,* $k \in \mathbb{N}_+$*, are recursively defined by*

$$\mathbf{\Omega}_\tau^{(k)} = \mathbf{\Omega}_{\tau+k\Delta}^{(0)} - \sum_{\ell=0}^{k-1} \mathbf{\Omega}_\tau^{(\ell)} \mathbf{\Omega}_{(k-\ell)\Delta}^{(0)}. \tag{B.6}$$

Proof. From Lemma B.1,

$$\mathbf{T}(\tau + k\Delta) = \Omega_{\tau+k\Delta}^{(0)} \mathbf{T}(0), \tag{B.7}$$

and the recursive relationship (B.6) states

$$\mathbf{\Omega}_{\tau+k\Delta}^{(0)} = \sum_{\ell=0}^{k} \mathbf{\Omega}_\tau^{(\ell)} \mathbf{\Omega}_{(k-\ell)\Delta}^{(0)}, \tag{B.8}$$

and thus

$$\mathbf{T}(\tau + k\Delta) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\tau}^{(\ell)} \mathbf{\Omega}_{(k-\ell)\Delta}^{(0)} \mathbf{T}(0) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\tau}^{(\ell)} \mathbf{T}((k-\ell)\,\Delta). \tag{B.9}$$

■

Corollary B.3. *Given the operators* (B.6), $\mathbf{T}\left((k+1)\,\Delta\right)$ *can be expressed as a linear combination of the past snapshots* $\mathbf{T}\left(\ell\Delta\right)$, $\ell = 0, 1, \ldots, k$*:*

$$\mathbf{T}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \mathbf{T}((k-\ell)\,\Delta). \tag{B.10}$$

Corollary B.4. *Because the correlation matrix* $\mathbf{C}(t)$ *satisfies* (3.25)*, which is of the form* (B.1)*, given the operators* (B.6)*, the snapshots of the correlation matrix at discrete times satisfy*

$$\mathbf{C}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \mathbf{C}((k-\ell)\,\Delta). \tag{B.11}$$

Corollary B.5. *Because the projected image* $\mathcal{P}\mathbf{g}(t)$ *satisfies* (3.27)*, which is of the form* (B.1)*, given the operators* (B.6)*, the snapshots of the projected image at discrete times satisfy*

$$\mathcal{P}\mathbf{g}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \mathcal{P}\mathbf{g}((k-\ell)\,\Delta). \tag{B.12}$$

Theorem B.6 (discretized generalized Langevin equation). *Given the GLE* (3.19) *and the associated operators* (B.6)*, the snapshots* $\mathbf{g}(t)$ *at discrete times* $(k+1)\Delta$, $k \in \mathbb{N}$*, satisfy*

$$\mathbf{g}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \mathbf{g}((k-\ell)\,\Delta) + \mathbf{W}_k, \tag{B.13}$$

where $\mathbf{W}_k$ *is the discrete-time orthogonal dynamics, which is a linear function of the orthogonal dynamics* $\mathbf{F}(t)$, $t \le (k+1)\Delta$*, and* $\mathbf{W}_{k+1}$ *is orthogonal to* $\mathcal{H}_{\mathbf{g}}$*.*

Proof. As illustrated in section 3.3, the solution of the GLE can be written as a general solution $\mathbf{g}_{\|}(t)$ and a particular solution $\mathbf{g}_{\perp}(t)$ satisfying (3.20) and (3.21), respectively. Note that $\mathbf{g}_{\|}(t)$ is just the projected image $\mathcal{P}\mathbf{g}(t)$, and from Corollary B.5, we conclude that

$$\mathbf{g}_{\|}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{\Omega}_{\Delta}^{(\ell)} \cdot \mathbf{g}_{\|}\left((k-\ell)\,\Delta\right). \tag{B.14}$$

Because $\mathbf{g}(t) = \mathbf{g}_{\|}(t) + \mathbf{g}_{\perp}(t)\ \forall t \geq 0$, at time $t = (k+1)\,\Delta$, we can equate

$$
\begin{aligned}
\mathbf{g}\left((k+1)\,\Delta\right) &= \mathbf{g}_{\|}\left((k+1)\,\Delta\right) + \mathbf{g}_{\perp}\left((k+1)\,\Delta\right) \\
&= \sum_{\ell=0}^{k} \Omega_{\Delta}^{(\ell)} \cdot \mathbf{g}_{\|}\left((k-\ell)\,\Delta\right) + \mathbf{g}_{\perp}\left((k+1)\,\Delta\right) \\
&= \sum_{\ell=0}^{k} \Omega_{\Delta}^{(\ell)} \cdot \left[\mathbf{g}_{\|}\left((k-\ell)\,\Delta\right) + \mathbf{g}_{\perp}\left((k-\ell)\,\Delta\right)\right] + \mathbf{g}_{\perp}\left((k+1)\,\Delta\right) \\
&\quad - \sum_{\ell=0}^{k} \Omega_{\Delta}^{(\ell)} \cdot \mathbf{g}_{\perp}\left((k-\ell)\,\Delta\right) \\
&= \sum_{\ell=0}^{k} \Omega_{\Delta}^{(\ell)} \cdot \mathbf{g}\left((k-\ell)\,\Delta\right) \\
&\quad + \left[\mathbf{g}_{\perp}\left((k+1)\,\Delta\right) - \sum_{\ell=0}^{k} \Omega_{\Delta}^{(\ell)} \cdot \mathbf{g}_{\perp}\left((k-\ell)\,\Delta\right)\right].
\end{aligned} \tag{B.15}
$$

