J. Fluid Mech. (2024), *vol.* 997, A2, doi:10.1017/jfm.2024.657

Representing turbulent statistics with partitions of state space. Part 2. The compressible Euler equations

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 second part of a two-part paper. We apply the methodology of the first paper (Souza, *J. Fluid Mech.*, vol. 997, 2024, A1) to construct a data-driven finite-volume discretization of the Liouville/Fokker–Planck equation of a high-dimensional dynamical system, i.e. the compressible Euler equations with gravity and rotation evolved on a thin spherical shell. We show that the method recovers a subset of the statistical properties of the underlying system, steady-state distributions of observables and autocorrelations of particular observables, as well as revealing the global Koopman modes of the system. We employ two different strategies for the partitioning of a high-dimensional state space, and explore their consequences.

Key words: low-dimensional models, big data, atmospheric flows

1. Introduction

In Part 1 (Souza 2024) we introduced a data-driven methodology for discretizing the continuity equation associated with a dynamical system, and demonstrated the methodology on the Lorenz equations. The methodology is similar to other data-driven methods for constructing Koopman and Perron–Frobenius operators – see Ulam (1964), Rowley *et al.* (2009), Schmid (2010), Klus & Péter (2016) and Colbrook (2023) – but accounts for finite sampling effects and assumes that the underlying data come from a continuous time dynamical system.

In the language of Williams, Kevrekidis & Rowley (2015), we make a particular choice of nonlinear dictionary that allows for a scalable computation of the Koopman/Perron–Frobenius operator. Two ideas are combined for the nonlinear

† Email address for correspondence: andrenogueirasouza@gmail.com

dictionary: the use of indicator functions for particular regions of state space as in Ulam (1964), along with a 'classifier' for flow states that implicitly partitions state space. The latter allows for geometric flexibility that frees one from the use of the 'boxes' in Ulam's method and provides scalability in any number of dimensions. Using indicator functions for the nonlinear dictionary allows for scalability in temporal data volume since there is no need to construct a pseudo-inverse. The operator is constructed directly. This specialization enables the method to be used in a 'streaming' fashion, where the operator is built as a simulation progresses.

We apply the method to a high-dimensional dynamical system: the compressible Euler equations with rotation and gravity. We modify the set-up described in Held & Suarez (1994), a benchmark for atmospheric dynamical cores used in climate modelling. The mean and variance statistics are robust across many numerical discretizations and equation formulations, and mimic statistics of the Earth system. A flux-differencing discontinuous Galerkin method is employed for the numerical discretization, and total energy is the chosen prognostic thermodynamic variable. The details of the set-up and discretization are in Souza, Lutz & Flierl (2023*b*). The discretization is viewed as a 1 000 000+ degrees of freedom dynamical system. The perspective taken herein is akin to that of Cvitanović *et al.* (2016).

The purpose of the Held–Suarez setup is twofold: the first goal is to have a stringent test on the methodology of Part 1 through a high-dimensional dynamical system without any discernible 'meta-stable' states. The method of Part 1 is expected to perform best when there are few meta-stable states with semi-rare transitions between them, thus the lack of 'meta-stable' states of the present test case should be considered a 'worst-case' scenario; however, the lack of clearly distinguishable states is typical with regard to turbulence. The second goal is to make a connection between an operator-theoretic approach to dynamical systems and climate science, such as the work of Froyland *et al.* (2021) and the idealized system in Geogdzhayev, Souza & Ferrari (2024). In this way, questions in climate science are placed on a rigorous theoretical foundation. We return to this point at the end of the paper.

The paper proceeds as follows. Section 2 reviews the methodology outlined in Part 1. In § 3, we delve into an application to the Euler equations in the Held and Suarez setup. We discuss partitioning strategies, global Koopman modes, statistically steady-states, and temporal autocorrelations. Finally, in § 4, we discuss future work and implications.

2. Methodology review

In Part 1, we developed a general method for using trajectory information of a dynamical system given by

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

to discretize the Liouville equation

$$\partial_t \mathcal{P} + \boldsymbol{\nabla} \boldsymbol{\cdot} (\boldsymbol{U}\mathcal{P}) = 0. \tag{2.2}$$

Here, $\boldsymbol{s} : \mathbb{R} \to \mathbb{R}^d$ is the mapping from a time t to state $\boldsymbol{s}(t)$, $\boldsymbol{U} : \mathbb{R}^d \to \mathbb{R}^d$ is the evolution rule for the dynamics, and $\mathcal{P}(\boldsymbol{\mathscr{s}}, t) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is the joint probability density for each state variable $\boldsymbol{\mathscr{s}} \in \mathbb{R}^d$ as a function of time.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

The primary ingredient to construct a discretization is a ‘classifier’

$$\mathcal{C} : \mathbb{R}^d \to \{1, \ldots, N\}, \tag{2.3}$$

which maps an arbitrary state $\mathscr{s}$ to an integer. The classifier implicitly defines cells, i.e. a partition of state space, for the underlying dynamical system; however, since chaotic trajectories are (in this work) assumed to belong to an attracting subset $\mathcal{M} \subset \mathbb{R}^d$, the ‘cells’ $\mathcal{M}_n$ also serve as a method of partitioning $\mathcal{M}$.

Applying the classifier to a time trajectory given by (2.1),

$$e(t) \equiv \mathcal{C}(\boldsymbol{s}(t)), \tag{2.4}$$

generates sequences of integer jumps, which we interpret as a continuous Markov process. Under this interpretation, we can define

(i) holding times, i.e. the average amount of time spent in a partition of state space,
(ii) exit probabilities, i.e. the probability of transitioning from partition i to j,

to construct the generator of the process. This is an $N \times N$ matrix, which we denote by Q. In Part 1, we showed how to construct Q from data and quantify the uncertainty of the matrix entries due to finite sampling effects. The uncertainty quantification is consistent with the continuous-time Markov formulation. We represented uncertainty using conjugate priors so that the prior/posterior distribution for the holding times is a gamma distribution $\Gamma(\alpha, \beta)$, and the prior/posterior distribution for the exit probabilities is a Dirichlet distribution $D(\boldsymbol{\alpha})$.

Furthermore, we showed that the matrix Q structure is the same structure that one would get through a finite-volume discretization of the Liouville equation (2.2). The same data-driven construction applies to stochastic dynamical systems and thus the Fokker–Planck equation. We use the convention that the steady-state probabilities are right eigenvectors of the operator since we view Q as arising from a numerical discretization.

To perform statistical calculations, we additionally need ‘cell centres’ (also called ‘Markov states’) $\boldsymbol{\sigma}^{[n]} \in \mathbb{R}^d$ associated with each embedding such that

$$\mathcal{C}(\boldsymbol{\sigma}^{[n]}) = n. \tag{2.5}$$

Using the generator Q and the Markov states $\boldsymbol{\sigma}^{[n]}$, we showed how to reconstruct the statistics induced by the dynamics of (2.1) in Part 1. In particular, we showed how to construct statistically steady probability distributions, temporal autocorrelations and Koopman modes. We approximated the density in partition $\mathcal{M}_n$ by a delta function centred on the state, i.e. $\mathcal{I}_n(\mathscr{s}) \approx \delta(\mathscr{s} - \boldsymbol{\sigma}^{[n]})$, and approximated the action of the transfer operator – the operator exponential of the generator – $\mathcal{T}^{\tau}$ on a partition as a weighted sum of delta functions centred on states, e.g. $\mathcal{T}^{\tau}\mathcal{I}_n(\mathscr{s}) \approx \sum_m [\exp(Q\tau)]_{mn}\, \delta(\mathscr{s} - \boldsymbol{\sigma}^{[m]})$. The detailed formulas in the continuous and discrete settings are given in Part 1.

A Koopman eigenvector is an eigenvector of the Koopman operator (the adjoint to the Perron–Frobenius operator) and a functional that acts on a state vector. Thus the Koopman eigenvector serves as a mapping from a state vector to a complex number, i.e. $g_\lambda : \mathbb{R}^d \to \mathbb{C}$, where λ denotes the associated eigenvalue. The numerical approximation to a Koopman eigenvector, $\boldsymbol{g}_\lambda \in \mathbb{R}^N$, is given by a left eigenvector of the generator $Q \in \mathbb{R}^{N\times N}$ (recall our convention of representing the steady-state probability as the right eigenvector of Q) with associated eigenvalue λ. Given that we classify our flow into N distinct states, the action

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

of the functional on an arbitrary state is approximated by first classifying an arbitrary state to an integer label n and then picking out the associated component, e.g. $\boldsymbol{g} \cdot \hat{e}_n$. In other words, we use a piecewise-constant approximation to the functional where subsets of state space are assigned the same value. Thus we take our nonlinear dictionary (see Williams *et al.* 2015; Colbrook 2023) to be indicator functions for different cells in a partition of state space.

In what follows, we make choices for the classifier, and Markov states as applied to the compressible Euler equations. We compare statistics given by the generator Q, and Markov states $\boldsymbol{\sigma}^{[n]}$ to those given by the temporal statistics $\boldsymbol{s}(t)$. In other words, we compare ensemble averaging of the discretized statistics, as encapsulated by Q and Markov states $\boldsymbol{\sigma}^{[n]}$, to time averaging given by dynamical trajectories. We compute Koopman eigenvectors as well as modes, and enact two different strategies of partitioning state space to extract different levels of information.

