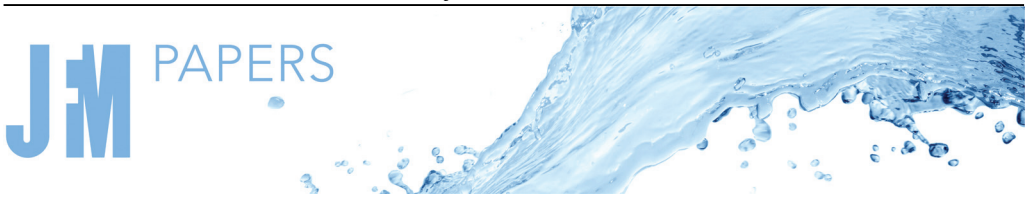
J. Fluid Mech. (2024), *vol*. 997, A1, doi:10.1017/jfm.2024.658



Representing turbulent statistics with partitions of state space. Part 1. Theory and methodology

Andre N. Souza†

Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

(Received 5 September 2023; revised 17 April 2024; accepted 6 June 2024)

This is the first of a two-part paper. We formulate a data-driven method for constructing finite-volume discretizations of an arbitrary dynamical system's underlying Liouville/Fokker–Planck equation. A method is employed that allows for flexibility in partitioning state space, generalizes to function spaces, applies to arbitrarily long sequences of time-series data, is robust to noise and quantifies uncertainty with respect to finite sample effects. After applying the method, one is left with Markov states (cell centres) and a random matrix approximation to the generator. When used in tandem, they emulate the statistics of the underlying system. We illustrate the method on the Lorenz equations (a three-dimensional ordinary differential equation) saving a fluid dynamical application for Part 2 (Souza, *J. Fluid Mech.*, vol. 997, 2024, A2).

Key words: chaos, low-dimensional models, big data

1. Introduction

Often, the goal of modelling a complex system is not to determine the dynamical equations but to construct models that converge in distribution to relevant statistics. In the context of turbulence modelling, this can be viewed as one of the goals of a large-eddy simulation, where subsets of statistics (often the kinetic energy spectra) are compared with that of direct numerical simulation. Similarly, in Earth systems modelling, the unpredictability of weather patterns over long time scales necessitates the development of nonlinear models that are queried for relevant statistics. Thus, the models are not meant to converge to dynamical trajectories but rather converge in distribution to target observables.

The present work is motivated by the need to construct simplified statistical models of complex physical phenomena such as turbulence. We take on a dynamical systems view of turbulence original to Hopf (1948), complemented by Lorenz (1963) and found

† Email address for correspondence: andrenogueirasouza@gmail.com

in its modern form in Cvitanović *et al.* (2016). Thus, the approach is to develop a direct discretization of the statistics associated with chaotic or turbulent dynamics, which we assume to be mixing and associated with a fractal in state space.

There exist many types of discretizations that directly target the statistics, which here means a discretization of the underlying Liouville equation (deterministic dynamics), Fokker–Planck equation (stochastic dynamics), Perron–Frobenius/transfer operator (discrete-time dynamics) or Koopman operator (adjoint of the Perron–Frobenius/transfer operator). Discretizations methods include that of Ulam (1964), Dellnitz, Froyland & Junge (2001), Dellnitz *et al.* (2005) or, for the stochastic Lorenz equations, Allawala & Marston (2016). Furthermore, there exist efficient extensions of Ulam's method for multidimensional systems such as the box-refinement methods of Dellnitz & Junge (1999), Dellnitz *et al.* (2001) or the sparse Haar tensor basis of Junge & Koltai (2009). Modern methods take on an operator theoretic plus data-driven approach, leading to Koopman operators that are measure preserving Colbrook (2023), which build off of earlier work such as Schmid (2010) and Rowley *et al.* (2009). Data-driven construction of the Perron–Frobenius operator is reviewed by Klus *et al.* (2016), Giannakis (2019) and Fernex, Noack & Semaan (2021). Convergence guarantees under various assumptions are found in, for example, Froyland (1997), Das, Giannakis & Slawinska (2021), Colbrook & Townsend (2023) and Schütte, Klus & Hartmann (2022). Furthermore, modern methods are beginning to use deep learning in order to learn optimal nonlinear dictionaries for observables, see Constante-Amores, Linot & Graham (2023) and Bittracher *et al.* (2023).

The method presented here can be viewed as a subset of the general method for constructing Koopman operators using the extended dynamic mode decomposition method (Williams, Kevrekidis & Rowley 2015); however, we make a particular choice of a nonlinear dictionary that allows for computational expedience (we do not need to explicitly calculate pseudoinverses), geometric interpretability (which allows for a natural method to increase the size of the nonlinear dictionary) and the quantification of uncertainty due to finite sampling effects. The first two changes allow for much larger-scale computations than what has been previously achieved. Furthermore, we take an approach that closely mirrors a combination of Froyland *et al.* (2013) and Fernex *et al.* (2021). When addressing statistics of a partial differential equation, we take a field-theoretic perspective such as that of Hopf (1952). The goal is to construct a discretization of the generator (continuity/Fokker–Planck operator), similar to Rosenfeld *et al.* (2021). The method of constructing the generator herein is motivated by finite-volume discretizations of advection–diffusion equations and yields an approximation that can be interpreted as a continuous-time Markov process with finite state space.

The rest of the paper is organized as follows. In § 2, we discuss the underlying theory and approximations. The approach is heavily inspired by Hopf (1952) and Cvitanović *et al.* (2016). The primary idea is to discretize the equations for the statistics (an 'Eulerian' quantity) by using the equation for the dynamics (a 'Lagrangian' quantity). Further approximations are then made to calculate observables of interest.

In § 3, we introduce a data-driven method with quantified uncertainties for calculating the approximate generator. The technique can be applied to arbitrarily long time-series data, dynamical systems with a large state space, and further provides uncertainty estimates on the entries of the discretized generator.

Section 4 gives an example of utilizing the method on Lorenz (1963). One sees that even a coarse discretization of statistics captures statistical features of the original system.

For those simply interested in what the calculations enable, §§ 2–3 are safely skipped in favour of § 4 or Part 2 of this series (Souza 2024). Furthermore, Appendices A, B, C

and D expand the text by discussing symmetries, matrix properties of the generator, an algebraic interpretation of constructing the generator alongside the connection to dynamic mode decomposition and convergence of the data-driven method with respect to the simple harmonic oscillator, respectively.

2. Theory

In the following sections, we outline the framework for calculations involving the Liouville equation of a dynamical system. The focus is not on the 'Lagrangian' view (as given by the dynamics) but rather an 'Eulerian' one (given by statistics). Consequently, we introduce notation to help distinguish the perspectives inherent in the two viewpoints. We show how to go from the continuous to the discrete and provide formulae.

2.1. *Finite-dimensional dynamical system*

We start with a generic continuous time dynamical system in d-dimensions given by

$$\dot{\boldsymbol{s}} = \boldsymbol{U}(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{s}(t) : \mathbb{R} \to \mathbb{R}^d$ is the state of the system and $\boldsymbol{U} : \mathbb{R}^d \to \mathbb{R}^d$ is the evolution equation. Equation (2.1) provides a succinct rule for determining the evolution of a dynamical system; however, uncertain initial conditions stymie future predictions in the presence of chaos (Lorenz 1963). It is, therefore, more natural to study the statistical evolution of probability densities as in Hopf (1952). Thus, we focus not on the d-dimensional ordinary differential equation given by (2.1) but rather the d-dimensional partial differential equation that governs the evolution of probability densities in state space.

We denote fixed vector in state space by $\boldsymbol{\mathfrak{s}} \in \mathbb{R}^d$, the components of the state $\boldsymbol{\mathfrak{s}}$ by $\boldsymbol{\mathfrak{s}} = (\mathfrak{s}_1, \mathfrak{s}_2, \ldots, \mathfrak{s}_d)$ and the components of the evolution rule by $\boldsymbol{U} = (U_1, U_2, \ldots, U_d)$. The evolution equation for the statistics of (2.1), as characterized by a probability density function,

$$\mathcal{P} = \mathcal{P}(\mathfrak{s}_1, \mathfrak{s}_2, \ldots, \mathfrak{s}_d, t) = \mathcal{P}(\boldsymbol{\mathfrak{s}}, t), \tag{2.2}$$

is given by the Liouville equation

$$\partial_t \mathcal{P} + \sum_{i=1}^{d} \frac{\partial}{\partial \mathfrak{s}_i} \left(U_i(\boldsymbol{\mathfrak{s}}) \mathcal{P} \right) = 0. \tag{2.3}$$

The above equation is a statement of probability conservation. It is precisely analogous to the mass conservation equation from the compressible Navier–Stokes equations. However, the 'mass' is being interpreted as a probability density. The distribution, $\mathcal{P}$, is guided by the flow dynamics $\boldsymbol{U}$ to likely regions of state space.

The presence of stochastic white noise, $\boldsymbol{\xi}$ in a dynamical system,

$$\dot{\boldsymbol{s}} = \boldsymbol{U}(\boldsymbol{s}) + \boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}} \boldsymbol{\xi} \tag{2.4}$$

where the ensemble average of $\boldsymbol{\xi}$ satisfies $\langle \xi_i(t) \xi_j(t') \rangle = \delta_{ij} \delta(t - t')$ and $\boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}}$ is the matrix square root of the covariance matrix $\boldsymbol{\mathcal{D}}$, e.g. $\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}} \boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}}$, modifies (2.3) through the inclusion of a diffusion term

$$\partial_t \mathcal{P} + \sum_{i=1}^{d} \frac{\partial}{\partial \mathfrak{s}_i} \left(U_i(\boldsymbol{\mathfrak{s}}) \mathcal{P} + \frac{1}{2} \sum_{i=j}^{d} \mathcal{D}_{ij} \frac{\partial}{\partial \mathfrak{s}_j} \mathcal{P} \right) = 0, \tag{2.5}$$

i.e. the Fokker–Planck equation. The presence of the diffusion tensor $\boldsymbol{\mathcal{D}}$ regularizes (2.3).

2.2. *Infinite-dimensional dynamical system*

When the underlying dynamical system is a partial differential equation, we assume a suitably well-defined discretization exists to reduce it to a formally $\mathbb{R}^d$-dimensional dynamical system. We contend ourselves to the study of the statistics of the $\mathbb{R}^d$ approximation. One hopes that different discretizations lead to similar statistical statements of the underlying partial differential equation; thus, it is worth introducing notation for the analogous Liouville equation for a partial differential equation, as was done by Hopf (1952). The calculations and notation that follow are formal, and there are mathematical difficulties in assigning them rigorous meaning; however, we find value in the approach, as it often allows for expedient calculations as was done in Souza, Lutz & Flierl (2023*b*) and Giorgini *et al.* (2024). The overall (heuristic) recipe for the transition to function spaces is to replace sums with integrals, derivatives with variational derivatives, discrete indices with continuous indices and Kronecker deltas with Dirac deltas.

The d-dimensional vector from before now becomes a vector in function space whose components are labelled by a continuous index, $\boldsymbol{x}$, a position in a domain Ω, and discrete index j, the index for the field of interest. Thus, the component choice $s_i(t)$ for a fixed index i is analogous to $s_{(\boldsymbol{x},j)}(t)$ for a fixed position $\boldsymbol{x}$ and field index $j \in \{1, \ldots, d_s\}$, e.g. $d_s = 3$ for the three velocity components of the incompressible Navier–Stokes. In the discrete case, the single index i loops over all velocity components and all simulation grid points.

Specifically, we consider a partial differential equation for a state $\boldsymbol{s}$ defined over a domain Ω, with suitable boundary conditions,

$$\partial_t \boldsymbol{s} = \boldsymbol{\mathcal{U}}[\boldsymbol{s}], \tag{2.6}$$

where the operator $\boldsymbol{\mathcal{U}} : \mathcal{X} \to \mathcal{X}$ characterizes the evolution of system and $\mathcal{X}$ is a function space. The component of $\boldsymbol{\mathcal{U}}$ at position $\boldsymbol{x}$ and field index j is denoted by $\mathcal{U}_{(\boldsymbol{x},j)}$.

The analogous evolution for the probability density functional,

$$\mathcal{P} = \mathcal{P}[\mathfrak{s}_{(\boldsymbol{x},1)}, \mathfrak{s}_{(\boldsymbol{x},2)}, \ldots, \mathfrak{s}_{(\boldsymbol{x},d_s)}, t] = \mathcal{P}[\boldsymbol{\mathfrak{s}}, t], \tag{2.7}$$

is denoted by

$$\partial_t \mathcal{P} + \sum_{j=1}^{d_s} \int_\Omega \mathrm{d}\boldsymbol{x} \, \frac{\delta}{\delta \mathfrak{s}_{(\boldsymbol{x},j)}} \left(\mathcal{U}_{(\boldsymbol{x},j)}[\boldsymbol{\mathfrak{s}}] \mathcal{P} \right) = 0. \tag{2.8}$$

The sum in (2.3) is replaced by an integral over position indices and a sum over field indices in (2.8). Furthermore, the partial derivatives are replaced by variational derivatives. The variational derivative is being used in the physicist's sense, that is to say,

$$\frac{\delta \mathfrak{s}_{(\boldsymbol{x},i)}}{\delta \mathfrak{s}_{(\boldsymbol{y},j)}} = \delta(\boldsymbol{x} - \boldsymbol{y}) \delta_{ij} \Leftrightarrow \frac{\partial \mathfrak{s}_{i'}}{\partial \mathfrak{s}_{j'}} = \delta_{i'j'}, \tag{2.9}$$

where $\delta(\boldsymbol{x} - \boldsymbol{y})$ is the Dirac delta function and δ_{ij} is the Kronecker delta function, $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, $i, j \in \{1, \ldots, d_s\}$, and $i', j' \in \{1, \ldots, d\}$. In the typical physics notation, it is common to drop the dependence on the position $\boldsymbol{x}$ and explicitly write out the field variable in terms of its components (as opposed to the indexing that we do here), e.g.

$$\frac{\delta}{\delta \mathfrak{s}_{(\boldsymbol{y},1)}} \sum_{i=1}^{d_s} \int_\Omega \mathrm{d}\boldsymbol{x} (\mathfrak{s}_{(\boldsymbol{x},i)})^2 = 2\mathfrak{s}_{(\boldsymbol{y},1)} \Rightarrow \frac{\delta}{\delta u} \int_\Omega (u^2 + v^2 + w^2) = 2u \tag{2.10}$$

for the prognostic variables u, v, w of the incompressible Navier–Stokes equations. See § 35 of Zinn-Justin (2021) for examples of the Fokker–Planck equation in the field-theoretic context.

To derive (2.8), we suppose that (2.1) is a discretization of (2.6). Starting from (2.3), first introduce a control volume at index i as $\Delta \boldsymbol{x}_i$ to rewrite the equation as

$$\partial_t \mathcal{P} + \sum_{i=1}^{d} \Delta \boldsymbol{x}_i \frac{1}{\Delta \boldsymbol{x}_i} \frac{\partial}{\partial \boldsymbol{s}_i} (U_i(\boldsymbol{s})\mathcal{P}) = 0. \tag{2.11}$$

In the 'limit' as the grid is refined and $|\Delta \boldsymbol{x}_i| \to 0$, we have

$$\sum_{i=1}^{d} \Delta \boldsymbol{x}_i \frac{1}{\Delta \boldsymbol{x}_i} \frac{\partial}{\partial \boldsymbol{s}_i} \to \sum_{j=1}^{d_s} \int_{\Omega} \mathrm{d}\boldsymbol{x} \, \frac{\delta}{\delta \boldsymbol{s}_{(\boldsymbol{x},j)}} \quad \text{and} \quad U_i \to \mathcal{U}_{(\boldsymbol{x},j)}. \tag{2.12a,b}$$

A stochastic partial differential equation is given by

$$\partial_t \boldsymbol{s} = \boldsymbol{\mathcal{U}}[s] + \boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}}[\boldsymbol{\xi}], \tag{2.13}$$

where $\boldsymbol{\xi}$ is space–time (and state) noise with covariance

$$\langle \xi_{(\boldsymbol{x},i)}(t) \xi_{(\boldsymbol{y},j)}(t') \rangle = \delta_{ij} \delta(\boldsymbol{x} - \boldsymbol{y}) \delta(t - t'). \tag{2.14}$$

The noise $\boldsymbol{\mathcal{D}}^{\boldsymbol{1/2}}[\boldsymbol{\xi}]$ is interpreted as a Gaussian process with a spatial and field covariance given by properties of $\boldsymbol{\mathcal{D}}$. Furthermore, the action of the symmetric positive definite linear operator $\boldsymbol{\mathcal{D}}$, is expressed as

$$\mathcal{D}_{(\boldsymbol{x},i)}[\boldsymbol{\xi}] = \sum_{j=1}^{d_s} \int_{\Omega} \mathrm{d}\boldsymbol{y} \, \mathcal{K}_{ij}(\boldsymbol{x}, \boldsymbol{y}) \xi_{(\boldsymbol{y},j)}, \tag{2.15}$$

where $\mathcal{K}_{ij}$ is the kernel of the integral operator. Formally, (2.13) has the corresponding Fokker–Planck equation

$$\partial_t \mathcal{P} + \sum_{j=1}^{d_s} \int_{\Omega} \mathrm{d}\boldsymbol{x} \, \frac{\delta}{\delta \boldsymbol{s}_{(\boldsymbol{x},j)}} \left(\mathcal{U}_{(\boldsymbol{x},j)}[\boldsymbol{s}]\mathcal{P} + \frac{1}{2} \sum_{k=1}^{d_s} \int_{\Omega} \mathrm{d}\boldsymbol{y} \, \mathcal{K}_{jk}(\boldsymbol{x}, \boldsymbol{y}) \frac{\delta \mathcal{P}}{\delta \boldsymbol{s}_{(\boldsymbol{y},k)}} \right) = 0. \tag{2.16}$$

Making sense of (2.13) is an area of active research (see Hairer (2014) and Corwin & Shen (2020)), where one must confront defining probability distributions over function spaces. Defining integrals over function spaces has met with considerable difficulties, although progress has been made (see Daniell (1919), DeWitt (1972) and Albeverio & Mazzucchi (2016)).

