Learning Forecasts of Rare Stratospheric Transitions from Short Simulations

JUSTIN FINKEL,[a] ROBERT J. WEBBER,[b] EDWIN P. GERBER,[c] DORIAN S. ABBOT,[d] AND JONATHAN WEARE[c]

[a] *Committee on Computational and Applied Mathematics, University of Chicago, Chicago, Illinois*
[b] *Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California*
[c] *Courant Institute of Mathematical Sciences, New York University, New York, New York*
[d] *Department of Geophysical Sciences, University of Chicago, Chicago, Illinois*

(Manuscript received 10 February 2021, in final form 14 August 2021)

ABSTRACT: Rare events arising in nonlinear atmospheric dynamics remain hard to predict and attribute. We address the problem of forecasting rare events in a prototypical example, sudden stratospheric warmings (SSWs). Approximately once every other winter, the boreal stratospheric polar vortex rapidly breaks down, shifting midlatitude surface weather patterns for months. We focus on two key quantities of interest: the probability of an SSW occurring, and the expected lead time if it does occur, as functions of initial condition. These *optimal forecasts* concretely measure the event's progress. Direct numerical simulation can estimate them in principle but is prohibitively expensive in practice: each rare event requires a long integration to observe, and the cost of each integration grows with model complexity. We describe an alternative approach using integrations that are *short* compared to the time scale of the warming event. We compute the probability and lead time efficiently by solving equations involving the transition operator, which encodes all information about the dynamics. We relate these optimal forecasts to a small number of interpretable physical variables, suggesting optimal measurements for forecasting. We illustrate the methodology on a prototype SSW model developed by Holton and Mass and modified by stochastic forcing. While highly idealized, this model captures the essential nonlinear dynamics of SSWs and exhibits the key forecasting challenge: the dramatic separation in time scales between a single event and the return time between successive events. Our methodology is designed to fully exploit high-dimensional data from models and observations, and has the potential to identify detailed predictors of many complex rare events in meteorology.

KEYWORDS: Stratospheric circulation; Extreme events; Stratosphere; Classification; Differential equations; Regression analysis; Risk assessment; Statistical techniques; Statistics; Uncertainty; Diagnostics; Nonlinear models; Parameterization; Stochastic models; Anomalies; Internal variability; Intraseasonal variability; Clustering; Model interpretation and visualization; Other artificial intelligence/machine learning

1. Introduction

As computing power increases and weather models grow more intricate and capable of generating a vast wealth of realistic data, the goal of extreme weather event prediction appears less distant (Vitart and Robertson 2018). To take full advantage of the increased computing power, we must develop new approaches to efficiently manage and parse the data we generate (or observe) to derive physically interpretable, actionable insights. Extreme weather events are worthy targets for simulation owing to their destructive potential to life and property. Rare events have attracted significant simulation efforts recently, including hurricanes (e.g., Zhang and Sippel 2009; Webber et al. 2019; Plotkin et al. 2019), heat waves (e.g., Ragone et al. 2018), rogue waves (e.g., Dematteis et al. 2018), and space weather events (e.g., coronal mass ejections; Ngwira et al. 2013). These are very difficult to characterize and predict, being exceptionally rare and pathological outliers in the spectrum of weather events. Ensemble forecasting in numerical weather prediction is best suited to estimate statistics of the average or most likely scenarios, and specialized methods are needed to examine the more extreme outlier scenarios.

In this study, we advance an alternative computational approach to predicting and understanding general rare events without sacrificing model fidelity.Our method relies on data generated by a high-fidelity model with a state space with many degrees of freedom d, representing, for example, spatial resolution of the primitive equations. In this way, our method is similar to recently introduced reduced order modeling techniques using statistical and machine learning (e.g., Kashinath et al. 2021 and references therein). However, in contrast to other data-driven techniques, our approach focuses on directly computing key quantities of interest that characterize the essential predictability of the rare event, rather than trying to capture the full detailed evolution of the system. In particular, we will compute estimators of *statistically optimal forecasts* that are useful for initial conditions somewhere between a "typical" configuration A and an "anomalous" configuration B that defines the rare event, where typical and anomalous are user-defined. We focus on two forecasts in particular to quantify risk. The *committor* is the probability that a given initial condition evolves directly into B rather than A. Given that it does reach B first, the *conditional mean first passage time*, or *lead time*, is the expected time that it takes to get there. The committor appears prominently in the molecular dynamics literature, with some recent applications in geoscience including Tantet et al. (2015), Lucente et al. (2019), and Finkel et al. (2020), which compute the committor for low-dimensional atmospheric models.