3. The compressible Euler equations and a reduced-order statistical model

We consider a flux-differencing discontinuous Galerkin discretization of the compressible Euler equations on the sphere with rotation and gravity. The prognostic variables of choice are density ρ, momentum $\rho \boldsymbol{u}$, and total energy ρe. The dynamics are given by the following equations:

$$\partial_t \rho = -\boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}), \tag{3.1}$$

$$\partial_t(\rho \boldsymbol{u}) = -\boldsymbol{\nabla} \cdot (\boldsymbol{u} \otimes \rho \boldsymbol{u} + p\mathbb{I}) - \rho \boldsymbol{\nabla} \varPhi + \boldsymbol{S}_{\rho \boldsymbol{u}}(\rho, \rho \boldsymbol{u}, \rho e), \tag{3.2}$$

$$\partial_t(\rho e) = -\boldsymbol{\nabla} \cdot (\boldsymbol{u}(p + \rho e)) + S_{\rho e}(\rho, \rho \boldsymbol{u}, \rho e), \tag{3.3}$$

where $\boldsymbol{S}_{\rho \boldsymbol{u}}$ and $S_{\rho e}$ are source terms, $\varPhi$ is the geopotential, and p is pressure. Details are given in Appendix A. The corresponding Liouville equation is

$$\begin{aligned} \partial_t \mathcal{P} &+ \int_\Omega \frac{\delta}{\delta \varrho} [-\mathrm{div}(\varrho \boldsymbol{u}) \mathcal{P}] \\ &+ \int_\Omega \frac{\delta}{\delta \varrho u} [(-\mathrm{div}(\boldsymbol{u} \varrho u + p\hat{x}) - \varrho \hat{x} \cdot \mathrm{grad}(\varPhi) + S_{\rho u})\mathcal{P}] \\ &+ \int_\Omega \frac{\delta}{\delta \varrho v} [(-\mathrm{div}(\boldsymbol{u} \varrho v + p\hat{y}) - \varrho \hat{y} \cdot \mathrm{grad}(\varPhi) + S_{\rho v})\mathcal{P}] \\ &+ \int_\Omega \frac{\delta}{\delta \varrho w} [(-\mathrm{div}(\boldsymbol{u} \varrho w + p\hat{z}) - \varrho \hat{z} \cdot \mathrm{grad}(\varPhi) + S_{\rho w})\mathcal{P}] \\ &+ \int_\Omega \frac{\delta}{\delta \varrho e} [(-\mathrm{div}(\boldsymbol{u}(p + \varrho e)) + S_{\rho e})\mathcal{P}] = 0, \end{aligned} \tag{3.4}$$

where we make the correspondence $\mathscr{s}_{(\boldsymbol{x},1)} = \varrho$, $\mathscr{s}_{(\boldsymbol{x},2)} = \varrho u$, $\mathscr{s}_{(\boldsymbol{x},2)} = \varrho v$, $\mathscr{s}_{(\boldsymbol{x},2)} = \varrho w$, $\mathscr{s}_{(\boldsymbol{x},5)} = \varrho e$ and $\boldsymbol{u} = (u, v, w)$ with notation established in Part 1. The source term $\boldsymbol{S}_{\rho \boldsymbol{u}}$ is broken up into three terms $S_{\rho u}$, $S_{\rho v}$ and $S_{\rho w}$. Furthermore, we use 'grad' and 'div' for the gradient and divergence, respectively.

Equation (3.4) is viewed as a formal expression and reviewed in Part 1. It is possible to manipulate such equations to derive helpful relations, which may be checked *a posteriori* such as in the conditional averaging procedure in Souza *et al.* (2023*b*) to derive a turbulent diffusivity operator or Giorgini *et al.* (2024) to analytically determine a 'score function' for particular classes of stochastic partial differential equations. Rigorous foundations for the

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

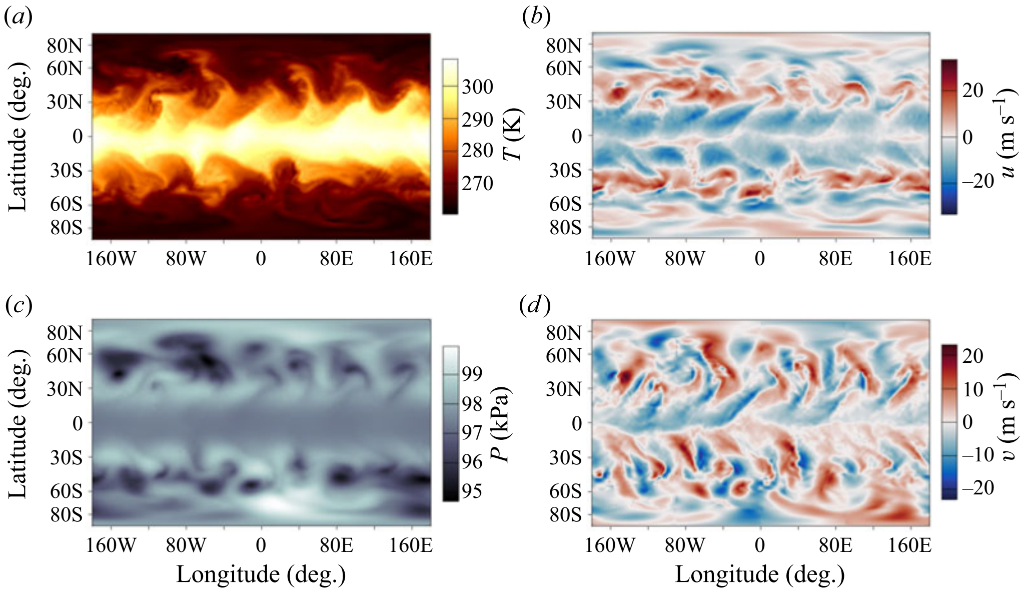

Figure 1. Surface fields of the Held–Suarez atmospheric test case. We show (*a*) surface temperature, (*b*) zonal velocity, (*c*) pressure, and (*d*) meridional velocity.

statistics of partial differential equations have been established in simpler contexts (see the theory in Hairer (2014) and the review of Corwin & Shen (2020)), but still require further development to apply to the case considered here.

The numerical discretization of the compressible Euler equations is outlined in Appendix A but is irrelevant for the present purposes. Instead, we consider the system a finite but high-dimensional space, with $d = 1\,481\,760$ in our specific case.

We choose the Held–Suarez test case for our analysis because it exhibits turbulence, has been extensively studied by the atmospheric community, and is a geophysically relevant configuration that produces wind and temperature patterns similar to those observed on Earth. Moreover, its statistics are robust across multiple discretization strategies, dissipation mechanisms and equation formulations. It does not exhibit meta-stable states and thus serves as a stringent test of the methodology.

Figure 1 shows a typical surface snapshot of the prognostic variables in the Held–Suarez simulation. The zonal velocity is the wind speed that flows in the east–west direction, and the meridional velocity flows in the north–south direction. Temperature and pressure are given by an equation of state and are nonlinear algebraic combinations of the state variables. With the Held–Suarez forcing, the temperature is hottest near the equator and cools towards the poles.

We now apply the methodology from § 2 to the Held–Suarez atmospheric test case. The first partition is chosen to provide insight into the topological structure of the turbulent attractor. The latter demonstrates that targeted partitioning strategies enable data-driven statistical modelling for observables of interest. The first strategy uses 400 cells, and the second uses 100 cells. The choice of classifiers $\mathcal{C}$ and Markov states $\boldsymbol{\sigma}^{[n]}$ is described in the relevant sections.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

3.1. *Sampling partition and steady-state statistics*

We start with an initial simulation run to reach a turbulent state, as detailed in Appendix A. A simulated 'day' is the unit of time and corresponds to one rotation of the planet based on its angular velocity. In the atmosphere, the weather's decorrelation time is approximately two weeks. Our first partition thus gathers Markov states every 15 simulated days until 400 states have been accumulated. This choice corresponds to random samples of the attractor.

The classifier $\mathcal{C}$, as before, corresponds to the index of the 'closest' Markov state. Our notion of 'close' is based on the distance function

$$d(\boldsymbol{\mathscr{s}}^1, \boldsymbol{\mathscr{s}}^2) = \sqrt{\int_{\Omega} \mathrm{d}\boldsymbol{x} \sum_i (\alpha_i)^{-2} \left(\mathscr{s}^1_{(\boldsymbol{x},i)} - \mathscr{s}^2_{(\boldsymbol{x},i)} \right)^2}, \tag{3.5}$$

which is a weighted L^2 norm between the different fields of the system (so that we add fields together in a dimensionless way). As a reminder: $\mathscr{s}_{(x,1)} = \varrho$, $\mathscr{s}_{(x,2)} = \varrho u$, $\mathscr{s}_{(x,2)} = \varrho v$, $\mathscr{s}_{(x,2)} = \varrho w$, $\mathscr{s}_{(x,5)} = \varrho e$. The α_i are

$$\left.\begin{array}{c} \alpha_1 = 1.3\ \mathrm{kg\ m^{-3}}, \quad \alpha_2 = \alpha_3 = \alpha_4 = 60\ \mathrm{m\ s^{-1}}, \\ \alpha_5 = 2.3 \times 10^6\ \mathrm{kg\ m^{-1}\ s^{-2}}. \end{array}\right\} \tag{3.6}$$

The reference values are chosen as pointwise maximum densities, speed and total energy. In total, the classifier is

$$\mathcal{C}(\boldsymbol{\mathscr{s}}) = n \text{ if } d(\boldsymbol{\mathscr{s}}, \boldsymbol{\sigma}^{[n]}) < d(\boldsymbol{\mathscr{s}}, \boldsymbol{\sigma}^{[m]}) \text{ for each } m \neq n, \tag{3.7}$$

i.e. we calculate the distance of the current state $\boldsymbol{\mathscr{s}}$ to all the Markov states $\boldsymbol{\sigma}^{[m]}$, and pick the integer corresponding to the Markov state with the smallest distance.

We evolve the system through an additional 200 simulated years, and apply the classifier $\mathcal{C}$ every $\Delta t = 0.03$ simulated days to the instantaneous state. This is a total of 2 000 000+ snapshots of time, but none are saved since this would have amounted to over 20 terabytes of data. The classifier is applied 'on the fly', and only an integer sequence is recorded. The first 30 simulated days of this process are shown in figure 2. We have ordered the indices *a posteriori* so that the most probable cell is assigned index 1, and the least probable cell is assigned index 400.

Given that we use a Bayesian method in estimating the generator entries, we need to assign a prior distribution for the entries of the generator Q. We assume that the holding time of every cell is Δt, and that each cell is equally connected to all others, but assigning very little weight to this initial guess. Specifically for our prior distribution on the generator entries, we take the initial parameters for the gamma distribution $\Gamma(\alpha, \beta)$ to be $\alpha = 1$ and $\beta = \Delta t$, where Δt is the sampling time interval for the time series. For each column of the matrix, we take the initial parameters of the Dirichlet distribution $D(\boldsymbol{\alpha})$ to be $\boldsymbol{\alpha} = 10^{-4}\mathbf{1}$, where $\mathbf{1}$ is the vector of all 1s. The combination of the two prior distributions is interpreted as follows. If a cell is not observed, then it is assumed that the holding time is below the sampling threshold given by Δt days. Furthermore, an unobserved cell is assumed to be connected to every other state uniformly in its exit probabilities. We take this precaution because it is unclear *a priori* if every cell is revisited over a finite sampling period. That being said, 200 simulated years sufficed for revisiting every cell, and results are sampled sufficiently so that the initial prior distribution has little effect on the end posterior distribution. The present results are relatively unchanged upon using an uninformative prior, i.e. $\boldsymbol{\alpha} = 0\mathbf{1}$ and $\alpha = \beta = 0$; however, there is one cell that was not revisited in the second half of the 200 simulated year period. Thus the Bayesian method