We identify

$$
\mathbf{W}_k = \mathbf{g}_{\perp}\left((k+1)\,\Delta\right) - \sum_{\ell=0}^{k} \Omega_{\Delta}^{(k-\ell)} \mathbf{g}_{\perp}\left(\ell\Delta\right). \tag{B.16}
$$

Note that $\mathbf{W}_k$ is a linear function of the snapshots of $\mathbf{g}_{\perp}$, which are linear functions of $\mathbf{F}(s)$, $t \leq (k+1)\,\Delta$, and thus $\mathbf{W}_k$ is orthogonal to $\mathcal{H}_{\mathbf{g}}$. ∎

B.2. Mori–Zwanzig equations of the generic discrete-time dynamics. We present an intuitive derivation of the discrete-time Mori–Zwanzig formalism that is analogous to section 3.2, noting that a more general derivation can be found in reference [24]. We begin with (3.16) and integrate the equation to the discrete-time grid $t = 0, \Delta, 2\Delta, \ldots$ to obtain the discrete mapping in the full Hilbert space $\mathcal{H}$:

$$
\begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t+\Delta) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t+\Delta) \end{bmatrix} = e^{\Delta \mathbf{L}} \cdot \begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t) \end{bmatrix} := \begin{bmatrix} \mathbf{U}_{\mathcal{M}\mathcal{M}} & \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}} \\ \mathbf{U}_{\bar{\mathcal{M}}\mathcal{M}} & \mathbf{U}_{\bar{\mathcal{M}}\bar{\mathcal{M}}} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{g}_{\mathcal{M}}(t) \\ \mathbf{g}_{\bar{\mathcal{M}}}(t) \end{bmatrix}. \tag{B.17}
$$

Given the initial condition $\mathbf{g}_{\mathcal{M}}(0)$ and $\mathbf{g}_{\bar{\mathcal{M}}}(0)$, the solution of the orthogonal component at the discrete times, $\mathbf{g}_{\bar{\mathcal{M}}}(k\Delta)$, can be expressed in terms of the historical snapshots of the resolved components, $\mathbf{g}_{\mathcal{M}}(\ell\Delta)$, $\ell = 0, 1, \ldots, k$, and the initial condition $\mathbf{g}_{\bar{\mathcal{M}}}(0)$:

$$
\mathbf{g}_{\bar{\mathcal{M}}}\left((k+1)\,\Delta\right) = \sum_{\ell=0}^{k} \mathbf{U}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}^{\ell} \mathbf{U}_{\bar{\mathcal{M}}\mathcal{M}} \mathbf{g}_{\mathcal{M}}\left((k-\ell)\,\Delta\right) + \mathbf{U}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}^{k} \mathbf{g}_{\bar{\mathcal{M}}}(0). \tag{B.18}
$$

Thus,

$$\begin{aligned}\mathbf{g}_{\mathcal{M}}((k+1)\,\Delta) &= \mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(k\Delta) + \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(k\Delta)\\ &= \mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(k\Delta) + \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(0)\\ &\quad + \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\sum_{\ell=0}^{k-1}\mathbf{U}^{\ell}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\bar{\mathcal{M}}\mathcal{M}}\mathbf{g}_{\mathcal{M}}((k-\ell-1)\,\Delta).\end{aligned} \tag{B.19}$$

Now we define $\mathbf{\Lambda}^{(0)}_{\Delta} := \mathbf{U}_{\mathcal{M}\mathcal{M}}$, $\mathbf{\Lambda}^{(\ell)}_{\Delta} := \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{(\ell-1)}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\bar{\mathcal{M}}\mathcal{M}}$, $\ell = 1, 2, \ldots$, and $\mathbf{V}_k := \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(0)$ and obtain the discrete-time GLE

$$\mathbf{g}((k+1)\,\Delta) = \sum_{\ell=0}^{k}\mathbf{\Lambda}^{(\ell)}_{\Delta}\mathbf{g}((k-\ell)\,\Delta) + \mathbf{V}_k. \tag{B.20}$$

A more general derivation first transforms the dynamics equation (2.1) to a discrete map of the solutions evaluated at $t = 0, \Delta, 2\Delta, \ldots$,

$$\mathbf{\Phi}(t+\Delta) = \mathbf{U}_{\Delta}(\mathbf{\Phi}(t)). \tag{B.21}$$

Here, $\mathbf{U}_{\Delta}$ is the nonlinear operator defined as the solution of the continuous-time equation:

$$\mathbf{U}_{\Delta}(\mathbf{\Phi}_0) := \int_0^{\Delta}\mathbf{R}(\mathbf{\Phi}(t))\,\mathrm{d}t + \mathbf{\Phi}_0. \tag{B.22}$$

With the transformed discrete-time map, $\mathbf{U}_{\Delta}$, we apply the generic Mori–Zwanzig formulation for the discrete-time dynamics [24] to obtain the discrete-time GLE (B.20).

