Data-driven transition path analysis yields a statistical understanding of sudden stratospheric warming events in an idealized model
Supplementary information

Justin Finkel, Edwin P. Gerber, Dorian S. Abbot, Jonathan Weare

October 4, 2022

This document has three sections. Section 1 spells out transition path theory (TPT) formally, with definitions and equations for all quantities of interest without regard to their numerical approximation. Section 2 describes the numerical method, dynamical Galerkin Approximation (DGA), and provides some numerical benchmarks. Section 3 gives details about the optimization method we used to find a minimum-action path. Finally, section 4 gives a detailed derivation of the enstrophy budget presented in section 5 of the main manuscript.

1 Transition path theory formalism

We begin a quantitative description of transition paths by formalizing the notion of the transition path ensemble. The theoretical development parallels [Vanden-Eijnden, 2006], but expands on it in several ways. Consider the stratosphere, or any other stochastic ergodic dynamical system, evolving through a very long time interval $(-T, T)$, during which it crosses from A to B and back a number M_T of times. As $T \to \infty$, ergodicity guarantees that $M_T \to \infty$ as well. The mth transition path begins at time τ_m^- (so $\mathbf{X}(\tau_m^-) \in A$[1]) and ends at time τ_m^+ (so $\mathbf{X}(\tau_m^+) \in B$). Each τ_m^- marks the beginning of an orange segment in Fig. 2, and τ_m^+ marks the end of it.

In principle, any statistical average over the transition path ensemble can be found by "ergodic simulation" (ES), in which we integrate the system for a long enough time to collect a large number of transition path samples. Although ES is simple and general, it is expensive for high-dimensional models, particularly for rare event simulation. The DGA method, explained below in section 2, circumvents ES by using only short trajectories (20 days long in our implementation). These are short not only compared to the return time $\tau_{m+1}^- - \tau_m^+$ (~ 1700 days for the Holton-Mass model), but even compared to the $(A \to B)$ transit time $\tau_m^+ - \tau_m^-$ (~ 80 days for the Holton-Mass model). DGA can be used to compute an important class of quantities including the committor functions, reactive density, and reactive current. The primitive ingredients of this calculation are *forecast functions*, defined below.

1.1 Forecast functions

The essential insight of TPT is to express the quantities of interest in terms of a set of *forecast functions*. We consider a probabilistic forecast to be an estimate of the future conditioned on the present, of the general form

$$F^+(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[Q(\{(t, \mathbf{X}(t)) : t \geq 0\})]. \tag{1}$$

Here, $\mathbb{E}_{\mathbf{x}}$ indicates a conditional expectation given a fixed initial condition $\mathbf{X}(0) = \mathbf{x}$ (we can set $t_0 = 0$ when assuming autonomous dynamics). Q is a generic functional of the future evolution of the state $\mathbf{X}(t)$. It is explicitly a random variable under the stochastic forcing we impose here, but even in a deterministic model, uncertainty from initial conditions and model error lead to effective randomness. For example, Q could return 1 if $\mathbf{X}(t)$ next hits B before A, and 0 if $\mathbf{X}(t)$ next hits A before B. This makes F^+ simply the forward committor, as introduced in section 3 of the main text:

$$F^+(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_B(\mathbf{X}(\tau_{A\cup B}^+))\right] \tag{2}$$

$$= \mathbb{P}_{\mathbf{x}}\{\mathbf{X}(\tau_{A\cup B}^+) \in B\} =: q_B^+(\mathbf{x}) \tag{3}$$

We might also wish to forecast the time it takes to get there, by defining $Q = \tau_{A\cup B}^+ \mathbb{1}_B(\mathbf{X}(\tau_{A\cup B}^+))$, which then gives us the expected lead time $\eta_B^+(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[Q]/q_B^+(\mathbf{x})$.

[1] Technically, we assume $\mathbf{X}(t)$ is right-continuous with left limits, meaning $\mathbf{X}(\tau_m^-) \notin A$ but $\lim_{t\uparrow\tau_m^-} \mathbf{X}(t) \in A$. This detail is not important for us here.

As explained in section 3 of the main text, the forward committor only looks to the future, and the backward committor is used to estimate where the system came from in the past:

$$q_A^-(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\,\mathbb{1}_A(\mathbf{X}(\tau_{A\cup B}^-))] = \mathbb{P}_{\mathbf{x}}\{\mathbf{X}(\tau_{A\cup B}^-) \in A\} \tag{4}$$

This is a backward-in-time forecast, or *aftcast*.

Forward and backward committors are central components in the existing transition path theory laid out in [E and Vanden-Eijnden, 2006, Vanden-Eijnden, 2006, Metzner et al., 2009, Metzner et al., 2006, Vanden-Eijnden and E, 2010], and elsewhere. Here, we generalize committors to forecast not only where the trajectory ends up, but what happens along the way. We consider forecast/aftcast functions of the form

$$F_\Gamma^+(\mathbf{x};\lambda) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_B(\mathbf{X}(\tau_{A\cup B}^+))\exp\left(\lambda\int_0^{\tau_{A\cup B}^+}\Gamma(\mathbf{X}(r))\,dr\right)\right] \tag{5}$$

$$F_\Gamma^-(\mathbf{x};\lambda) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A(\mathbf{X}(\tau_{A\cup B}^-))\exp\left(\lambda\int_{\tau_{A\cup B}^-}^0\Gamma(\mathbf{X}(r))\,dr\right)\right] \tag{6}$$

where λ is a real free parameter. For certain extreme weather events, Γ might be chosen to measure accumulated damage of some kind, say, the total rainfall deposited over an area (in the case of hurricanes) or total time with surface temperatures above a certain threshold (in the case of heat waves). In a downward-coupled SSW model, one could define Γ to reflect the human impact of extreme cold spells. However, in this paper we only ever set $\Gamma = 1$, so the integral is the expected lead time.

Everything we say about transition paths stems originally from the functions F_Γ^+ and F_Γ^- for various Γ, as well as the steady-state distribution π. Thus, we will now express the quantities of interest in terms of π, F_Γ^+, F_Γ^-, and their λ-derivatives. Section 2 will then explain how to compute them using short simulation data.

1.2 Ergodic averages

The key to transforming forecasts into ensemble averages (at either level) is the *ergodic assumption*, which goes as follows. Let $\mathbf{Y}(t)$ denote all the hidden variables of the system responsible for apparent randomness, such as unresolved turbulence, so that the joint process $(\mathbf{X}(t), \mathbf{Y}(t))$ is completely deterministic. (In simulation, $\mathbf{Y}(t)$ is the state of a pseudo-random number generator.) The hypothesis then states that there exists a probability density $p(\mathbf{x},\mathbf{y})$ such that for any bounded observable $G(\mathbf{X}(t),\mathbf{Y}(t))$,

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T G(\mathbf{X}(t),\mathbf{Y}(t))\,dt \tag{7}$$

$$= \int\left(\int G(\mathbf{x},\mathbf{y})p(\mathbf{x},\mathbf{y})\,d\mathbf{y}\right)\pi(\mathbf{x})\,d\mathbf{x}$$

$$=: \int\Gamma(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} =: \langle\Gamma\rangle_\pi. \tag{8}$$

As a prototypical example, define

$$G(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & (\mathbf{x},\mathbf{y}) \text{ comes from } A \text{ and goes to } B \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where knowledge of $\mathbf{y}$ lets us run the system forward and backward deterministically to evaluate the source and destination of $(\mathbf{x},\mathbf{y})$. In this case, Eq. (7) becomes

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{1}_A(\mathbf{X}(\tau_{A\cup B}^-(t)))\mathbb{1}_B(\mathbf{X}(\tau_{A\cup B}^+(t))\,dt$$

$$= \int q_A^-(\mathbf{x})q_B^+(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} =: \langle q_A^- q_B^+\rangle_\pi. \tag{10}$$

The left-hand side is the time fraction spent on the way from A to B, and can be estimated from ES. On the right hand side, $\Gamma(\mathbf{x}) = q_A^+(\mathbf{x})q_B^+(\mathbf{x})$. Substituting different combinations of A and B in Eq. (10) gives us the time fraction spend in the phases $B\to A$ ($\Gamma = q_B^- q_A^+$), $A\to A$ ($\Gamma = q_A^- q_A^+$), and $B\to B$ ($\Gamma = q_B^- q_B^+$). Fig. 1a shows these estimates from ES and DGA, which is further described in section 2 below.

1.3 Transition path averages and currents

In this section we use forecast functions (5) and (6) to express the transition rate and reactive current. In fact, we will go slightly further and define a *generalized rate*:

$$R_\Gamma(\lambda) := \lim_{T\to\infty} \frac{1}{T} \sum_{m=1}^{M_T} \exp\left(\lambda \int_{\tau_m^-}^{\tau_m^+} \Gamma(\mathbf{X}(r))\, dr\right) \tag{11}$$

The notation emphasizes that R_Γ depends on the observable Γ and the real parameter λ. To unpack this formula, first set $\lambda = 0$ and observe that $R_\Gamma(0) = \frac{M_T}{T}$ is the number of transitions per unit time, or ordinary rate, whose inverse is the average period of the full SSW life cycle.[2]

The ordinary rate has been studied extensively with TPT and preceding theories. A novel idea that we introduce here is to include the exponential factor $\exp\left(\lambda \int \Gamma(\mathbf{X}(t))\, dt\right)$, though we do not present these results in this paper. The theoretical development below therefore reduces to classical TPT by replacing Γ with 0.

Returning to (11), we divide through by $R_\Gamma(0)$:

$$\begin{aligned}\frac{R_\Gamma(\lambda)}{R_\Gamma(0)} &= \lim_{T\to\infty} \frac{1}{M_T} \sum_{m=1}^{M_T} \exp\left(\lambda \int_{\tau_m^-}^{\tau_m^+} \Gamma(\mathbf{X}(r))\, dr\right) \\ &= \mathbb{E}_{\text{paths}}\left[\exp\left(\lambda \int_{\tau_A^-}^{\tau_B^+} \Gamma(\mathbf{X}(r))\, dr\right)\right] \end{aligned} \tag{13}$$

where the subscript "paths" distinguishes the expectation as over all transition paths. The right side of (13) is a moment-generating function for the integral $\int \Gamma(\mathbf{X}(t))\, dt$ over the transition path. Differentiating in λ yields the moments of the distribution of the integral, allowing one to calculate variance, skew, and kurtosis:

$$\frac{\partial_\lambda^k R_\Gamma(0)}{R_\Gamma(0)} = \mathbb{E}_{\text{paths}}\left[\left(\int_{\tau_A^-}^{\tau_B^+} \Gamma(\mathbf{X}(r))\, dr\right)^k\right], \tag{14}$$

Thus, $R_\Gamma(\lambda)$ contains much information about the transition ensemble.

We now express R_Γ in terms of the forecast functions F_Γ^+ and F_Γ^-, again using the key assumption of ergodicity. We must convert Equation (11), a sum over transition paths $\sum_{m=1}^{M_T}(\cdot)$, into an integral over time $\int_{-T}^{T}(\cdot)\, dt$ and then (by ergodicity) into an integral over space $\int_{\mathbb{R}^d}(\cdot)\pi(\mathbf{x})\, d\mathbf{x}$. This approach extends the rate derivation in [Vanden-Eijnden, 2006] and [Strahan et al., 2021] to generalized rates.