With the formalism now set, the focus of this work is on methods for discretizing (2.3) and (2.8) on subsets of state space that are typically thought of as chaotic or turbulent given only trajectory information from (2.1) and (2.6), respectively. The data-driven methods developed herein apply without change to the stochastic analogues.

2.3. *Finite-volume discretization*

To focus our discussion, we use the finite-dimensional and deterministic setting. We first assume that the underlying dynamics are on a chaotic attractor associated with a compact subset of state space $\mathcal{M} \subset \mathbb{R}^d$. We further assume that the dynamical system is robust to noise in order to regularize the attractor and probability densities as in Young (2002) and Cowieson & Young (2005). Statements about deterministic chaos in this manuscript implicitly use a noise regularization where the zero noise limit is always the last limit

taken in any computation, similar to Giannakis (2019). Whether or not such a limit generally agrees with a Sinai–Ruelle–Bowen measure (see for example Young (2002), Blank (2017) and Araujo (2023)) is unknown. Furthermore, in the presence of correlated or state-space-dependent noise, one must be careful how a zero-noise limit is taken.

We introduce a partition of $\mathcal{M}$ into N cells which we denote by $\mathcal{M}_n$ for $n = 1, \ldots, N$. The coarse-grained discretization variables P_n are

$$\int_{\mathcal{M}_n} \mathrm{d}\boldsymbol{\mathfrak{s}}\, \mathcal{P} = \int_{\mathcal{M}_n} \mathcal{P} = P_n \tag{2.17}$$

as is common in finite-volume methods. When unambiguous, we drop the infinitesimal state space volume $\mathrm{d}\boldsymbol{\mathfrak{s}}$. Here $P_n(t)$ is the probability in time of being found in the subset of state space $\mathcal{M}_n$ at time t. Integrating (2.3) with respect to the cells yields

$$\frac{\mathrm{d}}{\mathrm{d}t} P_n = \int_{\mathcal{M}_n} \left[\sum_{i=1}^{d} \frac{\partial}{\partial \boldsymbol{\mathfrak{s}}_i} \left(U_i(\boldsymbol{\mathfrak{s}}) \mathcal{P} \right) \right] = \int_{\partial \mathcal{M}_n} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \mathcal{P}, \tag{2.18}$$

where $\partial \mathcal{M}_n$ is the boundary of the cell and $\hat{n}$ is a normal vector. The art of finite-volume methods comes from expressing the right-hand side of (2.18) in terms of the coarse-grained variables P_n through a suitable choice of numerical flux.

We go about calculating the numerical flux in a roundabout way. We list some desiderata for a numerical discretization.

(i) The discrete equation is expressed in terms of the instantaneous coarse-grained variables P_n.
(ii) The discrete equation is linear, in analogy to the infinite-dimensional one.
(iii) The equation must conserve probability.
(iv) Probability must be positive at all times.

The first two requirements state

$$\int_{\partial \mathcal{M}_n} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \mathcal{P} \approx \sum_m Q_{nm} P_m \tag{2.19}$$

for some matrix Q. Thus, we want an equation of the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \hat{P}_n = \sum_m Q_{nm} \hat{P}_m. \tag{2.20}$$

We introduced a 'hat' to distinguish the numerical approximation, $\hat{P}_n$, with the exact solution P_n. The third requirement states

$$\sum_n \hat{P}_n = 1 \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} \sum_n \hat{P}_n = 0 = \sum_{mn} Q_{nm} \hat{P}_m \tag{2.21}$$

for each $\hat{P}_m$, thus

$$\sum_n Q_{nm} = 0 \tag{2.22}$$

for each m, i.e. the columns of the matrix must add to zero. Moreover, the last requirement states that the off-diagonal terms of Q_{nm} must all be positive. To see this last point, we do a proof by contradiction. Suppose there is a negative off-diagonal entry, without loss of

generality, component Q_{21}. Then if at time zero our probability vector starts at $\hat{P}_1(0) = 1$ and $\hat{P}_n(0) = 0$ for $n > 1$, an infinitesimal time step dt later we have

$$\hat{P}_2(\mathrm{d}t) = \mathrm{d}t \sum_m Q_{2m}\hat{P}_m(0) = Q_{21} < 0, \tag{2.23}$$

a contradiction since probabilities must remain positive at all times. Of all the requirements, the combination of the second and the fourth ones restricts us to a lower-order discretization since it is not possible to have a higher-order discretization that is both positivity preserving and linear (Zhang & Shu 2011).

Motivation for (2.20) yielding an approximation of the underlying partial differential equation comes directly from the ability of finite-volume methods to converge (under suitable assumptions) to an underlying linear partial differential equation. An example of a matrix, Q, with such a structure, can be seen in the appendix of Hagan, Doering & Levermore (1989), where an approximation to an Ornstein–Uhlenbeck process is constructed. See appendices C.2 and C.3 of Souza *et al.* (2023*b*) for examples on analytically constructing coarse-grained operators of stochastic systems with these properties and Appendix D for an example with the simple harmonic oscillator.

These four requirements, taken together, are enough to identify the matrix Q as the generator of a continuous-time Markov process with finite state space. This observation forms the backbone of the data-driven approach towards discretizing (2.3) and (2.8). The diagonal entries of the matrix are related to the average amount of time spent in a cell, and the off-diagonal entries within a column are proportional to the probabilities of entering a given cell upon exit of a cell. The implication is that we construct the numerical fluxes on a boundary through Monte Carlo integration of the equations of motion.

Intuitively, as a state trajectory enters passes through a cell, it becomes associated with the cell. The time a trajectory spends within a cell is called the holding time from whence it will eventually exit to some other cell in state space. A sufficiently long integration of the equations of motion constructs the holding time distributions of a cell and exit probabilities to different cells in state space. Furthermore, to perform calculations, we associate each region of state space with a 'cell centre', which, in this paper, we call a Markov state. The Markov state will serve as a centre of a delta function approximation to the steady distribution in that region of state space. With the transition matrix Q and the Markov states associated with a partition, we can perform calculations of moments, steady-state distributions and autocorrelations of any variable of interest.

2.4. *Time versus ensemble calculations*

At this point, we have discussed the equation for statistics of a dynamical system, the notation for the infinite-dimensional case and how to associate a continuous time Markov process with a finite-volume discretization of the Liouville equation. We now discuss how to perform statistical calculations from the discretization and how we will confirm that the discretization captures statistics of the underlying Liouville equation. In short, we compare temporal averages with ensemble averages and analogous calculations for autocorrelations.

We must introduce additional notation. As stated before, we assume that a chaotic attractor exists and that there is an appropriately regularized invariant measure, which we denote by $\mathcal{I}(\mathfrak{s})$, associated with it (Cowieson & Young 2005). We further assume that the zero noise limit is the last limit taken in all computations in order to relate computations to deterministic systems. With regards to a turbulent flow, we are assuming that a statistically steady-state exists and denote its probability density (formally) by $\mathcal{I}(\mathfrak{s})$. The conditional

invariant measure with respect to a cell $\mathcal{M}_n$ is denoted by $\mathcal{I}_n(\mathfrak{s})$ and the probability of a state being found in a cell $\mathcal{M}_n$ is $\mathbb{P}(\mathcal{M}_n) = \int_{\mathcal{M}_n} \mathcal{I}$ so that the invariant measure is decomposed as

$$\mathcal{I}(\mathfrak{s}) = \sum_n \mathcal{I}_n(\mathfrak{s})\mathbb{P}(\mathcal{M}_n). \tag{2.24}$$

In addition, we introduce notation for the transfer operator $\mathcal{T}^\tau$, which is defined through the relation

$$\mathcal{P}(\mathfrak{s}, t+\tau) = \mathcal{T}^\tau \mathcal{P}(\mathfrak{s}, t), \tag{2.25}$$

where $\mathcal{P}$ is a solution to (2.3). Thus, the transfer operator is an instruction to evolve the density, $\mathcal{P}$, via (2.3) to a time τ in the future. Furthermore,

$$\lim_{\tau\to\infty} \mathcal{T}^\tau \mathcal{P}(\mathfrak{s}, t) = \mathcal{I}(\mathfrak{s}) \tag{2.26}$$

for arbitrary densities $\mathcal{P}(\mathfrak{s}, t)$, including $\delta(\mathfrak{s} - \mathfrak{s}')$ for an initial state $\mathfrak{s}' \in \mathcal{M}$ from our assumption of ergodicity. (We are being sloppy with limits here, but this should be understood where the limit to a delta function density is the last limit taken.)

For an observable $g : \mathcal{M} \to \mathbb{R}$, we calculate long time averages

$$\langle g\rangle_T = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(\boldsymbol{s}(t))\,\mathrm{d}t \tag{2.27}$$

and compare with ensemble averages

$$\langle g\rangle_E = \int_{\mathcal{M}} g(\mathfrak{s})\mathcal{I}(\mathfrak{s}). \tag{2.28}$$

Furthermore, we compare time-correlated observables. The time-series calculation is

$$R_T(g,\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(\boldsymbol{s}(t+\tau))g(\boldsymbol{s}(t))\,\mathrm{d}t, \tag{2.29}$$

whence we obtain the autocovariance, C_T, and autocorrelation, $\tilde{C}_T$,

$$C_T(g,\tau) \equiv R_T(g,\tau) - \langle g\rangle_T^2 \quad \text{and} \quad \tilde{C}_T(g,\tau) \equiv C_T(g,\tau)/C_T(g,0). \tag{2.30a,b}$$

The ensemble average version requires more explanation. We correlate a variable $g(\boldsymbol{s}(t))$ with $g(\boldsymbol{s}(t+\tau))$, which involves the joint distribution of two variables. We first review a fact about random variables X, Y with joint density $\rho(x, y)$, conditional density $\rho(x|y)$ and marginal density $\rho_y(y)$. The correlation of two observables is calculated as

$$\langle g(X)g(Y)\rangle = \iint \mathrm{d}x\,\mathrm{d}y\, g(x)g(y)\rho(x,y) = \iint \mathrm{d}x\,\mathrm{d}y\, g(x)g(y)\rho(x|y)\rho_y(y) \tag{2.31}$$

$$= \int \mathrm{d}y\, g(y)\rho_y(y)\left[\int \mathrm{d}x\, g(x)\rho(x|y)\right]. \tag{2.32}$$

To translate the above calculation to the present case, we consider ρ_y as the invariant measure, $\mathcal{I}$. The conditional distribution $\rho(x|y)$ is thought of as the probability density at a time τ in the future, given that we know that it is initially at state $\mathfrak{s}$ at $\tau = 0$.

Thus, in our present case, $\rho(x|y)$ becomes $\mathcal{T}^\tau\delta(\boldsymbol{s}-\boldsymbol{s}')$ where the δ function density is a statement of the exact knowledge of the state at time $\tau=0$. In total, the ensemble time-autocorrelation is calculated as

$$R_E(g,\tau)=\int_{\mathcal{M}}\mathrm{d}\boldsymbol{s}'\,g(\boldsymbol{s}')\mathcal{I}(\boldsymbol{s}')\left[\int_{\mathcal{M}}\mathrm{d}\boldsymbol{s}\,g(\boldsymbol{s})\mathcal{T}^\tau\delta(\boldsymbol{s}-\boldsymbol{s}')\right]. \tag{2.33}$$

The autocovariance, C_E, and autocorrelation, $\tilde{C}_E$, are

$$C_E(g,\tau)\equiv R_E(g,\tau)-\langle g\rangle_E^2 \quad\text{and}\quad \tilde{C}_E(g,\tau)\equiv C_E(g,\tau)/C_E(g,0). \tag{2.34a,b}$$

These calculations summarize the exact relations we wish to compare. However, first, we will approximate the temporal averages via long-time finite trajectories and the ensemble averages via the finite-volume discretization from § 2.3.

2.5. *Approximations to time versus ensemble calculations*

The prior section represents the mathematical ideal with which we would like to perform calculations; however, given that we use a data-driven construction, we are faced with performing calculations in finite-dimensional spaces and over finite-dimensional time.

Given the time series of a state at evenly spaced times at times t_n for $n=1$ to N_t with time spacing Δt, we approximate the mean and long time averages of an observable g as

$$\langle g\rangle_T\approx\frac{1}{N_t}\sum_{n=1}^{N_t}g(\boldsymbol{s}(t_n)), \tag{2.35}$$

$$R_T(g,\tau)\approx\frac{1}{N_t'}\sum_{n=1}^{N_t'}g(\boldsymbol{s}(t_n+\mathrm{round}(\tau/\Delta t)\Delta t))g(\boldsymbol{s}(t_n)), \tag{2.36}$$

where the round function computes the closest integer and $N_t'=N_t-\mathrm{round}(\tau/\Delta t)$.

We use the construction from § 2.3 to calculate ensemble averages. Recall that in the end, we had approximated the generator of the process with a matrix Q, which described the evolution of probabilities associated with cells of state space. In addition, we select a state $\boldsymbol{\sigma}^{[n]}$ associated with a cell $\mathcal{M}_n$ as the 'cell centre' to perform calculations. We use superscripts to denote different states since subscripts are reserved for the evaluation of the component of a state. Furthermore, we do not require the Markov state $\boldsymbol{\sigma}^{[n]}$ to be a member of the cell $\mathcal{M}_n$. For example, we could choose the $\boldsymbol{\sigma}^{[n]}$ as fixed points of the dynamical system or a few points along a periodic orbit within the chaotic attractor $\mathcal{M}$.