Both quantities depend on the initial condition, defining functions over d-dimensional state space that encode important information regarding the fundamental causes and precursors of the rare event. However, "decoding" the physical

Corresponding author: Justin Finkel, jfinkel@uchicago.edu

DOI: 10.1175/MWR-D-21-0024.1

insights is not automatic. With real-time measurement constraints, the risk metrics must be estimated from low-dimensional proxies. Even visualizing them requires projecting down to one or two dimensions. This calls for a principled selection of low-dimensional coordinates that are both physically meaningful and statistically informative for our chosen risk metrics. We address this problem using sparse regression, a simple but easily extensible solution with the potential to inform optimal measurement strategies to estimate risk as precisely as possible under constraints.

Estimation of the committor and lead time is a challenge. We employ a method that uses a large dataset of short-time independent simulations. We represent the committor and lead time as solutions to Feynman–Kac formulas (Oksendal 2003), which relate long-time forecasts to instantaneous tendencies. These equations are elegant and general, but computationally daunting: in the continuous time and space limit, they become partial differential equations (PDE) with d independent variables—the same as the model state space dimension. It is therefore hopeless to solve the equations using any standard spatial discretization. But, as we demonstrate, the equations can be solved with remarkable accuracy by expanding in a basis of functions informed by the dataset.

We illustrate our approach on the highly simplified Holton–Mass model (Holton and Mass 1976; Christiansen 2000) with stochastic velocity perturbations in the spirit of Birner and Williams (2008). The Holton–Mass model is well-understood dynamically in light of decades of analysis and experiments, yet complex enough to present the essential computational difficulties of probabilistic forecasting and test our methods for addressing them. In particular, this system captures the key difficulty in sampling rare events. The vast majority of the time, the system sits in one of two metastable states, characterizing a strong or weak vortex respectively. Extreme events are the infrequent jumps from one state to another. Our computational framework can accurately characterize these rare transitions using only a dataset of "short" model simulations: short not only compared to the long periods the system sits in one state or the other, but also relative to the time scale of the transition events themselves. In the future, the same methodology could be applied to query the properties of more complex models, such as GCMs, where less theoretical understanding is available.

In section 2, we review the dynamical model and define the specific rare event of interest. In section 3, we formally define the risk metrics introduced above and visualize the results for the Holton–Mass model, including a discussion of physical and practical insights gleaned from our approach. In section 4 we identify an optimal set of reduced coordinates for estimating risk using sparse regression. These results will provide motivation for the computational method, which we present afterward in section 5 along with accuracy tests. We then lay out future prospects and conclude in section 6.

2. Holton–Mass model

Holton and Mass (1976) devised a simple model of the stratosphere aimed at reproducing observed intraseasonal oscillations of the polar vortex, which they termed "stratospheric vacillation cycles." Earlier sudden stratospheric warming (SSW) models, originating with that of Matsuno (1971), proposed upward-propagating planetary waves as the major source of disturbance to the vortex. While Matsuno (1971) used impulsive forcing from the troposphere as the source of planetary waves, Holton and Mass (1976) suggested that even stationary tropospheric forcing could lead to an oscillatory response, suggesting that the stratosphere can self-sustain its own oscillations. While the Holton–Mass model is meant to represent internal stratospheric dynamics, Sjoberg and Birner (2014) point out that the stationary boundary condition does not lead to stationary wave activity flux, meaning that even the Holton–Mass model involves some dynamic interaction between the troposphere and stratosphere. Isolating internal from external dynamics is a subtle modeling question, but in the present paper we adhere to the original Holton–Mass framework for simplicity. Our methodology applies equally well to other formulations.

Radiative cooling through the stratosphere and wave perturbations at the tropopause are the two competing forces that drive the vortex in the Holton–Mass model. Altitude-dependent cooling relaxes the zonal wind toward a strong vortex in thermal wind balance with a radiative equilibrium temperature field. Gradients in potential vorticity along the vortex, however, can allow the propagation of Rossby waves. When conditions are just right, a Rossby wave emerges from the tropopause and rapidly propagates upward, sweeping heat poleward and stalling the vortex by depositing a burst of negative momentum. The vortex is destroyed and begins anew the rebuilding process.

Yoden (1987a) found that for a certain range of parameter settings, these two effects balance each other to create two distinct stable regimes: a strong vortex with zonal wind close to the radiative equilibrium profile, and a weak vortex with a possibly oscillatory wind profile. We focus our study on this bistable setting as a prototypical model of atmospheric regime behavior. The transition from strong to weak vortex state captures the essential dynamics of an SSW.