To write the rate as a time integral, we introduce a dividing surface between A and B (such as a committor level surface) and use the fact that a transition path crosses such a surface an odd number of times. A mask is applied to the time integral to select only the time segments when a reactive trajectory segment is crossing this surface (+1 for positive crossings and −1 for negative crossings), resulting in unit weight for each transition path. To be more explicit, let S be a region of state space that contains A and excludes B, so that its boundary $C = \partial S$ is a dividing surface between A and B. The generalized rate (11) can then be written as the following time integral:

$$R_\Gamma(\lambda) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \lim_{T\to\infty} \frac{1}{T} \int_0^T \tag{15}$$

$$\exp\left(\lambda \int_{\tau_{A\cup B}^-(t)}^{\tau_{A\cup B}^+(t+\Delta t)} \Gamma(\mathbf{X}(r))\, dr\right)\times \tag{16}$$

$$\mathbb{1}_A(\mathbf{X}(\tau_{A\cup B}^-(t)))\, \mathbb{1}_B(\mathbf{X}(\tau_{A\cup B}^+(t+\Delta t)))\times \tag{17}$$

$$\Big[\mathbb{1}_S(\mathbf{X}(t))\, \mathbb{1}_{S^c}(\mathbf{X}(t+\Delta t)) \tag{18}$$

$$-\mathbb{1}_{S^c}(\mathbf{X}(t))\, \mathbb{1}_S(\mathbf{X}(t+\Delta t))\Big]\, dt \tag{19}$$

[2]This is not to be confused with the asymmetric forward and backward rates,

$$k_{AB} = \frac{R_\Gamma(0)}{\langle q_A^-\rangle_\pi}, \qquad k_{BA} = \frac{R_\Gamma(0)}{\langle q_B^-\rangle_\pi} \tag{12}$$

which distinguish the $A \to B$ and $B \to A$ directions by how fast they occur. The factor $\langle q_A^-\rangle_\pi$ is the time fraction spent *having last been in A* rather than B, and $\langle q_B^-\rangle$ is the opposite. For example, if A were very stable and B very unstable, the system would spend most of its time in the basin of attraction of A, making $\langle q_A^-\rangle_\pi$ large and $k_{AB} \ll k_{BA}$. Asymmetric rates (or "rate constants") are very important for chemistry applications, but the symmetric rate is more useful to us presently.

The idea is to restrict the interval $(0,T)$ to the collection of time intervals $(t,t+\Delta t)$ during which the path crosses the surface ∂S. Line (17) applies a mask picking out transition path segments, which are those that come from A and next go to B. Line (18) applies a further mask picking out the narrow time intervals when $\mathbf{X}(t)$ exits the region from S to S^{c}, while line (19) subtracts the backward crossings from S^{c} to S. Using ergodicity, we can replace the time integral with a space integral and insert conditional expectations inside. For example, the part of the integrand

$$\exp\left(\lambda \int_{t+\Delta t}^{\tau^+_{A\cup B}(t+\Delta t)} \Gamma(\mathbf{X}(r))\, dr\right)\times \\ \mathbb{1}_B(\mathbf{X}(\tau^+_{A\cup B}(t+\Delta t)))\,\mathbb{1}_{S^{\mathrm{c}}}(\mathbf{X}(t+\Delta t)) \tag{20}$$

becomes, after taking conditional expectations,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{S^{\mathrm{c}}}(\mathbf{X}(t+\Delta t))F^+_\Gamma(\mathbf{X}(t+\Delta t))\middle|\mathbf{X}(t)=\mathbf{x}\right] \\ &\quad =: \mathcal{T}^{\Delta t}\left[\mathbb{1}_{S^{\mathrm{c}}}F^+_\Gamma\right](\mathbf{x})\end{aligned} \tag{21}$$

Where the *transition operator* is defined as $\mathcal{T}^{\Delta t} f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{X}(\Delta t))]$. Applying similar logic to all terms in the integrand, we have the following generalized rate formula:

$$R_\Gamma(\lambda) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\mathbb{R}^d} F^-_\Gamma(\mathbf{x};\lambda)\times \\ \left\{\mathbb{1}_S\mathcal{T}^{\Delta t}\left[\mathbb{1}_{S^{\mathrm{c}}}F^+_\Gamma\right] - \mathbb{1}_{S^{\mathrm{c}}}\mathcal{T}^{\Delta t}\left[\mathbb{1}_S F^+_\Gamma\right]\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{22}$$

which holds for any S enclosing A and disjoint from B. This is a form estimable from short simulation data, which the next section will explain.

The rate formula (22) is suggestive of a surface integral, counting hopping events across the surface ∂S. In fact, the reactive current $\mathbf{J}_{AB}$ is defined as the vector field whose surface integral is equal to the symmetric rate:

$$R_\Gamma(0) = \int_C \mathbf{J}_{AB}\cdot\mathbf{n}\, d\sigma \tag{23}$$

We have visualized $\mathbf{J}_{AB}$ in sections 3 and 5 of the main text using a discretization of the integrand in (22).

We have now completely described the mathematics of TPT, and our extensions to it. Exact knowledge of π, F^+_Γ, F^-_Γ, and their λ-derivatives is enough to generate all of the figures shown so far. The next section explains both how to compute these fundamental ingredients from data and assemble them into generalized rates.

2 Numerical method: dynamical Galerkin approximation (DGA)

2.1 Feynman-Kac formulae

We now sketch the numerical method, following [Thiede et al., 2019, Strahan et al., 2021], and [Finkel et al., 2021]. Equations (5) and (6) involve an integral in time all the way from $t=0$ to $t=\tau^+_{A\cup B}$ or (in backward time) to $\tau^-_{A\cup B}$, when $\mathbf{X}(t)$ hits either A or B after wandering through state space for an indeterminate period. This would seem to require long trajectories to estimate. However, below we write $F^\pm_\Gamma$ as solutions to partial differential equations called Feynman-Kac formulae [Oksendal, 2003, Karatzas and Shreve, 1998, E et al., 2019], which read

$$\begin{cases}(\mathcal{L}+\lambda\Gamma(\mathbf{x}))F^+_\Gamma(\mathbf{x};\lambda)=0 & \mathbf{x}\in D\\ F^+_\Gamma(\mathbf{x})=0 & \mathbf{x}\in A\\ F^+_\Gamma(\mathbf{x})=1 & \mathbf{x}\in B\end{cases} \tag{24}$$

$$\text{where } \mathcal{L}\phi(\mathbf{x}) := \lim_{\Delta t\to 0}\frac{\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{X}(\Delta t))]-\phi(\mathbf{x})}{\Delta t} \tag{25}$$

$$\begin{cases}(\widetilde{\mathcal{L}}+\lambda\Gamma(\mathbf{x}))F^-_\Gamma(\mathbf{x};\lambda)=0 & \mathbf{x}\in D\\ F^-_\Gamma(\mathbf{x})=1 & \mathbf{x}\in A\\ F^-_\Gamma(\mathbf{x})=0 & \mathbf{x}\in B\end{cases} \tag{26}$$

$$\text{where } \widetilde{\mathcal{L}}\phi(\mathbf{x}) := \lim_{\Delta t\to 0}\frac{\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{X}(-\Delta t))]-\phi(\mathbf{x})}{\Delta t} \tag{27}$$

The linear operators $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ are known as the forward and backward infinitesimal generators, pushing observable functions ϕ forward or backward in time analogously to a material derivative in fluid mechanics. The first term in the numerator of (25) is the transition operator. The backward-in-time expectations are defined specifically for the *equilibrium* process, leading to $\widetilde{\mathcal{L}}\phi(\mathbf{x}) = \frac{1}{\pi(\mathbf{x})}\mathcal{L}^*[\pi\phi](\mathbf{x})$, where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$ with respect to the reference (Lebesgue) measure $d\mathbf{x}$. Equivalently, $\widetilde{\mathcal{L}}$ is the adjoint of $\mathcal{L}$ with respect to the steady-state measure $\pi(\mathbf{x})\,d\mathbf{x}$. In addition, we have the stationary Fokker-Planck equation for π itself:

$$\begin{cases} \mathcal{L}^*\pi(\mathbf{x}) = 0 & \mathbf{x} \in \mathbb{R}^d \\ \int_{\mathbb{R}^d} \pi(\mathbf{x})\,d\mathbf{x} = 1 \end{cases} \tag{28}$$

We can further obtain equations for the derivatives of $F_\Gamma^\pm$ with respect to λ, using the Kac moment method [Fitzsimmons and Pitman, 1999] as also described in [Finkel et al., 2021]. Differentiating Eq. (24) in λ and setting $\lambda = 0$ yields a recursive sequence of equations for the higher derivatives of F_Γ^+:

$$\mathcal{L}\left[\partial_\lambda^k F_\Gamma^+\right](\mathbf{x};0) = -k\Gamma\partial_\lambda^{k-1}F_\Gamma^+(\mathbf{x};0), \qquad k \geq 1, \tag{29}$$

with boundary conditions $\partial_\lambda^k F_\Gamma^+\big|_{A\cup B} = 0$. The same procedure can be applied to the aftcast F_Γ^-. Thus, our entire numerical pipeline boils down to solving equations of the form (24), (26), and (28), as well as the inhomogeneous version (29). In particular, the expected lead time η_B^+ is equal to $\frac{1}{q_B^+(\mathbf{x})}\partial_\lambda F_\Gamma^+(\mathbf{x};0)$ where $\Gamma(\mathbf{x}) \equiv 1$. In other words,

$$\mathcal{L}\left[q_B^+\eta_B^+\right](\mathbf{x}) = -q_B^+(\mathbf{x}) \tag{30}$$

and the expected lead time η_B^+ can only be computed after first computing the committor q_B^+.

Section 5 additionally defined the *stochastic tendency* $\mathcal{L}_{AB}\phi$ for any observable ϕ. The idea is to turn the expectations in Eqs. (25) and (27) into conditional expectations, given that a transition is underway through $\mathbf{x}$. As such, $\mathcal{L}_{AB}$ is defined so that

$$\mathcal{L}_{AB}\phi(\mathbf{x}) = \lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\frac{\phi(\mathbf{X}(\Delta t)) - \phi(\mathbf{X}(-\Delta t))}{2\Delta t}\middle|\mathbf{X}(\tau_{A\cup B}^-) \in A \text{ and } \mathbf{X}(\tau_{A\cup B}^+) \in B\right] \tag{31}$$

To reveal its connection with the reactive current, and to justify the discretization to follow, we manipulate this formula to express $\mathcal{L}_{AB}$ in terms of committors. Above, $\tau_{A\cup B}^+$ is shorthand for the first hitting time to $A\cup B$ after $t = 0$, and $\tau_{A\cup B}^-$ is shorthand for the most recent hitting time to $A\cup B$ before $t = 0$. In the following, we abbreviate $\tau_{A\cup B}^\pm$ by $\tau^\pm$ to reduce clutter, and use the fact that $\tau^+(0) = \tau^+(\Delta t)$ and $\tau^-(0) = \tau^-(-\Delta t)$ for small enough Δt and for $\mathbf{x}$ sufficiently far from $A\cup B$.