Because the samples collected from this picture involve only discrete-time snapshots separated by a finite time Δ, we need to replace the inner product from averaging over a continuous-time trajectory (equation (3.24)) by a discrete-time sum:

$$\langle f, h\rangle_{\Delta} := \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\int_{\Omega} f(\mathbf{\Phi}_0)\,h(\mathbf{\Phi}_0)\,\mathrm{d}\mu(\mathbf{\Phi}_0). \tag{B.23}$$

We put a subscript under the inner product $\langle\cdot,\cdot\rangle_{\Delta}$ to denote the difference between (B.23) and its continuous-time counterpart (3.24).

Finally, a similar analysis to that presented in section 3.5 results in the discrete-time recursive relationship between the correlation matrix $\mathbf{C}_{\Delta}(k\Delta)$ and the projected image $\mathcal{P}_{\Delta}\mathbf{g}(k\Delta)$, $k = 0, 1, 2\ldots$,

$$\mathbf{C}_{\Delta}((k+1)\,\Delta) = \sum_{\ell=0}^{k}\mathbf{\Lambda}^{(\ell)}_{\Delta}\mathbf{C}_{\Delta}((k-\ell)\,\Delta), \tag{B.24}$$

$$\mathcal{P}_{\Delta}\mathbf{g}((k+1)\,\Delta) = \sum_{\ell=0}^{k}\mathbf{\Lambda}^{(\ell)}_{\Delta}\mathcal{P}_{\Delta}\mathbf{g}((k-\ell)\,\Delta). \tag{B.25}$$

Again, we put the subscript to the correlation matrix $\mathbf{C}_{\Delta}$ and the projection operator $\mathcal{P}_{\Delta}$ to differentiate them from their counterparts computed with continuous-time inner product, (3.24).

B.3. Difference between the two formulations. The discrete-time dynamics in the above two formulations coincide, and it is tempting to equate the operators $\mathbf{\Omega}_\Delta^{(\ell)}$ to $\mathbf{\Lambda}_\Delta^{(\ell)}$, and the discrete-time noise $\mathbf{W}_k$ to $\mathbf{V}_k$. Nevertheless, there is a subtle difference between these two formulations, and in general they do not have to be the same. The subtlety is that the inner product is defined by integrating over a continuous domain of time in the first formulation, but by summing over a discrete domain of the time in the second derivation. The invariant measure of the former does not have to be equal to the latter. Consequently, the projection operator, which depends on the definition of the inner product, (3.22), does not have to be identical in these two formulations.

In our proposed Algorithm 5.2, the discrete-time operators (either $\mathbf{\Omega}_\Delta^{(\ell)}$ or $\mathbf{\Lambda}_\Delta^{(\ell)}$) and noise (either $\mathbf{W}_k$ or $\mathbf{V}_k$) are extracted from the correlation matrices. In the first formulation, the correlation matrix was computed with the inner product equation (3.24),

$$\mathbf{C}(k\Delta) := \lim_{T\to\infty} \frac{1}{T} \int_0^{k\Delta} \mathbf{g} \circ \mathbf{\Phi}(k\Delta + s) \cdot \mathbf{g}^T \circ \mathbf{\Phi}(s)\, \mathrm{d}s, \tag{B.26}$$

and in the second approach, it is computed with the inner product equation (B.23),

$$\mathbf{C}_\Delta(k\Delta) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{g} \circ \mathbf{\Phi}((k+i)\Delta) \cdot \mathbf{g}^T \circ \mathbf{\Phi}(i\Delta). \tag{B.27}$$

Our analysis shows that a sufficient condition for $\mathbf{\Omega}_\Delta^{(\ell)} = \mathbf{\Lambda}_\Delta^{(\ell)}$ and $\mathbf{W}_k = \mathbf{V}_k$ is $\mathbf{C}(k\Delta) = \mathbf{C}_\Delta(k\Delta)$. The correlation matrices computed by two approaches are not necessarily identical. A simple harmonic oscillator $\dot{x} = p$ and $\dot{p} = -x$ with a unit amplitude $x^2(t)+p^2(t) = 1$ can be a counterexample. It has a period 2π, and if we choose Δ to be $2\pi/3$, these two formulations can be quite different: the continuous correlation matrix (equation (B.26)) has $\langle x, x\rangle = \pi$, but the discrete correlation matrix (equation (B.27)) has $\langle x, x\rangle = 1 + \cos^2(2\pi/3) + \cos^2(4\pi/3) = 3/2$ if $x(0) = 1$. Note that in the prediction (cf. section 6.2), the correlation matrix obtained from the first approach can be used to project any configuration satisfying $x^2(0) + p^2(0) = 1$, but the correlation matrix obtained from the second approach can be only used on a subset of possible initial conditions: $x(0) \in \{0, 2\pi/3, 4\pi/3\}$.