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

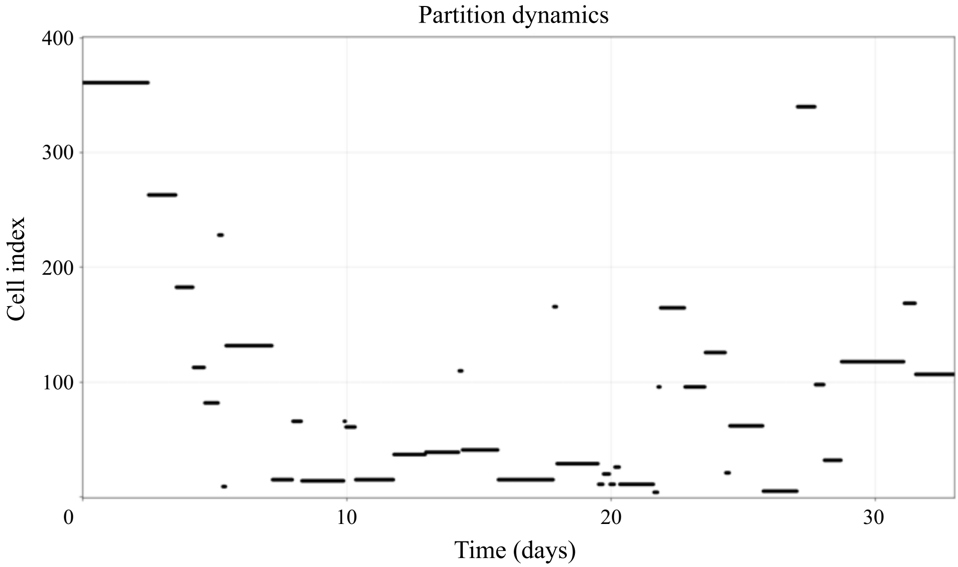

Figure 2. Held–Suarez partition dynamics. The dynamics are reduced to a sequence of integers. We order the indices by the steady-state probability of being within a cell, so that index 1 corresponds to the most probable cell, and index 400 to the least probable cell.

serves to regularize the matrix so that the 400-cell partition can be compared across the first 100 simulated years gathering period and the latter 100 simulated years.

Displaying the mean and variance of a 400 × 400 random matrix is not particularly illuminating. Thus we summarize four properties of the mean generator in figure 3: the real part of the inverse eigenvalues, the steady-state probability values associated with a cell, the connectivity of a given cell to every other cell, and the average holding time of a cell. The inverse eigenvalues' real part is associated with the slowest decaying autocorrelations of the system as captured by the partition choice. We see that there is a clustering of eigenvalues between $1/2$ and 1 simulated day. Furthermore, we see an apparent spectral gap between the first few eigenvalues (red) and the bulk (blue). This may imply the existence of continuous and discrete spectra in the limit of ever-refined partitions; however, it is unclear if there is a unique limit upon refining a coarse-grained state space or if the data-driven method of Part 1 even converges to such a limit. See § A.3 for corresponding oscillatory time scales.

The steady-state probability vector is not uniform (figure 3*a*), yet the amount of time spent in each state (figure 3*d*) is roughly the same for each cell. The reason for non-uniform probabilities is explained by looking at the connectivity of a given cell (figure 3*c*). The connectivity is the empirical number of exits from or entrances to a given cell. The more probable cells are more connected to the rest of state space than the rest. The connectivity of a cell can be thought of as the effective dynamical predictability associated with a cell. For example, sufficiently sampled cells of a periodic solution are connected to only one other cell since the future is precisely predictable from the past.

A priori, there is no reason to expect any cell to differ from another cell, given that Markov states were sampled uniformly in time; however, figure 3 suggests otherwise. The most probable regions of state space act as central hubs, connecting the various regions of state space together. These are perhaps associated with coherent structures such as fixed points or periodic orbits with few unstable directions.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

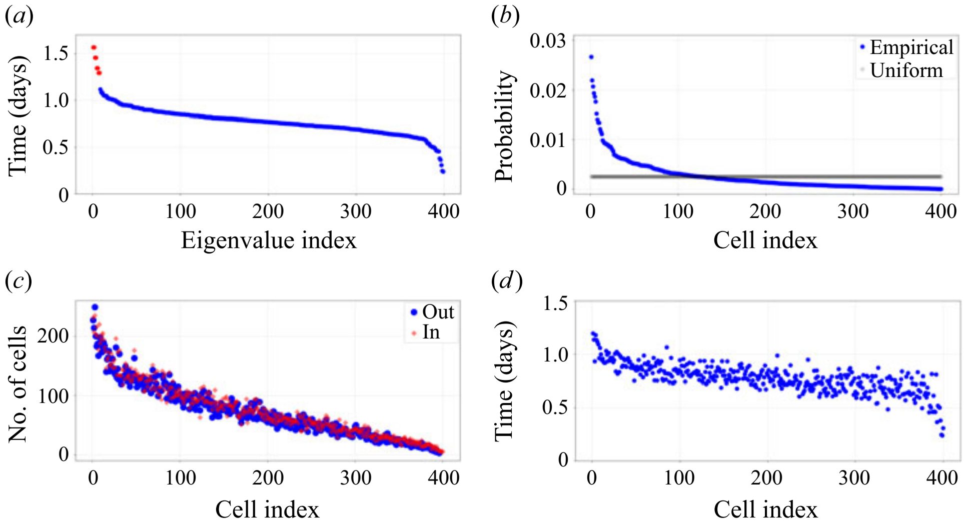

Figure 3. Generator properties. (*a*) The inverse real part of the eigenvalues of the generator, corresponding to the decorrelation time scales associated with the partition. (*b*) The steady-state cell probabilities associated with a partition. (*c*) Summary of the connectivity of a cell to other cells based on the empirically observed transitions. (*d*) The average holding time for a given cell.

We have discussed the topological characteristics of the partition choice as reflected by the generator. Additional details on finite sampling effects and the holding time distributions are explored in § A.3. We now move on to the calculation of statistical quantities.

In figure 4, we examine the histogram of the observable

$$g(\boldsymbol{\mathscr{s}}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\varphi}, \tag{3.8}$$

where $\hat{\varphi}$ is the unit vector along the zonal direction, and $\boldsymbol{x}$ is a point on the inner shell (surface) at latitude $\theta = 35^\circ$S and longitude $\varphi = 135^\circ$E. We show two overlapping histograms. One histogram is calculated from the generator steady-state probabilities and the Markov states, whereas the other comes from a time series. The time series of the observable was accumulated over a 30-year time span disjoint from the data used to construct the generator. The purple region is where the two histograms overlap, the red region is where the Markov model overpredicts the probability, and the blue region is where the Markov model underpredicts the probability. We show several bins, as before, to capture the notion of ‘convergence in quantile’. When we have as many bins as Markov states (400 bins in the present case), the delta function approximation begins to reveal itself. The height of the delta functions is associated with the steady-state probability distribution of the generator.

We now calculate the mean for a continuum of observables. We use the zonal average of the zonal velocity field for each latitude and each height:

$$g^{(\theta,r)}(\boldsymbol{\mathscr{s}}) = \frac{1}{2\pi}\int_0^{2\pi} (\boldsymbol{u}_{(\theta,\varphi,r)} \cdot \hat{\varphi})\, \mathrm{d}\varphi. \tag{3.9}$$

A fixed latitude θ and height r constitute one observable, and we expanded a position $\boldsymbol{x} = \theta\hat{\theta} + \varphi\hat{\varphi} + r\hat{r}$ in terms of its components in a spherical basis. We calculate each observable’s ensemble and temporal mean, and visualize the result as a heat map

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

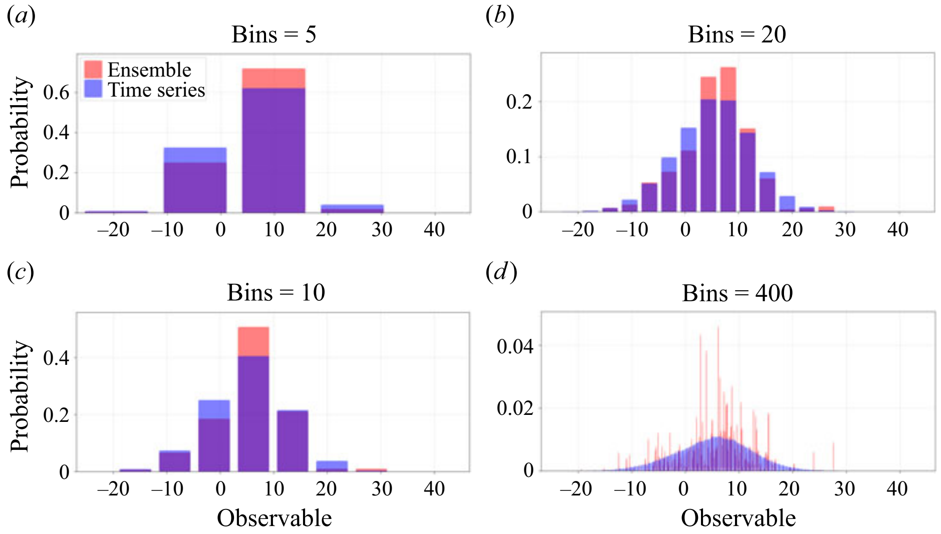

Figure 4. Steady state distribution of an observable. Here, we use the delta function approximation to the probability densities within a cell and look at the inferred distributions based on different coarse-grainings of the distribution. The overlap region is in purple; red bars correspond to 'overpredicting' probabilities, and blue bars correspond to 'underpredicting' probabilities. The temporal and ensemble means are 5.3 and 5.7, respectively. The temporal and ensemble standard deviations are 7.1 and 6.5, respectively.

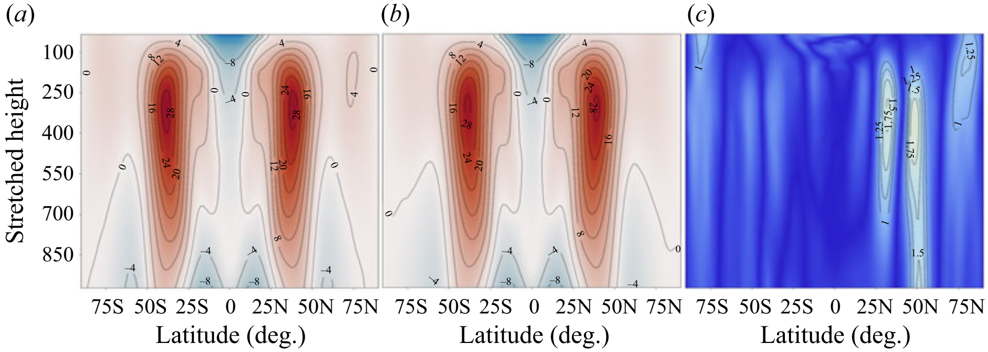

Figure 5. Mean value for a continuum of observables. The (*a*) ensemble average, and (*b*) temporal average, zonal mean zonal winds display a mean for a continuum of observables. (*c*) The pointwise absolute difference between the two means.

in figure 5. The ensemble mean uses the probability weights given in figure 3 and the 400 Markov states. The temporal mean is gathered over three simulated years. To make a connection with how this field is usually visualized, see Held & Suarez (1994), we rescale the height of the axis according to the zonal average of pressure at the equator. This rescaling mimics the effect of using 'pressure coordinates' in the atmospheric literature. The ensemble and temporal means differ by less than 2 m s^{-1} on the right-hand half of the zonal wind 'butterfly' wing.