$$\mathcal{L}_{AB}\phi(\mathbf{x}) = \lim_{\Delta t\to 0}\frac{1}{2\Delta t}\frac{\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A(\mathbf{X}(\tau^-))\mathbb{1}_B(\mathbf{X}(\tau^+))\big(\phi(\mathbf{X}(\Delta t)) - \phi(\mathbf{X}(-\Delta t))\big)\right]}{\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A(\mathbf{X}(\tau^-))\mathbb{1}_B(\mathbf{X}(\tau^+))\right]} \tag{32}$$

$$= \frac{1}{q_A^-(\mathbf{x})q_B^+(\mathbf{x})}\lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A(\mathbf{X}(\tau^-))\mathbb{1}_B(\mathbf{X}(\tau^+))\frac{\phi(\mathbf{X}(\Delta t)) - \phi(\mathbf{X}(-\Delta t))}{2\Delta t}\right] \tag{33}$$

$$= \frac{1}{q_A^-(\mathbf{x})q_B^+(\mathbf{x})}\frac{1}{2}\left\{\lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A\Big(\mathbf{X}\big(\tau^-(0)\big)\Big)\mathbb{1}_B\Big(\mathbf{X}\big(\tau^+(\Delta t)\big)\Big)\frac{\phi(\mathbf{X}(\Delta t)) - \phi(\mathbf{x})}{\Delta t}\right]\right. \tag{34}$$

$$\left. + \lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A\Big(\mathbf{X}\big(\tau^-(-\Delta t)\big)\Big)\mathbb{1}_B\Big(\mathbf{X}\big(\tau^+(0)\big)\Big)\frac{\phi(\mathbf{x}) - \phi(\mathbf{X}(-\Delta t))}{\Delta t}\right]\right\}$$

$$= \frac{1}{q_A^-(\mathbf{x})q_B^+(\mathbf{x})}\frac{1}{2}\left\{\lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_A\Big(\mathbf{X}\big(\tau^-(0)\big)\Big)\frac{\mathbb{1}_B(\mathbf{X}(\tau^+(\Delta t)))\phi(\mathbf{X}(\Delta t)) - \mathbb{1}_B(\mathbf{X}(\tau^+(0)))\phi(\mathbf{x})}{\Delta t}\right]\right. \tag{35}$$

$$\left. + \lim_{\Delta t\to 0}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_B\Big(\mathbf{X}\big(\tau^+(0)\big)\Big)\frac{\mathbb{1}_A(\mathbf{X}(\tau^-(0)))\phi(\mathbf{x}) - \mathbb{1}_A(\mathbf{X}(\tau^-(-\Delta t)))\phi(\mathbf{X}(-\Delta t))}{\Delta t}\right]\right\}$$

$$= \frac{1}{q_A^-(\mathbf{x})q_B^+(\mathbf{x})}\frac{1}{2}\left\{q_A^-(\mathbf{x})\mathcal{L}\left[q_B^+\phi\right](\mathbf{x}) - q_B^+\widetilde{\mathcal{L}}\left[q_A^-\phi\right](\mathbf{x})\right\} \tag{36}$$

$$= \frac{1}{2}\left\{\frac{\mathcal{L}[q_B^+\phi](\mathbf{x})}{q_B^+(\mathbf{x})} - \frac{\widetilde{\mathcal{L}}[q_A^-\phi](\mathbf{x})}{q_A^-(\mathbf{x})}\right\} \tag{37}$$

Now consider the average of $\mathcal{L}_{AB}\phi(\mathbf{x})$ with respect to the reactive density π_{AB} restricted to a region $\mathscr{R}$ of state space, e.g., near a certain level set of the committor. We can express such an average as

$$\langle \mathcal{L}_{AB}\phi \rangle_{\pi_{AB}|\mathscr{R}} = \frac{\int_{\mathscr{R}} \mathcal{L}_{AB}\phi(\mathbf{x}) q_A^-(\mathbf{x}) q_B^+(\mathbf{x}) \pi(\mathbf{x})\, d\mathbf{x}}{\int_{\mathscr{R}} q_A^-(\mathbf{x}) q_B^+(\mathbf{x}) \pi(\mathbf{x})\, d\mathbf{x}} = \frac{1}{2} \frac{\int_{\mathscr{R}} \left(q_A^- \mathcal{L}[q_B^+\phi] - q_B^+ \mathcal{L}[q_A^-\phi] \right) \pi\, d\mathbf{x}}{\int_{\mathscr{R}} q_A^- q_B^+ \pi\, d\mathbf{x}} \tag{38}$$

In [Strahan et al., 2021], Eq. S12, it is shown that

$$\mathbf{J}_{AB} \cdot \nabla\phi = \frac{1}{2}\Big(q_A^- \mathcal{L}\big[q_B^+\phi\big] - q_B^+ \widetilde{\mathcal{L}}\big[q_A^-\phi\big] \Big)\pi \tag{39}$$

which means that

$$\langle \mathcal{L}_{AB}\phi \rangle_{\pi_{AB}|\mathscr{R}} = \frac{\int_{\mathscr{R}} \mathbf{J}_{AB} \cdot \nabla\phi\, d\mathbf{x}}{\int_{\mathscr{R}} q_A^- q_B^+ \pi\, d\mathbf{x}} \tag{40}$$

In words, this ratio conveys the total flow of reactive current across contours of ϕ in a region $\mathscr{R}$, adjusted for the likelihood for transition paths to be in $\mathscr{R}$ to begin with. The denominator ensures that regions with slow and fast movement of transition paths are compared fairly.

2.2 Discretization of Feynman-Kac formulae

We will now describe how to discretize and solve these three equations, which requires three similar but distinct procedures.

First we attack (24). The generator of a diffusion processes can be expressed as a partial differential operator, and so the above equations are PDEs over state space. PDEs cannot be practically discretized in high dimensions, but the essential property of spatial locality allows for data-driven approximation with short trajectories, using the probabilistic definition in (25) and (27). This is how we use our large data set of short trajectories,

$$\{\mathbf{X}_n(t) : 0 \le t \le \Delta t\}_{n=1}^N, \tag{41}$$

where the initial points $\mathbf{X}_n(0)$ are drawn from a *sampling measure* μ, which we will define in the following subsection. To discretize Eq. (24), we first eliminate the numerically problematic limit $\Delta t \to 0$ and integrate the equation using Dynkin's Formula [Oksendal, 2003, E et al., 2019]: for any stopping time $\theta > 0$,

$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{X}(\theta))] = f(\mathbf{x}) + \mathbb{E}_{\mathbf{x}}\left[\int_0^\theta \mathcal{L} f(\mathbf{X}(t))\, dt \right] \tag{42}$$

Here we take $\theta = \min(\Delta t, \tau_{A\cup B})$. In other words, we artificially halt the nth trajectory $\mathbf{X}_n(t)$ if it wanders into $A \cup B$ before the terminal time Δt. The nth stopping time from the data set is called θ_n. The operator on the left-hand side of (42) is known as the *stopped* transition operator $\mathcal{T}^\theta$. Applying it to the unknown forecast function F_Γ^+ in (24) and using the fact $\mathcal{L}F_\Gamma^+ = -\lambda\Gamma F_\Gamma^+$, we get

$$\mathcal{T}^\theta F_\Gamma^+(\mathbf{x};\lambda) = F_\Gamma^+(\mathbf{x};\lambda) - \lambda \mathbb{E}_{\mathbf{x}}\left[\int_0^\theta \Gamma(\mathbf{X}(t)) F_\Gamma^+(\mathbf{X}(t);\lambda)\, dt \right] \tag{43}$$

To be more concise, we define the integral operator $\mathcal{K}_\Gamma^\theta f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\big[\int_0^\theta \Gamma(\mathbf{X}(t)) f(\mathbf{X}(t))\, dt \big]$, and write an integrated version of (24):

$$\begin{cases} (\mathcal{T}^\theta - 1 + \lambda\mathcal{K}_\Gamma^\theta) F_\Gamma^+(\mathbf{x};\lambda) = 0 & \mathbf{x} \in D \\ F_\Gamma^+(\mathbf{x};\lambda) = 0 & \mathbf{x} \in A \\ F_\Gamma^+(\mathbf{x};\lambda) = 1 & \mathbf{x} \in B \end{cases} \tag{44}$$

To discretize this equation and impose regularity on the solution, we approximate F_Γ^+ as a finite linear combination with coefficients $w_j(F_\Gamma^+(\mathbf{x};\lambda))$, which we abbreviate $w_j(\lambda)$ for simplicity:

$$F_\Gamma^+(\mathbf{x};\lambda) \approx \hat{F}_\Gamma^+(\mathbf{x};\lambda) + \sum_{j=1}^M w_j(\lambda)\phi_j(\mathbf{x};\lambda) \tag{45}$$

where $\hat{F}^+_\Gamma$ is a guess function obeying the boundary conditions on $A \cup B$, and $\{\phi_j\}_{j=1}^M$ is a collection of basis functions that are zero on $A \cup B$, which will be defined in the following subsection. The task is now to solve for the coefficients $w_j(\lambda)$. Equation (44) becomes a system of linear equations in $w_j(\lambda)$:

$$\sum_{j=1}^{M} w_j(\lambda)(\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\phi_j(\mathbf{x};\lambda) = -(\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\hat{F}^+_\Gamma(\mathbf{x};\lambda) \tag{46}$$

Since the transfer and integral operators are expectations over the future state of the system beginning at $\mathbf{x}$, we can estimate their action at $\mathbf{x} = \mathbf{X}_n(0)$ (a short-trajectory starting point) as

$$\begin{aligned}(\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\phi_j(\mathbf{X}_n(0);\lambda) \approx \phi_j(\mathbf{X}_n(\theta_n);\lambda) - \phi_j(\mathbf{X}_n(0);\lambda) \\ + \lambda\int_0^{\theta_n} \Gamma(\mathbf{X}_n(t))\phi_j(\mathbf{X}_n(t);\lambda)\,dt\end{aligned} \tag{47}$$

or, if multiple independent trajectories are launched from $\mathbf{x}$, we can average over them. Applying this to every short trajectory and plugging into Eq. (46), we obtain a system of N equations in M unknowns. In practice, $N \gg M$, meaning we have many more trajectories than basis functions, and the system is overdetermined. A unique, and regularized, solution is obtained by casting it into weak form: we multiply both sides by $\phi_i(\mathbf{x})$ and integrate over state space:

$$\sum_{j=1}^{M} w_j(\lambda)\left\langle \phi_i, (\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\phi_j \right\rangle_\zeta = -\left\langle \phi_i, (\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\hat{F}^+_\Gamma \right\rangle_\zeta \tag{48}$$

where the inner products are defined with respect to a measure ζ:

$$\langle f, g\rangle_\zeta = \int f(\mathbf{x})g(\mathbf{x})\zeta(\mathbf{x})\,d\mathbf{x} \tag{49}$$

With our finite data set, we approximate the inner product by a sum over pairs of points. Given that $\mathbf{X}_n(0) \sim \mu$, the law of large numbers ensures that for any bounded function $H(\mathbf{x})$,

$$\frac{1}{N}\sum_{n=1}^{N} H(\mathbf{X}_n(0)) \approx \int H(\mathbf{x})\mu(\mathbf{x})\,d\mathbf{x} \tag{50}$$

becomes more accurate as $N \to \infty$. Thus we set $H(\mathbf{x}) = \phi_i(\mathbf{x})(\mathcal{T}^\theta - 1 + \lambda\mathcal{K}^\theta_\Gamma)\phi_j(\mathbf{x})$ as estimated by (47), approximate the inner products with $\zeta = \mu$, plug them into (48), and solve the $M \times M$ system of linear equations for $w_j(\lambda)$.

Next we address (28). The integrated version of (28) is found by observing that for any bounded function H,

$$\mathbb{E}_{\mathbf{X}(0)\sim\pi}[H(\mathbf{X}(\Delta t))] = \mathbb{E}_{\mathbf{X}(\Delta t)\sim\pi}[H(\mathbf{X}(\Delta t))] \tag{51}$$

where $\mathbf{X}(0) \sim \pi$ means the initial condition is drawn from equilibrium, and thus so is $\mathbf{X}(\Delta t)$ since π is stationary. Of course, our initial data is *not* distributed according to π, but rather by μ; the goal is to solve for the *change of measure* $\frac{d\pi}{d\mu}(\mathbf{x})$. Writing (51) as an integral,

$$\int \mathcal{T}^{\Delta t}H(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} = \int H(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} \tag{52}$$

$$0 = \int (\mathcal{T}^{\Delta t} - 1)H(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} \tag{53}$$

$$= \int (\mathcal{T}^{\Delta t} - 1)H(\mathbf{x})\frac{d\pi}{d\mu}(\mathbf{x})\mu(\mathbf{x})\,d\mathbf{x} \tag{54}$$

$$= \left\langle (\mathcal{T}^{\Delta t} - 1)H, \frac{d\pi}{d\mu}\right\rangle_\mu \tag{55}$$

As this holds for every bounded H, we enforce the equation for $H = \phi_i$, $i = 1, \ldots, M$ and approximate $\frac{d\pi}{d\mu} = \sum_{j=1}^M w_j\left(\frac{d\pi}{d\mu}\right)\phi_j$, resulting in a homogeneous linear system for the coefficients w_j similar to (48). The matrix elements are

$$\langle(\mathcal{T}^{\Delta t} - 1)\phi_i, \phi_j\rangle_\mu = \langle\phi_i, (\mathcal{T}^{\Delta t} - 1)^*_\mu\phi_j\rangle_\mu \tag{56}$$

where $(\cdot)^*_\mu$ denotes the adjoint operator with respect to μ. We solve this homogeneous system by QR decomposition. Note that there are no boundary conditions, and the trajectories need not be stopped early. Instead there is a normalization condition, which we enforce as $\sum_{n=1}^N \frac{d\pi}{d\mu}(\mathbf{X}_n(0)) = 1$. If we were to divide by Δt and take the limit $\Delta t \to 0$, we would recover the strong form of the Fokker-Planck equation, (28).