The subtlety between the two formulations serves as a caution to practitioners to pay attention to formulating *what we aim to learn*. On the one hand, with the first formulation, presented in section B.1, , one uses the simulation data to learn the discrete-time operators $\mathbf{\Omega}_\Delta^{(\ell)}$, which have a direct connection to the continuous-time operators because they are defined in terms of the continuous-time Markov transition $\mathbf{M}$ and the memory kernel $\mathbf{K}(t)$. Despite the fact that the correlation matrix is discretized and stored with a coarser resolution, the overall learned operators $\mathbf{\Omega}_\Delta^{(\ell)}$ serve a similar role to the continuous-time Mori–Zwanzig operators. For example, the operators predict the same $\mathcal{P}\mathbf{g}(t)$ in $\mathcal{H}_\mathbf{g}$. Thus, if one aims to learn about the continuous-time operators, the first approach is more desirable. The cost of the first approach is that even though the discretization Δ can be finite, one still needs a very finely sampled temporal grid to evaluate the continuous-time correlation matrix $\mathbf{C}(t)$ if an analytical expression of $\mathbf{C}(t)$ is not possible. We remark that an additional online computation of $\mathbf{C}$ in the fine-scale simulation can make computations efficient. On the other hand, a discretization with a finite

Δ is beneficial because it provides another coarser resolution in which the correlation matrix is stored. With the second formulation, presented in section B.2, one learns the Mori–Zwanzig operators of the *generic discrete maps*. In those cases where the inner product defined by the discrete-time statistics (equation (B.27)) is not identical to the continuous-time statistics (equation (B.26)), the projection operator in the second formulation is not the same as the one in the continuous-formulation. Consequently, the operators would predict a different projected image $\mathcal{P}_\Delta \mathbf{g}(t)$ in $\mathcal{H}_\mathbf{g}$. The above simple harmonic oscillator provides an intuitive example.

Finally, we point out that despite the subtle differences, if the two measures (statistics from discrete-time snapshots and continuous-time dynamics) are identical, these two formulations converge. In this scenario, there is no difference between the formulations in the theoretical sense. Nevertheless, computationally, the first approach—choosing a small δ to collect the snapshot data, computing the correlation matrix $\mathbf{C}$, and then discretizing $\mathbf{C}$ to multiples of δ—possesses two advantages. First, because the computation of the correlation matrix is agnostic to the discretization parameter Δ, one does not need to resimulate the discrete-time simulation when we change Δ. Second, recall that we need to compute the correlation matrix from the snapshots, which were recorded from a high-fidelity simulation which is generally computationally expensive. Thus, given a fixed amount of computational resources, there is generally an upper bound of the physical time which we are allowed to simulate. Because the first approach uses an equal or finer temporal resolution in comparison to the second ($\delta \leq \Delta$), we will collect more snapshots before the high-fidelity simulation ends. Thus, computationally, the convergence of the correlation matrix is generally better due to more samples.

B.4. Generalized fluctuation-dissipation relationship in the discrete-time formulations. In each of the frameworks presented in sections B.1 and B.2, with respect to the corresponding inner product (see discussion in section B.3), there exists a GFD relationship between the discrete-time noise ($\mathbf{W}_k$ or $\mathbf{V}_k$) and the discrete-time operators ($\mathbf{\Omega}_\Delta^{(\ell)}$ or $\mathbf{\Lambda}_\Delta^{(\ell)}$):

$$\mathbf{\Omega}_\Delta^{(k)} = -\left\langle \mathbf{W}_k, \mathbf{W}_0^T \right\rangle \cdot \mathbf{C}^{-1}(-\Delta), \quad k \in \mathbb{N}, \tag{B.28}$$

$$\mathbf{\Lambda}_\Delta^{(k)} = -\left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle_\Delta \cdot \mathbf{C}_\Delta^{-1}(-\Delta), \quad k \in \mathbb{N}. \tag{B.29}$$

Here, the first inner product in (B.28) is an integration against the continuous-time trajectory (e.g., (3.24)), and the second inner product in (B.29) is a discrete sum over the snapshots (e.g., (B.23)). Note that $\mathbf{C}(-\Delta) \equiv \mathbf{C}^T(\Delta)$ and $\mathbf{C}_\Delta(-\Delta) = \mathbf{C}_\Delta^T(\Delta)$ by the definitions (B.26) and (B.27). In this section, we present the proof to the GFD relationship, equations (B.28) and (B.29).