The ensemble average of an observable is calculated by making use of the decomposition of the invariant measure (2.24), but then approximating

$$\mathcal{I}_n(\boldsymbol{s})\approx\delta(\boldsymbol{s}-\boldsymbol{\sigma}^{[n]}) \tag{2.37}$$

which is a simple but crude approximation. Thus, the ensemble averages are calculated as

$$\langle g\rangle_E=\int_{\mathcal{M}}g(\boldsymbol{s})\left[\sum_n\mathcal{I}_n(\boldsymbol{s})\mathbb{P}(\mathcal{M}_n)\right]=\sum_n\left[\int_{\mathcal{M}}g(\boldsymbol{s})\mathcal{I}_n(\boldsymbol{s})\right]\mathbb{P}(\mathcal{M}_n), \tag{2.38}$$

$$\approx\sum_n\left[\int_{\mathcal{M}}g(\boldsymbol{s})\delta\left(\boldsymbol{s}-\boldsymbol{\sigma}^{[n]}\right)\right]\mathbb{P}(\mathcal{M}_n)=\sum_n g(\boldsymbol{\sigma}^{[n]})\mathbb{P}(\mathcal{M}_n). \tag{2.39}$$

For the ensemble average version of time autocorrelations, we must, in addition to approximating the invariant measure, approximate the transfer operator acting delta

function density of the state, $\mathcal{T}^\tau \delta(\mathfrak{s} - \mathfrak{s}')$. We calculate

$$R_E(g,\tau) = \int_{\mathcal{M}}\int_{\mathcal{M}} \mathrm{d}\mathfrak{s}'\,\mathrm{d}\mathfrak{s}\; g(\mathfrak{s}')\mathcal{I}(\mathfrak{s}')g(\mathfrak{s})\mathcal{T}^\tau\delta(\mathfrak{s}-\mathfrak{s}') \tag{2.40}$$

$$\approx \sum_n \int_{\mathcal{M}}\int_{\mathcal{M}} \mathrm{d}\mathfrak{s}'\,\mathrm{d}\mathfrak{s}\; g(\mathfrak{s}')\delta(\mathfrak{s}'-\boldsymbol{\sigma}^{[n]})\mathbb{P}(\mathcal{M}_n)g(\mathfrak{s})\mathcal{T}^\tau\delta(\mathfrak{s}-\mathfrak{s}') \tag{2.41}$$

$$= \sum_n g(\boldsymbol{\sigma}^{[n]})\mathbb{P}(\mathcal{M}_n)\int_{\mathcal{M}} \mathrm{d}\mathfrak{s}\; g(\mathfrak{s})\mathcal{T}^\tau\delta(\mathfrak{s}-\boldsymbol{\sigma}^{[n]}) \tag{2.42}$$

then additionally approximate

$$\mathcal{T}^\tau\delta(\mathfrak{s}-\boldsymbol{\sigma}^{[n]}) \approx \sum_{m=1}^{N}\delta(\boldsymbol{\sigma}^{[m]}-\mathfrak{s})[\exp(Q\tau)]_{mn}. \tag{2.43}$$

The matrix exponential $\exp(Q\tau)$ is the discrete Perron–Frobenius/transfer operator. We also directly consider the Perron–Frobenius operator at time scale τ and denote this construction by $\mathcal{F}^{[\tau]}$. The matrix is a (column) stochastic matrix whose entries sum to one. The intuition behind the approximation is to treat the forward evolution of the transfer operator for delta distribution centred at state $\boldsymbol{\sigma}^{[n]}$ as a weighted sum of delta functions centred at state $\boldsymbol{\sigma}^{[m]}$. The weights are given by the probability of being in cell $\mathcal{M}_m$ a time τ in the future, given that we started in cell $\mathcal{M}_n$. Putting together the pieces results in

$$R_E(g,\tau) \approx \sum_{n=1}^{N} g(\boldsymbol{\sigma}^{[n]})\mathbb{P}(\mathcal{M}_n)\left[\sum_{m=1}^{N} g(\boldsymbol{\sigma}^{[m]})[\exp(Q\tau)]_{mn}\right]. \tag{2.44}$$

For finite state space ergodic Markov processes in statistical equilibrium, continuous or discrete in time, the above equation is exact. In practice, we use (2.44) as an *a posteriori* check on the fidelity of a partition by computing the autocorrelation of a few select observables. When eigenvectors and eigenvalues of the underlying continuous operator exist, the ability of (2.44) to properly represent autocorrelations relies on the data-driven method's ability to represent the underlying operator's eigenvectors and eigenvalues.

In addition, all covariances and correlations are calculated using the above approximations. This review completes the discussion of how to approximate ensemble averages and covariances from the finite-volume discretization of the generator. However, it remains to be shown how to construct the matrix, and Markov states, from data. The construction of the generator is the subject of § 3.

2.6. *Eigenvalues and eigenvectors of the generator*

The generator has eigenvalues as well as (left and right) eigenvectors. The simplest eigenvectors are the ones associated with the eigenvalue 0. Recall the requirement that the column sum of matrix entries should add up to zero, for each column. This requirement is a statement about the left eigenvector of Q. To see this denote $\mathbf{1}$ as the eigenvector of all 1s; then

$$\mathbf{1}^T Q = \mathbf{0}^T = 0\mathbf{1}^T. \tag{2.45}$$

The first equality is exactly the statement that the column sum of the matrix should add up to zero for each column and the second equality rewrites the zero vector as the zero scalar

times the original vector. The latter inequality shows that it is an eigenvector of the system. The assumption that the Q matrix is ergodic then implies that there is only one eigenvector corresponding to the eigenvalue 0. We shall now discuss the other eigenpairs.

2.7. *Global Koopman eigenvectors and modes*

We do not ask 'can we predict an observable of interest?', but rather, 'what can we predict?'. The latter question is an emergent property of the system and captured by the Koopman eigenvectors of the underlying system. Those Koopman eigenvectors whose decorrelation time scales are long-lived constitute the most predictable features of the system.

We use the same terminology introduced in Williams *et al.* (2015) to discuss the Koopman operator and its eigenvectors. Koopman eigenvectors are observables as well as left eigenvectors of the transition probability operator $\mathcal{T}^\tau$. For example, if g_λ is a left eigenvector of $\mathcal{T}^\tau$ with eigenvalue $\mathrm{e}^{\lambda\tau}$ then we have the following:

$$R_E(g_\lambda,\tau) = \int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}'\, g_\lambda(\boldsymbol{\mathfrak{s}}')\mathcal{I}(\boldsymbol{\mathfrak{s}}')\left[\int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}\, g_\lambda(\boldsymbol{\mathfrak{s}})\mathcal{T}^\tau\delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}')\right] \tag{2.46}$$

$$= \int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}'\, g_\lambda(\boldsymbol{\mathfrak{s}}')\mathcal{I}(\boldsymbol{\mathfrak{s}}')\left[\int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}\, g_\lambda(\boldsymbol{\mathfrak{s}})\mathrm{e}^{\lambda\tau}\delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}')\right] \tag{2.47}$$

$$= \mathrm{e}^{\lambda\tau}\int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}'\, g_\lambda(\boldsymbol{\mathfrak{s}}')^2\mathcal{I}(\boldsymbol{\mathfrak{s}}') \tag{2.48}$$

$$= \mathrm{e}^{\lambda\tau}\langle g_\lambda^2\rangle_E. \tag{2.49}$$

Thus, the most useful Koopman eigenvectors, from a predictability standpoint, are those such that they decorrelate slowly in time, i.e. $\mathrm{real}(\lambda)\approx 0$, but additionally have an oscillatory component so that the ratio $\mathrm{real}(\lambda)/\mathrm{imaginary}(\lambda)\approx 0$ holds.

If $\mathrm{real}(\lambda)=0$ on a chaotic attractor, then we expect this to be the 'trivial' observable $g_\lambda(\boldsymbol{\mathfrak{s}})=c$ for a constant c. (The presence of pure-imaginary eigenvalues would imply the existence of observables that are predictable for arbitrary times in the future on a chaotic attractor.) Otherwise, we expect that $\mathrm{real}(\lambda)<0$ for all eigenvalues of Q, i.e. we expect that all non-trivial observables will eventually decorrelate. This implies $\langle g_\lambda\rangle_E = 0$ since

$$\langle g_\lambda\rangle_E = \int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}\, g_\lambda(\boldsymbol{\mathfrak{s}})\mathcal{I}(\boldsymbol{\mathfrak{s}}) = \lim_{\tau\to\infty}\int_{\mathcal{M}} \mathrm{d}\boldsymbol{\mathfrak{s}}\, g_\lambda(\boldsymbol{\mathfrak{s}})\mathcal{T}^\tau\delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}') = \lim_{\tau\to\infty}\mathrm{e}^{\lambda\tau}g_\lambda(\boldsymbol{\mathfrak{s}}') = 0, \tag{2.50}$$

where $\boldsymbol{\mathfrak{s}}'$ is an arbitrary state on the attractor $\mathcal{M}$.

The following four statements about a Koopman eigenvectors g_λ cannot hold simultaneously:

(i) the Koopman eigenvector satisfies the relation $g_\lambda(\boldsymbol{s}(t+\tau)) = \mathrm{e}^{\lambda\tau}g_\lambda(\boldsymbol{s}(t))$;
(ii) the Koopman eigenvector g_λ is a continuous function of state space;
(iii) there exist an arbitrary number of near recurrences on the dynamical trajectory;
(iv) the eigenvalue associated with the Koopman eigenvector satisfies $\mathrm{real}(\lambda)<0$.

The proof is as follows. Suppose that all four criteria are satisfied. Let $\boldsymbol{\mathfrak{s}}'_n$ be near-recurrences of $\boldsymbol{\mathfrak{s}}$ some sequence of times $\{\tau_n\}$, where $\tau_n\to\infty$, in the future so that $\|\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}'_n\|<\epsilon$ for some norm and all n uniformly. Continuity of g_λ with respect to the

norm implies

$$|g_\lambda(\mathfrak{s}) - g_\lambda(\mathfrak{s}'_n)| < \delta \tag{2.51}$$

but $g_\lambda(\mathfrak{s}'_n) = \mathrm{e}^{\lambda\tau_n} g_\lambda(\mathfrak{s})$ by assumption, hence

$$|g_\lambda(\mathfrak{s})||1 - \mathrm{e}^{\lambda\tau_n}| < \delta \tag{2.52}$$

which is a contradiction since τ_n can be made arbitrarily large and δ arbitrarily small. The non-existence of Koopman eigenvectors satisfying $g_\lambda(\boldsymbol{s}(t+\tau)) = \mathrm{e}^{\lambda\tau} g_\lambda(\boldsymbol{s}(t))$ over all of state space is corroborated by numerical evidence (Parker & Page 2020). In so far as a turbulent attractor is mixing one does not expect a finite-dimensional linear subspace for the Koopman eigenvectors (except for the constant observable). See Arbabi & Mezić (2017) for a similar statement with regards to the Lorenz attractor. We take the above proof as a plausible argument for the use of a piecewise constant basis in the representation of Koopman eigenvectors.

For stochastic dynamical systems, one expects that the Koopman eigenvectors (the Koopman operator is defined in terms of the adjoint of the Fokker–Planck operator in that context) are more likely to be continuous functions of the state but no longer obey the relation $g_\lambda(\boldsymbol{s}(t+\tau)) = \mathrm{e}^{\lambda\tau} g_\lambda(\boldsymbol{s}(t))$. Heuristically, we expect better regularity because the stochastic noise in the dynamics acts as a diffusion in probability space, which smooths out non-smooth fields. For a stochastic differential equation

$$\dot{\boldsymbol{s}} = \boldsymbol{U}(\boldsymbol{s}) + \epsilon\boldsymbol{\xi}, \tag{2.53}$$

where ϵ is the noise variance and $\boldsymbol{\xi}$ is a d-dimensional white noise, the Koopman eigenvector evolves according to

$$\dot{g}_\lambda(\boldsymbol{s}(t)) = \lambda g_\lambda(\boldsymbol{s}(t)) + \epsilon\boldsymbol{\nabla} g_\lambda(\boldsymbol{s}(t)) \cdot \boldsymbol{\xi}, \tag{2.54}$$

see (6.55) of Maćešić & Črnjarić-Žic (2020). The formula above follows from Ito's lemma and using that g_λ as an eigenvector with eigenvalue λ. The composition with the state variable is necessary because $g_\lambda : \mathbb{R}^d \to \mathbb{R}$, but one cannot consider the evolution of g_λ independently from where it is being evaluated in state space. In the limit that the noise goes to zero, $\epsilon \to 0$, the gradient term, $\boldsymbol{\nabla} g_\lambda$, can go to infinity at particular points in state space, as would be expected in a one-dimensional stochastic dynamical system with a two-well potential. Consequently, pathologies are unexpected in linear systems.

As another point, although they are often called Koopman eigenfunctions when the dynamical system is a partial differential equation, they should perhaps more appropriately be called Koopman eigenoperators (in analogy to eigenvectors and eigenfunctions in lower-dimensional contexts), i.e. functionals that act on a state. The right eigenvectors of the transfer operator act as projection operators of an observable to Koopman eigenvectors. When a continuum of observables describe a field the result of the projection of this continuum is called a Koopman mode, see Williams *et al.* (2015). Part 2 goes through a concrete example, but we remain abstract here.

A continuum of observables indexed by spatial index $\boldsymbol{x}$ define statistical modes as

$$G_\lambda(\boldsymbol{x}) \equiv \int_{\mathcal{M}} \mathrm{d}\mathfrak{s}\, g^{\boldsymbol{x}}(\mathfrak{s}) v_\lambda(\mathfrak{s}), \tag{2.55}$$

where v_λ is a right eigenvector of the transfer operator, i.e.

$$\mathcal{T}^\tau v_\lambda = \mathrm{e}^{\lambda\tau} v_\lambda. \tag{2.56}$$

Equation (2.55) projects the observable $g^{\boldsymbol{x}}$ onto the appropriate Koopman eigenvector, due to biorthogonality of left and right eigenvectors. The object $G_\lambda(\boldsymbol{x})$, for a fixed λ is a

Koopman mode. Whether or not the set of Koopman eigenvectors form a complete basis so that $g^{\boldsymbol{x}} = \sum_\lambda G_\lambda(\boldsymbol{x}) g_\lambda$ is unclear. The implication is that an arbitrary observable $g^{\boldsymbol{x}}$ could be fundamentally unpredictable if it cannot be expressed as a sum of Koopman eigenvectors.

In the next section, we discuss the numerical approximation to Koopman eigenvectors.

2.8. *Numerical approximation*

The numerical Koopman eigenvectors are the left eigenvectors of the matrix Q, denoted by $\boldsymbol{g}_\lambda$ and their approximation as functionals acting on the state $\boldsymbol{\mathfrak{s}} \in \mathcal{M}_n$ is given by

$$g_\lambda(\boldsymbol{\mathfrak{s}}) \approx [\boldsymbol{g}_\lambda]_n, \tag{2.57}$$

where $[\boldsymbol{g}_\lambda]_n$ is the nth component of the eigenvector $\boldsymbol{g}_\lambda$. Thus, we first determine which cell the state $\boldsymbol{\mathfrak{s}}$ belongs to and then use the integer label to pick out the component of the eigenvector $\boldsymbol{g}_\lambda$. Thus, we use a piecewise constant approximation to the Koopman eigenvector.

3. Methodology

In this section, we outline the general approach to constructing the approximate generator in terms of trajectory data. The most critical component of a discretization comes from defining a classifier $\mathcal{C} : \mathbb{R}^d \to \{1, 2, \ldots, N\}$ which maps an arbitrary state $\boldsymbol{\mathfrak{s}} \in \mathbb{R}^d$ to an integer $n \in \{1, 2, \ldots, N\}$. This function implicitly defines a cell through the relation

$$\mathcal{M}_j = \{\boldsymbol{\mathfrak{s}} : \mathcal{C}(\boldsymbol{\mathfrak{s}}) = j \text{ for each } \boldsymbol{\mathfrak{s}} \in \mathbb{R}^d\} \cap \mathcal{M}, \tag{3.1}$$

and thus we identify an integer j with a cell $\mathcal{M}_j$. The intersection with the manifold $\mathcal{M}$ is critical to the methodology's success.

Furthermore, the Markov states (cell centres) are chosen to satisfy $\mathcal{C}(\boldsymbol{\sigma}^{[n]}) = n$ for each $n \in \{1, \ldots, N\}$. There is an extraordinary amount of freedom in defining the classifier, and we will go through examples in § 4. We also comment on practical considerations and generalizations in § 5. One can simultaneously solve for a classifier and Markov states using a K-means algorithm (see Lloyd (1982)), but we do not wish to restrict ourselves to that choice here. The classifier 'classifies' (in the machine learning sense) different flow states with integers as the category labels. The classifier serves as a particular choice of the nonlinear dictionary (basis) for observables of the system; see Appendix C and Klus *et al.* (2016) for a discussion on this point. Choosing a good basis for observables is difficult and may not even exist, but deep learning methods combined with novel loss functions hold promise for their discovery (Bittracher *et al.* 2023; Constante-Amores *et al.* 2023). For now, we will assume that such a function is given and focus on constructing the generator Q.

The classifier $\mathcal{C}$ transforms dynamical trajectories into sequences of integers, which we interpret as the realization of a Markov process with finite state space. At this stage, traditional methods can be employed to construct a transfer/Perron–Frobenius operator from data, see Klus *et al.* (2016) and Fernex *et al.* (2021). Given that our interest is in constructing a continuous time Markov process and quantifying the uncertainty of matrix entries due to finite sampling effects, the algorithm will modify the traditional approaches.

To construct Q, two quantities must be calculated for each cell:

(i) the holding times – the amount of time a dynamical trajectory stays in cell $\mathcal{M}_n$ before exiting;

(ii) the exit probabilities – the probability of moving from cell $\mathcal{M}_j$ to $\mathcal{M}_i$ upon exiting the cell $\mathcal{M}_j$.