It is not necessary to use the methodology of Part 1 to compute steady-state statistics, since 400 snapshots that are uniformly spaced in time would typically be averaged according to a uniform weight between them to calculate a statistic of interest; however, as we have seen from figure 3, the weighting between snapshots is far from uniform in the

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

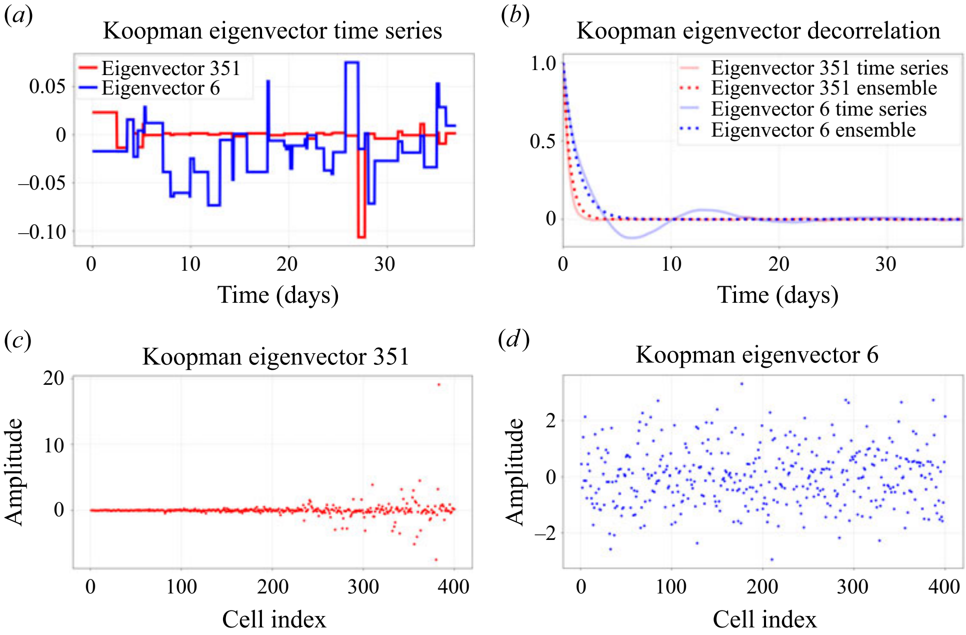

Figure 6. Held–Suarez Koopman eigenvectors. We show two numerical Koopman eigenvectors, in red and blue. (*a*) The numerical Koopman eigenvector as a function of time. (*b*) The decorrelation time scales for the numerical Koopman eigenvector computed by the generator (dashed) and the time series (solid). (*c*,*d*) The real parts of the Koopman eigenvectors as a function of the state index and thus implicitly as a functional of the state.

present case. Despite the highly non-uniform weighting, the statistics are well-captured. We next proceed to a more stringent test: global Koopman modes and temporal autocorrelations.

3.2. *Sampling partition: global Koopman modes and temporal autocorrelations*

The numerical Koopman eigenvectors are the left eigenvectors of the matrix Q, which we denote by $\boldsymbol{g}_\lambda$, and their approximation as functionals acting on the state is given by

$$g_\lambda(\mathscr{s}) \approx [\boldsymbol{g}_\lambda]_{\mathcal{C}(\mathscr{s})}, \tag{3.10}$$

where $[\boldsymbol{g}_\lambda]_n$ is the nth component of the eigenvector $\boldsymbol{g}_\lambda$, which has the property $\boldsymbol{g}_\lambda^{\mathrm{T}} Q = \lambda \boldsymbol{g}_\lambda^{\mathrm{T}}$, where T denotes the transpose of the vector, and λ is a eigenvalue. Hence we first apply the classifier to the state, and then use the integer label to pick out the component of the eigenvector $\boldsymbol{g}_\lambda$.

At each moment in time, we plot the approximate Koopman eigenvector

$$g_\lambda(\boldsymbol{s}(t)) \approx [\boldsymbol{g}_\lambda]_{\mathcal{C}(\boldsymbol{s}(t))}, \tag{3.11}$$

where the component of the vector $\boldsymbol{g}$ is given by $\mathcal{C}(\boldsymbol{s}(t))$. We show these dynamics for the first 30 simulated days of the Held–Suarez set-up in figure 6. Furthermore, we compute autocorrelations in two ways to check the fidelity of the numerical Koopman eigenvectors. The first uses the generator, and the second uses the Koopman eigenvector time series. This calculation is shown in figure 6.

We select two of the 400 modes to illustrate these points: modes 6 and 351 are associated with the 7th and 352nd eigenvalues when ordering the real part part of the spectrum from

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

least to most negative. Figure 6(*b*) shows that the two calculation methods agree for mode 351 (red) at all times, but only for the first 5 days for mode 6 (blue). We see that the estimate for the eigenvalue of mode 6 is overly dissipative. This situation is typical for the data-driven approximation of the generator. Given the near-exponential decay structure using the time series, this suggests a perturbation to the eigenvalues of the generator that could align the time series and ensemble calculation as was done in Giorgini, Souza & Schmid (2023).

The peak at approximately 14 days may be synonymous with the usual decorrelation time assumed for the atmosphere. The real component of the Koopman eigenvector as a function of cell index are shown in figures 6(*c*,*d*). We see that mode 351 picks up on unlikely cells of state space, whereas mode 6 is distributed amongst all states.

We also calculate Koopman modes associated with a particular eigenvalue. As an example, we will take the observable of interest to be the surface temperature field

$$g^{[\boldsymbol{x}_s]}(\boldsymbol{\mathcal{s}}) = \mathcal{T}_{\boldsymbol{x}_s}, \tag{3.12}$$

where $\boldsymbol{x}_s$ is a surface position, and $\mathcal{T}_{\boldsymbol{x}_s}$ is the temperature observable defined by

$$\mathcal{T}_{\boldsymbol{x}} \equiv \frac{\gamma - 1}{R_d \varrho_{\boldsymbol{x}}} \left(\varrho e_{\boldsymbol{x}} - \frac{1}{2} \varrho_{\boldsymbol{x}} \left\| \boldsymbol{u}_{\boldsymbol{x}} \right\|^2 - \varrho_{\boldsymbol{x}} \Phi_{\boldsymbol{x}} \right), \tag{3.13}$$

where $R_d = 287$ is the ideal gas constant, $\gamma = 1.4$ is the specific heat ratio of air, and Φ is the geopotential.

The Koopman modes are computed utilizing the right eigenvectors of Q, which we denote by v_λ, according to the formula

$$G_\lambda(\boldsymbol{x}_s) = \sum_n g^{[\boldsymbol{x}_s]}(\boldsymbol{\sigma}^{[n]}) \, [v_\lambda]_n, \tag{3.14}$$

where v_λ is the eigenvector associated with eigenvalue λ, i.e. $Qv_\lambda = \lambda v_\lambda$. The choice $\lambda = 0$ in (3.14) reproduces the ensemble mean. Furthermore, $[v_\lambda]_n$ denotes the nth component of the right eigenvector. We refer to G_{λ_m} as 'Koopman mode m' where the eigenvalues are ordered λ_m for $m = 0, \ldots, 399$. To further explain the connection to an 'amplitude', first think of $g^{[\boldsymbol{x}_s]}(\boldsymbol{\sigma}^{[n]})$, at a fixed $\boldsymbol{x}_s$, as a 400-dimensional vector whose entries are given by evaluating the observable $g^{[\boldsymbol{x}_s]}$ on each Markov state $\boldsymbol{\sigma}^{[n]}$ for $n = 1, \ldots, 400$. We now expand this 400-dimensional vector with respect to the basis of left eigenvectors generator. The 'amplitudes' G_{λ_n} are the coefficients in the expansion

$$g^{[\boldsymbol{x}_s]}(\boldsymbol{\sigma}^{[n]}) = \sum_m G_{\lambda_m}(\boldsymbol{x}_s) \, [\boldsymbol{g}_{\lambda_m}]_n, \tag{3.15}$$

where we re-emphasize that $\boldsymbol{x}_s$ is thought of as a fixed value, and the formula holds for each n. That is, we expand an observable as a linear combination of Koopman eigenvectors and then pick out the amplitude. We do this for each $\boldsymbol{x}_s$ individually, and then aggregate results afterwards.

We show the real and imaginary components of Koopman modes associated with the surface temperature field in figure 7. All mode amplitudes are linear combinations of the temperature Markov states. Mode 0 is the ensemble mean, and the other modes are associated with wavelike patterns. The general shapes and structures of the Koopman modes are robust in terms of the number of cells and amount of data collected. For example, the same structures emerge utilizing either the first 100 simulated years or the latter 100 simulated years during the 200-simulated-years data collection period.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

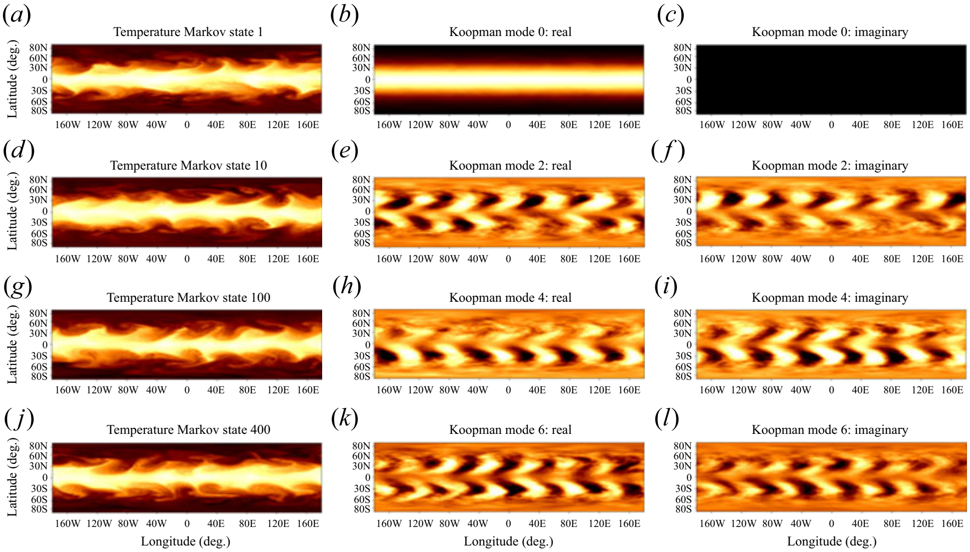
(a)
Temperature Markov state 1
(b)
Koopman mode 0: real
(c)
Koopman mode 0: imaginary
(d)
Temperature Markov state 10
(e)
Koopman mode 2: real
(f)
Koopman mode 2: imaginary
(g)
Temperature Markov state 100
(h)
Koopman mode 4: real
(i)
Koopman mode 4: imaginary
(j)
Temperature Markov state 400
(k)
Koopman mode 6: real
(l)
Koopman mode 6: imaginary
Latitude (deg.)
80N
60N
30N
0
30S
60S
80S
160W 120W 80W 40W 0 40E 80E 120E 160E
Longitude (deg.)