The Holton–Mass model takes the linearized quasi-geostrophic potential vorticity (QGPV) equation for a perturbation streamfunction $\psi'(x, y, z, t)$ on top of a zonal mean flow $\overline{u}(y, z, t)$, and projects these two fields onto a single zonal wavenumber $k = 2/(a\cos 60°)$ and a single meridional wavenumber $\ell = 3/a$, where a is Earth's radius. This notation is consistent with Holton and Mass (1976) and Christiansen (2000), and we refer the reader to these earlier papers for complete description of the equations and parameters. The resulting ansatz is

$$\begin{aligned} \overline{u}(y,z,t) &= U(z,t)\sin(\ell y), \\ \psi'(x,y,z,t) &= \mathrm{Re}\{\Psi(z,t)e^{ikx}\}\, e^{z/2H}\sin(\ell y), \end{aligned} \tag{1}$$

which is fully determined by the reduced state space $U(z, t)$, and $\Psi(z, t)$, the latter being complex valued; H is a scale height, 7 km. Inserting this into the linearized QGPV equations yields the coupled PDE system

$$\left\{-\left[\mathcal{G}^2(k^2+\ell^2)+\frac{1}{4}\right]+\frac{\partial^2}{\partial z^2}\right\}\frac{\partial \Psi}{\partial t} = \left[\left(\frac{\alpha}{4}-\frac{\alpha_z}{2}-i\mathcal{G}^2k\beta\right)-\alpha_z\frac{\partial}{\partial z}-\alpha\frac{\partial^2}{\partial z^2}\right]\Psi + \left\{ik\varepsilon\left[\left(k^2\mathcal{G}^2+\frac{1}{4}\right)-\frac{\partial}{\partial z}+\frac{\partial^2}{\partial z^2}\right]U\right\}\Psi - ik\varepsilon\frac{\partial^2\Psi}{\partial z^2}U, \quad (2)$$

for $\Psi(z, t)$, and

$$\left(-\mathcal{G}^2\ell^2-\frac{\partial}{\partial z}+\frac{\partial^2}{\partial z^2}\right)\frac{\partial U}{\partial t} = [(\alpha_z-\alpha)U_z^R-\alpha U_{zz}^R] - \left[(\alpha_z-\alpha)\frac{\partial}{\partial z}+\alpha\frac{\partial^2}{\partial z^2}\right]U+\frac{\varepsilon k\ell^2}{2}e^z\,\mathrm{Im}\left\{\Psi\frac{\partial^2\Psi^*}{\partial z^2}\right\}, \quad (3)$$

for $U(z, t)$. Here, $\varepsilon = 8/(3\pi)$ is a coefficient for projecting $\sin^2(\ell y)$ onto $\sin(\ell y)$. We have nondimensionalized the equations with the parameter $\mathcal{G}^2 = H^2N^2/(f_0^2L^2)$, where $N^2 = 4\times10^{-4}\,\mathrm{s}^{-2}$ is a constant stratification (Brunt–Väisälä frequency), f_0 is the Coriolis parameter, and $L = 2.5\times10^5\,\mathrm{m}$ is a horizontal length scale, selected in order to create a homogeneously shaped dataset more suited to our analysis. See Holton and Mass (1976), Yoden (1987a), and Christiansen (2000) for details on parameters. Boundary conditions are prescribed at the bottom of the stratosphere, which in this model corresponds to $z = 0\,\mathrm{km}$, and the top of the stratosphere $z_{\mathrm{top}} = 70\,\mathrm{km}$.

$$\Psi(0,t)=\frac{gh}{f_0},\qquad \Psi(z_{\mathrm{top}},t)=0,$$
$$U(0,t)=U^R(0),\qquad \partial_z U(z_{\mathrm{top}},t)=\partial_z U^R(z_{\mathrm{top}}). \quad (4)$$

The vortex-stabilizing influence is represented by $\alpha(z)$, the altitude-dependent cooling coefficient, and the radiative wind profile $U^R(z) = U^R(0) + (\gamma/1000)z$ (with z in m), which relaxes the vortex toward radiative equilibrium. Here $\gamma = \mathcal{O}(1)$ is the vertical wind shear ($\mathrm{m\,s^{-1}\,km^{-1}}$). The competing force of wave perturbation is encoded through the lower boundary condition $\Psi(0, t) = gh/f_0$.

Detailed bifurcation analysis of the model by both Yoden (1987a) and Christiansen (2000) in (γ, h) space revealed the bifurcations that lead to bistability, vacillations, and ultimately quasiperiodicity and chaos. Here we will focus on an intermediate parameter setting of $\gamma = 1.5\,\mathrm{m\,s^{-1}\,km^{-1}}$ and $h = 38.5\,\mathrm{m}$, where two stable states coexist: a strong vortex with U closely following U^R and an almost barotropic stationary wave, as well as a weak vortex with U dipping close to zero at an intermediate altitude and a stationary wave with strong westward phase tilt. The two stable equilibria, which we call **a** and **b**, are illustrated in Figs. 1a and 1b by their z-dependent zonal wind and perturbation streamfunction profiles.