B.4.1. Discretization of the continuous-time Mori–Zwanzig framework. From (B.16) and noting that $\mathbf{g}_\perp(0) = 0$ (because $\mathbf{g}(0) = \mathbf{g}_\parallel(0)$), we obtain

$$\begin{aligned}\left\langle \mathbf{W}_k, \mathbf{W}_0^T \right\rangle &= \left\langle \mathbf{g}_\perp((k+1)\Delta) - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \mathbf{g}_\perp(\ell\Delta), \mathbf{g}_\perp^T(\Delta) \right\rangle \\ &= \left\langle \mathbf{g}((k+1)\Delta) - \mathbf{g}_\parallel((k+1)\Delta), \mathbf{g}^T(\Delta) - \mathbf{g}_\parallel^T(\Delta) \right\rangle \\ &\quad - \left\langle \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left[\mathbf{g}(\ell\Delta) - \mathbf{g}_\parallel(\ell\Delta)\right], \mathbf{g}^T(\Delta) - \mathbf{g}_\parallel^T(\Delta) \right\rangle.\end{aligned} \tag{B.30}$$

This is because $\mathbf{g}_\perp(t) = \mathbf{g}(t) - \mathbf{g}_\parallel$. Then we can express $\langle \mathbf{W}_k, \mathbf{W}_0^T \rangle$ in terms of pairs of observables,

$$\begin{aligned}\langle \mathbf{W}_k, \mathbf{W}_0^T \rangle = & \left\langle \mathbf{g}\left((k+1)\Delta\right), \mathbf{g}^T(\Delta)\right\rangle - \left\langle \mathbf{g}\left((k+1)\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle \\ & - \left\langle \mathbf{g}_\parallel\left((k+1)\Delta\right), \mathbf{g}^T(\Delta)\right\rangle + \left\langle \mathbf{g}_\parallel\left((k+1)\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle \\ & - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}\left(\ell\Delta\right), \mathbf{g}^T(\Delta)\right\rangle + \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}\left(\ell\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle \\ & + \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}_\parallel\left(\ell\Delta\right), \mathbf{g}^T(\Delta)\right\rangle - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}_\parallel\left(\ell\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle. \end{aligned} \tag{B.31}$$

The above expression can be simplified by

$$\left\langle \mathbf{g}(t), \mathbf{g}_\parallel^T(s)\right\rangle = \left\langle \mathbf{g}_\parallel(t), \mathbf{g}_\parallel^T(s)\right\rangle = \left\langle \mathbf{g}_\parallel(t), \mathbf{g}^T(s)\right\rangle, \tag{B.32}$$

which leads to

$$\begin{aligned}\langle \mathbf{W}_k, \mathbf{W}_0^T \rangle = & \left\langle \mathbf{g}\left((k+1)\Delta\right), \mathbf{g}^T(\Delta)\right\rangle - \left\langle \mathbf{g}_\parallel\left((k+1)\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle \\ & - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}\left(\ell\Delta\right), \mathbf{g}^T(\Delta)\right\rangle + \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}_\parallel\left(\ell\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle \\ = & \left\langle \mathbf{g}\left((k+1)\Delta\right), \mathbf{g}^T(\Delta)\right\rangle - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \left\langle \mathbf{g}\left(\ell\Delta\right), \mathbf{g}^T(\Delta)\right\rangle. \end{aligned} \tag{B.33}$$

The last equality comes from (B.12),

$$\begin{aligned} & \mathbf{g}_\parallel\left((k+1)\Delta\right) - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \mathbf{g}_\parallel\left(\ell\Delta\right) = 0 \\ \Rightarrow & \left\langle \mathbf{g}_\parallel\left((k+1)\Delta\right) - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \mathbf{g}_\parallel\left(\ell\Delta\right), \mathbf{g}_\parallel^T(\Delta)\right\rangle = 0. \end{aligned} \tag{B.34}$$

By definition, $\left\langle \mathbf{g}\left((k+1)\Delta\right), \mathbf{g}^T(\Delta)\right\rangle = \mathbf{C}(k\Delta)$ and $\left\langle \mathbf{g}\left(\ell\Delta\right), \mathbf{g}^T(\Delta)\right\rangle = \mathbf{C}\left((\ell-1)\Delta\right)$. Using (B.11), we can further simplify the two-time correlation of the discrete-time noise $\mathbf{W}$,

$$\begin{aligned}\langle \mathbf{W}_k, \mathbf{W}_0^T \rangle &= \mathbf{C}\left(k\Delta\right) - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \mathbf{C}\left((\ell-1)\Delta\right) \\ &= \sum_{\ell=0}^{k-1} \Omega_\Delta^{(k-1-\ell)} \mathbf{C}\left(\ell\Delta\right) - \sum_{\ell=0}^{k} \Omega_\Delta^{(k-\ell)} \mathbf{C}\left((\ell-1)\Delta\right) \\ &= \sum_{\ell=0}^{k-1} \Omega_\Delta^{(k-1-\ell)} \mathbf{C}\left(\ell\Delta\right) - \sum_{\ell'=-1}^{k-1} \Omega_\Delta^{(k-1-\ell')} \mathbf{C}\left(\ell'\Delta\right) \\ &= -\Omega_\Delta^{(k)} \mathbf{C}(\Delta), \end{aligned} \tag{B.35}$$

and establish (B.28) by multiplying $\mathbf{C}^{-1}(-\Delta)$ to both sides of the above equation.