We use a particularly simple basis consisting of indicator functions,

$$\phi_j(\mathbf{x}) = \mathbb{1}_{S_j}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_j \\ 0 & \mathbf{x} \notin S_j \end{cases} \tag{57}$$

where $\{S_1, \ldots, S_M\}$ is a disjoint partition of state space. We obtain these regions by clustering data points with a hierarchical version of the K-means algorithm implemented in `scikit-learn` [Pedregosa et al., 2011], with $M = 1500$. With such a basis, the inner products yield the entries of the Markov matrix P_{ij} described in the main text. This guarantees a null space automatically. In a more general basis, one can add a constant vector to ensure a nontrivial null vector.

Given the weights $\frac{d\pi}{d\mu}(\mathbf{X}_n(0))$, we can take any ergodic average Γ by inserting the change of measure:

$$\langle \Gamma \rangle_\pi = \int_{\mathbb{R}^d} \Gamma(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^d} \Gamma(\mathbf{x})\, d\mathbf{x} \approx \sum_{n=1}^N \Gamma(\mathbf{X}_n(0)) \frac{d\pi}{d\mu}(\mathbf{X}_n(0)) \tag{58}$$

For the specific case of a Markov state model, we can estimate the π-weighted state-space integral by decomposing it over clusters:

$$\begin{aligned}
\langle \Gamma \rangle_\pi &= \sum_{j=1}^M \int_{S_j} \Gamma(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} && (59)\\
&\approx \sum_{j=1}^M w_j(\pi) \times (\text{average of } \Gamma \text{ over } S_j) && (60)\\
&\approx \sum_{j=1}^M w_j(\pi) \frac{\sum_{n=1}^N \Gamma(\mathbf{X}_n(0)) \mathbb{1}_{S_j}(\mathbf{X}_n(0))}{\sum_{n'=1}^N \mathbb{1}_{S_j}(\mathbf{X}_{n'}(0))} && (61)\\
&= \sum_{n=1}^N \Gamma(\mathbf{X}_n(0)) \frac{w_{j(n)}(\pi)}{\#\{n' : j(n') = j(n)\}} && (62)
\end{aligned}$$

where $j(n)$ is defined as the cluster that $\mathbf{X}_n(0)$ is assigned to, i.e., $\mathbf{X}_n(0) \in S_{j(n)}$. Hence the change of measure for a Markov state model is the coefficient of $\Gamma(\mathbf{X}_n(0))$ in the last equation. Note that it sums to one over all data points, as it must.

Finally, we address the time-reversed Kolmogorov equation (26). The only modification from (24) is that the inner products in (48) are interpreted in backward time, i.e., with all trajectories reversed, $\mathbf{X}_n(\Delta t)$ becoming the beginning and $\mathbf{X}_n(0)$ becoming the end. The problem is that $\mathbf{X}_n(\Delta t)$ is not distributed according to μ, and so we cannot use the same Monte Carlo inner product as in Eq. (50) with reference measure $\zeta = \pi$. However, we can solve the problem by reweighting with the change of measure as follows, leading to $\zeta = \pi$. We let the trajectory be discrete in time, i.e.,

$$\mathbf{X}_n = \left[\mathbf{X}_n(0), \mathbf{X}_n\left(\frac{\Delta t}{K}\right), \mathbf{X}_n\left(\frac{2\Delta t}{K}\right), \ldots, \mathbf{X}_n(\Delta t)\right] \tag{63}$$

and consider functionals $H[\mathbf{X}_n]$ of the whole trajectory. Defining the transition density $p(\mathbf{x}, \mathbf{y})$ for each step of size Δt, the expectation of H with $\mathbf{X}_n(0) \sim \pi$ is given by

$$\mathbb{E}_{\mathbf{X}(0)\sim\pi} H[\mathbf{X}] = \int d\mathbf{x}_0 \pi(\mathbf{x}_0) \int d\mathbf{x}_1 p(\mathbf{x}_0, \mathbf{x}_1) \int \ldots \int d\mathbf{x}_K\, p(\mathbf{x}_{K-1}, \mathbf{x}_K) H[\mathbf{x}_0, \ldots, \mathbf{x}_K] \tag{64}$$

The time reversal step explicitly assumes the *equilibrium* backward process, leading to a backward transition kernel $\widetilde{p}(\mathbf{y}, \mathbf{x}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{y})} p(\mathbf{x}, \mathbf{y})$. Inserting this throughout converts the expectation over $\mathbf{X}(0)$ into an expectation over $\mathbf{X}(\Delta t)$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}(0)\sim\pi} H[\mathbf{X}] &= \int d\mathbf{x}_0 \pi(\mathbf{x}_0) \int d\mathbf{x}_1 \frac{\pi(\mathbf{x}_1)}{\pi(\mathbf{x}_0)} \widetilde{p}(\mathbf{x}_1, \mathbf{x}_0) \int && (65)\\
&\quad \ldots \int d\mathbf{x}_K \frac{\pi(\mathbf{x}_K)}{\pi(\mathbf{x}_{K-1})} \widetilde{p}(\mathbf{x}_K, \mathbf{x}_{K-1}) H[\mathbf{x}_0, \ldots, \mathbf{x}_K] && (66)\\
&= \int d\mathbf{x}_K \pi(\mathbf{x}_K) \int d\mathbf{x}_{K-1} \widetilde{p}(\mathbf{x}_K, \mathbf{x}_{K-1}) \int \ldots \int d\mathbf{x}_0 \widetilde{p}(\mathbf{x}_1, \mathbf{x}_0) H[\mathbf{x}_0, \ldots, \mathbf{x}_K] && (67)\\
&= \widetilde{\mathbb{E}}_{\mathbf{X}(\Delta t)\sim\pi} H[\mathbf{X}] && (68)
\end{aligned}$$

where $\widetilde{\mathbb{E}}$ denotes backward-in-time expectation. This is precisely what we need to apply (50) to the time-reversed process, namely, define H such that

$$\phi_i(\mathbf{X}(\Delta t))(\mathcal{T}^\theta - 1 + \lambda \mathcal{K}_\Gamma^\theta)\phi_j(\mathbf{X}(\Delta t)) =: H[\mathbf{X}] \tag{69}$$

and then integrate over state space weighted by π, turning the left-hand side into an inner product:

$$\langle \phi_i, (\widetilde{\mathcal{T}}^\theta - 1 + \lambda \widetilde{\mathcal{K}}_\Gamma^\theta)\phi_j \rangle_\pi = \widetilde{\mathbb{E}}_{\mathbf{X}(\Delta t)\sim\pi} H[\mathbf{X}] \tag{70}$$

$$= \mathbb{E}_{\mathbf{X}(0)\sim\pi} H[\mathbf{X}] \approx \sum_{n=1}^{N} H[\mathbf{X}_n] \frac{d\pi}{d\mu}(\mathbf{X}_n(0)) \tag{71}$$

The right-hand side of Eq. (50) can be estimated similarly, also with $\zeta = \pi$.

In both forward- and backward-time estimates, we never solve for $F_\Gamma^+(\mathbf{x};\lambda)$ or $F_\Gamma^-(\mathbf{x};\lambda)$ with nonzero λ; rather, we repeat the recursion process with Eq. (29). This is equivalent to implicitly differentiating the discretized system Eq. (48).

2.3 Rate estimate and numerical benchmarking

To estimate generalized rates (in particular, the ordinary rate), we reproduce here the rate estimate from [Strahan et al., 2021] for reference, which is an almost-direct implementation of the formula (22), repeated here:

$$R_\Gamma(\lambda) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\times \tag{72}$$
$$\left\{ \mathbb{1}_S \mathcal{T}^{\Delta t}\left[\mathbb{1}_{S^c} F_\Gamma^+\right] - \mathbb{1}_{S^c}\mathcal{T}^{\Delta t}\left[\mathbb{1}_S F_\Gamma^+\right]\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x}$$

In principle, the integral could be estimated directly with any choice of dividing surface S, but the sum would only use data either exiting S (first term) or entering S (second term). We can use all the data at once and improve numerical stability by averaging over multiple such surfaces, and furthermore converting the transition operator $\mathcal{T}^{\Delta t}$ into the generator $\mathcal{L}$. However, we cannot simply take the limit under the integral due to the discontinuity in $\mathbb{1}_S$. Instead we get a smooth function into the integrand with the following steps. First, replace $\mathbb{1}_S$ with $1 - \mathbb{1}_{S^c}$ everywhere:

$$R_\Gamma(\lambda) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\times \tag{73}$$
$$\left\{ (1-\mathbb{1}_{S^c})\mathcal{T}^{\Delta t}\left[\mathbb{1}_{S^c} F_\Gamma^+\right] - \mathbb{1}_{S^c}\mathcal{T}^{\Delta t}\left[(1-\mathbb{1}_{S^c}) F_\Gamma^+\right]\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{74}$$
$$= \lim_{\Delta t\to 0} \frac{1}{\Delta t} \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\left\{ \mathcal{T}^{\Delta t}\left[\mathbb{1}_{S^c} F_\Gamma^+\right] - \mathbb{1}_{S^c}\mathcal{T}^{\Delta t} F_\Gamma^+\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{75}$$
$$\tag{76}$$

Next, add and subtract $\mathbb{1}_{S^c} F_\Gamma^+$ inside the integrand.

$$R_\Gamma(\lambda) = \lim_{\Delta t\to 0} \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\left\{ \frac{\mathcal{T}^{\Delta t}-1}{\Delta t}\left[\mathbb{1}_{S^c} F_\Gamma^+\right] - \mathbb{1}_{S^c}\frac{\mathcal{T}^{\Delta t}-1}{\Delta t} F_\Gamma^+\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{77}$$

At this point it is tempting to take the limit inside the integral, as $(\mathcal{T}^{\Delta t}-1)/\Delta t$ formally approaches $\mathcal{L}$. But the first term acts on a discontinuous function, which won't have a well-defined time derivative. We first replace $\mathbb{1}_{S^c}$ with a smooth function (on D), as follows.

Let $K : \mathbb{R}^d \to [0,1]$ be a function that increases from 0 on set A to 1 on set B (for instance, the committor). Let $S_\zeta = \{\mathbf{x} : K(\mathbf{x}) \le \zeta\}$ for $\zeta \in (0,1)$, and integrate both sides over ζ, noting that $\int_0^1 \mathbb{1}_{S_\zeta^c}(\mathbf{x})\, d\zeta = \int_0^1 \mathbb{1}\{K(\mathbf{x}) > \zeta\}\, d\zeta = K(\mathbf{x})$.

$$\int_0^1 R_\Gamma(\lambda)\, d\zeta = \lim_{\Delta t\to 0} \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\left\{ \frac{\mathcal{T}^{\Delta t}-1}{\Delta t}\left[K F_\Gamma^+\right] - K\mathcal{L} F_\Gamma^+\right\}(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{78}$$

Now we can move the limit inside and use the PDE to find

$$R_\Gamma(\lambda) = \int_{\mathbb{R}^d} F_\Gamma^-(\mathbf{x};\lambda)\left\{\mathcal{L}[K F_\Gamma^+](\mathbf{x}) + \lambda K(\mathbf{x})\Gamma(\mathbf{x}) F_\Gamma^+(\mathbf{x})\right]\right\}\pi(\mathbf{x})\, d\mathbf{x} \tag{79}$$

This formula can be estimated directly from knowledge of F_Γ^-, F_Γ^+, and π, using the ergodic assumption and with a discrete finite difference in time to estimate $\mathcal{L}[KF_\Gamma^+]$, i.e.,

$$\mathcal{L}[KF_\Gamma^+](\mathbf{X}_n(0)) \approx \frac{K(\mathbf{X}_n(\Delta t))F_\Gamma^+(\mathbf{X}_n(\Delta t)) - K(\mathbf{X}_n(0))F_\Gamma^+(\mathbf{X}_n(0))}{\Delta t} \tag{80}$$

Derivatives with respect to λ can be found by iterating the product rule, as we have solved for the derivatives of F_Γ^+ and F_Γ^-.