Figure 7. Held–Suarez Koopman modes. We show four representative surface temperature fields constructed from Markov states as well as their Koopman modes by projecting the fields onto the real and imaginary parts of the associated Koopman eigenvector. The statistically steady state is associated with mode 0.

The Koopman modes are related to the emergent time scales and autocorrelations captured from a given partition choice. We now switch to analysing autocorrelation for observables unrelated to the cells. Specifically, the autocorrelations of four observables,

$$g^{[1]}(\boldsymbol{s}) = \varrho_{\boldsymbol{x}}, \quad g^{[2]}(\boldsymbol{s}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\varphi}, \quad g^{[3]}(\boldsymbol{s}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\theta}, \quad g^{[4]}(\boldsymbol{s}) = \mathcal{T}_{\boldsymbol{x}}, \qquad (3.16a\text{–}d)$$

are shown in figure 8 for the same position $\boldsymbol{x}$ as before. We show the empirically obtained autocorrelation from the time series in black, and the generator in purple. Since most of the eigenvalues of the generator are clustered (as seen from figure 3), we expect observables uncorrelated with a partition choice to have similar decorrelation time scales. We see that this is a poor approximation for observable $g^{[3]}$, but not so for the other variables.

In the next subsection, we choose a partition that targets an observable associated with extreme events.

3.3. *Extreme statistics partition*

We now partition the turbulent attractor differently to target statistics of a particular observable: temperature extremes at a particular point in the domain. Specifically, we choose

$$g(\boldsymbol{s}) = \begin{cases} 1 & \text{if } \mathcal{T}_{\boldsymbol{x}} > 290\ \text{K}, \\ 0 & \text{otherwise}. \end{cases} \qquad (3.17)$$

Here, $\boldsymbol{x}$ is a point on the inner shell of the sphere at latitude $\theta = 35^\circ$S, longitude $\varphi = 135^\circ$E. The choice of 290 K came from the 95 % quantile of temperature at that point over a short simulation run.

We gather the Markov states by first partitioning an arbitrary state into two classifications: $g(\boldsymbol{s}) = 1$ and $g(\boldsymbol{s}) = 0$. The former is representative of an ‘extreme state’,

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

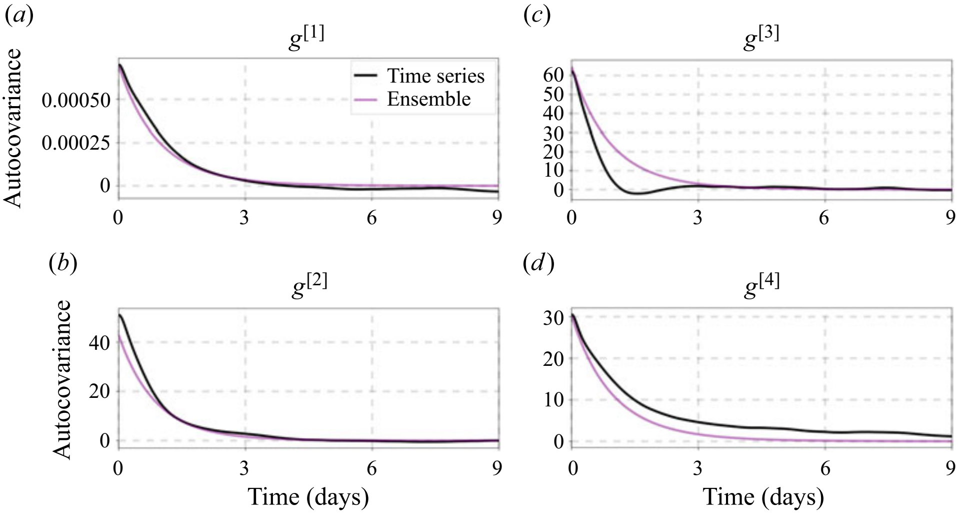

Figure 8. Autocovariance for several observables in the Held–Suarez atmospheric test. The autocovariances for several observables based on the time series (black) and generator (purple) are shown. The observables are uncorrelated with the partition; thus the autocorrelation predicted from the Markov model is similar for all four cases.

and the latter of a ‘benign state’. We then gathered 10 representative extreme states and 90 representative benign states. Specifically, a simulation is run, and the states are checked every two weeks. We apply the observable (i.e. classifier) g to determine whether or not the state is extreme. The process continues until at least 10 extreme and 90 benign states are gathered. We only keep 100 total states. Thus any extra states are discarded. The extreme states are assigned indices 1–10, while the benign states are assigned indices 11–100.

The choice of partition is arbitrary, and many choices would yield similar insight. For example, the steady-state probability of an extreme state would be, by construction, well captured with even one state; however, one would not be able to assess statistics within the extreme state. Using ten extreme states serves as a compromise on gathering information about the system when such an event is occurring. An analogy between the present method and local grid-refinement methods from numerical methods can be made. Here, we are choosing to refine a particular subset of state space with respect to a criterion of interest. If we refine too much of a given region (in this case, include more states in the extreme states category), then we will not improve the overall representation of other statistics. Furthermore, one runs into a data problem if statistics are too rare to gather enough extreme states.

With these Markov states in place, the classifier is defined as

$$\mathcal{C}(\mathcal{s}) = \begin{cases} n_1 & \text{if } g(\mathcal{s}) = 1 \text{ and } d(\mathcal{s}, \boldsymbol{\sigma}^{[n_1]}) \leq d(\mathcal{s}, \boldsymbol{\sigma}^{[m_1]}) \text{ for each } m_1, \\ n_2 & \text{if } g(\mathcal{s}) = 0 \text{ and } d(\mathcal{s}, \boldsymbol{\sigma}^{[n_2]}) \leq d(\mathcal{s}, \boldsymbol{\sigma}^{[m_2]}) \text{ for each } m_2, \end{cases} \tag{3.18}$$

where m_1, n_1 are indices associated with extreme cells, i.e. $m_1, n_1 \in \{1, \ldots, 10\}$, and m_2, n_2 are indices associated with benign cells, i.e. $m_2, n_2 \in \{11, \ldots, 100\}$. To elaborate, we first classify the state according to the observable g, and then calculate the closest Markov state within the respective category. Finally, we run the model for 100 simulated years, and construct the Markov chain embedding.

Subgraph for extreme states

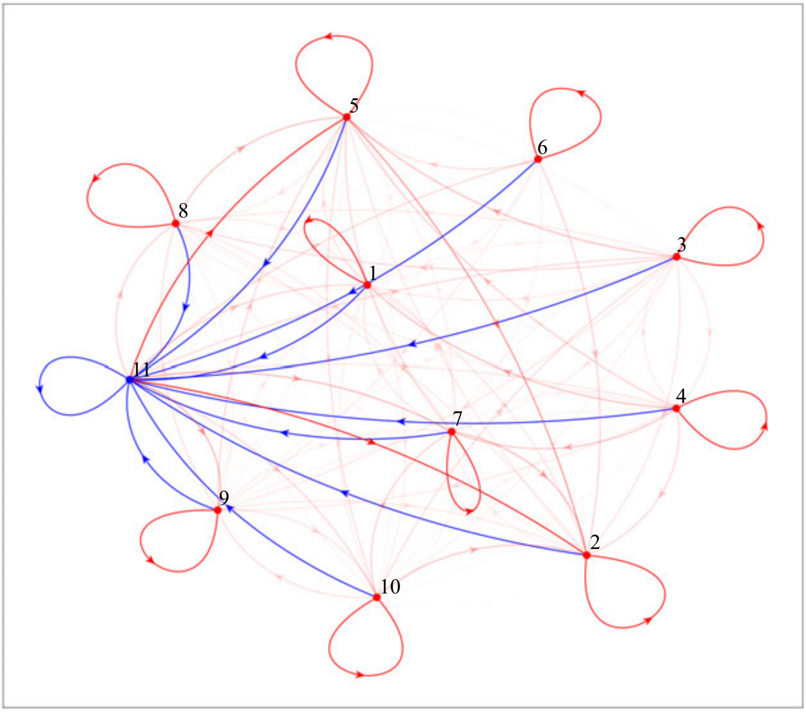

Figure 9. Network structure of extreme transitions. Eleven states are shown, where 1–10 correspond to extreme states, and 11 corresponds to the other 90 states, lumped together as a single node for visualization purposes. The blue lines correspond to transitions to a benign state, and the red lines are transitions to extreme states. The opacity of the lines is proportional to the probability of transitioning between states. There exist transitions between extreme states.

We first focus on the holding time distribution of being in a cell associated with an extreme event. This distribution is taken as a proxy for the duration of a heatwave. Given that we have ten possible states corresponding to an extreme event, we also account for transitions between states within an extreme event duration. Heuristically, transitions between different global states during an extreme event are rare since the duration is short compared to the holding time of being in a state.

Nevertheless, they do occur in this simulation. Figure 9 summarizes the transition pathways between cells associated with extreme states. In this figure, states 11–100 have been lumped together as a single state, and the graph structure of the transition pathways is shown. The transparency of the red lines corresponds to the probability of transitioning between the different extreme state cells, and the blue lines correspond to the transition probability of leaving an extreme state.

We see that an extreme state has many 'microstates' corresponding to the macrostate (defined by the cell induced from $g(\mathfrak{s}) = 1$ in (3.17)), and there are non-zero transitions between the microstates during a given macrostate configuration. This nuance partially explains the complexity of an extreme event prediction: the atmospheric state changes considerably during an extreme event, and more information may be required to make predictive statements about the duration and intensity of a local extreme event.