The two equilibria can be interpreted as two different winter climatologies, one with a strong vortex and one with a weak vortex susceptible to vacillation cycles. To explore transitions between these two states, we follow Birner and Williams (2008) and modify the Holton–Mass equations with small additive noise in the U variable to mimic momentum perturbations by smaller scale Rossby waves, gravity waves, and other unresolved sources. The form of noise will be specified in Eq. (7).

While the details of the additive noise are ad hoc, the general approach can be more rigorously justified through the Mori–Zwanzig formalism (Zwanzig 2001). Because many hidden degrees of freedom are being projected onto the low-dimensional space of the Holton–Mass model, the dynamics on small observable subspaces can be considered stochastic. This is the perspective taken in stochastic parameterization of turbulence and other high-dimensional chaotic systems (Hasselmann 1976; DelSole and Farrell 1995; Franzke and Majda 2006; Majda et al. 2001; Gottwald et al. 2016). In general, unobserved deterministic dynamics can make the system non-Markovian, which technically violates the assumptions of our methodology. However, with sufficient separation of time scales the Markovian assumption is not unreasonable. Furthermore, memory terms can be ameliorated by lifting data back to higher-dimensional state space with time-delay embedding (Berry et al. 2013; Thiede et al. 2019; Lin and Lu 2021).

We follow Holton and Mass (1976) and discretize the equations using a finite-difference method in z, with 27 vertical levels (including boundaries). After constraining the boundaries, there are $d = 3\times(27-2) = 75$ degrees of freedom in the model. Christiansen (2000) investigated higher resolution and found negligible differences. The full discretized state is represented by a long vector

$$\begin{aligned}\mathbf{X}(t) = [&\mathrm{Re}\{\Psi\}(\Delta z,t),\ldots,\mathrm{Re}\{\Psi\}(z_{\mathrm{top}}-\Delta z,t),\\ &\mathrm{Im}\{\Psi\}(\Delta z,t),\ldots,\mathrm{Im}\{\Psi\}(z_{\mathrm{top}}-\Delta z,t),\\ &U(\Delta z,t),\ldots,U(z_{\mathrm{top}}-\Delta z,t)]\in\mathbb{R}^d=\mathbb{R}^{75}\end{aligned} \quad (5)$$

The deterministic system can be written $d\mathbf{X}(t)/dt = \mathbf{v}[\mathbf{X}(t)]$ for a vector field $\mathbf{v}\colon \mathbb{R}^d \to \mathbb{R}^d$ specified by discretizing (2) and (3). Under deterministic dynamics, $\mathbf{X}(t) \to \mathbf{a}$ or $\mathbf{X}(t) \to \mathbf{b}$ as $t \to \infty$ depending on initial conditions. The addition of white noise changes the system into an Itô diffusion

$$d\mathbf{X}(t) = \mathbf{v}[\mathbf{X}(t)]dt + \boldsymbol{\sigma}[\mathbf{X}(t)]d\mathbf{W}(t), \quad (6)$$

where $\boldsymbol{\sigma}\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$ imparts a correlation structure to the vector $\mathbf{W}(t) \in \mathbb{R}^m$ of independent standard white noise processes. As discussed above, we design $\boldsymbol{\sigma}$ to be a low-rank, constant matrix that adds spatially smooth stirring to only the zonal wind U (not the streamfunction Ψ) and which respects boundary conditions at the bottom and top of the stratosphere. Its structure is defined by the following Euler–Maruyama scheme: in a time step $\delta t = 0.005$ days, after a deterministic forward Euler step we add the stochastic perturbation to zonal wind on large vertical scales

$$\delta U(z) = \sigma_U\sum_{k=0}^{m}\eta_k\sin\left[\left(k+\frac{1}{2}\right)\pi\frac{z}{z_{\mathrm{top}}}\right]\sqrt{\delta t}, \quad (7)$$

where $\eta_k (k = 0, 1, 2)$ are independent unit normal samples, $m = 2$, and σ_U is a scalar that sets the magnitudes of entries in $\boldsymbol{\sigma}$. In terms of physical units,

$$\sigma_U^2 = \frac{\mathbb{E}[(\delta U)^2]}{\delta t}\approx(1\,\mathrm{m\,s^{-1}})^2\,\mathrm{day}^{-1}, \quad (8)$$

σ_U has units of $(L/T)/T^{1/2}$, where the square root of time comes from the quadratic variation of the Wiener process. It is best