To validate DGA numerically, we can compare to the results of ES. In [Finkel et al., 2021] (Fig. 7), we saw convergence of the DGA committor to the ES committor across state space as sample size and lag time were increased. Here, we turn our attention to summary statistics of interest for full transition paths, not just forecasting. This will benchmark our current DGA implementation for comparison with future algorithmic developments.

Fig. 1a displays the time fractions spent in each phase of the SSW lifecycle: $A \to B$, $B \to A$, $A \to A$, and $B \to B$, including estimates from ES (cyan) and DGA (red) and their uncertainties. The DGA estimate of the $A \to B$ time fraction is a π-weighted average of $q_A^-(\mathbf{x})q_B^+(\mathbf{x})$ over state space,

$$\langle q_A^- q_B^+ \rangle_\pi = \int q_A^-(\mathbf{x})q_B^+(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \tag{81}$$

and similarly for the other phases (section 1 above justifies this formula rigorously, and section 2 above details the numerical computation of the integral). The DGA error bars are generated by repeating the entire pipeline three times with different short trajectory realizations. The bar height shows the mean, and the error bars show the minimum and maximum. The ES error bars are generated by bootstrap resampling (with replacement) 500 times from the control simulation, treating an entire SSW lifecycle as a single unit (from the beginning of one $A \to B$ transition until the beginning of the next one). This assumes no memory between successive events, which we have found to be reasonable; there is insignificant autocorrelation between consecutive return periods. The bars extend two root-mean-squared errors in both directions, enclosing a 95% confidence interval. To first order, DGA agrees well with ES on the fraction of time spent in each phase. A is the more stable of the two regimes, accounting for $\sim 50\%$ of the time compared to the $\sim 40\%$ of time spent in the orbit of B. The transition events are both an order of magnitude shorter, with $B \to A$ taking slightly longer on average. DGA ranks the $A \to B$ and $B \to A$ time fractions correctly, despite a bias in the absolute magnitudes.

The numbers in Fig. 1a are only relative durations; they do not tell us how long a full life cycle takes. That number is given by (one over) the rate. Fig. 1b shows three different rate estimates (that is, the generalized rate with $\Gamma = 0$) using the formulas above. The cyan bars come from ES, counting the number of $A \to B$ transitions per unit time. Of course, this equals the number of $B \to A$ transitions per unit time, so the $A \to B$ and $B \to A$ cyan bars are identical. Error bars come from bootstrapping, as with the relative durations. The red bars come from DGA, and these estimates are not technically symmetric. The DGA estimate labeled $A \to B$ integrates $\mathbf{J}_{AB} \cdot \mathbf{n}$ over dividing surfaces with $\mathbf{n}$ pointing away from A toward B, while the estimate labeled $B \to A$ integrates $\mathbf{J}_{BA} \cdot \mathbf{n}$ over surfaces with $\mathbf{n}$ pointing away from B toward A. Numerical and sampling errors cause slight differences between them, but Fig. 1b shows them both to come within 20% of the ES estimate.

DGA estimates should converge with increasing M (cluster number) and N (short-trajectory ensemble size). Larger M makes the approximation space $\{\phi_1, \ldots, \phi_M\}$ more expressive, making finer estimates possible. However, as M grows, we need more short trajectories N to robustly estimate the entries of the expanding matrix (48). Conversely, as M shrinks, P_{ij} will become closer to diagonal, because trajectories will escape from their starting cluster less frequently. Thus Δt would have to increase when M decreases. The optimal choice for a given model will depend on the relative costs of integrating the model, building basis sets, and solving large linear systems on different computer architectures. With our choice of $M = 1500$, increasing N from 5×10^4 to 3×10^6 does not change the DGA point estimates very much, but shrinks the error bars by a factor of ~ 4. To further reduce the bias in Fig. 1, we would likely need more refined basis functions, perhaps using nonlinear features as input to K-means. We do not yet have theoretical guarantees or optimal prescriptions for DGA parameters, but given the flexibility and parallelizability of the method, we believe it has much room for growth.

2.4 Visualization method

The two-dimensional projections in the main paper are generated with the following procedure. Let $\mathbf{y} = \mathbf{Y}(\mathbf{x})$ be an observable subspace, typically with dimension much less than that of $\mathbf{x}$ (usually two). Any scalar field $F(\mathbf{x})$, such as the committor, has a projection $F^{\mathbf{Y}}(\mathbf{y})$ onto this subspace by

$$F^{\mathbf{Y}}(\mathbf{y}) = \int F(\mathbf{x})\pi(\mathbf{x})\delta(\mathbf{Y}(\mathbf{x}) - \mathbf{y})\, d\mathbf{x} \tag{82}$$

In practice, the $\mathbf{y}$ space is partitioned into grid boxes $d\mathbf{y}$, and the integral is estimated from the dataset, yielding

$$F^{\mathbf{Y}}(\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N} F(\mathbf{X}_n(0))\frac{d\pi}{d\mu}(\mathbf{X}_n(0))\mathbb{1}_{d\mathbf{y}}(\mathbf{Y}(\mathbf{X}_n(0))) \tag{83}$$

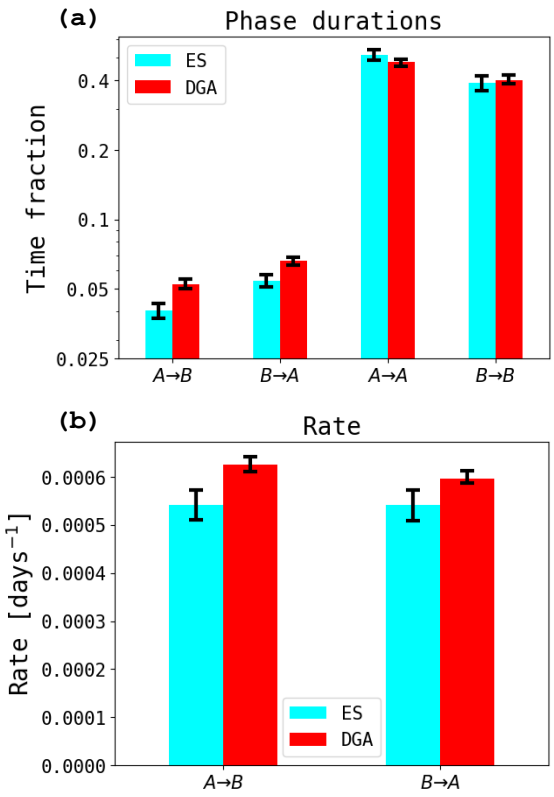

Figure 1: **DGA benchmarks and comparison to ES.** (a) Time fractions spent in each phase, as calculated by DGA (red) and ES (cyan). (b) Total SSW rate estimated using both $\mathbf{J}_{AB}$ and $\mathbf{J}_{BA}$; the two cyan columns are identical, as the counting method includes one $B \to A$ transition for every $A \to B$ transition.

where $\mathbb{1}_{d\mathbf{y}}(\mathbf{Y}(\mathbf{x})) = 1$ if $\mathbf{Y}(\mathbf{x}) \in d\mathbf{y}$ and zero otherwise. In words, we simply take a weighted average over all data points $\mathbf{X}_n(0)$ that project onto the grid box $d\mathbf{y}$, with weights given by the change of measure. In Fig. 3a-c, we use q_B^+ and q_A^- for F. In Fig. 3d-i, we use $F = \pi q_A^- q_B^-$ to generate the background colors. The same choices are used in Fig. 6, though in different coordinates.

To display the overlaid vector field, however, requires a more involved formula. We use the exact same reactive current formula as in the supplement of [Strahan et al., 2021], but repeat it here for reference. The projected current is defined as

$$\mathbf{J}_{AB}^{\mathbf{Y}}(\mathbf{y}) = \int \mathbf{J}_{AB}(\mathbf{x}) \cdot \nabla \mathbf{Y}(\mathbf{x}) \delta\big(\mathbf{Y}(\mathbf{x}) - \mathbf{y}\big)\, d\mathbf{x} \tag{84}$$

In the discretized $\mathbf{y}$ space, this leads to the discretized projected current:

$$\mathbf{J}_{AB}^{\mathbf{Y}}(\mathbf{y}) \approx \frac{1}{2\Delta t} \sum_{n=1}^{N} \frac{d\pi}{d\mu}(\mathbf{X}_n(0)) \Bigg[\mathbb{1}_{d\mathbf{y}}(\mathbf{X}_n(0)) q_A^-(\mathbf{X}_n(0)) q_B^+(\mathbf{X}_n(\theta_n)) \frac{\mathbf{Y}(\mathbf{X}_n(\theta_n)) - \mathbf{Y}(\mathbf{X}_n(0))}{\theta_n} \tag{85}$$

$$+ \mathbb{1}_{d\mathbf{y}}(\mathbf{X}_n(\Delta t)) q_A^-(\mathbf{X}_n(\widetilde{\theta}_n)) q_B^+(\mathbf{X}_n(\Delta t)) \frac{\mathbf{Y}(\mathbf{X}_n(\Delta t)) - \mathbf{Y}(\mathbf{X}_n(\widetilde{\theta}_n))}{\Delta t - \widetilde{\theta}_n} \Bigg] \tag{86}$$

where θ_n and $\widetilde{\theta}_n$ are the "first-entry times" to $D = (A \cup B)^c$ in the nth trajectory with time running forward and backward, respectively.

We use a similar formula to display the composites in Figs. 4, 5 and the stochastic tendencies in Fig. 7. Suppose we have a progress coordinate f evaluated on all the data points. The composite SSW at a given level $f_0 \pm df$ is approximated

$$\left\{ \mathbf{X}_n(t) : |f(\mathbf{X}_n(t)) - f_0| < df, \text{ weighted by } q_A^-(\mathbf{X}_n(t)) q_B^+(\mathbf{X}_n(t)) \frac{d\pi}{d\mu}(\mathbf{X}_n(0)) \right\} \tag{87}$$

where t can range from 0 to 20 days, the length of the short trajectory. In Fig. 4b, we use $-\eta_B^+$ as the progress coordinate f, spanning the range from -80 days to zero with a step size of 5 days and a tolerance of $df = 7.5$ days, except for the final level where the tolerance is reduced to 1.5 days. Smaller tolerances should be possible with a larger dataset, but in our case produced artifacts in the plot due to limited data. In Fig. 4c, we use q_B^+ as the progress coordinate f, spanning the range from 0 to 1 with a step size of 0.05 and a tolerance of 0.075, except for the final level where the tolerance is reduced to 0.015.