Let T_j be the distribution of holding times associated with cell j and E_{ij} denote entries of the exit probability matrix for $i \neq j$. By convention we take $E_{ii} = -1$ so that $\sum_i E_{ij} = 0$ and for each j. The entries of the matrix Q_{ij} are constructed as follows:

$$Q_{ij} = E_{ij}/\langle T_j \rangle, \tag{3.2}$$

where $\langle T_j \rangle$ denotes the expected value of the holding time distribution of cell j. The matrix Q is decomposed into the product of the exit probability matrix E and inverse holding time matrix $R \equiv \text{Diagonal}(\boldsymbol{T})^{-1}$, $Q = ER$ when numerically constructing the matrix entries. The data-driven construction of the Perron–Frobenius operator $\mathcal{F}^{[\Delta t]}$ is similar and one simply keeps track of the transition probabilities from cell $\mathcal{M}_j$ to $\mathcal{M}_i$. A Markovian assumption implies that the holding time is exponentially distributed in the generator case and geometrically distributed for the discrete in time case. We re-examine the Markovian assumption when applying the method in § 4.

In the subsections, we outline an empirical construction of the matrix from finite data and a Bayesian approach incorporating uncertainty due to finite sampling effects. With the latter approach, we do not treat the entries of the Q_{ij} matrix as deterministic numbers but as distributions. The result is a random matrix representation of the generator that incorporates uncertainty. We emphasize that our focus on using the generator of the process is critical to the incorporation uncertainty since we assume that the data comes from a continuous-time dynamical system.

3.1. *Empirical construction*

We start with an empirical construction of the generator. It suffices to focus on cell j associated with the jth column of the matrix Q_{ij}. To calculate the empirical holding time distribution and empirical mean, we count how often we see cell j before transitioning to cell $i \neq j$. For example, suppose that we have three cells, $j = 1$, and consider the following sequence of integers given by the classifier applied to a time series with Δt spacing in time:

$$1, 1, 1, 2, 2, 1, 1, 3, 1, 2, 1, 1. \tag{3.3}$$

We group the sequence as follows:

$$(1, 1, 1), 2, 2, (1, 1), 3, (1), 2, (1, 1) \tag{3.4}$$

to determine the holding times. Thus, the holding times for cell 1 would be

$$3\Delta t, 2\Delta t, \Delta t, 2\Delta t \tag{3.5}$$

whose empirical expected value is $2\Delta t$ implying a transition rate $1/(2\Delta t)$.

To calculate exit probabilities for cell j, we count how often we see transitions to cells i and divide by the total number of transitions. In the example, to calculate the exit probabilities for cell 1 into cell 2 or 3, we group them together as follows:

$$1, 1, (1, 2), 2, 1, (1, 3), (1, 2), 1, 1. \tag{3.6}$$

Thus, we observed three exits, two of which went to state 2 and one to state 3; hence, the exit probabilities are $E_{21} = 2/3$ and $E_{31} = 1/3$.

The rest are constructed analogously to produce the matrix

$$Q = ER = \begin{bmatrix} -1 & 1 & 1 \\ 2/3 & -1 & 0 \\ 1/3 & 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2\Delta t} & 0 & 0 \\ 0 & \frac{2}{3\Delta t} & 0 \\ 0 & 0 & \frac{1}{\Delta t} \end{bmatrix} = \frac{1}{\Delta t} \begin{bmatrix} -1/2 & 2/3 & 1 \\ 1/3 & -2/3 & 0 \\ 1/6 & 0 & -1 \end{bmatrix}. \tag{3.7}$$

We give an alternative method of constructing Q in Appendix C where we interpret operations from an algebraic perspective. As described, the generator is only accurate to order Δt since we do not interpolate in time to find the 'exact' holding time. We do not preoccupy ourselves with improving this since we believe that the primary source of error comes from finite sampling effects.

In certain cases, the data-driven construction given above can be obtained directly from a data-driven construction of the Perron–Frobenius operator. For example, when a time series is sampled with a time step of Δt and a trajectory only spends one time step Δt within each cell, then the Q matrix is discretely related to the data-driven construction Perron–Frobenius operator, $\mathcal{F}^{[\Delta t]}$, through the relation $\mathcal{F}^{[\Delta t]} = \mathbb{I} + \Delta t Q$ where $\mathbb{I}$ is the identity matrix. This happens because the only transition probabilities observed are exit probabilities. Furthermore, the relation $\mathcal{F}^{[\Delta t]} \approx \mathbb{I} + \Delta t Q$ also holds in the limit that a trajectory spends sufficient time within each cell j, i.e. the average holding times are much greater than the time step size. See Appendix D for concrete examples using the simple harmonic oscillator. The advantage of the continuous-time formulation comes from uncertainty-quantification which yields sensible uncertainty estimates in both limits.

In the following section, we augment the empirical construction with uncertainty estimates based on finite sampling and a Bayesian framework.

3.2. *Bayesian construction*

Our goal is to quantify the uncertainty of the discrete generator's matrix entries due to finite sampling effects from data-driven methods. To this end, we use a Bayesian method, which requires choices for likelihood functions and prior distributions. In the case of a generator matrix, we require four ingredients:

(i) a likelihood distribution for the holding times of each cell;
(ii) a prior distribution for the transition rates associated with the holding times;
(iii) a likelihood distribution for the exit probabilities of each cell;
(iv) a prior distribution for the probability values associated with the exit probabilities.

The likelihood distribution is the distribution we believe we are sampling from, and the prior distribution encapsulates our uncertainty with respect to the parameters of the likelihood function. Bayesian methods are computationally efficient under certain likelihood and prior distribution assumptions (Gelman *et al.* 2013). In particular, if the prior distribution and the posterior distribution are from the same family of distributions, then it is only necessary to update the parameters of the prior/posterior distribution. For a fixed likelihood function, in special cases, it is possible to determine a prior distribution such that the posterior distribution comes from the same family as the prior. Such distributions are called conjugate priors (conjugate with respect to the likelihood), and we shall use them.

We make choices for our likelihood function compatible with that of a continuous-time Markov process:

(i) the likelihood distribution for the holding times is exponentially distributed with rate parameter λ_i for each cell independently;
(ii) the likelihood distribution for exit probabilities is a multinomial distribution with parameters $\boldsymbol{p} \in [0, 1]^{N-1}$ satisfying the relation $\sum_{i=1}^{N-1} p_i = 1$ for each cell independently.

Both the exponential distribution and the multinomial distribution have known conjugate priors:

(i) the conjugate prior distribution for the rate parameter of the exponential distribution is distributed according to the Gamma distribution with parameters (α, β), denoted by $\Gamma(\alpha, \beta)$;
(ii) the conjugate prior distribution for the exit probabilities of the multinomial distribution comes from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$ of length $N - 1$, which we denote by Dirichlet $(\boldsymbol{\alpha})$.

Thus, the posterior distributions then come from the same family as the prior distributions, e.g. posterior distributions are $\Gamma(\alpha', \beta')$ and Dirichlet $(\boldsymbol{\alpha}')$ upon the observation of data; see Gelman *et al.* (2013) and the section below for an example. Under the construction of this section, a 3×3 matrix will always be of the form

$$Q = \begin{bmatrix} -1 & [\boldsymbol{D}_2]_1 & [\boldsymbol{D}_3]_1 \\ [\boldsymbol{D}_1]_1 & -1 & [\boldsymbol{D}_3]_2 \\ [\boldsymbol{D}_1]_2 & [\boldsymbol{D}_2]_2 & -1 \end{bmatrix} \begin{bmatrix} G_1 & 0 & 0 \\ 0 & G_2 & 0 \\ 0 & 0 & G_3 \end{bmatrix} \tag{3.8}$$

where $G_i \sim \Gamma(\alpha_i, \beta_i)$, $\boldsymbol{D}_i \sim \text{Dirichlet}(\boldsymbol{\alpha}_i)$ and $[\boldsymbol{D}_i]_j$ denotes the jth component of the random vector $\boldsymbol{D}_i$.

As mentioned in the previous paragraphs, conjugate priors greatly expedite the Bayesian update procedure; only the parameters of the Gamma and Dirichlet distributions need to be updated to construct a posterior distribution. The parameters (α_i, β_i) and $\boldsymbol{\alpha}_i$ are updated according to Bayes' rule for each column upon data acquisition. For example, suppose that we have observed the following empirical counts associated with cell i:

(i) M exits from cell i;
(ii) $[\boldsymbol{M}]_j$ exits from cell i to cell j;
(iii) $\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_M$ empirically observed holding times;

and that we start with (α^0, β^0) and $\boldsymbol{\alpha}^0$ as the parameters for our prior distributions. The relation $\sum_j [\boldsymbol{M}]_j = M$ holds. (There can be an 'off-by-one' error here which we ignore for presentation purposes.) The posterior distribution parameters (α^1, β^1) and $\boldsymbol{\alpha}^1$ are

$$\alpha^1 = \alpha^0 + M, \quad \beta^1 = \beta^0 + \sum_i^M \hat{T}_i \quad \text{and} \quad \boldsymbol{\alpha}^1 = \boldsymbol{\alpha}^0 + \boldsymbol{M}. \tag{3.9a–c}$$

In the limit that α^0, β^0 and $|\boldsymbol{\alpha}^0|$ go to zero, then the empirical approach from the previous section agrees with the expected value from the Bayesian approach.

The current approach is one of many approaches to constructing matrices with quantified uncertainties. See Singhal & Pande (2005) and Trendelkamp-

Schroer *et al.* (2015) for other Bayesian examples of calculating uncertainties and Scherer *et al.* (2015) for numerical implementations. We comment that the present methodology also allows for uncertainty quantification of generators that satisfy detailed balance, see Appendix B. Other probabilistically inspired methods of estimating generator entries consist of methods such as the expectation maximization method in Otto, Peitz & Rowley (2023), which yield maximum likelihood estimates.

The current uncertainty quantification is imperfect (e.g. when holding times do not follow an exponential distribution or the system is not Markovian over infinitesimal steps). Still, we hold the position that some quantification of uncertainty is better than none. One of the benefits of the construction of this section is that it is robust to infinite temporal resolution over a fixed period and, hence, is consistent with data that comes from a continuous time process. We use uncertainty quantification to dismiss spurious results rather than increase confidence in the correctness of an inference. Furthermore, the assumptions that we made for the posterior and prior distribution still yield the same empirical construction from (3.1) upon using uninformative priors. In the large data limit, the Bayesian update procedure eventually yields a sharply peaked distribution that converges to a Gaussian.

An example now follows.

4. Illustration of the methodology with the Lorenz equations

We apply the methodology from the previous section to the Lorenz equations. The dynamics are given by

$$\dot{x} = -\sigma(x - y), \tag{4.1}$$

$$\dot{y} = -y + (r - z)x, \tag{4.2}$$

$$\dot{z} = -bz + xy, \tag{4.3}$$

where we identify $x = s_1$, $y = s_2$, $z = s_3$. The corresponding Liouville equation is given by

$$\partial_t \mathcal{P} + \partial_x \left(\left[-\sigma(x - \mathit{y}) \right] \mathcal{P} \right) + \partial_{\mathit{y}} \left(\left[-\mathit{y} + (r - \mathit{z})x \right] \mathcal{P} \right) + \partial_{\mathit{z}} \left(\left[-b\mathit{z} + x\mathit{y} \right] \mathcal{P} \right) = 0 \tag{4.4}$$

where we use the notation x for s_1, y for s_2 and z for s_3. We shall examine three different methods of partitioning the chaotic attractor.

4.1. *Fixed point partition*

We choose the classic parameter values $r = 28$, $\sigma = 10$ and $b = 8/3$ for the Lorenz system, which is known to exhibit chaotic solutions. Construction of the generator is automated through the methodology of § 3 upon choosing the Markov states $\boldsymbol{\sigma}^{[n]}$ and a classifier $\mathcal{C}$. We use the following fiction to guide our choices.

It is said that the coherent structures of a flow organize and guide the dynamics of chaos (Cvitanović 2013). As a trajectory wanders through state space, it spends a disproportionate time near coherent structures and inherits their properties. The coherent structures then imprint their behaviour on the chaotic trajectory, manifesting in ensemble averages. Thus, chaotic averages are understood in terms of transitions between simpler structures.

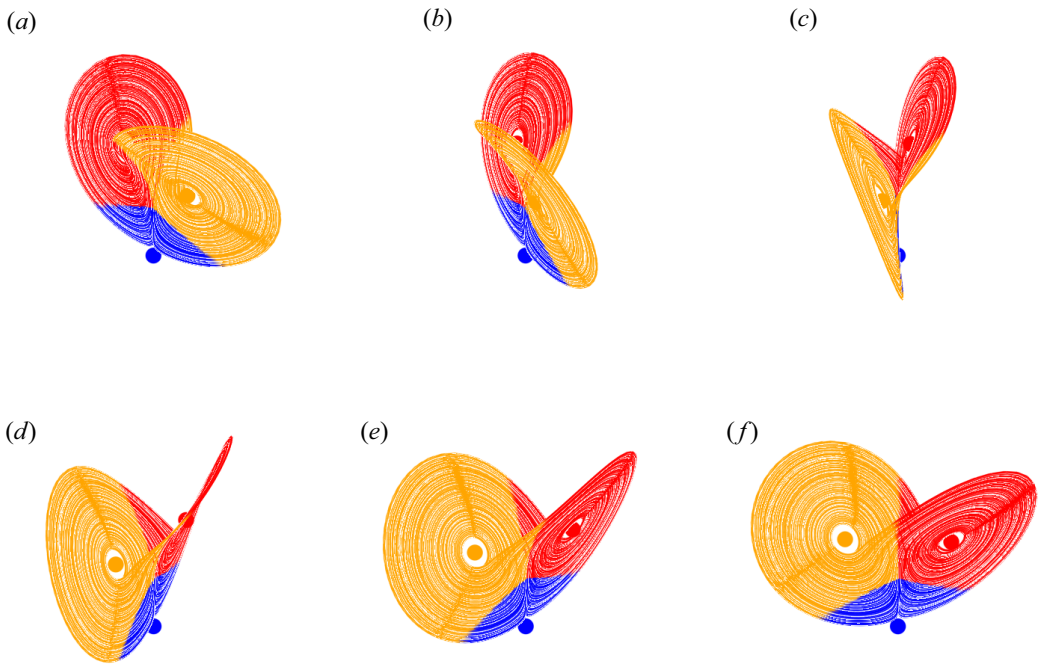


Figure 1. Lorenz fixed point partition. Here, we show the emerging partition from several angles. The colours correspond to the different partitions associated with trajectories that are 'closest' to a given fixed point.

This picturesque story motivates the use of fixed points, as Markov states,

$$\boldsymbol{\sigma}^{[1]} = [-\sqrt{72}, -\sqrt{72}, 27], \tag{4.5}$$

$$\boldsymbol{\sigma}^{[2]} = [0, 0, 0], \tag{4.6}$$

$$\boldsymbol{\sigma}^{[3]} = [\sqrt{72}, \sqrt{72}, 27], \tag{4.7}$$

and partitioning state space according to the closest fixed point,

$$\mathcal{C}(\boldsymbol{\mathfrak{s}}) = \begin{cases} 1 & \text{if } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[1]}\| < \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[2]}\| \text{ and } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[3]}\| \\ 2 & \text{if } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[2]}\| < \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[3]}\| \text{ and } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[1]}\| \\ 3 & \text{if } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[3]}\| < \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[1]}\| \text{ and } \|\boldsymbol{\mathfrak{s}} - \boldsymbol{\sigma}^{[2]}\| \end{cases} \tag{4.8}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. The classifier determines the partition by associating a trajectory with the closest fixed point. This partitioning strategy is the intersection of the chaotic attractor, $\mathcal{M}$, with a Voronoi tessellation over the full state space, $\mathbb{R}^3$. Figure 1 shows the partition induced by this choice. The regions are colour-coded according to the closest fixed points.

We construct a time series from the Lorenz equations using a fourth-order Runge–Kutta time stepping scheme with time step $\Delta t = 5 \times 10^{-3}$. We take the initial condition to be $(x(0), y(0), z(0)) = (14, 20, 27)$ and integrate to time $T = 10^5$, leading to 2×10^7 time snapshots. At each moment in time, we apply the classifier to create a sequence of integers representing the partition dynamics. (One can think of this as defining a symbol sequence.) Figure 2 visualizes this process.