B.4.2. Generalized fluctuation-dissipation relationship in the time-discretized dynamics. A parallel analysis to the one presented in section B.4 can be carried out to prove (B.29) with the inner product (e.g., (B.27)), the GLE (B.20), and the evolutionary equations of the correlation matrix equation (B.24) and projected image equation (B.25).

An alternative derivation, parallel to the derivation in section 3.6, can be carried out. We use the intuitive notation presented in section B.2. By definition, the discrete-time noise after k snapshots, starting at the ith snapshot along the long trajectory, is $\mathbf{V}_{k|i} := \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta)$. Then the two-time correlation between the noise with respect to the inner product is

$$\left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta)\,\mathbf{g}^T_{\bar{\mathcal{M}}}(i\Delta)\,\mathbf{U}^T_{\mathcal{M}\bar{\mathcal{M}}}. \tag{B.36}$$

From the discrete-time mapping equation (B.17),

$$\mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta) = \mathbf{g}_{\mathcal{M}}((i+1)\Delta) - \mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(i\Delta), \tag{B.37}$$

and thus

$$\begin{aligned}
\left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle &= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta)\left[\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta) - \mathbf{g}^T_{\mathcal{M}}(i\Delta)\,\mathbf{U}^T_{\mathcal{M}\mathcal{M}}\right] \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta)\,\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta) \\
&\quad - \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}^T_{\bar{\mathcal{M}}}\right\rangle \mathbf{U}^T_{\mathcal{M}\mathcal{M}} \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^k_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta)\,\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta),
\end{aligned} \tag{B.38}$$

because by construction $\left\langle \mathbf{g}_{\mathcal{M}}, \mathbf{g}^T_{\bar{\mathcal{M}}}\right\rangle = 0$. Again, using (B.17), $\mathbf{U}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}(i\Delta) = \mathbf{g}_{\bar{\mathcal{M}}}((i+1)\Delta) - \mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(i\Delta)$,

$$\begin{aligned}
\left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle &= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{g}_{\bar{\mathcal{M}}}((i+1)\Delta)\,\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta) \\
&\quad - \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(i\Delta)\,\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta) \\
&= \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\left\langle \mathbf{g}_{\bar{\mathcal{M}}}, \mathbf{g}^T_{\mathcal{M}}\right\rangle \\
&\quad - \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{g}_{\mathcal{M}}(i\Delta)\,\mathbf{g}^T_{\mathcal{M}}((i+1)\Delta) \\
&= -\mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\mathcal{M}\mathcal{M}}\mathbf{C}_{\Delta}(-\Delta).
\end{aligned} \tag{B.39}$$

By definition, $\mathbf{U}_{\mathcal{M}\bar{\mathcal{M}}}\mathbf{U}^{k-1}_{\bar{\mathcal{M}}\bar{\mathcal{M}}}\mathbf{U}_{\mathcal{M}\mathcal{M}} = \mathbf{\Lambda}^{(k)}_{\Delta}$, and thus we prove (B.29):

$$\left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle = -\mathbf{\Lambda}^{(k)}_{\Delta}\mathbf{C}_{\Delta}(-\Delta) \Rightarrow \mathbf{\Lambda}^{(k)}_{\Delta} = \left\langle \mathbf{V}_k, \mathbf{V}_0^T \right\rangle \mathbf{C}^{-1}_{\Delta}(-\Delta). \tag{B.40}$$

Acknowledgment. YTL sincerely appreciates many insightful discussions with Danny Perez.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, Benjamin/Cummings, Reading, MA, 1978.

[2] A. Ale, P. Kirk, and M. P. H. Stumpf, *A general moment expansion method for stochastic kinetic models*, J. Chem. Phys., 138 (2013), 174101, https://doi.org/10.1063/1.4802475.

[3] H. Arbabi and I. Mezic, *Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator*, SIAM J. Appl. Dyn. Syst., 16 (2017), pp. 2096–2126, https://doi.org/10.1137/17M1125236.

[4] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, *Chaos as an intermittently forced linear system*, Nature Commun., 8 (2017), pp. 1–9.