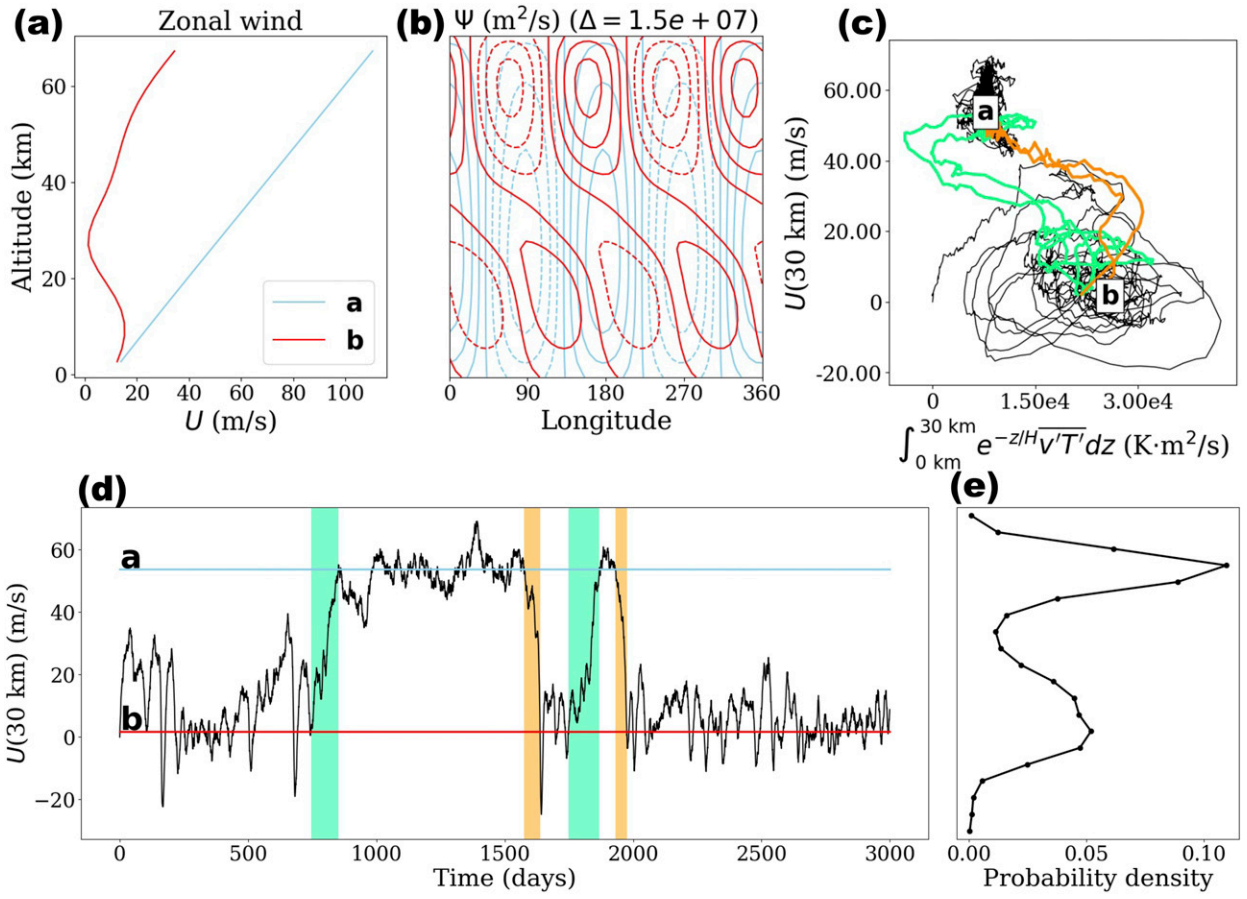

FIG. 1. Illustration of the two stable states of the Holton–Mass model and transitions between them. (a) Zonal wind profiles of the radiatively maintained strong vortex (the fixed point **a**, blue), which increases linearly with altitude, and the weak vortex (the fixed point **b**, red), which dips close to zero in the midstratosphere. (b) Streamfunction contours are overlaid for the two equilibria **a** and **b**. (c) Parametric plot of a control simulation in a two-dimensional state space projection, including two transitions from A to B (orange) and from B to A (green). (d) Time series of $U(30\,\mathrm{km})$ from the same simulation. (e) The steady-state density projected onto $U(30\,\mathrm{km})$.

interpreted in terms of the daily root-mean-square velocity perturbation of $1.0\,\mathrm{m\,s^{-1}}$. We have experimented with this value, and found that reducing the noise level below 0.8 dramatically reduces the frequency of transitions, while increasing it past 1.5 washes out metastability. We keep σ_U constant going forward as a favorable numerical regime to demonstrate our approach, while acknowledging that the specifics of stochastic parameterization are important in general to obtain accurate forecasts. The resulting matrix $\boldsymbol{\sigma}$ is 75×3, with nonzero entries only in the last 25 rows as forcing only applies to $U(z)$.