The holding time distribution of an extreme state (as calculated by the Markov embedding) accounts for transitions between the different states. Furthermore, we gather

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

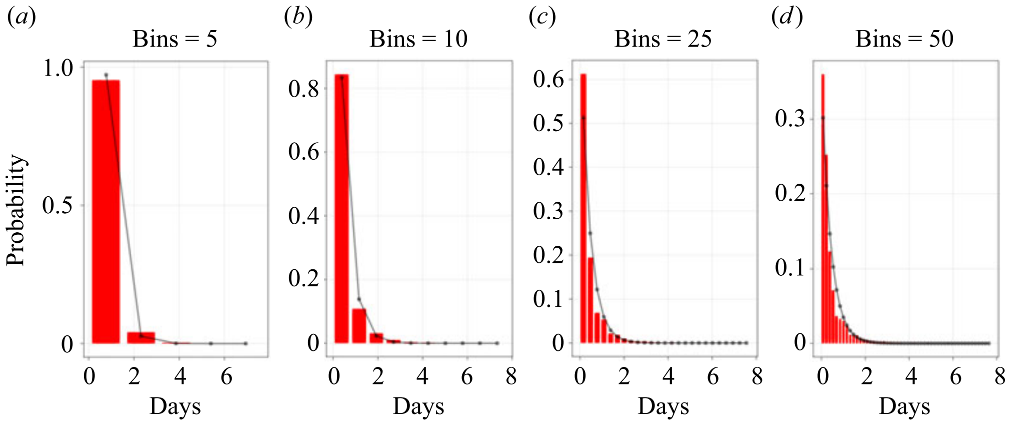

Figure 10. Held–Suarez holding time extreme. Several quantiles for the duration of extreme states, as calculated from the time series, are shown in red. For simplicity, we show the exponential distribution holding time as black dots, where the decorrelation time is approximately half a day as calculated by looking at the holding time for a single extreme state cell.

statistics from the temperature observable at a disjoint set in time, and show its holding time distribution in figure 10. The Markov state representation shows that the holding time distribution is well-captured.

The presence of an extreme state can be viewed as an exit time problem from the point of view of stochastic processes. An extreme event corresponds to a particular subset of state space, characterized by ten cells. The average amount of time spent in an extreme state must incorporate the transitions within the duration of an extreme event.

We also compare the temperature distribution at $\boldsymbol{x}$ as calculated by the 400-state system, the 100-state system, and the time series in figure 11. In particular, the 100-state system better captures the 95 % tail distribution. Thus selecting a partition that targets an observable of interest is feasible and is of increased fidelity compared to the naive generic partitioning. This procedure is equivalent to local grid refinement from numerical methods.

As a final comment for this subsection, we note that the choice of partition does not need to be binary. As one runs a simulation, the same Markov states can be used to compute several different partitioning strategies simultaneously. If partitioning strategies make use of independent observables, then it is natural to construct tensor product generators. We did not pursue any of these strategies here.

3.4. *Discussion: subtleties of high-dimensional discretizations*

Due to the high dimensionality of the system, we reflect on subtleties encountered thus far in our computations. It is intuitive to assume that if a particular observable is uncorrelated with a partitioning strategy, then it is unlikely that the autocorrelations are well captured by the generator. Furthermore, one would assume that only observables constant within a partition will have faithful representations of their ensemble mean statistics.

However, we have seen that both intuitions are accurate only sometimes. In the high-dimensional setting, we must distinguish between two classes of observables: those highly correlated with a given partitioning strategy, and those not. We rely on Monte Carlo sampling to compute ensemble statistics for uncorrelated observables with a partitioning strategy. Effectively, an observable is a random vector with respect to the partitions.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

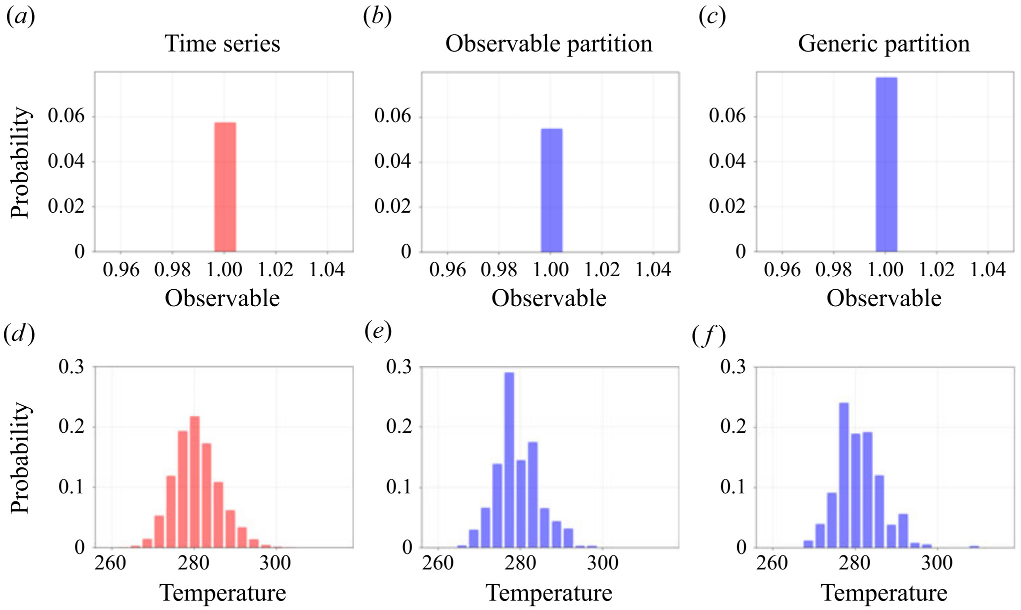

Figure 11. Held–Suarez observable comparison. We show the effect of grid refinement on an observable of interest. (*a–c*) The refinement strategy better captures the tail probability quantile than the generic 400-cell partition. (*d–f*) The temperature distribution is shown as a point of comparison.

For example, suppose that the approximate generator yields a uniform distribution for the steady-state distribution. In that case, it is expected that as long as the Markov states are ‘independent’ of one another, one can do at least as well as Monte Carlo sampling. If a partitioning strategy is well correlated with an observable of interest, then we expect to do better than ‘random’ sampling of the Markov states.

For autocorrelations, a similar phenomenon occurs. An observable that is uncorrelated with the partitions is a random vector concerning the partitions. Insofar as there are many observables with similar autocorrelations, this strategy will do well to capture those observables.

There are other issues when attempting to coarse-grain a high-dimensional space. The first issue comes from the representation of high-dimensional operators. For example, consider d copies of an Ornstein–Uhlenbeck process to yield a d-dimensional state space. In this case, the eigenvalues and eigenvectors of the joint process can be written analytically. Without exploiting the structure of the resulting generator (in this case, the generator can be written as a Kronecker sum of d generators), the high dimensionality of the system yields a similar ‘central hub’ structure that was observed in the Held–Suarez system (in § 3.1), especially when using a distance metric and taking one of the cells to be centred at the origin. The solution, in this case, is to exploit the structure of the problem and consider each process separately. For a concrete example, see Appendix A of Souza *et al.* (2023*b*), where a tensor product decomposition of a Gaussian process was enacted. One hopes that for turbulence, there exists a way to decompose the full generator as combinations of smaller generators. In this way, one can express high-dimensional behaviour compactly. Similar considerations apply to the Koopman operator/Perron–Frobenius operator (for these, one wants a Kronecker-product-like structure).

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

A second issue is related to the necessity of time history as in Lin *et al.* (2023), which is based on the Mori–Zwanzig formalism or, similarly, the absence of a 'semigroup' property of the data-driven generator/Perron–Frobenius operators. Normally, one gets around such issues by augmenting state space to include past-time information, e.g. Takens embedding. Another option is to instead construct a different operator for each time scale of interest. The failure of the semigroup property implies that the data-driven construction of a Perron–Frobenius operator at time scale τ, denoted by $\mathcal{F}^{[\tau]}$, and a data-driven construction of $\mathcal{F}^{[2\tau]}$ does not satisfy the identity $\mathcal{F}^{[2\tau]} = \mathcal{F}^{[\tau]}\mathcal{F}^{[\tau]}$. Thus to reconstruct statistics, one could instead compute several operators $\mathcal{F}^{[n\tau]}$ for $n \in \{1, \ldots, N_t\}$, and some final time scale $N_t\tau$. In this way, one can represent autocorrelations/autocovariances several times for an observable interest. Thus at least two strategies exist: (1) augment state space with past-time information, or (2) change the operator depending on the time scale of interest. See Giorgini *et al.* (2023) for a compromise on the latter strategy where one uses an alternative data-driven construction to the generator to retain the semigroup property. This latter work yields a third option.

4. Conclusion

In summary, we have applied two partitions of state space and used them to compute various statistical quantities of interest: steady-state statistics, global Koopman modes, and temporal autocorrelations. The methodology was implemented in an 'online' manner allowing for significant savings in memory. The first partitioning strategy defined a distance function centred on random decorrelated snapshots from a turbulent simulation. The second partitioning strategy targeted an observable of interest, temperature 'extremes' on the inner radius of the spherical shell (meant to represent heat waves at a fixed location).

The latter is more akin to what is done in statistical mechanics, where one defines a 'macrostate', but we also picked out a few 'microstates' corresponding to the macrostate. In general, one can create partitions of the entire state space by partitioning according to one (or several) observables, as is commonly done when performing dimensionality reduction; however, we contend that we always want a representative state associated with a partition to calculate ensemble mean statistics and correlations associated with the total state space.

Taken together, we see that the most critical component in the statistical representation of a system is the choice of partitioning strategy. Future directions necessitate the development of novel partitioning strategies, e.g. partitioning according to modal amplitudes given by dynamic mode decomposition, or using machine-learning methods such as auto-encoders to reduce the dimensionality of state space. Consistency between short-time computations and long-term statistics are likely to yield benefits. Typical Koopman mode expansions are local (in state space) expansions, thus creating a patchwork of local linearizations that could yield further improvement. Incorporating partial temporal coherence in the Markov state partitioning also seems promising, e.g. in the Held–Suarez case, choosing Markov states that are one day apart for a month, then skipping a few months, and repeatedly gathering Markov states.

Since the method has been formulated as a numerical discretization, there are straightforward generalizations to consider. For example, in addition to discretizing space using a finite-volume method, one can discretize time using a discontinuous Galerkin method. In this way, time trajectories are represented as piecewise-polynomial instead of piecewise-constant. Furthermore, the probability flux to a different region of state space would now (in discrete time) depend on the history.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

A more radical departure from the methods proposed here is to use generative models, similar to Ho, Jain & Abbeel (2020), to represent distributions within a partition. Partitions of state space may be more amenable to representation than the entirety of the manifold. Furthermore, using nonlinear models for the generator to account for all the different physical features to which one must assign 'attention' could yield a better overall representation (e.g. Vaswani *et al.* 2017).

Computing Koopman modes and eigenvalues allows one to determine an effective diffusivity operator; see the summary in Thiffeault (2023) and the derivation in Souza *et al.* (2023*b*). In that work, it was shown that the spectrum of the Koopman/Perron–Frobenius/Liouville/Fokker–Planck operators could be linked to a rigorous definition of turbulent transport. This theoretical link allows for another assessment of the fidelity of a partition.