From the sequence of integers, we apply the method from § 3.2 to construct the data-driven approximation to the generator with quantified uncertainty. For our prior

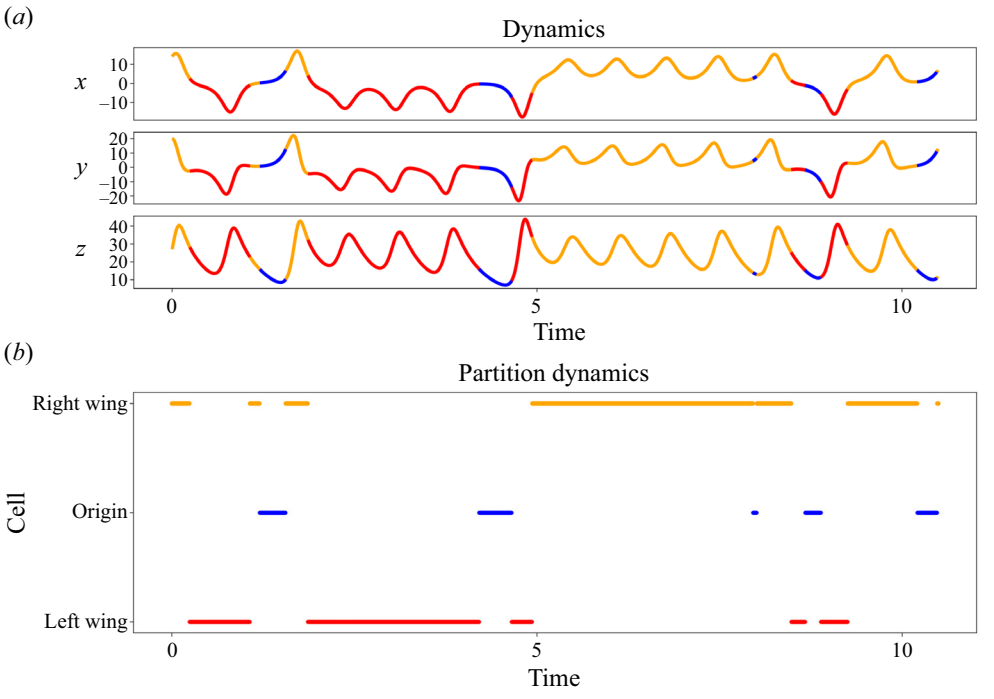


Figure 2. Lorenz fixed point partition Markov chain. The dynamics of the x, y, z variables are shown in (a), and the associated partition dynamics is shown in (b). As a dynamical trajectory moves through state space, it is labelled according to its proximity to the closest fixed point.

distribution, we use an uninformative prior – i.e. initial parameters $\alpha = \beta = 0$ for the Gamma distribution and $\boldsymbol{\alpha} = \mathbf{0}$ for the Dirichlet distribution – so that the mean of the random matrix agrees with the empirical construction from 3.1. The mean for each entry of the generator (reported to two decimal places) is

$$\langle Q \rangle = \begin{bmatrix} -1.17 & 1.93 & 0.65 \\ 0.52 & -3.86 & 0.52 \\ 0.65 & 1.93 & -1.17 \end{bmatrix} \approx \begin{bmatrix} -1.0 & 0.5 & 0.55 \\ 0.45 & -1.0 & 0.45 \\ 0.55 & 0.5 & -1.0 \end{bmatrix} \begin{bmatrix} 1.17 & 0.0 & 0.0 \\ 0.0 & 3.86 & 0.0 \\ 0.0 & 0.0 & 1.17 \end{bmatrix}, \tag{4.9}$$

where we have decomposed the matrix into the exit probability matrix and the diagonal rate matrix on the right-hand side.

In the matrix decomposition, the off-diagonals of the left matrix correspond to the exit probabilities, and the right matrix is the rate matrix, whose entries are the inverse of the time spent within a partition. From the latter matrix, we see that trajectories spend less time in the partition associated with the zero fixed point (blue) since $1/3.86 < 1/1.17$. The apparent symmetry in the matrix results from the truncation to two decimal places and the abundance of data. In Appendix A, we show how to incorporate symmetries of the Lorenz equation and report ensemble mean statistics.

The utility of using a random matrix to represent uncertainty is summarized in figure 3. The distribution of each matrix entry for various subsets of time is displayed. Using fewer data (represented by a shorter gathering time, T) results in significant uncertainty for the matrix entries. Additionally, using unconnected subsets of time demonstrates an apparent convergence of matrix entries.

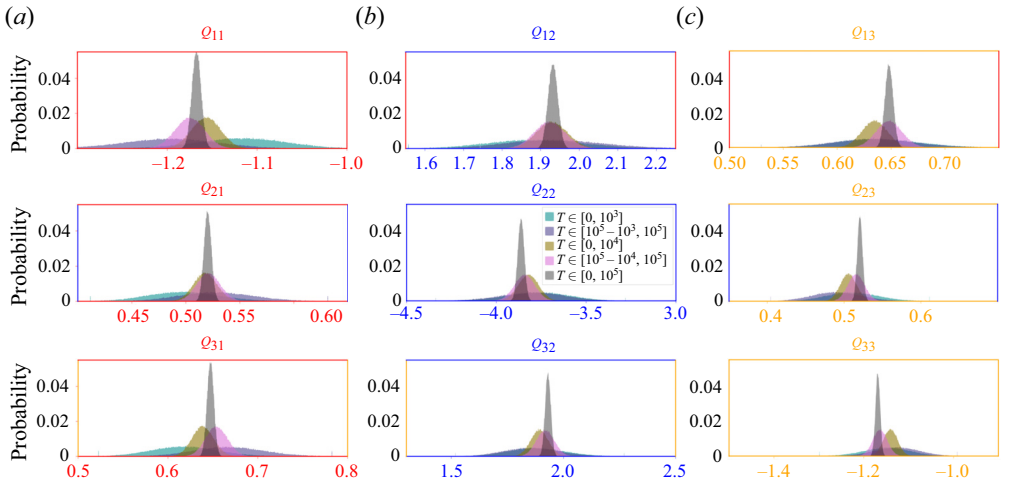


Figure 3. Lorenz fixed point partition distributions of the generator. The uncertainty estimates for the entries of the 3×3 generator are shown in the above figure. A one-to-one correspondence exists between the distributions in the panel and the matrix entries. The different coloured distributions within a panel represent different estimates of the entries based on the amount of available data, here presented in terms of the simulation time of the Lorenz system. We see that as we increase the time interval of the simulation and thus have more data, we become more confident about the matrix entries. Furthermore, the distributional spreads overlap with one another.

We are now in a position to calculate statistical quantities. For simplicity, we only report first-, second- and third-order moments calculated from the mean value of the generator, $\langle Q \rangle$. The steady-state distribution of $\langle Q \rangle$, corresponding to eigenvalue $\lambda = 0$, is reported to two decimal places as

$$[\mathbb{P}(\mathcal{M}_1), \mathbb{P}(\mathcal{M}_2), \mathbb{P}(\mathcal{M}_3)] \approx [0.44, 0.12, 0.44] \tag{4.10}$$

from whence we calculate the steady state statistics for any observable using the approximations in § 2.4 and the Markov states $\boldsymbol{\sigma}^{[n]}$ for $n = 1, 2, 3$. Explicitly, the ensemble average of the observables,

$$g^{[1]}(\mathfrak{z}) = \mathfrak{z}_3 = \mathfrak{x}, \quad g^{[2]}(\mathfrak{z}) = (\mathfrak{z}_3)^2 = \mathfrak{x}^2, \quad \text{or} \quad g^{[3]}(\mathfrak{z}) = (\mathfrak{z}_1)^2 \mathfrak{z}_3 = \mathfrak{x}^2 \mathfrak{x} \tag{4.11}$$

is approximated via (2.39), repeated here for convenience,

$$\langle g^{[j]} \rangle_E = g^{[j]}(\boldsymbol{\sigma}^{[1]}) \mathbb{P}(\mathcal{M}_1) + g^{[j]}(\boldsymbol{\sigma}^{[2]}) \mathbb{P}(\mathcal{M}_2) + g^{[j]}(\boldsymbol{\sigma}^{[3]}) \mathbb{P}(\mathcal{M}_3) \quad \text{for each } j \tag{4.12}$$

to yield

$$\langle \mathfrak{x} \rangle_E \approx 27 \times 0.44 + 0 \times 0.12 + 27 \times 0.44 \approx 24, \tag{4.13}$$

$$\langle \mathfrak{x}^2 \rangle_E \approx 27^2 \times 0.44 + 0^2 \times 0.12 + 27^2 \times 0.44 \approx 642, \tag{4.14}$$

$$\langle x^2 \mathfrak{x} \rangle_E \approx \left(-\sqrt{72}\right)^2 \times 27 \times 0.44 + 0^3 \times 0.12 + \left(\sqrt{72}\right)^2 \times 27 \times 0.44 \approx 1711. \tag{4.15}$$

Table 1 shows the result from both the temporal and ensemble average (using full machine precision for computations). There is a correspondence for all averages, with the most significant discrepancy being those involving y^2 terms, for which the relative error is within 25 %. The fixed points of a dynamical system are unique in that they satisfy all the

	$\langle x\rangle$	$\langle y\rangle$	$\langle z\rangle$	$\langle xx\rangle$	$\langle xy\rangle$	$\langle xz\rangle$	$\langle yy\rangle$	$\langle yz\rangle$	$\langle zz\rangle$
Ensemble	−0.0	−0.0	23.8	63.5	63.5	−0.1	63.5	−0.1	642.4
Temporal	−0.0	−0.0	23.5	62.8	62.8	−0.2	81.2	−0.2	628.9
	$\langle xxy\rangle$	$\langle xxz\rangle$	$\langle xyy\rangle$	$\langle xyz\rangle$	$\langle xzz\rangle$	$\langle yyy\rangle$	$\langle yyz\rangle$	$\langle yzz\rangle$	$\langle zzz\rangle$
Ensemble	−0.3	1713.2	−0.3	1713.2	−3.4	−0.3	1713.2	−3.4	17346.1
Temporal	−0.4	1879.7	−0.4	1677.2	−4.1	−0.4	1997.2	−4.2	18446.3

Table 1. Empirical moments of the Lorenz attractor. A comparison between ensemble averaging and time averaging.

same dynamical balances of a statistically steady state, e.g. chaotic trajectories, periodic orbits and fixed points of the Lorenz equation satisfy the relation

$$\langle x y\rangle = b\langle z\rangle, \tag{4.16}$$

and more generally

$$0 = \langle \boldsymbol{U}(\mathscr{A}) \cdot \nabla g\rangle, \tag{4.17}$$

for any bounded and differentiable observable g, where $\boldsymbol{U}$ is vector field defined by the right-hand side of the Lorenz equations, and the averaging brackets $\langle\cdot\rangle$ are defined over the trajectory. Thus, an accurate estimation of $\langle x y z^{n-1}\rangle$ only depends on an accurate estimate for $\langle z^n\rangle$ for the fixed points of the Lorenz equation (using $g(\mathscr{A}) = z^n$). In this case the 'closure problem' in fluid mechanics is a boon rather than a curse. A good representation of a lower-order moment automatically yields a good representation of a higher-order moment through the closure 'problem'.

Although we focused on moments, one can compare the statistics of any observable, e.g.

$$\langle z\log(z)\rangle_E \approx 78.4 \quad \text{and} \quad \langle z\log(z)\rangle_T \approx 76.0, \tag{4.18a,b}$$

where we used $z\log(z) \to 0$ as $z \to 0$. By symmetry, one expects

$$\langle x\rangle = \langle y\rangle = \langle xz\rangle = \langle yz\rangle = \langle yyy\rangle = \langle xxy\rangle = \langle xyy\rangle = \langle xzz\rangle = \langle yzz\rangle = 0 \tag{4.19}$$

but finite sampling effects prevent this from happening. As done in Appendix A, incorporating the symmetries allows ensemble calculations to achieve this to machine precision.

In addition to containing information about steady-state distributions, the generator Q provides temporal information: autocorrelations and the average holding time within a given cell. We show the autocorrelation of six observables,

$$g^{[1]}(\mathscr{A}) = x, \quad g^{[2]}(\mathscr{A}) = y, \quad g^{[3]}(\mathscr{A}) = z, \quad g^{[4]}(\mathscr{A}) = \begin{cases} 1 & \text{if } \mathcal{C}(\mathscr{A}) = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{4.20a–d}$$

$$g^{[5]}(\mathscr{A}) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad g^{[6]}(\mathscr{A}) = \begin{cases} 1 & \text{if } \mathcal{C}(\mathscr{A}) = 2 \\ 0 & \text{otherwise} \end{cases} \tag{4.21a,b}$$

in figure 4, which are calculated via (2.36) and (2.44), with appropriate modifications accounting for means and normalizing the height to one. Here, we see both the success and limitations of the method at capturing autocorrelations. In general, the decorrelation

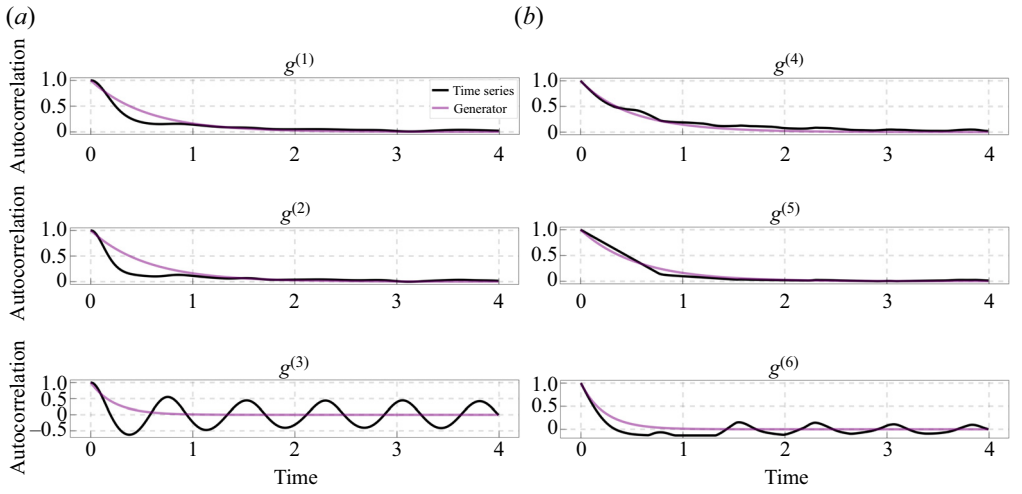


Figure 4. Lorenz autocorrelations generator versus time series. Six autocorrelations of observables are shown. The transparent purple line is calculated from the generator, and the black line is calculated from the time series. Even a coarse partition can capture observables $g^{[4]}$ and $g^{[5]}$ but struggles with oscillatory correlations.

of an observable is captured by the Markov model if it is approximately constant within a given cell, e.g. the observables $g^{[4]}$ and $g^{[5]}$. However, sometimes it is possible to do 'well', such as $g^{[1]}$ or $g^{[2]}$, despite not being constant within a region.

The inability to capture the autocorrelation of $g^{[6]}$, which is constant within $\mathcal{M}_2$, is partially due to the holding time distribution being far from exponentially distributed. To see this mode of failure, we plot the holding time distribution of the cells in figure 5. We show several binning strategies of the distribution to demonstrate the ability of an exponential distribution to capture quantiles of the empirical holding time distribution.

Depending on the time scale of interest, the $\mathcal{M}_1$ and $\mathcal{M}_3$ cells are approximately exponentially distributed, although they become fractal-like in terms of the distribution of holding times. In contrast, the holding time distribution of cell $\mathcal{M}_2$ is far from exponentially distributed upon refining the bins of the histogram. This calls into question using a Markovian, i.e. 'memoryless' model. There is an inherent assumption in the construction of the generator that transition probabilities are independent of the amount of time spent in a particular subset of state space. A better statistical model would incorporate exit probabilities conditioned on the time spent in a cell. Stated differently, memory is necessary to correctly reproduce the autocorrelation of a coarse discretization, see Lin *et al.* (2023); however, the approach taken here is to view a given discrete representation as inherently imperfect and subject to improvement upon refinement of a partition.

Figure 6 summarizes the resulting statistical dynamics, where the generator and transition probabilities define a graph structure. The graph structure contains information about the topological connectivity between different regions of state space and the 'strength' of connectivity over different time scales as encapsulated by the transition probabilities.

In the next section we consider two different partitioning strategies of the Lorenz equations, over the same dataset.