A long simulation of the model reveals metastability, with the system tending to remain close to one fixed point for a long time before switching quickly to the other, as shown by the time series of $U(30\,\mathrm{km})$ in Fig. 1d. Figure 1e shows a projection of the steady-state distribution, also known as the equilibrium/invariant distribution, of U as a function of z. We call this density $\pi(\mathbf{x})$, which is a function over the full d-dimensional state space. We focus on the zonal wind U at 30 km following Christiansen (2000), because this is where its strength is minimized in the weak vortex. While the two regimes are clearly associated with the two fixed points, they are better characterized by extended *regions* of state space with strong and weak vortices. We thus define the two metastable subsets of $\mathbb{R}^d$:

$$A = \{\mathbf{X}: U(\mathbf{X})(30\,\mathrm{km}) \geq U(\mathbf{a})(30\,\mathrm{km}) = 53.8\,\mathrm{m\,s^{-1}}\},$$
$$B = \{\mathbf{X}: U(\mathbf{X})(30\,\mathrm{km}) \leq U(\mathbf{b})(30\,\mathrm{km}) = 1.75\,\mathrm{m\,s^{-1}}\}.$$

This straightforward definition roughly follows the convention of Charlton and Polvani (2007), which defines an SSW as a reversal of zonal winds at 10 hPa. We use 30 km for consistency with Christiansen (2000); this is technically higher than 10 hPa because $z = 0$ in the Holton–Mass model represents the tropopause. Our method is equally applicable to any definition, and the results are not qualitatively dependent on this choice. Incidentally, the analysis tools we present may be helpful in distinguishing predictability properties between different definitions. In fact, we will show that the height neighborhood of 20 km is actually more salient for predicting the event than wind at the 30-km level, even when the event is defined by wind at 30 km! This emerges from statistical analysis alone, and gives us confidence that essential SSW

properties are stable with respect to reasonable changes in definition.

The orange highlights in Fig. 1d begin when the system exits the A region bound for B, and end when the system enters B. The green highlights start when the system leaves B bound for A, and end when A is reached. Note that $A \to B$ transitions, SSWs, are much shorter in duration than $B \to A$ transitions. Figure 1c shows the same paths but viewed parametrically in a two-dimensional state space consisting of integrated heat flux (IHF) $\int_{0\,\mathrm{km}}^{30\,\mathrm{km}} e^{-z/H}\overline{v'T'}\,dz$, and zonal wind $U(30\,\mathrm{km})$. IHF is an informative number because it captures both magnitude and phase information of the streamfunction in the Holton–Mass model:

$$\mathrm{IHF} = \int_{0\,\mathrm{km}}^{30\,\mathrm{km}} e^{-z/H}\overline{v'T'}\,dz \propto \int_{0\,\mathrm{km}}^{30\,\mathrm{km}} |\Psi|^2 \frac{\partial \varphi}{\partial z} dz, \tag{9}$$

where φ is the phase of Ψ. The $A \to B$ and $B \to A$ transitions are again highlighted in orange and green respectively, showing geometrical differences between the two directions. We will refer to the $A \to B$ transition as an SSW event, even though it is more accurately a transition between climatologies according to the Holton–Mass interpretation. The $B \to A$ transition is a vortex restoration event. Our focus in this paper is on predicting these transition events (mainly the $A \to B$ direction) and monitoring their progress in a principled way. In the next section we explain the formalism for doing so.

3. Forecast functions: The committor and lead time statistics

a. *Defining risk and lead time*

We will introduce the quantities of interest by way of example. First, suppose the stratosphere is observed in an initial state $\mathbf{X}(0) = \mathbf{x}$ that is neither in A nor B, so $U(\mathbf{b})(30\,\mathrm{km}) < U(\mathbf{x})(30\,\mathrm{km}) < U(\mathbf{a})(30\,\mathrm{km})$ and the vortex is somewhat weakened, but not completely broken down. We call this intermediate zone $D = (A \cup B)^c$ (the complement of the two metastable sets). Because A and B are attractive, the system will soon find its way to one or the other at the *first-exit time* from D, denoted

$$\tau_{D^c} = \min\{t \geq 0 : \mathbf{X}(t) \in D^c\}. \tag{10}$$

Here, D^c emphasizes that the process has left D, that is, gone to A or B. The first-exit location $\mathbf{X}(\tau_{D^c})$ is itself a random variable that importantly determines how the system exits D: either $\mathbf{X}(\tau_{D^c}) \in A$, meaning the vortex restores to radiative equilibrium, or $\mathbf{X}(\tau_{D^c}) \in B$, meaning the vortex breaks down into vacillation cycles. A fundamental goal of forecasting is to determine the probabilities of these two events, which naturally leads to the definition of the (forward) committor function