The primary reason for undertaking the perspective in this paper was to gain a foothold in understanding climate change from an operator-theoretic approach, similar to Froyland *et al.* (2021). Climate change is often characterized as 'statistics changing over time' and thus requires a precise definition. We focused on a high-dimensional measure that is invariant with respect to time. This trait is not valid for the climate system, whose statistics are non-stationary. The predominant signal for a 'stationary' climate is not stationary but rather time-periodic due to the diurnal and seasonal cycles. Thus the first simplification is to consider a generator whose entries are periodic functions of time and whose Markov states are also periodic functions of time; see Wang & Schütte (2015) for similar considerations in molecular dynamics. Climate change is then characterized as deviations from this time-periodic (high-dimensional) flow.

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 US National Science Foundation and the US Office of Naval Research.

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

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

Appendix A. Held–Suarez model

Isaac Held and Max Suarez introduced a simplified atmospheric model test in Held & Suarez (1994). The test case purposely did not specify dissipation mechanisms, and was meant to be flexible as to which prognostic variables or coordinate systems were employed in its calculation. Its primary purpose was as a robust 'physics test' to be compared across different numerical schemes and equations of motion. In § A.1, we specify the equations, and in § A.2, the numerical discretization that was used. Finally, we conclude in § A.3 with a follow-up to some of the points made in § 3 about holding times, the convergence of matrix entries, and eigenvalue sensitivities.

A.1. *Partial differential equation set-up*

The model is described in Souza *et al.* (2023*a*), but here we give a summary. We choose to use an equation set that retains fully compressible dynamics and is formulated in terms

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

Parameter	Value	Unit	Description
$\mathcal{X}$	80	—	Scaling parameter
z_{top}	3×10^4	m	Atmosphere height
r_{planet}	$6.371 \times 10^6/\mathcal{X}$	m	Planetary radius
R_d	287	$\mathrm{m^2\ s^{-2}\ K^{-1}}$	Gas constant for dry air
$\mathcal{W}$	$2\pi/86\,400 \times \mathcal{X}$	$\mathrm{s^{-1}}$	Coriolis magnitude
p_0	1×10^5	$\mathrm{kg\ m^{-1}\ s^{-2}}$	Reference sea-level pressure
T_{min}	200	K	Minimum equilibrium temperature
$T_{equator}$	315	K	Equatorial equilibrium temperature
h_b	0.7	—	Dimensionless damping height
c_v	717.5	$\mathrm{J\ kg^{-1}\ K^{-1}}$	Specific heat capacity of dry air at constant volume
c_p	1004.5	$\mathrm{J\ kg^{-1}\ K^{-1}}$	Specific heat capacity of dry air at constant pressure
k_f	$\mathcal{X}/86\,400$	$\mathrm{s^{-1}}$	Damping scale for momentum
k_a	$\mathcal{X}/(40 \times 86\,400)$	$\mathrm{s^{-1}}$	Polar relaxation scale
k_s	$\mathcal{X}/(4 \times 86\,400)$	$\mathrm{s^{-1}}$	Equatorial relaxation scale
ΔT_y	60	K	Latitudinal temperature difference
$\Delta \theta_z$	10	K	Vertical temperature difference
G	6.67408×10^{-11}	$\mathrm{kg^{-1}\ m^3\ s^{-2}}$	Gravitational constant
M_P	$5.9722/\mathcal{X}^2 \times 10^{24}$	kg	Planetary mass

Table 1. Parameter values for the Held–Suarez test case. The value $\mathcal{X} = 1$ corresponds to the standard test case, and $\mathcal{X} = 80$ is the version that we use here.

of density, total energy, and Cartesian momentum as the prognostic variables, yielding the equations

$$\partial_t \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0, \tag{A1}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla \cdot (\boldsymbol{u} \otimes \rho \boldsymbol{u} + p\mathbb{I}) = -\rho \nabla \varPhi - 2(\boldsymbol{\mathcal{W}} \cdot \hat{r})\hat{r} \times \rho \boldsymbol{u} - k_v(\mathbb{I} - \hat{r} \otimes \hat{r})\rho \boldsymbol{u}, \tag{A2}$$

$$\partial_t (\rho e) + \nabla \cdot (\boldsymbol{u}(p + \rho e)) = -k_T \rho c_v (T - T_{equilibrium}), \tag{A3}$$

where $\varPhi = 2GM_P r_{planet}^{-1} - GM_P r^{-1}$ is the geopotential, $\boldsymbol{\mathcal{W}} = \mathcal{W}\hat{z}$ is the planetary angular velocity, $\hat{z}$ is the direction of the planetary axis of rotation, and r is the radial direction in spherical coordinates. The Coriolis force is projected to the radial component so that small-planet analogues (which we use for the simulation in § 3) have a climatology similar to that of Earth. Furthermore, the variable $T_{equilibrium}$ is the radiative equilibrium temperature depending on latitude, φ, and pressure, $h = p/p_0$,

$$T_{equilibrium}(\varphi, h) = \max(T_{min}, [T_{equator} - \Delta T_y \sin^2(\varphi) - \Delta\theta_z \ln(h) \cos^2(\varphi)]h^{R_d/c_p}), \tag{A4}$$

and the parameters k_v, k_T are the inverse time scales for momentum damping and temperature relaxation, respectively, with

$$k_v = k_f\, \Delta h \quad \text{and} \quad k_T = k_a + (k_s - k_a)\, \Delta h \cos^4(\varphi), \tag{A5a,b}$$

and $\Delta h = \max\{0, (h - h_b)/(1 - h_b)\}$. The temperature and pressure are

$$T = \frac{1}{c_v \rho}\left(\rho e - \frac{1}{2}\rho \,\|\boldsymbol{u}\|^2 - \rho \varPhi\right) \quad \text{and} \quad p = \rho R_d T. \tag{A6a,b}$$

The parameter values for the simulation set-up are in table 1.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

We use no-flux boundary conditions for density and total energy, free-slip boundary conditions for the horizontal momenta, and no-penetration boundary conditions for the vertical momentum. The initial condition is a fluid that starts from rest, $\rho \boldsymbol{u} = 0$, in an isothermal atmosphere,

$$p(r) = p_0 \exp\left(-\frac{\Phi(r) - \Phi(r_{planet})}{R_d T_I}\right) \quad \text{and} \quad \rho(r) = \frac{1}{R_d T_I} p(r), \qquad \text{(A7}a,b\text{)}$$

where we use $T_I = 285$ K.

A.2. *Numerical method*

To approximate the equation of the previous section, we use the flux-differencing discontinuous Galerkin method outlined in Souza *et al.* (2023*a*) and formulated precisely in Waruszewski *et al.* (2022). We choose numerical fluxes that are kinetic and potential energy preserving to help to ensure the flow's nonlinear stability and Roe fluxes for dissipation. In addition, the low-storage fourth-order 14-stage Runge–Kutta method of Niegemann, Diehl & Busch (2012) is used for time stepping and induces a form of numerical dissipation. All simulations were run on an Nvidia Titan V graphics processing unit.

The domain is a piecewise-polynomial approximation to a thin spherical shell of radius r_{planet} and height z_{top}. The thin spherical domain is partitioned into curved elements and uses an isoparametric representation of the domain, and the cubed sphere mapping by Ronchi, Iacono & Paolucci (1996). In essence, this choice represents the domain as a piecewise-polynomial function where the order of the polynomial corresponds to the order of the discretization (Winters *et al.* 2021). The metric terms are treated as in Kopriva (2006) and satisfy the discrete property that the divergence of a constant vector field is zero, i.e. the metric terms are free-stream preserving.

We use 4 elements in the vertical direction, 6×6^2 elements for the sphere's surface (6^2 elements per cubed sphere panel), and order 6 polynomials within each element. Given that we have 5 prognostic states (density, the three components of the Cartesian momenta, and total energy), this leads to a total of $5 \times 4 \times (6 \times 6^2) \times 7^3 = 1\,481\,760$ degrees of freedom – the horizontal acoustic CFL limits time steps.

A.3. *Partition properties and uncertainty quantification*

In this subsection, we examine additional properties of the generator from § 3.1. We examine the oscillatory time scales associated with the eigenvectors of the generator, quantify the uncertainty with the Bayesian approach from Part 1, and examine the holding times of three partitions.

We define the oscillatory time scale of an eigenvalue λ as $2\pi/|\text{imag}(\lambda)|$, where 'imag' signifies the imaginary part. In figure 12, we show the oscillatory time scales associated with the eigenvalues of the generator. These correspond directly with figure 3(*a*). In general, the imaginary component of the eigenvalue λ is much smaller than the real component, leading to longer oscillatory time scales. For example, the decorrelation time scales are roughly of the order of 1 day, whereas the oscillatory time scales are of the order of 100 days. The time scales come in pairs since the generator is a real matrix, thus any complex eigenvalue must have an accompanying complex conjugate.

We show two figures for investigating convergence with data volume. In figure 13, we show the generator's inverse holding times (diagonal entries) for the first 16 most

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

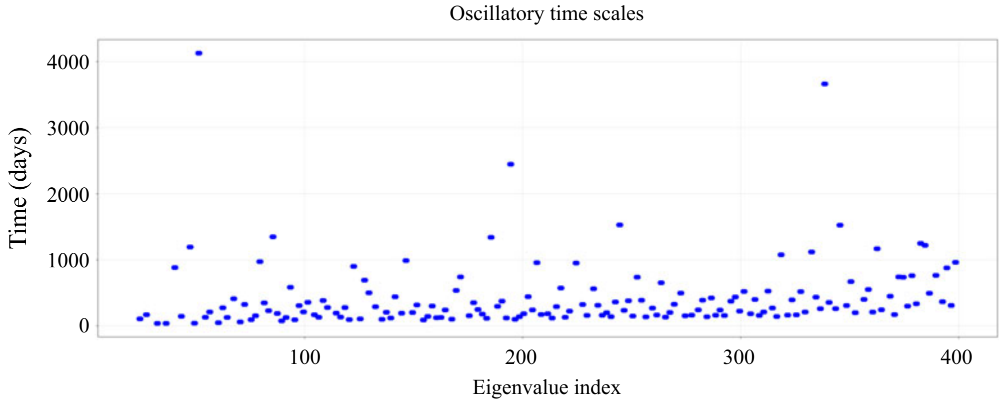

Figure 12. Held–Suarez generator oscillatory time scales. Here, we visualize the time scales associated with oscillatory motion of generator eigenvalues. The ranges of oscillatory time scales range from approximately 38 days to 11 years.