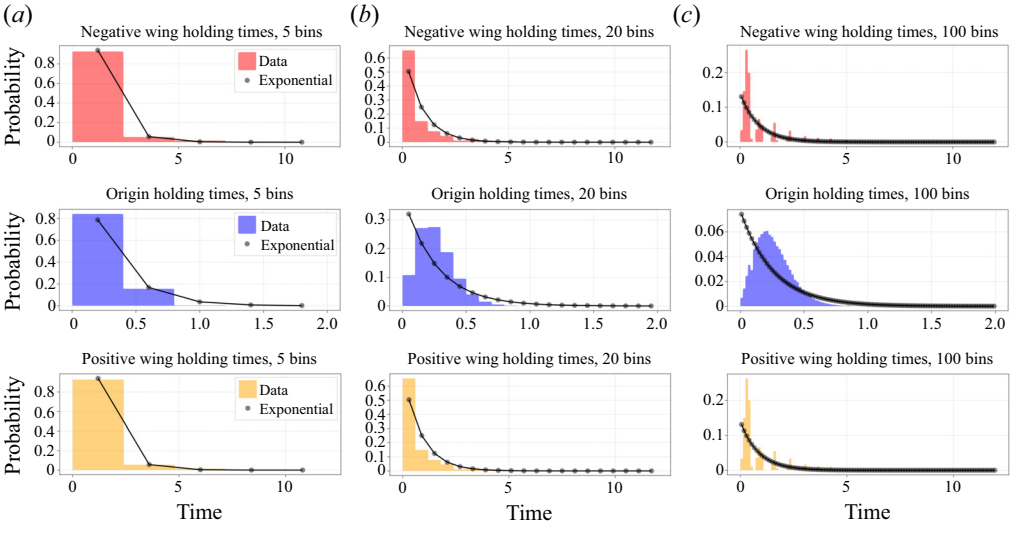


Figure 5. Lorenz fixed point partition holding times. An underlying assumption of using a generator for a given partitioning strategy is that the time spent in a cell is exponentially distributed. Here, we examine quantiles of the holding time distribution for a cell as given by the different binning numbers. The black dots correspond to the equivalent exponential distribution quantile, where the generator gives the rate parameter.

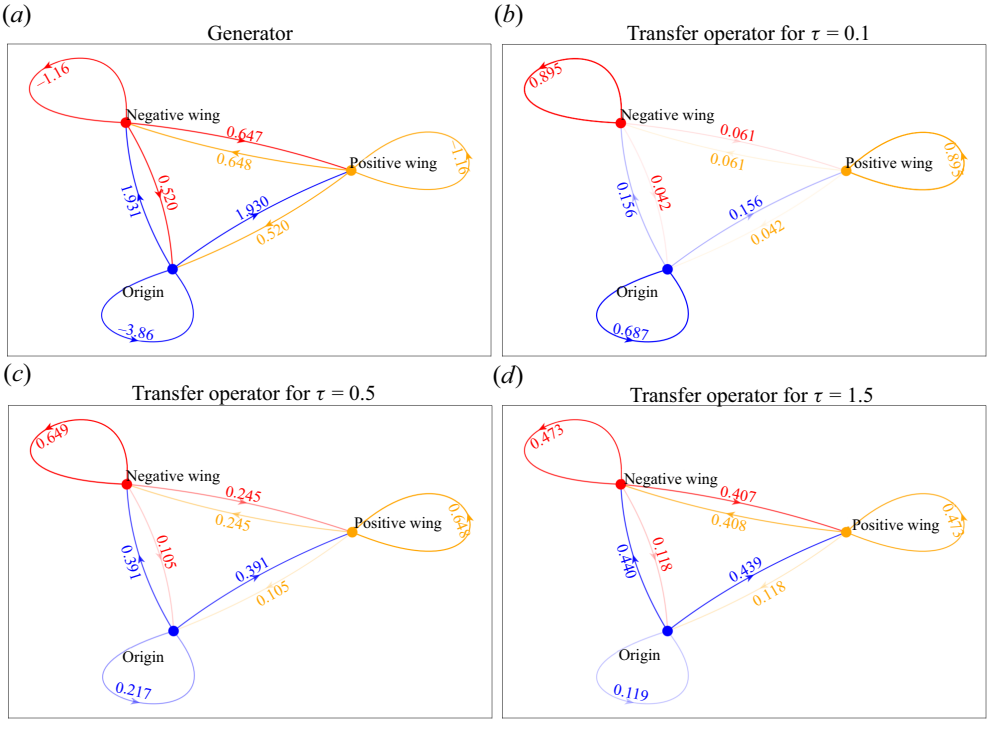


Figure 6. Lorenz fixed point partition graph. The generator (*a*) and transition probabilities over several time scales are visualized as a graph. The transition probabilities change depending on the time scale.

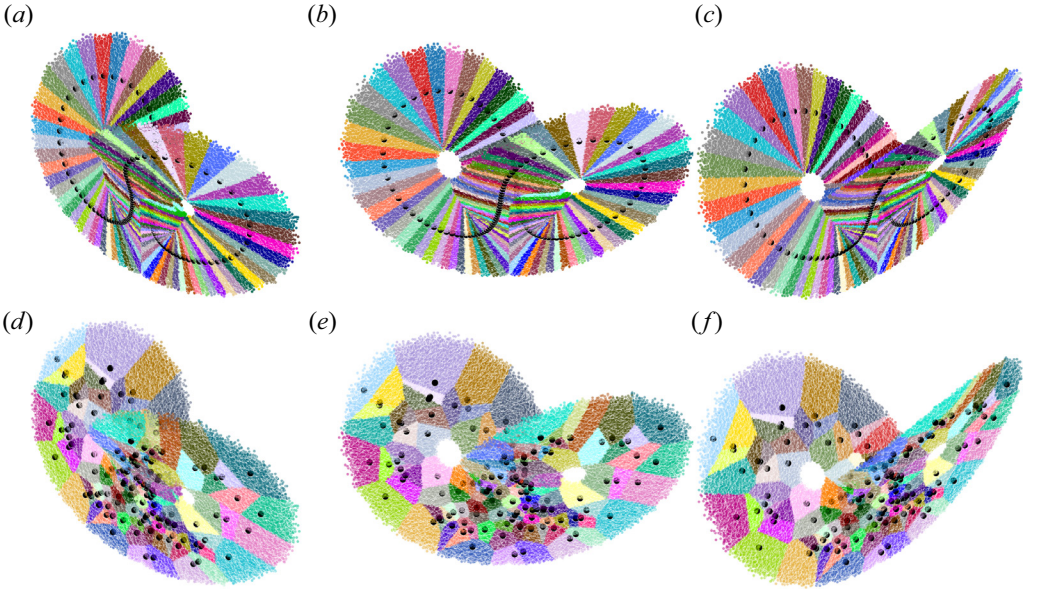


Figure 7. Lorenz AB and sampling partitions. In (*a–c*), we show the partition of the Lorenz attractor according to 128 points on the AB periodic orbit of Lorenz, and in (*d–f*), we show the sampling partition associated with 128 randomly chosen points. The black dots are the Markov states (cell centres) associated with the cells of each partition.

4.2. *The AB and sampling partitions*

We now show how the partition choice affects steady-state statistics and temporal autocorrelations. We choose two partitions, both using 128 cells. The first partition uses 128 evenly spaced (in time) points on the AB periodic orbit of the Lorenz equation, see Viswanath (2003), to define the cells of the partition, and our second partition uses 128 points sampled on the attractor. We call the former the 'AB partition' and the latter the 'sampling partition'. Both methods define cells by first labelling the cell centres with integers within 1 to 128, computing the distance to all 128 points, and then assigning a cell by picking out the cell centre with the smallest distance, similar to what was done for the three fixed points in (4.8). More succinctly, the points to define a Voronoi tessellation of the domain using the same time series from § 4.1. In both cases, we use an uninformative prior to construct the generator. See figure 7 for a visualization of the two partitioning strategies.

The AB periodic yields partitions that are thin wedges. In this case, further refinement by using more points on the periodic orbit does not yield a refinement strategy that converges to the statistics of the Lorenz equations. For example, the maximum value of z can never be approximated by points on the AB periodic orbit. On the other hand, the sampling partition yields a partition with disproportionate cell sizes and clusters in regions of high probability on the attractor. We expect that further refinement by choosing more random points yields a higher fidelity partition, but have no proof on this matter. This latter strategy is readily employable on any data set and can be thought of as a 'go-to' strategy in the absence of system knowledge.

We show two statistical measures to assess the quality of the partition. The first is in table 2. We see that the statistics of the AB partition are farther from the attractor for many of the variables than the sampling partition. This discrepancy is because the AB

	$\langle x\rangle$	$\langle y\rangle$	$\langle z\rangle$	$\langle xx\rangle$	$\langle xy\rangle$	$\langle xz\rangle$	$\langle yy\rangle$	$\langle yz\rangle$	$\langle zz\rangle$
AB	0.0	0.0	23.6	65.9	66.7	−0.1	91.4	−0.1	638.7
Sampling	0.0	0.1	23.5	61.7	62.3	1.3	79.7	−2.8	621.8
Temporal	−0.0	−0.0	23.5	62.8	62.8	−0.2	81.2	−0.2	628.9
	$\langle xxy\rangle$	$\langle xxz\rangle$	$\langle xyy\rangle$	$\langle xyz\rangle$	$\langle xzz\rangle$	$\langle yyy\rangle$	$\langle yyz\rangle$	$\langle yzz\rangle$	$\langle zzz\rangle$
AB	−0.9	2007.0	−1.2	1773.7	−4.5	−1.7	2215.2	−3.6	19203.6
Sampling	−3.6	1831.8	−4.3	1673.6	69.0	−9.2	1981.2	−68.8	18018.4
Temporal	−0.4	1879.7	−0.4	1677.2	−4.1	−0.4	1997.2	−4.2	18446.3

Table 2. Empirical moments of the Lorenz attractor. A comparison between ensemble averaging with two different partitions, 'AB' and 'sampling', and time averaging.

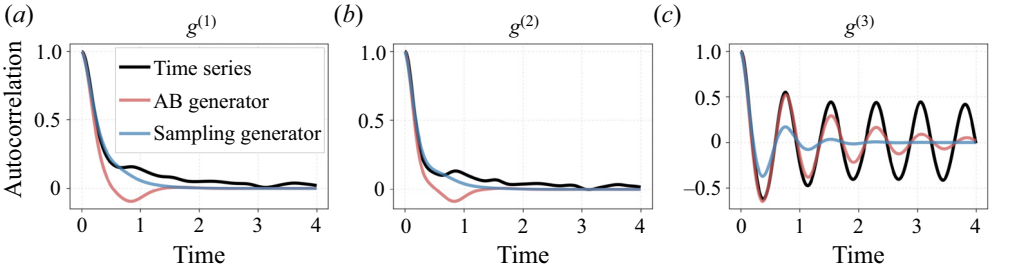


Figure 8. Autocorrelation for observables of the Lorenz system. The AB partition (red) captures oscillatory behaviour, while the sampling partition (blue) better captures the decay of the temporally obtained autocorrelation (black).

partition uses cell centres that are all on the periodic orbit and thus will represent a compromise between the periodic orbit statistics and the attractor statistics. In figure 8, we show the autocorrelations using the two partitions where the observables are defined in (4.20*a–d*). We see that the AB periodic orbit partition better captures oscillatory behaviour at the expense of the decay behaviour in the $g^{[1]}$ and $g^{[2]}$ autocorrelations. The $g^{[3]}$ autocorrelations, exhibiting less decay than its $g^{[1]}$ and $g^{[2]}$ counterparts are better captured by the periodic orbit partition. The oscillatory behaviour in $g^{[3]}$ is also captured by the sampling partition, albeit with more dissipation. We expect that using more points on the attractor for the sampling partition would yield better correspondence with the attractor. See Giorgini, Souza & Schmid (2023) for a data-driven method that uses the methods developed here as a baseline for improving convergence of autocorrelations in coarse settings.

5. Conclusion

In summary, we have done three things:

(i) § 2 – reviewed a theoretical formulation for transforming a dynamical system into a continuous time Markov process with finite state space;
(ii) § 3 – developed a Bayesian stream-based data-driven algorithm for constructing the generator of a continuous-time Markov process with finite state space;
(iii) § 4 – applied the methodology to the Lorenz equations.

We have seen that statistics are captured, even with a coarse discretization. In the Lorenz case, we used the fixed points of the dynamical system as the Markov states and the anchors for the partitioning strategy. Three states sufficed to capture the first and second moments. Furthermore, even some autocorrelations and residency times were well-captured with the coarse discretization, depending on the time scale of interest. Furthermore, we explored two further partitioning strategies (using a periodic orbit and random points of the attractor) and reported their resulting emulated statistics.

Future extensions of the present work include a detailed examination of convergence properties by varying the number of states and choosing different classifiers. When the number of cells becomes large, calculating the minimum distance of a state to a cell centre is computationally expensive. In such a case, using a tree structure for the classifier (e.g. a hierarchical k-means) is one option. Similar computationally expedient extensions include the use of a tensor product basis such as Junge & Koltai (2009) or box-refinement strategies used in Dellnitz & Junge (1999) and Dellnitz *et al.* (2001). Another option is to borrow ideas from thermodynamics and divide state space into 'macrostates' and 'microstates'. For example, we first divide a cell into energy shells and then only compute distances within each energy shell. A noteworthy example of a physics-based partition is found in Jiménez (2023), where the author constructed a Perron–Frobenius operator for wall-bounded flows.

The present work suggests the feasibility of extending this approach to more complex and high-dimensional systems. The only necessary step is to define a method for classifying states. Part 2 of this series focuses on applying the methodology to the compressible Euler equations with gravity and rotation (the model in Souza *et al.* (2023*a*)), a high-dimensional dynamical system exhibiting geophysical turbulence. We will detail the particular choices for the classifier and Markov states as applied to the Euler equations. The investigation will include an analysis of different strategies for partitioning the high-dimensional state space, a feature of the methodology that lends itself to adaptation to various dynamical systems. Ultimately, the hope is that the present method enables insights into high-dimensional spaces.

Acknowledgements. The author would like to thank G. Geogdzhayeva, L. Giorgini, F. Falasca, S. Silvestri, P. Hassanzadeh, P. Schmid, G. Froyland, G. Flierl, R. Ferrari and P. Cvitanović for many fruitful discussions. The author would also like to thank the 2022 Geophysical Fluid Dynamics Program, where much of this work was completed, and two anonymous reviewers for their feedback, which greatly improved the manuscript.

Funding. This work acknowledges support by Schmidt Sciences, LLC, through the Bringing Computation to the Climate Challenge, an MIT Climate Grand Challenge Project. The Geophysical Fluid Dynamics Program is supported by the National Science Foundation, USA, and the Office of Naval Research, USA.

Declaration of interests. The author reports no conflict of interest.

Author ORCIDs.
Andre N. Souza https://orcid.org/0000-0001-8025-3558.

Appendix A. Symmetries

In § 4.1, the symmetries of the Lorenz equations were not incorporated directly into the generator. We rectify this deficiency here and outline a method for incorporating symmetries. The Lorenz equations are invariant with respect to the transformation $(x, y, z) \mapsto (-x, -y, z)$. In so far as one chaotic attractor exists, this symmetry is expected to apply to chaotic trajectories. To incorporate this symmetry, we take two steps.

The first step is to verify that the Markov states also satisfy this symmetry. Since $\boldsymbol{\sigma}^{[1]} \mapsto \boldsymbol{\sigma}^{[3]}$, $\boldsymbol{\sigma}^{[3]} \mapsto \boldsymbol{\sigma}^{[1]}$ and $\boldsymbol{\sigma}^{[2]} \mapsto \boldsymbol{\sigma}^{[2]}$ under the symmetry operation, the Markov

	$\langle x \rangle$	$\langle y \rangle$	$\langle z \rangle$	$\langle xx \rangle$	$\langle xy \rangle$	$\langle xz \rangle$	$\langle yy \rangle$	$\langle yz \rangle$	$\langle zz \rangle$
Ensemble	−0.0	−0.0	23.8	63.5	63.5	−0.0	63.5	−0.0	642.4
	$\langle xxy \rangle$	$\langle xxz \rangle$	$\langle xyy \rangle$	$\langle xyz \rangle$	$\langle xzz \rangle$	$\langle yyy \rangle$	$\langle yyz \rangle$	$\langle yzz \rangle$	$\langle zzz \rangle$
Ensemble	0.0	1713.2	0.0	1713.2	0.0	0.0	1713.2	0.0	17346.1

Table 3. Empirical moments of the Lorenz attractor. A comparison between ensemble averaging and time averaging.

states, defined by the fixed points of the Lorenz equations, incorporate the symmetry. Generally, one must apply the symmetry operator to each Markov state and include the 'symmetry states' as necessary.