$$q^+(\mathbf{x}) = \begin{cases} \mathbb{P}_{\mathbf{x}}\{\mathbf{X}(\tau_{D^c}) \in B\} & \mathbf{x} \in D = (A \cup B)^c \\ 0 & \mathbf{x} \in A \\ 1 & \mathbf{x} \in B \end{cases}, \tag{11}$$

where the subscript $\mathbf{x}$ indicates that the probability is conditional on a fixed initial condition $\mathbf{X}(0) = \mathbf{x}$, that is, $\mathbb{P}_{\mathbf{x}}\{\cdot\} = \mathbb{P}\{\cdot\,|\mathbf{X}(0) = \mathbf{x}\}$. The superscript "+" distinguishes the forward committor from the *backward committor*, an analogous quantity for the time-reversed process that we do not use in this paper. Throughout, we will use capital $\mathbf{X}(t)$ to denote a stochastic process, and lower-case $\mathbf{x}$ to represent a specific point in state space, typically an initial condition, that is, $\mathbf{X}(0) = \mathbf{x}$. Both are $d = 75$-dimensional vectors.

The committor is the probability that the system will be in state B (the disturbed state) next rather than A (the strong vortex state). Hence $q^+(\mathbf{x}) = 0$ if you start in A, and is 1 if you are already in B. In between (i.e., when $\mathbf{x} \in D$), $q^+(\mathbf{x})$ tells you the probability that you will first go to B rather than to A. That is, $q^+(\mathbf{x})$ tells you the probability that an SSW will happen.

Another important forecasting quantity is the lead time to the event of interest. While the forward committor reveals the probability of experiencing vortex breakdown *before* returning to a strong vortex, it does not say how long either event will take. Furthermore, even if the vortex is restored first, how long will it be until the next SSW does occur? The time until the next SSW event is denoted τ_B, again a random variable, whose distribution depends on the initial condition $\mathbf{x}$. We call $\mathbb{E}_{\mathbf{x}}[\tau_B]$ the *mean first passage time* (MFPT) to B. Conversely, we may ask how long a vortex disturbance will persist before normal conditions return; the answer (on average) is $\mathbb{E}_{\mathbf{x}}[\tau_A]$, the mean first passage time to A. These same quantities have been calculated previously in other simplified models, for example, Birner and Williams (2008) and Esler and Mester (2019).

The $\mathbb{E}_{\mathbf{x}}[\tau_B]$ has an obvious shortcoming: it is an average over all paths starting from $\mathbf{x}$, including those that go straight into B (i.e., an orange trajectory in Figs. 1c,d) and the rest that return to A, that is, a green trajectory) and linger there, potentially for a very long time, before eventually recrossing back into B. It is more relevant for near-term forecasting to condition τ_B on the event that an SSW is coming before the strong vortex returns. For this purpose, we introduce the *conditional* mean first passage time, or lead time, to B:

$$\eta^+(\mathbf{x}) := \mathbb{E}_{\mathbf{x}}[\tau_B | \tau_B < \tau_A], \tag{12}$$

which quantifies the suddenness of SSW.

All of these quantities can, in principle, be estimated by direct numerical simulation (DNS). For example, suppose we observe an initial condition $\mathbf{X}(0) = \mathbf{x}$ in an operational forecasting setting, and wish to estimate the probability and lead time for the event of next hitting B. We would initialize an ensemble $\{\mathbf{X}_n(0) = \mathbf{x}, \quad n = 1, \ldots, N\}$ and evolve each member forward in time until it hits A or B at the random time τ_n. In an explicitly stochastic model, random forcing would drive each member to a different fate, while in a deterministic model their initial conditions would be perturbed slightly. To estimate the committor to B, we could calculate the fraction of members that hit B first. Averaging the arrival times (τ_n), over only those members, gives an estimate of the lead time to B. For a single initial condition $\mathbf{x}$ reasonably close to B, DNS may be the most economical. But how do we systematically compute $q^+(\mathbf{x})$ over all of state space (here 75 variables, but potentially billions of variables in a GCM or other state-of-the-art forecast system)?

For this more ambitious goal, DNS is prohibitively expensive. By definition, transitions between A and B are infrequent. Therefore, if starting from $\mathbf{x}$ far from B, a huge number of sampled trajectories (N) will be required to observe even a small number ending in B, and they may take a long time to get there. If instead we could precompute these functions offline over all of state space, the online forecasting problem would reduce to "reading off" the committor and lead time with every new observation. Achieving this goal is the key point of our paper, and we achieve this using the dynamical Galerkin approximation (DGA) recipe described by Thiede et al. (2019).