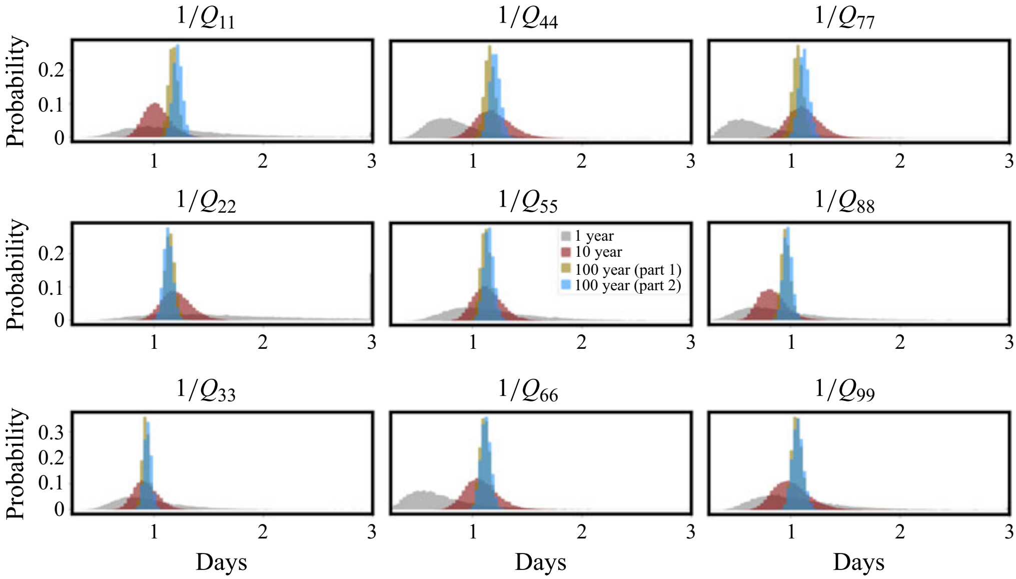

Figure 13. Held–Suarez generator rate entries. The uncertainty with respect to the inverse rates is shown for various time intervals. In grey, red, gold and blue, we show the uncertainty corresponding to 1 year, 10 years, 100 years and another (separate) 100-year simulation. We see that there is a significant overlap in the two 100-year estimates.

probable states. We see that there appears to be convergence to the matrix entries over disparate time intervals.

In figure 14, we show the real part of the inverse eigenvalue as distributions from random samples of the generator matrix. This variable corresponds to the decorrelation time scale as given by the partition. The Bayesian approach suggests that we cannot trust the slowest decorrelation scale obtained from the numerical solution since it varies between 1.5 days and 20 days. On the other hand, the other eigenvalues cannot be dismissed as meaningless since the probability distributions overlap one another for data collected over disjoint subsets of time. As a technical note, potentially, uncertainty propagation of the eigenvalues

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

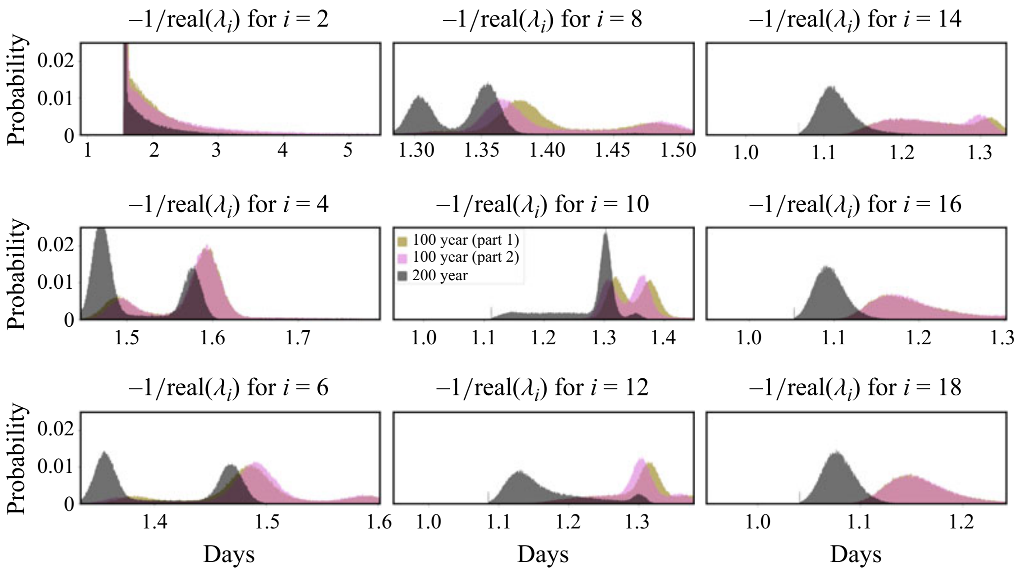

Figure 14. Held–Suarez generator decorrelation time scales. We propagate the uncertainty with respect to the random generator to look at the inferred distribution of eigenvalues. We propagate uncertainty for three cases: the entire 200 years, the first 100 years, and the second 100 years. Furthermore, we display the point estimate of the 200-year generator as calculated from the mean of the random matrix. The slowest decorrelation time scale is the most sensitive to perturbations, and the other eigenvalues are less so.

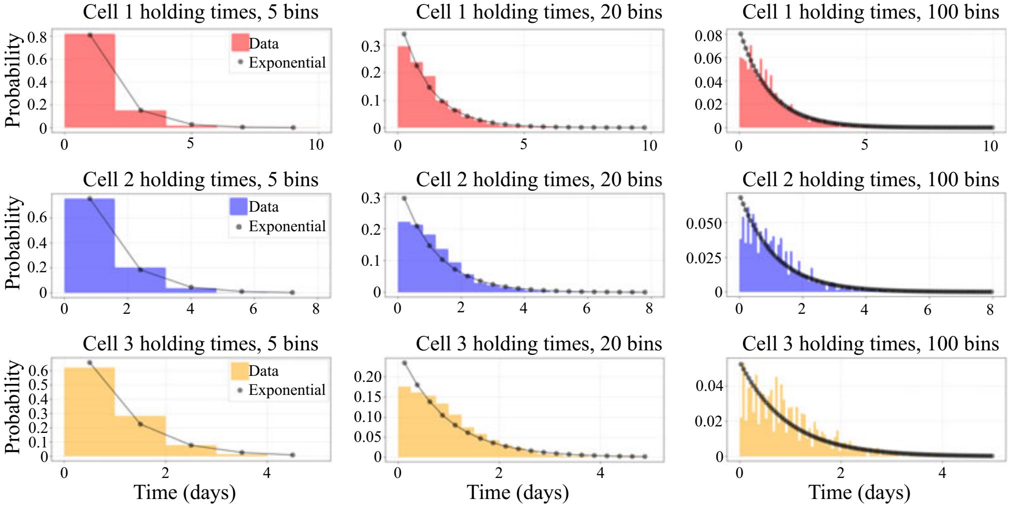

Figure 15. Held–Suarez holding times: continuous-time Markov model versus time series empirical distribution. The amount of time spent in a partition is approximately exponentially distributed.

can be accelerated by using the eigenvalue decomposition of the mean generator as a guess for an iterative procedure.

We show the holding times for the first three most probable partitions in figure 15. Quantiles are approximately exponentially distributed but become imperfect upon closer inspection.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press

References

Colbrook, M.J. 2023 The mpEDMD algorithm for data-driven computations of measure-preserving dynamical systems. *SIAM J. Numer. Anal.* **61** (3), 1585–1608.

Corwin, I. & Shen, H. 2020 Some recent progress in singular stochastic partial differential equations. *Bull. Am. Math. Soc.* **57** (3), 409–454.

Cvitanović, P., Artuso, R., Mainieri, R., Tanner, G. & Vattay, G. 2016 *Chaos: Classical and Quantum*. Niels Bohr Institute.

Froyland, G., Giannakis, D., Lintner, B.R., Pike, M. & Slawinska, J. 2021 Spectral analysis of climate dynamics with operator-theoretic approaches. *Nat. Commun.* **12** (1), 6570.

Geogdzhayev, G., Souza, A.N. & Ferrari, R. 2024 The evolving butterfly: statistics in a changing attractor. *Physica* D*: Nonlinear Phenom.*, 134107.

Giorgini, L.T., Deck, K., Bischoff, T. & Souza, A. 2024 Response theory via generative score modeling. arXiv:2402.01029

Giorgini, L.T., Souza, A.N. & Schmid, P.J. 2023 Clustering of dynamical systems. arXiv:2308.10864

Hairer, M. 2014 A theory of regularity structures. *Invent. Math.* **198** (2), 269–504.

Held, I.M. & Suarez, M.J. 1994 A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Am. Meteorol. Soc.* **75** (10), 1825–1830.

Ho, J., Jain, A. & Abbeel, P. 2020 Denoising diffusion probabilistic models. arxiv:2006.11239

Klus, S., Koltai, Péter & Schütte, C. 2016 On the numerical approximation of the Perron–Frobenius and Koopman operator. *J. Comput. Dyn.* **3** (1), 51–79.

Kopriva, D.A. 2006 Metric identities and the discontinuous spectral element method on curvilinear meshes. *J. Sci. Comput.* **26** (3), 301.

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.

Niegemann, J., Diehl, R. & Busch, K. 2012 Efficient low-storage Runge–Kutta schemes with optimized stability regions. *J. Comput. Phys.* **231**, 364–372.

Ronchi, C., Iacono, R. & Paolucci, P.S. 1996 The 'cubed sphere': a new method for the solution of partial differential equations in spherical geometry. *J. Comput. Phys.* **124** (1), 93–114.

Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P. & Henningson, D.S. 2009 Spectral analysis of nonlinear flows. *J. Fluid Mech.* **641**, 115–127.

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

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 1. Theory and methodology. *J. Fluid Mech.* **997**, A1.

Thiffeault, J.-L. 2023 Ineffective diffusivity. *J. Fluid Mech.* **972**, F1.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N, Kaiser, Ł. & Polosukhin, I. 2017 Attention is all you need. In *Advances in Neural Information Processing Systems* (ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett), vol. 30. Curran Associates.

Wang, H. & Schütte, C. 2015 Building Markov state models for periodically driven non-equilibrium systems. *J. Chem. Theory Comput.* **11** (4), 1819–1831.

Waruszewski, M., Kozdon, J.E., Wilcox, L.C., Gibson, T.H. & Giraldo, F.X. 2022 Entropy stable discontinuous Galerkin methods for balance laws in non-conservative form: applications to the Euler equations with gravity. *J. Comput. Phys.* **468**, 111507.

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.

Winters, A.R., Kopriva, D.A., Gassner, G.J. & Hindenlang, F. 2021 Construction of modern robust nodal discontinuous Galerkin spectral element methods for the compressible Navier–Stokes equations. In *CISM International Centre for Mechanical Sciences, Courses and Lectures,* vol. 602, pp. 117–196.

https://doi.org/10.1017/jfm.2024.657 Published online by Cambridge University Press