The second step is to incorporate symmetries into the resulting partition dynamics. For example, in the case of the Lorenz equations, if we observe the sequence

$$\text{first sequence} = 1, 1, 1, 2, 2, 3, 3, 1, 1. \tag{A1}$$

Then, applying the symmetry operation to the above sequence yields

$$\text{second sequence} = 3, 3, 3, 2, 2, 1, 1, 3, 3. \tag{A2}$$

We then apply the Bayesian matrix construction on the first sequence and calculate the posterior distributions. We then use these posterior distributions as the new prior for a Bayesian matrix construction for the second sequence. Doing so yields a matrix that incorporates symmetry through data augmentation. Note, one can apply the symmetry operation to the underlying state time series first, and then apply the classifier to generate a new sequence of integers.

We show the expected values of the generator under this symmetry augmentation in table 3. We see that the expected values of quantities that should be zero are now zero.

Similar considerations apply to other types of symmetries. For example, continuous symmetries can approximated as discrete symmetries, which can then use the methodology here; and detailed balance can be satisfied through data augmentation, e.g. reverse the partition dynamics and apply the Bayesian update procedure.

Appendix B. Matrix decomposition into reversible and irreversible dynamics

To further understand the time scales associated with the generator Q, we decompose the matrix into a negative semidefinite component and a component with purely imaginary eigenvalues. First, we assume that the generator Q is ergodic so that it has one zero eigenvalue, and all other eigenvalues have strictly negative real parts.

Let $\mathbb{P} = [\mathbb{P}(\mathcal{M}_1), \ldots, \mathbb{P}(\mathcal{M}_n)]$ be the normalized eigenvector corresponding to eigenvalue $\lambda = 0$, where we take the normalization to be

$$\mathbf{1}^T \mathbb{P} = 1, \tag{B1}$$

where $\mathbf{1}$ is the vector of all ones. We further assume that all entries of the vector $\mathbb{P}$ are strictly positive.

We split the matrix Q into a negative semidefinite and pure imaginary part as follows:

$$Q = \tfrac{1}{2}\underbrace{Q + PQ^TP^{-1}}_{\text{negative semidefinite}} + \tfrac{1}{2}\underbrace{Q - PQ^TP^{-1}}_{\text{imaginary eigenvalues}} \tag{B2}$$

where $P = \text{Diagonal}(\mathbb{P})$ is a diagonal matrix whose entries along the diagonal are the steady state distribution $\mathbb{P}$. The relation $P^{-1}\mathbb{P} = \mathbf{1}$ holds. The proof that the matrix $Q + PQ^TP^{-1}$ is negative semidefinite is as follows. First, recall that the autocovariance of an observable, $R_E(g,\tau)$, was defined in (2.44). We first need to show that $R_E(g,\tau) \leqslant R_E(g,0)$ for an arbitrary time $\tau > 0$. We introduce the notation $[\boldsymbol{g}]_n = g_n$, $[\exp(Q\tau)]_{mn} = w_{mn}$, and $[\mathbb{P}]_n = w_n$. We rewrite $R_E(g,\tau)$ and then use the Cauchy–Schwarz inequality,

$$\sum_{mn} g_n w_n g_m w_{mn} = \sum_{mn}\left(g_n\sqrt{w_n w_{mn}}\right)\left(g_m\sqrt{w_n w_{mn}}\right) \tag{B3}$$

$$\leqslant \sqrt{\sum_{mn} g_n^2 w_n w_{mn}}\sqrt{\sum_{mn} g_m^2 w_n w_{mn}}. \tag{B4}$$

The two terms in the square root are both individually $R_E(g,0)$ since

$$\sum_{mn} g_n^2 w_n w_{mn} = \sum_n g_n^2 w_n \sum_m w_{mn} = \sum_n g_n^2 w_n = R_E(g,0) \tag{B5}$$

and

$$\sum_{mn} g_m^2 w_n w_{mn} = \sum_m g_m^2 \sum_n w_{mn} w_n = \sum_m g_m^2 w_m = R_E(g,0). \tag{B6}$$

In the first line, we used the fact that the column sum of the operator is one, and in the second line, we used that w_n is an eigenvector with eigenvalue 1 of the w_{mn} matrix. More intuitively, the relation $R_E(g,\tau) \leqslant R_E(g,0)$ states that 'observables decorrelate in time'.

Then,

$$R_T(g,\mathrm{d}t) = \boldsymbol{g}^T\exp(Q\,\mathrm{d}t)P\boldsymbol{g} \approx \boldsymbol{g}^TP\boldsymbol{g} + \mathrm{d}t\boldsymbol{g}^TQP\boldsymbol{g} \leq \boldsymbol{g}^TP\boldsymbol{g} \Rightarrow \boldsymbol{g}^TQP\boldsymbol{g} \leqslant 0. \tag{B7}$$

Since QP is negative semidefinite and we can rescale $\boldsymbol{g}$ as $\boldsymbol{h} = P^{-1/2}\boldsymbol{g}$, the symmetric part of the matrix $\tilde{Q} = P^{-1/2}QP^{1/2}$ is negative semidefinite. Noting the similarity transformations

$$Q + PQ^TP^{-1} = P^{1/2}\left[P^{-1/2}QP^{1/2} + (P^{-1/2}QP^{1/2})^T\right]P^{-1/2}, \tag{B8}$$

$$Q - PQ^TP^{-1} = P^{1/2}\left[P^{-1/2}QP^{1/2} - (P^{-1/2}QP^{1/2})^T\right]P^{-1/2} \tag{B9}$$

completes the proof since similar matrices have equivalent eigenvalues.

The matrix $S = (Q + PQ^TP^{-1})/2$ and its matrix exponential $\exp(S\tau)$ satisfies detailed balance since

$$SP = PS^T \Leftrightarrow \exp(S\tau)P = P\exp(S\tau)^T. \tag{B10}$$

A way to directly show the latter is to use (B8) by first enacting the transformation $S = P^{1/2}\tilde{S}P^{-1/2}$ where $\tilde{S}$ is symmetric and then noting that the matrix exponential of a symmetric matrix is symmetric. This matrix has been commented on before by Froyland (2005) when constructing a Perron–Frobenius operator for a time-reversible Markov chain. The antisymmetric matrix is viewed as the appropriate model for a Hamiltonian system since its matrix exponential yields a unitary operator.

For a continuous time process with finite state space the equation for the evolution of probability densities is given by

$$\dot{p}_f = Qp_f \tag{B11}$$

and for the reverse time process (see Anderson (1982) for the continuous state space analogue), the evolution of probability densities is governed by

$$\dot{p}_r = -PQ^TP^{-1}p_r, \tag{B12}$$

where the above equation must be evolved backwards in time. The purpose of the second equation is to determine 'where I was, given that I know where I am'. Stated differently, the matrix

$$PQ^TP^{-1} = \tfrac{1}{2}\left(Q + PQ^TP^{-1}\right) - \tfrac{1}{2}\left(Q - PQ^TP^{-1}\right) \tag{B13}$$

is the generator that one would get if the partition dynamics time series is reversed. Both Q and PQ^TP^{-1} have the same steady state distribution $\mathbb{P}$.

Appendix C. Algebraic interpretation and connection to dynamic mode decomposition

In the main text we have specified a classifier $\mathcal{C}$ that maps state vectors to integers. Now consider the case where we have a data matrix whose columns are the state vector $\boldsymbol{\mathcal{s}}_n$ at time $t = t_n$ where we shall assume that the times are evenly spaced. Given m snapshots in time so that the data matrix $D \in \mathbb{R}^{n\times m}$, we assume the extreme case where every state is mapped to a different integer, e.g. $\mathcal{C}(\boldsymbol{\mathcal{s}}_n) = n$, so that the sequence of integers becomes

$$\text{state dynamics} = [1, 2, 3, 4, 5, \ldots, m]. \tag{C1}$$

In this case the generator becomes

$$Q = \frac{1}{\Delta t}\begin{bmatrix} -1 & 0 & 0 & 0 & \ldots \\ 1 & -1 & 0 & 0 & \ldots \\ 0 & 1 & -1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{C2}$$

where $Q \in \mathbb{R}^{m\times m}$. The last column is ambiguous since we have never observed a transition from the state, but we can simply take the last column to be the zero vector. An alternative is to artificially add the transition from $m \to 1$ so that the last column has a first entry as $1/\Delta t$ and the last entry as $-1/\Delta t$. The 'last column' issue becomes negligible in the large data limit.

Consider the following example where we have

$$\text{state dynamics} = 1, 2, 3, 4, 5 \quad \text{and} \quad Q = \frac{1}{\Delta t}\begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{C3a,b}$$

Now consider grouping together states according to

$$C^1 = (1, 4), \quad C^2 = (2, 5) \quad \text{and} \quad C^3 = (3) \tag{C4a–c}$$

to transform into the partition dynamics

$$\text{coarse partition dynamics} = C^1, C^2, C^3, C^1, C^2. \tag{C5}$$

For simplicity we will drop the 'C' and simply write

$$\text{partition dynamics} = 1, 2, 3, 1, 2. \tag{C6}$$

We can account for the partitioning of states algebraically through the introduction of a new matrix C. The rows of C are associated with a partition and the columns are Boolean values that assign each time column to a partition. For our example, the matrix $C \in \mathbb{R}^{3\times 5}$ is

$$C = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}. \tag{C7}$$

Upon choosing the pseudoinverse

$$C^{+} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{C8}$$

the reduced matrix

$$\tilde{Q} = CQC^{+} = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \tag{C9}$$

agrees with the data-driven construction from § 3.1 using the partition dynamics equation (C6).

In connection to dynamic mode decomposition, we comment that if we take the first $m-1$ columns of E and denote it by X and the last $m-1$ columns of C and denote it by Y, then the Perron–Frobenius operator is given by

$$\text{Perron–Frobenius} = YX^{+}, \tag{C10}$$

where X^{+} is the Moore–Penrose inverse of X. In the common parlance of the Koopman literature, for example Colbrook (2023), we take our nonlinear dictionary to be an indicator function for partitions, e.g. that is $\Psi_n(\mathfrak{z}) = 1$ if $\mathfrak{z} \in \mathcal{M}_n$ and 0 otherwise, i.e. a one-hot encoding of the partition label. See Klus *et al.* (2016) for a similar discussion on connecting the Perron–Frobenius operator to the Koopman operator.

In general introducing a matrix E and pseudoinverse C^{+} will not agree to within machine precision with the data-driven construction from § 3.1 and will differ in the column associated with the last partition in the partition dynamics sequence; however, the matrix method presented here suggests an extension of the methodology. The columns of the C matrix can instead be replaced by values that sum up to one. If in addition we require the entries to be positive we can interpret the columns as probabilistic classifications. Furthermore, the original Q matrix could be approximated with higher-order difference formulae. The choice of which pseudoinverse to use offers additional flexibility.

Appendix D. Simple harmonic oscillator analysis

Consider the system

$$\dot{x} = -y \quad \text{and} \quad \dot{y} = x, \tag{D1a,b}$$

and take the initial condition to be $x(t=0) = 1$ and $y(t=0) = 0$. The solution is

$$x(t) = \cos(t) \quad \text{and} \quad y(t) = \sin(t), \tag{D2a,b}$$

and the analytic generator for this trajectory is

$$\partial_t \mathcal{P} = -\partial_s (U\mathcal{P}), \tag{D3}$$

where $s \in [0, 2\pi)$ is the angle, the velocity $U = 1$, and the domain for the partial differential equation is periodic. The eigenvalues and eigenvectors for the right-hand side are $(\lambda)_k = ik$ for $k \in \mathbb{Z}$ and $v_k = \exp(\lambda_k s)$.

Now assume that the data matrix is sampled with frequency

$$\Delta t = \frac{2\pi}{M}, \tag{D4}$$

where M is a fixed natural number, i.e. $M \in \mathbb{N}$, and that the columns of the data matrix go from represent the solution from $t \in [0, 2\pi]$ so that discretely the entries of the data matrix X are

$$t_j = \frac{2\pi}{M}(j-1), \quad X_{1j} = \cos(t_j) \quad \text{and} \quad X_{2j} = \sin(t_j) \tag{D5a–c}$$

for $j = 1, \ldots, M+1$. First, consider partitioning the system into $N = M$ states where the classifier is given by which sector of the circle the state is in, for example. We take the classifier to be

$$\mathcal{C}(x, y) = \left[\text{floor}\left(\frac{\arctan(x, y)}{2\pi} N \right) \right] \%N + 1, \tag{D6}$$

where the floor function converts rounds down a real number to the closest integer and $\arctan(x, y)$ is a (non-standard) two-argument arctangent function whose range is $[0, 2\pi)$, e.g. $\arctan(1, 0) = 0$, $\arctan(1, 1) = \pi/4$, $\arctan(-1, -1) = \pi/4 + \pi$ and $\arctan(X_{1j}, X_{2j}) = t_j \mod 2\pi$, see figure 9. We now show that the Perron–Frobenius operator corresponding to time step Δt is consistent with the generator construction. Based on the data matrix the entries Perron–Frobenius operator $\mathcal{F}^{[\Delta t]}$ and generator Q would be given by

$$[\mathcal{F}^{[\Delta t]}]_{ij} = \delta_{i(j+1)\%N} \quad \text{and} \quad Q_{ij} = \frac{1}{\Delta t}\left(-\delta_{ij} + \delta_{i(j+1)\%N} \right) \tag{D7a,b}$$

for example; for $M = 4$ the matrices are

$$\left.\begin{aligned} X = \begin{bmatrix} 1 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix}, \quad \mathcal{F}^{[\Delta t]} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \\ Q = \tfrac{1}{\Delta t} \begin{bmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \end{aligned}\right\}. \tag{D8a–c}$$

The Perron–Frobenius operator $\mathcal{F}^{[\Delta t]}$ is exact for this time scale. The identity $\mathcal{F}^{[\Delta t]} = \mathbb{I} + \Delta t Q$ holds for all N. Applying the classifier to X yields the partition dynamics 1, 2, 3, 4, 1.

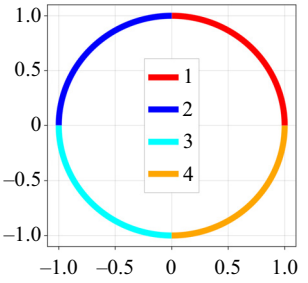


Figure 9. State space partition of the simple harmonic oscillator. Here, we partition the state space of the simple harmonic oscillator into four segments.

Both $\mathcal{F}^{[\Delta t]}$ and Q are circulant matrices, so the columns of the discrete Fourier transform operator are the eigenvectors. The eigenvalues of $\mathcal{F}^{[\Delta t]}$ and Q are then given by

$$(\lambda_{\mathcal{F}})_k = \exp(\mathrm{i}\Delta t k) \quad \text{and} \quad (\lambda_Q)_k = \frac{1}{\Delta t}\left(\exp(\mathrm{i}\Delta t k) - 1\right), \tag{D9a,b}$$

where $k \in \{0, \ldots, N-1\}$, respectively. The eigenvalues and eigenvectors of the continuous state-space Perron–Frobenius operator, formally the matrix exponential of (D3), e.g. $\exp(-\partial_x U\cdot)$, are aliased to the finite set of values given by $(\lambda_{\mathcal{F}})_k$ due to the time scale Δt of observation. The operator norm of the difference $\|\mathcal{F}^{[\Delta t]} - \exp(Q\Delta t)\|_2$ is bounded below by $1 + \exp(-2)$ for all even $N \geqslant 4$ (as can be seen by choosing $k = N/2 \Rightarrow \exp(\mathrm{i}\Delta t k) = -1$). Thus, the two operators do not converge in norm to one another; however, there is a sense in which the two operators converge to similar answers, as we will see shortly.