[5] T. Carleman, *Application de la théorie des équations intégrales linéaires aux systèmes d'équations différentielles non linéaires*, Acta Math., 59 (1932), pp. 63–87, https://doi.org/10.1007/BF02546499.

[6] M. Chen, X. Li, and C. Liu, *Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules*, J. Chem. Phys., 141 (2014), 064112, https://doi.org/10.1063/1.4892412.

[7] 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 (2000), pp. 2968–2973, https://doi.org/10.1073/pnas.97.7.2968.

[8] A. J. Chorin, O. H. Hald, and R. Kupferman, *Optimal prediction with memory*, Phys. D, 166 (2002), pp. 239–257, https://doi.org/10.1016/S0167-2789(02)00446-3.

[9] D. J. Evans and G. P. Morriss, *Statistical Mechanics of Nonequilibrium Liquids*, 2nd ed., Cambridge University Press, Cambridge, UK, 2008.

[10] S. K. J. Falkena, C. Quinn, J. Sieber, J. Frank, and H. A. Dijkstra, *Derivation of delay equation climate models using the Mori-Zwanzig formalism*, Proc. A, 475 (2019), 20190075, https://doi.org/10.1098/rspa.2019.0075.

[11] P. Frederickson, J. L. Kaplan, E. D. Yorke, and J. A. Yorke, *The Lyapunov dimension of strange attractors*, J. Differential Equations, 49 (1983), pp. 185–207, https://doi.org/10.1016/0022-0396(83)90011-6.

[12] G. A. Gottwald, D. T. Crommelin, and C. L. E. Franzke, *Stochastic climate theory*, in Nonlinear and Stochastic Climate Dynamics, Cambridge University Press, Cambridge, UK, 2017, pp. 209–240, https://doi.org/10.1017/9781316339251.009.

[13] I. Horenko, C. Hartmann, C. Schütte, and F. Noe, *Data-based parameter estimation of generalized multidimensional Langevin processes*, Phys. Rev. E, 76 (2007), 016706, https://doi.org/10.1103/PhysRevE.76.016706.

[14] T. Hudson and X. H. Li, *Coarse-graining of overdamped Langevin dynamics via the Mori–Zwanzig formalism*, Multiscale Model. Simul., 18 (2020), pp. 1113–1135, https://doi.org/10.1137/18M1222533.

[15] M. Kamb, E. Kaiser, S. L. Brunton, and J. N. Kutz, *Time-delay observables for Koopman: Theory and applications*, SIAM J. Appl. Dyn. Syst., 19 (2020), pp. 886–917, https://doi.org/10.1137/18M1216572.

[16] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, *Data-driven model reduction and transfer operator approximation*, J. Nonlinear Sci., 28 (2018), pp. 985–1010, https://doi.org/10.1007/s00332-017-9437-7.

[17] B. O. Koopman, *Hamiltonian systems and transformation in Hilbert space*, Proc. Natl. Acad. Sci. USA, 17 (1931), pp. 315–318, https://doi.org/10.1073/pnas.17.5.315.

[18] B. O. Koopman and J. von Neumann, *Dynamical systems of continuous spectra*, Proc. Natl. Acad. Sci. USA, 18 (1932), pp. 255–263, https://doi.org/10.1073/pnas.18.3.255.

[19] K. Kowalski and W.-H. Steeb, *Nonlinear Dynamical Systems and Carleman Linearization*, World Scientific, 1991, https://doi.org/10.1142/1347.

[20] S. Le Clainche and J. M. Vega, *Higher order dynamic mode decomposition*, SIAM J. Appl. Dyn. Syst.,