The stochastic tendencies shown in the bottom row of Fig. 7 are found in a similar way. For a given observable, such as Γ, its stochastic tendency $\mathcal{L}_{AB}\Gamma$ is estimated at a single point in each trajectory, in particular day $t_1 = 5$ days out of 20, as follows. First, we set $t_0 = 0$ days, or else the last-exit time from $A \cup B$ between days 0 and 5. Second, we set $t_2 = 10$ days, or else the first-entry time to $A \cup B$ between days 5 and 10. Then

$$\mathcal{L}_{AB}\Gamma(\mathbf{X}_n(t_1)) \approx \frac{1}{2q_A^-(t_1)q_B^+(t_1)} \left[\frac{\Gamma(\mathbf{X}_n(t_2)) - \Gamma(\mathbf{X}_n(t_1))}{t_2 - t_1} q_A^-(t_1) q_B^-(t_2) - \frac{\Gamma(\mathbf{X}_n(t_1)) - \Gamma(\mathbf{X}_n(t_0))}{t_1 - t_0)} q_A^-(t_0) q_B^-(t_1) \right] \tag{88}$$

and we estimate the mean stochastic tendency across a surface of constant q_B^+ by the usual method of estimating ergodic averages, i.e., Eq. (83)

2.5 Algorithmic parameters

Having sketched the general numerical procedure, we now provide the exact parameters used here, which are similar to those in [Finkel et al., 2021]. We use $N = 3 \times 10^5$ trajectories, each of length $\Delta t = 20$ days, with a sampling interval of 1 day. The initial conditions are resampled from a long ($\times 10^6$-day) control simulation to be uniformly distributed on the space $(|\Psi|(30\,\text{km}), U(30\,\text{km}))$. With a more complex, expensive model, we cannot rely on a control simulation to seed the initial points, but here we focus on TPT and DGA as a proof of concept rather than optimizing the numerical procedure. We use $M = 1500$ basis functions defined as indicators on a partition induced by K-means clustering on $\{\mathbf{X}_n(0)\}$. The clustering is hierarchical so that the cluster size does not become too imbalanced.

There are many potential directions for methodological improvement. In an expensive model without the ability to run a long control simulation, we should use a splitting and killing method to seed the initial trajectories across state space. Moreover, we could perform DGA repeatedly with new data seeded at each iteration in areas of high sensitivity. The choice of basis function can also powerfully affect DGA's performance. Indicator functions are advantageous in producing a *bona fide* Markov matrix and guaranteeing a maximum principle for the committor probabilities. However, smooth and/or global basis functions have in some cases found to be more efficient at capturing the structure of the committor with fewer basis elements [Thiede et al., 2019, Strahan et al., 2021].

3 Minimum-action method

To compute the minimum-action paths, we use a completely discrete approach for simplicity and to accommodate the low-rank nature of the stochastic forcing. Heuristically, we wish to find the most probable path connecting A and B, which we take as the mode of the (discretized) path density over the distribution of paths from A to B. For concreteness, fix $\mathbf{x}(0) = \mathbf{x}_0 \in A$ and a time horizon T discretized into K intervals, with a timestep $\delta t = T/K = 0.005$ days. The discretized dynamics evolve according to the Euler-Maruyama method as

$$\mathbf{x}(k\delta t) = \mathbf{x}((k-1)\delta t) + \boldsymbol{v}\big(\mathbf{x}((k-1)\Delta t)\big)\Delta t + \boldsymbol{\sigma}\boldsymbol{\eta}_k\sqrt{\delta t} \tag{89}$$

where $\boldsymbol{\eta}_k$ is a vector of i.i.d. unit normal samples, $\boldsymbol{v}$ is the deterministic drift, and $\boldsymbol{\sigma} \in \mathbb{R}^{d\times m}$ is the diffusion matrix, with a noise rank $m = 3$ and a spatially smooth structure as defined in [Finkel et al., 2021]. In the classical minimum-action approach, $\boldsymbol{\sigma}$ is assumed to be a $d \times d$ invertible matrix, and the probability density of a path $(\mathbf{x}_0, \ldots, \mathbf{x}_K)$ (where $\mathbf{x}_k = \mathbf{x}(k\delta t)$) is

$$\prod_{k=1}^{K} \mathcal{N}\Big(\mathbf{x}_k \Big| \mathbf{x}_{k-1} + \boldsymbol{v}(\mathbf{x}_{k-1})\delta t, \sigma\sigma^\top \delta t\Big) \tag{90}$$

$$= \prod_{k=1}^{K} \frac{1}{(2\pi\delta t)^{dK/2}(\det\sigma)^K} \times \tag{91}$$

$$\exp\left\{-\frac{1}{2}\Big(\mathbf{x}_k - \mathbf{x}_{k-1} - \boldsymbol{v}(\mathbf{x}_{k-1})\delta t\Big)^\top \frac{(\sigma\sigma^\top)^{-1}}{\delta t}\Big(\mathbf{x}_k - \mathbf{x}_{k-1} - \boldsymbol{v}(\mathbf{x}_{k-1})\delta t\Big)\right\} \tag{92}$$

$$\propto \exp\left\{-\frac{\delta t}{2}\sum_{k=1}^{K}\left(\frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\delta t} - \boldsymbol{v}(\mathbf{x}_{k-1})\right)^\top (\sigma\sigma^\top)^{-1}\left(\frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\delta t} - \boldsymbol{v}(\mathbf{x}_{k-1})\right)\right\} \tag{93}$$

$$\sim \exp\left\{-\frac{1}{2}\int_0^T \Big[\dot{\mathbf{x}}(t) - \boldsymbol{v}(\mathbf{x}(t))\Big](\sigma\sigma^\top)^{-1}\Big[\dot{\mathbf{x}}(t) - \boldsymbol{v}(\mathbf{x}(t))\Big]\,dt\right\} \text{ as } \delta t \to 0 \tag{94}$$

and the problem becomes to minimize the quadratic form in the argument of the exponential, which is the Freidlin-Wentzell action functional, subject to the constraint $\mathbf{x}_K \in B$. However, because we stir the wind field with smooth spatial forcing in only $m \ll d$ wavenumbers, σ is low-rank and thus $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top$ is singular. Given any realized path $(\mathbf{x}_0, \ldots, \mathbf{x}_K)$, there may be no possible underlying forcing $\boldsymbol{\eta}_k$ that could have produced it under our noise model. So the obvious optimization strategy of fixing $\mathbf{x}_0$ and $\mathbf{x}_K$ and varying the steps in between may lead to impossible paths. For this reason, we perform optimization in the space of perturbations, and ensure that every step of the optimization is realizable under our noise model. This is a strategy we adopt from the cyclogenesis model [Plotkin et al., 2019]. The result will be a simpler, convex objective function at the expense of a more complicated constraint. The probability density of a particular forcing sequence $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$ is given by

$$\prod_{k=1}^{K} \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right) = \frac{1}{(2\pi)^{mK/2}} \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right) \tag{95}$$

The objective inside the exponential is now a simple quadratic in perturbation space which can be easily differentiated with respect to those perturbations. The constraint, meanwhile, takes the form of a complicated iterated function. Define the flow map $F(\mathbf{x}) = \mathbf{x} + \boldsymbol{v}(\mathbf{x})\delta t$ as the deterministic part of the timestep, so $\mathbf{x}_k = F(\mathbf{x}_{k-1}) + \boldsymbol{\sigma}\boldsymbol{\eta}_k\sqrt{\delta t}$. In terms of F, the endpoint has to be written as a recursive function

$$\mathbf{x}_K = F(\mathbf{x}_{K-1}) + \sigma\boldsymbol{\eta}_K\sqrt{\delta t} \tag{96}$$

$$\mathbf{x}_{K-1} = F(\mathbf{x}_{K-2}) + \sigma\boldsymbol{\eta}_{K-1}\sqrt{\delta t} \tag{97}$$

$$\vdots \tag{98}$$

$$\mathbf{x}_1 = F(\mathbf{x}_0) + \sigma\boldsymbol{\eta}_1\sqrt{\delta t} \tag{99}$$

We impose the end constraint by adding to the action a penalty $\Phi(\mathbf{x}_K) = \mathrm{dist}(\mathbf{x}_K, B)$, a function which linearly increases with distance to B. The full optimization problem is

$$\min_{\boldsymbol{\eta}}\left\{\frac{1}{2K}\sum_{k=1}^{K}\boldsymbol{\eta}_k^\top\boldsymbol{\eta}_k + \alpha\Phi(\mathbf{x}_K)\right\} \tag{100}$$

$$\mathbf{x}_0 \in A \text{ is fixed} \tag{101}$$

$$\mathbf{x}_k = F(\mathbf{x}_{k-1}) + \boldsymbol{\sigma}\boldsymbol{\eta}_k\sqrt{\delta t} \text{ for } k = 1, \ldots, K \tag{102}$$

Here α is a weight which can be increased to harden the end constraint. We divide by K so that the path action does not overwhelm the endpoint penalty as $K \to \infty$. (This makes the sum converge to an integral.) We set $\mathbf{x}_0$ to be the fixed point $\mathbf{a} \in A$ when finding the least-action path from A to B and the fixed point $\mathbf{b} \in B$ when finding the least-action path from B to A. We used the L-BFGS method as implemented in `scipy`, with a maximum of 10 iterations. We differentiate $\Phi(x_K)$ with respect to $\boldsymbol{\eta}_k$ using knowledge of the adjoint model, with a backward pass through the path to compute each gradient. At each descent step, we refine the stepsize with backtracking line search. One way to guarantee the end constraint is ultimately satisfied is to gradually increase α and lengthen T; however, we found it sufficient to fix $\alpha = 1.0$ and $T = 100$, in keeping with the typical observed transit time. We have kept the algorithm simple, not devoting too much effort to finding the global optimimum over all time horizons, as we only care for a qualitative assessment to compare with results of TPT.

4 Enstrophy budget derivation

Here we derive an enstrophy budget from the Holton-Mass model. For completeness, we also provide a derivation of the original Holton-Mass model, starting from Eqs. (1) and (6) of [Holton and Mass, 1976] and deriving their Eqs. (9) and (10), equivalent to Eqs. (3) of our main text. This provides some helpful context for the enstrophy budget.

4.1 Holton-Mass projected model

This section inserts the ansatz (Eqs. 7 of [Holton and Mass, 1976])

$$\psi'(x,y,z,t) = \mathrm{Re}\{\Psi(z,t)e^{z/2H}e^{ikx}\sin\ell y\} \tag{103a}$$

$$u(x,y,z,t) = U(z,t)\sin\ell y \tag{103b}$$

into the wave-mean flow model. The equations are nonlinear, involving products of pairs of terms containing $\sin\ell y$, which projects onto wavenumber 2ℓ. Following [Holton and Mass, 1976], we approximate $\sin^2\ell y$ by $\varepsilon\sin\ell y$, where $\varepsilon = 8/(3\pi)$, to close the equations in terms of a single wavenumber. We will also make extensive use of the identity

$$e^{cz}\partial_z\left(e^{-cz}f(z)\right) = (\partial_z - c)f(z) \tag{104}$$

4.1.1 Wave equation

We start with the quasi-geostrophic potential vorticity equation, Eq. (1) of [Holton and Mass, 1976], which governs the tendency of the streamfunction Ψ:

$$0 = (\partial_t + \overline{u}\partial_x)q' + \partial_y\overline{q}\partial_x\psi' + \frac{f_0^2}{N^2}e^{z/H}\partial_z(e^{-z/H}\alpha\partial_z\psi') \tag{105a}$$

where

$$q' = \nabla^2\psi' + \frac{f_0^2}{N^2}e^{z/H}\partial_z(e^{-z/H}\partial_z\psi') \tag{105b}$$

$$\partial_y\overline{q} = \beta - \partial_y^2\overline{u} - \frac{f_0^2}{N^2}e^{z/H}\partial_z(e^{-z/H}\partial_z\overline{u}) \tag{105c}$$