To compare the two eigenvalues, we take the logarithm of $\lambda_{\mathcal{F}}$ and divide by Δt. Furthermore, we fix k and consider the limit as $N \to \infty$, i.e. the number of partitions goes to infinity and $\Delta t \to 0$. In this limit, we have

$$(\lambda_Q)_k - \frac{\ln[(\lambda_{\mathcal{F}})_k]}{\Delta t} = -\Delta t k^2/2 - \mathrm{i}\Delta t^2 k^3/6 + O(\Delta t^3 k^4). \tag{D10}$$

Thus, we see that as $\Delta t \to 0$, for a fixed k, the two constructions agree with one another; however, to first order, the generator construction induces an extra dissipation of order $-\Delta t k^2$. We obtain first-order convergence to the real part of the eigenvalue and second-order convergence to the imaginary part.

To see where this dissipation comes from, we make use of the decomposition from Appendix B to split the matrix Q into reversible and irreversible dynamics

$$Q_{ij} = \frac{1}{\Delta t}\left(-\delta_{ij} + \delta_{i(j+1)\%N}\right) \tag{D11}$$

$$= \frac{1}{2\Delta t}\left(-2\delta_{ij} + \delta_{i(j+1)\%N} + \delta_{i(j-1)\%N}\right) + \frac{1}{2\Delta t}\left(\delta_{i(j+1)\%N} - \delta_{i(j-1)\%N}\right); \tag{D12}$$

e.g. for $M = 4$

$$Q = \frac{1}{\Delta t}\begin{bmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{D13}$$

$$= \frac{1}{2\Delta t}\begin{bmatrix} -2 & 1 & 0 & 1 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 1 & -2 \end{bmatrix} + \frac{1}{2\Delta t}\begin{bmatrix} 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \end{bmatrix}. \tag{D14}$$

The first matrix is the discrete Laplacian on a periodic grid, and the latter is the central difference operator on a grid. Both matrices are circulant matrices and, therefore, commute with one another. (The matrix logarithm of $\mathcal{F}^{[\Delta t]}$ divided by Δt is a Fourier spectral differentiation matrix, see Trefethen (2000).) The Q matrix is precisely the form of a finite-volume upwinding approximation to the advection operator on a periodic grid. Furthermore, their eigenvalues are given by the real and imaginary parts of the eigenvalues of Q, and the contribution to the dissipative part of the spectrum comes purely from the discrete Laplacian part of the decomposition.

Concretely, we use that $\Delta s = U\Delta t$, where here the velocity is $U = 1$, Δs is the size of the state-space volume, and then multiply the discrete Laplacian by $1 = \Delta s/\Delta s$ to obviate the grid-scale-dependent diffusivity constant of $U\Delta s/2$ in front of the discrete Laplacian,

$$Q = \frac{U\Delta s}{2\Delta s^2}\begin{bmatrix} -2 & 1 & 0 & 1 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 1 & -2 \end{bmatrix} + \frac{U}{2\Delta s}\begin{bmatrix} 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \end{bmatrix} \tag{D15}$$

$$\approx \frac{U\Delta s}{2}\partial_s^2 - U\partial_s. \tag{D16}$$

Thus, the amount of dissipation automatically decreases upon cell refinement. In summary, the data-driven method of § 3 provides the implicit regularization common to upwinding schemes for this example. We expect that such an implicit regularization is generally true for other systems, and thus, the generator approximation will be overly dissipative.

Given that we know the simple harmonic oscillator is purely periodic, we can achieve higher-order convergence to the eigenvalues of $\mathcal{F}^{[\Delta t]}$ simply by removing the dissipative part of the spectrum. Doing so respects the algebraic structure of the continuous system.

We now describe another limit where the Perron–Frobenius operator and generator converge in norm to one another. We now consider the limit with a fixed number of partitions but finer Δt data resolution and infinite data. Here, we fix the number of partitions to N and examine the limit as $M = LN$ for $L \to \infty$. This refinement increases the temporal resolution of the data matrix for the same physical time. The classifier

$$\mathcal{C}(x, y) = \left[\text{floor}\left(\frac{\arctan(x, y)}{2\pi}N\right)\right] \%N + 1 \tag{D17}$$

divides the circle into N evenly spaced sectors, each with the same discrete number of points inside, see figure 9. A trajectory spends the same physical time $2\pi/N$ inside a

given partition. For all $L \geqslant 1$ the Q matrix is given by

$$Q_{ij} = \frac{N}{2\pi}\left(-\delta_{ij} + \delta_{i(j+1)\%N}\right) \tag{D18}$$

since the amount of time spent remains constant. We now examine the convergence of $\ln(\mathcal{F}^{[\Delta t]})/\Delta t$ where $\Delta t = 2\pi/(LN)$ and ln is the matrix logarithm, to the Q matrix. For $L = 1$, we have

$$[\mathcal{F}^{[\Delta t]}]_{ij} = \delta_{i(j+1)\%N} \tag{D19}$$

as before, and the general case is

$$[\mathcal{F}^{[\Delta t]}]_{ij} = \frac{L-1}{L}\delta_{ij} + \frac{1}{L}\delta_{i(j+1)\%N}. \tag{D20}$$

The intuition behind the above formula is that we observe $L-1$ transitions within a sector and 1 transitions out of a sector of the circle. We observe that the identity

$$(\mathcal{F}^{[\Delta t]} - \mathbb{I})/\Delta t = Q \tag{D21}$$

holds for all L and thus

$$\ln\left(\mathcal{F}^{[\Delta t]}\right) = \ln\left(\mathbb{I} + \Delta t Q\right) \approx \Delta t Q + O\left(\Delta t^2\right) \tag{D22}$$

as $\Delta t \to 0$. Hence $\ln(\mathcal{F}^{[\Delta t]})/\Delta t \to Q$ (in the operator norm) as $L \to \infty$ or equivalently $\Delta t \to 0$.

Note the difference between the two notions of convergence. In the first, we examine the infinite partition limit with the finest level of refinement, whereas the latter uses a finite partition but examines convergence to the Q matrix. We see two things in the simple harmonic oscillator case. First, the generator approximation is consistent with the Perron–Frobenius operator in the dual limit $\Delta t \to 0$ and $N \to \infty$. In the simple harmonic oscillator case, the Perron–Frobenius operator was exact (due to aliasing). Thus, all the errors are in the generator approximation, which is equivalent to a first-order upwinding scheme. In the fixed partition and $\Delta t \to 0$ limit, the generator approximation remains the same for all data resolutions for which a trajectory spends at least a Δt amount of time within a partition. In this case, the Perron–Frobenius operator converges to the generator approximation in the infinite temporal resolution limit. All cases were examined in the infinite data limit, e.g. $T \to \infty$. Thus, we always first took the $T \to \infty$ limit, followed by either the dual limit where $\Delta t \to 0$ and the number of states went to infinity $N \to \infty$ simultaneously, or kept the number of states N finite while refining Δt.

References

Albeverio, S. & Mazzucchi, S. 2016 A unified approach to infinite-dimensional integration. *Rev. Math. Phys.* **28** (02), 1650005.

Allawala, A. & Marston, J.B. 2016 Statistics of the stochastically forced Lorenz attractor by the Fokker–Planck equation and cumulant expansions. *Phys. Rev.* E **94**, 052218.

Anderson, B.D.O. 1982 Reverse-time diffusion equation models. *Stoch. Proc. Applics.* **12** (3), 313–326.

Araujo, V. 2023 On the number of ergodic physical/SRB measures of singular-hyperbolic attracting sets. *J. Differ. Equ.* **354**, 373–402.

Arbabi, H. & Mezić, I., 2017 Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. *SIAM J. Appl. Dyn. Syst.* **16** (4), 2096–2126.

Bittracher, A., Mollenhauer, M., Koltai, P. & Schütte, C. 2023 Optimal reaction coordinates: variational characterization and sparse computation. *Multiscale Model. Simul.* **21** (2), 449–488.

Blank, M. 2017 Ergodic averaging with and without invariant measures. *Nonlinearity* **30** (12), 4649.

COLBROOK, M.J. 2023 The mpedmd algorithm for data-driven computations of measure-preserving dynamical systems. *SIAM J. Numer. Anal.* **61** (3), 1585–1608.
COLBROOK, M.J. & TOWNSEND, A. 2023 Rigorous data-driven computation of spectral properties of Koopman operators for dynamical systems. *Commun. Pure Appl. Maths*. **77** (1), 221–283.
CONSTANTE-AMORES, C.R., LINOT, A.J. & GRAHAM, M.D. 2023 Enhancing predictive capabilities in data-driven dynamical modelling with automatic differentiation: Koopman and neural ode approaches. *Chaos*, **34** (4), 043119.
CORWIN, I. & SHEN, H. 2020 Some recent progress in singular stochastic partial differential equations. *Bull. Am. Math. Soc.* **57** (3), 409–454.
COWIESON, W. & YOUNG, L.-S. 2005 SRB measures as zero-noise limits. *Ergod. Theory Dyn. Sys.* **25** (4), 1115–1138.
CVITANOVIĆ, P., ARTUSO, R., MAINIERI, R., TANNER, G. & VATTAY, G. 2016 *Chaos: Classical and Quantum*. Niels Bohr Institute.
CVITANOVIĆ, P. 2013 Recurrent flows: the clockwork behind turbulence. *J. Fluid Mech.* **726**, 1–4.
DANIELL, P.J. 1919 Integrals in an infinite number of dimensions. *Ann. Maths* **20** (4), 281–288.
DAS, S., GIANNAKIS, D. & SLAWINSKA, J. 2021 Reproducing kernel Hilbert space compactification of unitary evolution groups. *Appl. Comput. Harmon. Anal.* **54**, 75–136.
DELLNITZ, M., FROYLAND, G. & JUNGE, O. 2001 The algorithms behind GAIO — set oriented numerical methods for dynamical systems. In *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* (ed. B. Fiedler), pp. 145–174. Springer.
DELLNITZ, M. & JUNGE, O. 1999 On the approximation of complicated dynamical behavior. *SIAM J. Numer. Anal.* **36** (2), 491–515.
DELLNITZ, M., JUNGE, O., KOON, W.S., LEKIEN, F., LO, M.W., MARSDEN, J.E., PADBERG, K., PREIS, R., ROSS, S.D. & THIERE, B. 2005 Transport in dynamical astronomy and multibody problems. *Intl J. Bifurcation Chaos* **15** (03), 699–727.
DEWITT, C.M. 1972 Feynman's path integral. *Commun. Math. Phys.* **28** (1), 47–67.
FERNEX, D., NOACK, B. & SEMAAN, R. 2021 Cluster-based network modeling—from snapshots to complex dynamical systems. *Sci. Adv.* **7**, eabf5006.
FROYLAND, G. 1997 Computer-assisted bounds for the rate of decay of correlations. *Commun. Math. Phys.* **189** (1), 237–257.
FROYLAND, G. 2005 Statistically optimal almost-invariant sets. *Physica* D **200** (3), 205–219.
FROYLAND, G., JUNGE, O. & KOLTAI, P. 2013 Estimating long-term behavior of flows without trajectory integration: the infinitesimal generator approach. *SIAM J. Numer. Anal.* **51** (1), 223–247.
GELMAN, A., CARLIN, J.B., STERN, H.S., DUNSON, D.B., VEHTARI, A. & RUBIN, D.B. 2013 *Bayesian Data Analysis*, 3rd edn. Chapman and Hall/CRC.
GIANNAKIS, D. 2019 Data-driven spectral decomposition and forecasting of ergodic dynamical systems. *Appl. Comput. Harmon. Anal.* **47** (2), 338–396.
GIORGINI, L.T., DECK, K., BISCHOFF, T. & SOUZA, A. 2024 Response theory via generative score modeling. *Phys. Rev. Lett.* (submitted) arXiv:2402.01029.
GIORGINI, L.T., SOUZA, A.N. & SCHMID, P.J. 2023 Reduced Markovian models of dynamical systems. *Physica* D (submitted) arXiv:2308.10864.
HAGAN, P.S., DOERING, C.R. & LEVERMORE, C.D. 1989 Mean exit times for particles driven by weakly colored noise. *SIAM J. Appl. Maths* **49** (5), 1480–1513.
HAIRER, M. 2014 A theory of regularity structures. *Invent. Math.* **198** (2), 269–504.
HOPF, E. 1948 A mathematical example displaying features of turbulence. *Commun. Pure Appl. Maths* **1** (4), 303–322.
HOPF, E. 1952 Statistical hydromechanics and functional calculus. *Indiana Univ. Math. J.* **1**, 87–123.
JIMÉNEZ, J. 2023 A Perron–Frobenius analysis of wall-bounded turbulence. *J. Fluid Mech.* **968**, A10.
JUNGE, O. & KOLTAI, P. 2009 Discretization of the Frobenius–Perron operator using a sparse Haar tensor basis: the sparse Ulam method. *SIAM J. Numer. Anal.* **47** (5), 3464–3485.
KLUS, S., KOLTAI, P. & CHRISTOF, S. 2016 On the numerical approximation of the Perron–Frobenius and Koopman operator. *J. Comput. Dyn.* **3** (1), 51–79.
LIN, Y.T., TIAN, Y., PEREZ, D. & LIVESCU, D. 2023 Regression-based projection for learning Mori–Zwanzig operators. *SIAM J. Appl. Dyn. Syst.* **22** (4), 2890–2926.
LLOYD, S. 1982 Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28** (2), 129–137.
LORENZ, E.N. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.
MAĆEŠIĆ, S. & ČRNJARIĆ-ŽIC, N. 2020 *Koopman Operator Theory for Nonautonomous and Stochastic Systems*, pp. 131–160. Springer.

OTTO, S.E., PEITZ, S. & ROWLEY, C.W. 2023 Learning bilinear models of actuated Koopman generators from partially-observed trajectories. *SIAM J. Appl. Dyn. Syst.* **23** (1), 885–923.

PARKER, J.P. & PAGE, J. 2020 Koopman analysis of isolated fronts and solitons. *SIAM J. Appl. Dyn. Syst.* **19** (4), 2803–2828.

ROSENFELD, J.A., KAMALAPURKAR, R., GRUSS, L.F. & JOHNSON, T.T. 2021 Dynamic mode decomposition for continuous time systems with the Liouville operator. *J. Nonlinear Sci.* **32** (1), 5.

ROWLEY, C.W., MEZIĆ, I., BAGHERI, S., SCHLATTER, P. & HENNINGSON, D.S. 2009 Spectral analysis of nonlinear flows. *J. Fluid Mech.* **641**, 115–127.

SCHERER, M.K., TRENDELKAMP-SCHROER, B., PAUL, F., PÉREZ-HERNÁNDEZ, G., HOFFMANN, M., PLATTNER, N., WEHMEYER, C., PRINZ, J.-H. & NOÉ, F. 2015 Pyemma 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **11** (11), 5525–5542.

SCHMID, P.J. 2010 Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **656**, 5–28.

SCHÜTTE, C., KLUS, S. & HARTMANN, C. 2022 Overcoming the timescale barrier in molecular dynamics: transfer operators, variational principles, and machine learning. *Tech. Rep.* 22-25. ZIB.

SINGHAL, N. & PANDE, V.S. 2005 Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **123** (20), 204909.

SOUZA, A.N., *et al.* 2023*a* The flux-differencing discontinuous galerkin method applied to an idealized fully compressible nonhydrostatic dry atmosphere. *J. Adv. Model. Earth Syst.* **15** (4), e2022MS003527.

SOUZA, A.N., LUTZ, T. & FLIERL, G.R. 2023*b* Statistical non-locality of dynamically coherent structures. *J. Fluid Mech.* **966**, A44.

SOUZA, A.N. 2024 Representing turbulent statistics with partitions of state space. Part 2. The compressible Euler equations. *J. Fluid Mech.* **997**, A2.

TREFETHEN, L.N. 2000 *Spectral Methods in MATLAB*. Society for Industrial and Applied Mathematics.

TRENDELKAMP-SCHROER, B., WU, H., PAUL, F. & NOÉ, F. 2015 Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* **143** (17), 174101.

ULAM, S.M. 1964 *Problems in Modern Mathematics*. Dover.

VISWANATH, D. 2003 Symbolic dynamics and periodic orbits of the Lorenz attractor. *Nonlinearity* **16** (3), 1035.

WILLIAMS, M.O., KEVREKIDIS, I.G. & ROWLEY, C.W. 2015 A data–driven approximation of the Koopman operator: extending dynamic mode decomposition. *J. Nonlinear Sci.* **25** (6), 1307–1346.

YOUNG, L.-S. 2002 What are SRB measures, and which dynamical systems have them? *J. Stat. Phys.* **108** (5), 733–754.

ZHANG, X. & SHU, C.-W. 2011 Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. Lond.* A **467** (2134), 2752–2776.

ZINN-JUSTIN, J. 2021 *Quantum Field Theory and Critical Phenomena: Fifth Edition*. Oxford University Press.