16 (2017), pp. 882–925, https://doi.org/10.1137/15M1054924.
[21] H. Lei, N. A. Baker, and X. Li, *Data-driven parameterization of the generalized Langevin equation*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 14183–14188, https://doi.org/10.1073/pnas.1609587113.
[22] Z. Li, X. Bian, X. Li, and G. E. Karniadakis, *Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism*, J. Chem. Phys., 143 (2015), 243128, https://doi.org/10.1063/1.4935490.
[23] Z. Li, H. S. Lee, E. Darve, and G. E. Karniadakis, *Computing the non-Markovian coarse-grained interactions derived from the Mori–Zwanzig formalism in molecular systems: Application to polymer melts*, J. Chem. Phys., 146 (2017), 014104, https://doi.org/10.1063/1.4973347.
[24] K. K. Lin and F. Lu, *Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism*, J. Comput. Phys., 424 (2021), 109864, https://doi.org/10.1016/j.jcp.2020.109864.
[25] E. N. Lorenz, *Predictability: A problem partly solved*, in Predictability of Weather and Climate, T. Palmer and R. Hagedorn, eds., Cambridge University Press, Cambridge, UK, 2006, pp. 40–58, https://doi.org/10.1017/CBO9780511617652.004.
[26] A. Mauroy and I. Mezic, *Global stability analysis using the eigenfunctions of the Koopman operator*, IEEE Trans. Automat. Control, 61 (2016), pp. 3356–3369, https://doi.org/10.1109/TAC.2016.2518918.
[27] I. Mezic, *Spectral properties of dynamical systems, model reduction and decompositions*, Nonlinear Dynam., 41 (2005), pp. 309–325.
[28] L. Molgedey and H. Schuster, *Separation of a mixture of independent signals using time delayed correlations*, Phys. Rev. Lett., 72 (1994), pp. 3634–3638.
[29] H. Mori, *Transport, collective motion, and Brownian motion*, Progr. Theoret. Phys., 33 (1965), pp. 423–455.
[30] F. Noé and F. Nüske, *A variational approach to modeling slow processes in stochastic dynamical systems*, Multiscale Model. Simul., 11 (2013), pp. 635–655, https://doi.org/10.1137/110858616.
[31] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, *Variational approach to molecular kinetics*, J. Chem. Theory Comput., 10 (2014), pp. 1739–1752, https://doi.org/10.1021/ct4009156.
[32] E. J. Parish and K. Duraisamy, *Non-Markovian closure models for large eddy simulations using the Mori-Zwanzig formalism*, Phys. Rev. Fluids, 2 (2017), 014604, https://doi.org/10.1103/PhysRevFluids.2.014604.
[33] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, *Identification of slow molecular order parameters for Markov model construction*, J. Chem. Phys., 139 (2013), 015102, https://doi.org/10.1063/1.4811489.
[34] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, *Spectral analysis of nonlinear flows*, J. Fluid Mech., 641 (2009), pp. 115–127, https://doi.org/10.1017/S0022112009992059.
[35] P. J. Schmid, *Dynamic mode decomposition of numerical and experimental data*, J. Fluid Mech., 656 (2010), pp. 5–28, https://doi.org/10.1017/S0022112010001217.
[36] P. J. Schmid, L. Li, M. P. Juniper, and O. Pust, *Applications of the dynamic mode decomposition*, Theoret. Comput. Fluid Dynam., 25 (2011), pp. 249–259.
[37] D. Schnoerr, G. Sanguinetti, and R. Grima, *Comparison of different moment-closure approximations for stochastic chemical kinetics*, J. Chem. Phys., 143 (2015), 185101, https://doi.org/10.1063/1.4934990.
[38] P. Stinis, *Stochastic optimal prediction for the Kuramoto–Sivashinsky equation*, Multiscale Model. Simul., 2 (2004), pp. 580–612, https://doi.org/10.1137/030600424.
[39] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, CRC press, 2000.
[40] F. Takens, *Detecting strange attractors in turbulence*, in Dynamical Systems and Turbulence (Warwick, 1980), D. Rand and L.-S. Young, eds., Springer, Berlin, Heidelberg, 1981, pp. 366–381.
[41] Y. Tian, Y. T. Lin, M. Anghel, and D. Livescu, *Data-driven learning of Mori-Zwanzig operators for isotropic turbulence*, Phys. Fluids, 33 (2021), 125118, https://doi.org/10.1063/5.0070548.
[42] L. Tong, V. Soon, Y. Huang, and R. Liu, *AMUSE: A new blind identification algorithm*, in IEEE International Symposium on Circuits and Systems (New Orleans, LA, 1990), IEEE, 1990, pp. 1784–1787, https://doi.org/10.1109/ISCAS.1990.111981.

[43] J. von Neumann, *Zur Operatorenmethode in der klassischen Mechanik*, Ann. of Math. (2), 33 (1932), pp. 587–642, https://doi.org/10.2307/1968537.
[44] S. Wang, Z. Li, and W. Pan, *Implicit-solvent coarse-grained modeling for polymer solutions via Mori-Zwanzig formalism*, Soft Matter, 15 (2019), pp. 7567–7582, https://doi.org/10.1039/C9SM01211G.
[45] M. Wanner and I. Mezić, *Robust Approximation of the Stochastic Koopman Operator*, preprint, https://arxiv.org/abs/2011.00078, 2020.
[46] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition*, J. Nonlinear Sci., 25 (2015), pp. 1307–1346.
[47] H. Wu and F. Noé, *Variational approach for learning Markov processes from time series data*, J. Nonlinear Sci., 30 (2020), pp. 23–66, https://doi.org/10.1007/s00332-019-09567-y.
[48] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, *Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations*, J. Chem. Phys., 146 (2017), 154104, https://doi.org/10.1063/1.4979344.
[49] E. Yeung, S. Kundu, and N. Hodas, *Learning deep neural network representations for Koopman operators of nonlinear dynamical systems*, in 2019 American Control Conference (ACC), 2019, pp. 4832–4839, https://doi.org/10.23919/ACC.2019.8815339.
[50] R. Zwanzig, *Nonlinear generalized Langevin equations*, J. Statist. Phys., 9 (1973), pp. 215–220.
[51] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, 2001.