We deal with the three terms one at a time. All three terms are linear in ψ', and so we can take the real part outside. The first term—the material derivative of PV following the mean flow—is

$$(\partial_t + \overline{u}\partial_x)q' = \mathrm{Re}\left\{(\partial_t + Uik\sin\ell y)\left[-(k^2+\ell^2)\Psi e^{z/2H} + \frac{f_0^2}{N^2}e^{z/H}\partial_z\left(e^{-z/H}\partial_z(\Psi e^{z/2H})\right)\right]e^{ikx}\sin\ell y\right\} \tag{106a}$$

$$\approx \mathrm{Re}\left\{(\partial_t + Uik\varepsilon)\left[-(k^2+\ell^2)\Psi e^{z/2H} + \frac{f_0^2}{N^2}e^{z/2H}e^{z/2H}\partial_z\left(e^{-z/2H}e^{-z/2H}\partial_z(e^{z/2H}\Psi)\right)\right]e^{ikx}\sin\ell y\right\} \tag{106b}$$

$$= \mathrm{Re}\left\{(\partial_t + Uik\varepsilon)\left[-(k^2+\ell^2)\Psi e^{z/2H} + \frac{f_0^2}{N^2}e^{z/2H}\left(\partial_z - \frac{1}{2H}\right)\left(\partial_z + \frac{1}{2H}\right)\Psi\right]e^{ikx}\sin\ell y\right\} \tag{106c}$$

$$= \mathrm{Re}\left\{(\partial_t + Uik\varepsilon)\left[-(k^2+\ell^2)\Psi + \frac{f_0^2}{N^2}\left(\partial_z^2 - \frac{1}{4H^2}\right)\Psi\right]e^{z/2H}e^{ikx}\sin\ell y\right\} \tag{106d}$$

$$\tag{106e}$$

The second term—meridional advection of PV due to eddies—is

$$\partial_y \overline{q} \partial_x \psi' = \mathrm{Re}\left\{ \left[\beta + \ell^2 U \sin \ell y - \frac{f_0^2}{N^2} \left(U_{zz} - \frac{1}{H} U_z \right) \sin \ell y \right] ik \Psi e^{z/2H} e^{ikx} \sin \ell y \right\} \tag{107}$$

$$\approx \mathrm{Re}\left\{ ik\Psi \left[\beta + \varepsilon \left(\ell^2 U + \frac{f_0^2}{N^2} \left(\frac{1}{H} U_z - U_{zz} \right) \right) e^{z/2H} e^{ikx} \sin \ell y \right\} \right. \tag{108}$$

The third term—radiative cooling—is

$$\frac{f_0^2}{N^2} e^{z/H} \partial_z (e^{-z/H} \alpha \partial_z \psi') = \mathrm{Re}\left\{ \frac{f_0^2}{N^2} \left[e^{z/2H} e^{z/2H} \partial_z \left(e^{-z/2H} \alpha e^{-z/2H} \partial_z \left(e^{z/2H} \Psi \right) \right) \right] e^{ikx} \sin \ell y \right\} \tag{109}$$

$$= \mathrm{Re}\left\{ \frac{f_0^2}{N^2} \left(\partial_z - \frac{1}{2H} \right) \left[\alpha \left(\partial_z + \frac{1}{2H} \right) \Psi \right] e^{z/2H} e^{ikx} \sin \ell y \right\} \tag{110}$$

We now combine the three terms. Stipulating they add to zero for all x actually means we can drop the "real part" on the outside. This is because for any complex function $f(x) = g(x) + ih(x)$ for real g and h,

$$\mathrm{Re}\{ f(x) e^{ikx} \} = \mathrm{Re}\{ \big(g(x) + ih(x) \big) (\cos kx + i \sin kx) \} = g(x) \cos kx - h(x) \sin kx \tag{111}$$

By orthogonality of $\sin(kx)$ and $\cos(kx)$ on the domain, both g and h must vanish identically if the left-hand side is zero. Hence, we recover the projected QGPV equation, Eq. (9) of [Holton and Mass, 1976]:

$$\begin{aligned} 0 = (\partial_t + ik\varepsilon U) \left[-\left(k^2 + \ell^2 + \frac{f_0^2}{N^2} \frac{1}{4H^2} \right) + \frac{f_0^2}{N^2} \partial_z^2 \right] \Psi \\ + ik\Psi \left[\beta + \varepsilon \ell^2 U + \varepsilon \frac{f_0^2}{N^2} \left(\frac{1}{H} U_z - U_{zz} \right) \right] \\ + \frac{f_0^2}{N^2} \left(\partial_z - \frac{1}{2H} \right) \left[\alpha \left(\partial_z + \frac{1}{2H} \right) \Psi \right] \end{aligned} \tag{112}$$

We take a further step and non-dimensionalize for a more compact expression. We select a horizontal length scale $L = 2.5 \times 10^5$ m, a vertical length scale $H = 7 \times 10^3$ m (the same as the H in the equation), and a time scale $T = 1$ day $= 86400$ s. Thus rescaling all variables,

$$\begin{aligned} 0 = \frac{1}{T} (\partial_t + ik\varepsilon U) \left[-\left(\frac{k^2 + \ell^2}{L^2} + \frac{f_0^2}{N^2} \frac{1}{4H^2} \right) + \frac{f_0^2}{N^2} \frac{1}{H^2} \partial_z^2 \right] \frac{L^2}{T} \Psi \\ + \frac{L}{T} ik\Psi \left[\frac{\beta}{LT} + \frac{\varepsilon \ell^2 U}{LT} + \varepsilon \frac{f_0^2}{N^2} \left(\frac{L}{H^2 T} U_z - \frac{L}{H^2 T} U_{zz} \right) \right] \\ + \frac{f_0^2}{N^2} \frac{1}{H} \left(\partial_z - \frac{1}{2} \right) \left[\frac{\alpha}{HT} \left(\partial_z + \frac{1}{2} \right) \frac{L^2}{T} \Psi \right] \end{aligned} \tag{113}$$

We now define the dimensionless group $\mathcal{G}^2 := H^2 N^2 / (L^2 f_0^2)$ and multiply through by $T^2 \mathcal{G}^2$ and obtain

$$\begin{aligned} (\partial_t + ik\varepsilon U) \left[-\mathcal{G}^2 (k^2 + \ell^2) - \frac{1}{4} + \partial_z^2 \right] \Psi \\ + ik\Psi \left[\mathcal{G}^2 \beta + \varepsilon \left(\mathcal{G}^2 \ell^2 U + U_z - U_{zz} \right) \right] \\ = -\left(\partial_z - \frac{1}{2} \right) \left[\alpha \left(\partial_z + \frac{1}{2} \right) \Psi \right] \end{aligned} \tag{114}$$

exactly as in Eq. (3) of the main paper.

4.1.2 Mean-flow equation

The PDE for $\overline{u}$ is specified in Eq. (6) of [Holton and Mass, 1976]:

$$\partial_t \left[\partial_y^2 \overline{u} + \frac{f_0^2}{N^2} e^{z/H} \partial_z \left(e^{-z/H} \partial_z \overline{u} \right) \right] \tag{115}$$

$$= -\frac{f_0^2}{N^2} e^{z/H} \partial_z \left[\alpha e^{-z/H} \partial_z (\overline{u} - u_R) \right] \tag{116}$$

$$+ \frac{f_0^2}{N^2} \partial_y^2 \left[e^{z/H} \partial_z \left(e^{-z/H} \overline{\partial_x \psi' \partial_z \psi'} \right) \right] \tag{117}$$

Putting in the ansatz for U renders the first two terms trivial. The left-hand side becomes

$$\partial_t \left[-\ell^2 U + \frac{f_0^2}{N^2} \left(U_{zz} - \frac{1}{H} U_z \right) \right] \sin \ell y \tag{118}$$

and setting $u_R(z) = U^R(z) \sin \ell y$, the first term on the right becomes

$$-\frac{f_0^2}{N^2} e^{z/H} \partial_z \left[\alpha e^{-z/H} \partial_z (U - U^R) \right] \tag{119}$$

The third term is more involved due to the zonal correlation term. However, we can simplify using the rule for zonal correlations from Eq. (25) of the main text, which is derived more fully here. Let $\psi_1' = \mathrm{Re}\{\Psi_1 e^{ikx}\}$ and $\psi_2' = \mathrm{Re}\{\Psi_2 e^{ikx}\}$, where Ψ_1, Ψ_2 are two complex numbers independent of x with real parts X_1, X_2 and imaginary parts Y_1, Y_2 respectively. Their zonal eddy correlation is

$$\overline{\psi_1' \psi_2'} = \overline{\mathrm{Re}\{\Psi_1 e^{ikx}\} \mathrm{Re}\{\Psi_2 e^{ikx}\}} \tag{120}$$

$$= \overline{(X_1 \cos kx - Y_1 \sin kx)(X_2 \cos kx - Y_2 \sin kx)} \tag{121}$$

$$= X_1 X_2 \overline{\cos^2 kx} - (X_1 Y_2 + Y_1 X_2) \overline{\cos kx \sin kx} + Y_1 Y_2 \overline{\sin^2 kx} \tag{122}$$

$$= \frac{1}{2} (X_1 X_2 + Y_1 Y_2) \tag{123}$$

$$= \frac{1}{2} \mathrm{Re}\{(X_1 - iY_1)(X_2 + iY_2)\} = \frac{1}{2} \mathrm{Re}\{\Psi_1^* \Psi_2\} \tag{124}$$

Using this simple rule, we have

$$\overline{\partial_x \psi' \partial_z \psi'} = \frac{1}{2} \mathrm{Re} \left\{ \left(ik \Psi e^{z/2H} \sin \ell y \right)^* \partial_z \left(\Psi e^{z/2H} \sin \ell y \right) \right\} \tag{125}$$

$$\approx \frac{1}{2} k \varepsilon \, \mathrm{Re} \left\{ -i \Psi^* e^{z/2H} \left(\Psi_z + \frac{1}{2H} \Psi \right) e^{z/2H} \right\} \sin \ell y \tag{126}$$

$$= \frac{1}{2} k \varepsilon \, \mathrm{Im}\{\Psi^* \Psi_z\} e^{z/H} \sin \ell y \tag{127}$$

where we have used that $\mathrm{Re}\{-i\Psi^* \Psi\} = 0$. The second term on the right-hand side of the mean-flow equation is then

$$-\frac{f_0^2}{N^2} \ell^2 e^{z/H} \partial_z \left(\frac{1}{2} k \varepsilon \, \mathrm{Im}\{\Psi^* \Psi_z\} \right) \sin \ell y = -\frac{f_0^2}{N^2} \frac{\varepsilon k \ell^2}{2} e^{z/H} \mathrm{Im}\{\Psi^* \Psi_{zz}\} \sin \ell y \tag{128}$$

where we have used $\mathrm{Im}\{\Psi_z^* \Psi_z\} = 0$. Putting these terms together, dropping the $\sin \ell y$, and negating each term, we reproduce Eq. (10) of [Holton and Mass, 1976]:

$$\partial_t \left[\ell^2 U + \frac{f_0^2}{N^2} \left(\frac{1}{H} U_z - U_{zz} \right) \right] = \frac{f_0^2}{N^2} e^{z/H} \partial_z \left[\alpha e^{-z/H} \partial_z (U - U^R) \right] + \frac{f_0^2}{N^2} \frac{\varepsilon k \ell^2}{2} e^{z/H} \mathrm{Im}\{\Psi^* \Psi_{zz}\} \tag{129}$$

Let us non-dimensionalize this equation as we did for the mean-flow equation with the scales L, H, and T.

$$\frac{1}{T} \partial_t \left[\frac{1}{LT} \ell^2 U + \frac{f_0^2}{N^2} \frac{L}{H^2 T} (U_z - U_{zz}) \right] = \frac{f_0^2}{N^2} e^z \frac{1}{H} \partial_z \left[\frac{\alpha}{T} e^{-z} \frac{L}{HT} \partial_z (U - U^R) \right] + \frac{f_0^2}{N^2} \frac{1}{L^3} \frac{\varepsilon k \ell^2}{2} e^z \frac{L^4}{T^2 H^2} \mathrm{Im}\{\Psi^* \Psi_{zz}\} \tag{130}$$

Multiplying through by $LT^2\mathcal{G}^2$,

$$\partial_t\left[\mathcal{G}^2\ell^2U+U_z-U_{zz}\right]=e^z\partial_z\left[\alpha e^{-z}\partial_z(U-U^R)\right]+\frac{\varepsilon k\ell^2}{2}e^z\mathrm{Im}\{\Psi^*\Psi_{zz}\} \tag{131}$$

Notice that the first term in brackets, $\mathcal{G}^2\ell^2U+U_z-U_{zz}$, also appears in the nonlinear coupling on the right-hand side of the QGPV equation (114). To obtain more symmetry between the wave and mean-flow equations, we add $(\mathcal{G}^2/\varepsilon)\partial_t\beta$ (which is zero) to the left-hand side of Eq. (131), and then multiply through by $2/(\varepsilon\ell^2)$:

$$\partial_t\left[\frac{\mathcal{G}^2}{\varepsilon}\beta+\mathcal{G}^2\ell^2U+U_z-U_{zz}\right]=e^z\partial_z\left[e^{-z}\alpha\partial_z(U-U^R)\right]+\frac{\varepsilon k\ell^2}{2}e^z\mathrm{Im}\{\Psi^*\Psi_{zz}\} \tag{132}$$

$$\frac{2}{(\varepsilon\ell)^2}\partial_t\left[\mathcal{G}^2\beta+\varepsilon\left(\mathcal{G}^2\ell^2U+U_z-U_{zz}\right)\right]=\frac{2}{\varepsilon\ell^2}e^z\partial_z\left[e^{-z}\alpha\partial_z(U-U^R)\right]+ke^z\,\mathrm{Im}\{\Psi^*\Psi_{zz}\} \tag{133}$$

This is Eq. (3)b of our main text.

In the next section, we derive an enstrophy budget directly from these projected equations.

4.2 Enstrophy budget

The enstrophy budget starts with the projected wave-mean flow interaction equations, repeated below with the following abbreviations:

- $\beta_e=\mathcal{G}^2\beta+\varepsilon(\mathcal{G}^2\ell^2U+U_z-U_{zz})$. [Holton and Mass, 1976] used the same symbol to denote the corresponding quantity in the PDE, $\beta-\partial_y^2\overline{u}-e^{z/H}\partial_z[e^{-z/H}\partial_z\overline{u}]$, which physically represents meridional gradient of zonal-mean PV. Here, we use $\partial_y\overline{q}$ for the PDE version, and β_e strictly for the projected version of the equation.
- $\delta=\mathcal{G}^2(k^2+\ell^2)+\frac{1}{4}$. This constant term appears in the expression for QGPV, arising from the three-dimensional Laplacian.
- $F_q=ke^z\mathrm{Im}\{\Psi^*\Psi_{zz}\}$. This term represents the eddy meridional PV flux.
- $R=\frac{2}{\varepsilon\ell^2}e^z\partial_z\left[e^{-z}\alpha\partial_z(U-U^R)\right]$. This term represents thermal relaxation of the wind field toward the radiative wind U^R.

With these abbreviations, the wave and mean-flow equations are

$$\frac{2}{(\varepsilon\ell)^2}\partial_t\beta_e-F_q=R \tag{134a}$$

$$(\partial_t+ik\varepsilon U)\left(-\delta+\partial_z^2\right)\Psi+ik\Psi\beta_e=-\left(\partial_z-\frac{1}{2}\right)\left[\alpha\left(\partial_z+\frac{1}{2}\right)\Psi\right] \tag{134b}$$

Three steps will give us the enstrophy budget.

Step (i) Multiply Eq. (134a) by β_e.

$$\frac{1}{(\varepsilon\ell)^2}\partial_t\left(\beta_e^2\right)-F_q\beta_e=R\beta_e \tag{135a}$$

$$\partial_t\Gamma-F_q\beta_e=R\beta_e \tag{135b}$$

$$\text{where }\Gamma:=\left(\frac{\beta_e}{\varepsilon\ell}\right)^2 \tag{135c}$$

Step (ii) Multiply Eq. (134b) by $(-\delta+\partial_z^2)\Psi^*$, and then take the real part. This corresponds to multiplying by q' and taking the zonal mean, as in the standard EP relation.

$$\mathrm{Re}\left\{\left[(-\delta+\partial_z^2)\Psi^*\right](\partial_t+ik\varepsilon U)(-\delta+\partial_z^2)\Psi\right\}+k\beta_e\mathrm{Re}\left\{i\left[(-\delta+\partial_z^2)\Psi^*\right]\Psi\right\} \tag{136a}$$

$$=-\mathrm{Re}\left\{\left[(-\delta+\partial_z^2)\Psi^*\right]\left(\partial_z-\frac{1}{2}\right)\left[\alpha\left(\partial_z+\frac{1}{2}\right)\Psi\right]\right\} \tag{136b}$$

Let us simplify the terms one at a time. The first term on the left can be easily simplified by applying the identity

$$\partial_t\left(\frac{1}{2}|f(t)|^2\right) = \frac{1}{2}\partial_t(f^*f) = \frac{1}{2}(f_t^*f + f^*f_t) = \mathrm{Re}\{f_t^*f\} \tag{137}$$

to $f = (-\delta + \partial_z^2)\Psi^*$:

$$\mathrm{Re}\left\{\left[(-\delta+\partial_z^2)\Psi^*\right](\partial_t + ik\varepsilon U)(-\delta+\partial_z^2)\Psi\right\} \tag{138}$$

$$= \mathrm{Re}\left\{\partial_t\left(\frac{1}{2}\left|(-\delta+\partial_z^2)\Psi\right|^2\right)\right\} + k\varepsilon U\,\mathrm{Re}\left\{i\left|(-\delta+\partial_z^2)\Psi\right|^2\right\} \tag{139}$$

$$= \partial_t\left(\frac{1}{2}\left|(-\delta+\partial_z^2)\Psi\right|^2\right) \tag{140}$$

The second term becomes

$$k\beta_e\mathrm{Re}\left\{-\delta i\Psi^*\Psi + i\Psi_{zz}^*\Psi\right\} = -k\beta_e\mathrm{Im}\{\Psi_{zz}^*\Psi\} = k\beta_e\mathrm{Im}\{\Psi^*\Psi_{zz}\} = F_q\beta_e e^{-z} \tag{141}$$

where in the last step we recognized the same term from the mean-flow equation. The right-hand side cannot be very simplified, but we expand term by term for an alternative expression:

$$-\mathrm{Re}\left\{\left[(-\delta+\partial_z^2)\Psi^*\right]\left(\partial_z - \frac{1}{2}\right)\left[\alpha\left(\partial_z+\frac{1}{2}\right)\Psi\right]\right\} \tag{142}$$

$$= -\mathrm{Re}\left\{\left(-\delta\Psi^* + \Psi_{zz}^*\right)\left(\partial_z - \frac{1}{2}\right)\left(\alpha\Psi_z + \frac{\alpha}{2}\Psi\right)\right\} \tag{143}$$

$$= -\mathrm{Re}\left\{\left(-\delta\Psi^* + \Psi_{zz}^*\right)\left[\alpha\Psi_{zz} + \alpha_z\Psi_z + \left(\frac{\alpha_z}{2} - \frac{\alpha}{4}\right)\Psi\right]\right\} \tag{144}$$

$$= -\alpha|\Psi_{zz}|^2 - \alpha_z\mathrm{Re}\{\Psi_{zz}^*\Psi_z\} - \left(\frac{\alpha_z}{2} - \frac{\alpha}{4}\right)\mathrm{Re}\{\Psi_{zz}^*\Psi\} \tag{145}$$

$$+ \delta\left[\alpha\,\mathrm{Re}\{\Psi^*\Psi_{zz}\} + \alpha_z\mathrm{Re}\{\Psi^*\Psi_z\} + \left(\frac{\alpha_z}{2} - \frac{\alpha}{4}\right)|\Psi|^2\right] \tag{146}$$

$$= \delta\left(\frac{\alpha_z}{2} - \frac{\alpha}{4}\right)|\Psi|^2 + \delta\alpha_z\mathrm{Re}\{\Psi^*\Psi_z\} \tag{147}$$

$$+ \left(\delta\alpha + \frac{\alpha}{4} - \frac{\alpha_z}{2}\right)\mathrm{Re}\{\Psi_{zz}^*\Psi\} - \alpha_z\mathrm{Re}\{\Psi_{zz}^*\Psi_z\} - \alpha|\Psi_{zz}|^2 \tag{148}$$

Finally, we recombine all three terms and multiply through by e^z for symmetry with the mean-flow equation.

$$\partial_t\mathcal{E} + F_q\beta_e = D \tag{149}$$

where (150)

$$\mathcal{E} = \frac{1}{2}e^z\left|(-\delta+\partial_z^2)\Psi\right|^2 \tag{151}$$

$$D = -\mathrm{Re}\left\{e^z\left[(-\delta+\partial_z^2)\Psi^*\right]\right\}\left(\partial_z - \frac{1}{2}\right)\left[\alpha\left(\partial_z+\frac{1}{2}\right)\Psi\right]\right\} \tag{152}$$

Step (iii) : Add together the evolution equations (135b) and (149) for Γ and $\mathcal{E}$.

$$\partial_t\mathcal{E} = D - F_q\beta_e \tag{153}$$

$$\partial_t\Gamma = R\beta_e + F_q\beta_e \tag{154}$$

$$\therefore\ \partial_t(\Gamma+\mathcal{E}) = R\beta_e + D \tag{155}$$

The advantage of this arrangement is the left-hand side terms are quadratic in Ψ and U respectively, while the right-hand side terms contain all the information about radiative damping via the Newtonian cooling coefficient $\alpha(z)$. In the absence of such dissipation, therefore, $\Gamma+\mathcal{E}$ would be a conserved quantity.

References

[E et al., 2019] E, W., Li, T., and Vanden-Eijnden, E. (2019). *Applied stochastic analysis*, volume 199. American Mathematical Soc.

[E and Vanden-Eijnden, 2006] E, W. and Vanden-Eijnden, E. (2006). Towards a Theory of Transition Paths. *Journal of Statistical Physics*, 123(3):503.

[Finkel et al., 2021] Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., and Weare, J. (2021). Learning forecasts of rare stratospheric transitions from short simulations. *Monthly Weather Review*, 149(11):3647 – 3669.

[Fitzsimmons and Pitman, 1999] Fitzsimmons, P. and Pitman, J. (1999). Kac's moment formula and the feynman–kac formula for additive functionals of a markov process. *Stochastic Processes and their Applications*, 79(1):117–134.

[Holton and Mass, 1976] Holton, J. R. and Mass, C. (1976). Stratospheric vacillation cycles. *Journal of the Atmospheric Sciences*, 33(11):2218–2225.

[Karatzas and Shreve, 1998] Karatzas, I. and Shreve, S. E. (1998). *Brownian Motion and Stochastic Calculus*. Springer.

[Metzner et al., 2006] Metzner, P., Schutte, C., and Vanden-Eijnden, E. (2006). Illustration of transition path theory on a collection of simple examples. *The Journal of Chemical Physics*, 125(8):1–2.

[Metzner et al., 2009] Metzner, P., Schutte, C., and Vanden-Eijnden, E. (2009). Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219.

[Oksendal, 2003] Oksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Plotkin et al., 2019] Plotkin, D. A., Webber, R. J., O'Neill, M. E., Weare, J., and Abbot, D. S. (2019). Maximizing simulated tropical cyclone intensity with action minimization. *Journal of Advances in Modeling Earth Systems*, 11(4):863–891.

[Strahan et al., 2021] Strahan, J., Antoszewski, A., Lorpaiboon, C., Vani, B. P., Weare, J., and Dinner, A. R. (2021). Long-time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein. *Journal of Chemical Theory and Computation*, 17(5):2948–2963. PMID: 33908762.

[Thiede et al., 2019] Thiede, E., Giannakis, D., Dinner, A. R., and Weare, J. (2019). Approximation of dynamical quantities using trajectory data. *arXiv:1810.01841 [physics.data-an]*, pages 1–24.

[Vanden-Eijnden, 2006] Vanden-Eijnden, E. (2006). *Transition Path Theory*, pages 453–493. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Vanden-Eijnden and E, 2010] Vanden-Eijnden, E. and E, W. (2010). Transition-path theory and path-finding algorithms for the study of rare events. *Annual Review of Physical Chemistry*, 61(1):391–420.