A brute force way to estimate these functions is to integrate the model for a long time until it reaches statistical steady state, meaning it has explored its attractor thoroughly according to the steady-state distribution. After long enough, it will have wandered close to every point $\mathbf{x}$ sufficiently often to estimate $q^+(\mathbf{x})$ and η^+ robustly as in DNS. We have performed such a "control simulation" of 5×10^5 days for validation purposes, but our main contribution in this paper is to compute the forecast functions using only *short* trajectories with DGA, allowing for massive parallelization. However, we will defer the methodological details to section 5, and first justify the effort with some results. We visualize the committor and lead time computed from short trajectories and elaborate on their interpretation, utility, and relationship to ensemble forecasting methods.

b. Steady-state distribution

Before visualizing the committor and lead time, it will be helpful to have a precise notion of the steady-state distribution, denoted $\pi(\mathbf{x})$, a probability density that describes the long-term behavior of a stochastic process $\mathbf{X}(t)$. Assuming the system is ergodic, averages over time are equivalent to averages over state space with respect to π. That is, for any well-behaved function $g\colon \mathbb{R}^d \to \mathbb{R}$,

$$\langle g \rangle_\pi := \lim_{T\to\infty} \frac{1}{T}\int_0^T g[\mathbf{X}(t)]\,dt = \int_{\mathbb{R}^d} g(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}. \tag{13}$$

For example, if $g(\mathbf{x}) = \mathbb{1}_S(\mathbf{x})$ (an indicator function, which is 1 for $\mathbf{x} \in S \subset \mathbb{R}^d$ and 0 for $\mathbf{x} \notin S$), Eq. (13) says that the fraction of time spent in S can be found by integrating the density over S. The density peaks in Fig. 1d indicates clearly that the neighborhoods of **a** and **b** are two such regions with especially large probability under π. Note that both sides of (13) are independent of the initial condition, which is forgotten eventually. Short-term forecasts are by definition out-of-equilibrium processes, depending critically on initial conditions; however, $\pi(\mathbf{x})$ is important to us here as a "default" distribution for missing information. If the initial condition is only partially observed, for example, in only one coordinate, we have no information about the other $d-1$ dimensions, and in many cases the most principled tactic is to assume those other dimensions are distributed according to π, conditional on the observation.

c. Visualizing committor and lead times

The forecasts $q^+(\mathbf{x})$ and $\eta^+(\mathbf{x})$ are functions of a high-dimensional space $\mathbb{R}^d$. However, these degrees of freedom may not all be "observable" in a practical sense, given the sparsity and resolution limits of weather sensors, and visualizing them requires projecting onto reduced-coordinate spaces of dimension 1 or 2. We call these "collective variables" (CVs) following chemistry literature (e.g., Noé and Clementi 2017), and denote them as vector-valued functions from the full state space to a reduced space, $\boldsymbol{\theta}\colon \mathbb{R}^d \to \mathbb{R}^k$, where $k = 1$ or 2. For instance, Fig. 1c plots trajectories in the CV space consisting of integrated heat flux and zonal wind at 30 km: $\boldsymbol{\theta}(\mathbf{x}) = \left[\int_{0\,\mathrm{km}}^{30\,\mathrm{km}} e^{-z/H}\overline{v'T'}dz,\ U(30\,\mathrm{km})\right]$. The first component is a nonlinear function involving products of $\mathrm{Re}\{\Psi\}$ and $\mathrm{Im}\{\Psi\}$, while the second component is a linear function involving only U at a certain altitude. For visualization in general, we have to approximate a function $F\colon \mathbb{R}^d \to \mathbb{R}$, such as the committor or lead time, as a function of reduced coordinates. That is, we wish to find $f\colon \mathbb{R}^k \to \mathbb{R}$ such that $F(\mathbf{x}) \approx f[\boldsymbol{\theta}(\mathbf{x})]$. Given a fixed CV space $\boldsymbol{\theta}$, an "optimal" f is chosen by minimizing some function-space metric between $f \circ \boldsymbol{\theta}$ and F.

A natural choice is the mean-squared error weighted by the steady-state distribution π, so the projection problem is to minimize over functions $f\colon \mathbb{R}^k \to \mathbb{R}$ the penalty

$$\begin{aligned} S[f;\boldsymbol{\theta}] &:= ||f\circ\boldsymbol{\theta} - F||^2_{L^2(\pi)} \\ &= \int_{\mathbb{R}^d} \{f[\boldsymbol{\theta}(\mathbf{x})] - F(\mathbf{x})\}^2 \pi(\mathbf{x})\,d\mathbf{x}. \end{aligned} \tag{14}$$

The optimal f for this purpose is the conditional expectation

$$\begin{aligned} f(\mathbf{y}) &= \mathbb{E}_{\mathbf{X}\sim\pi}[F(\mathbf{X})|\boldsymbol{\theta}(\mathbf{X}) = \mathbf{y}] \\ &= \lim_{|d\mathbf{y}|\to 0} \frac{\int f(\mathbf{x})\mathbb{1}_{d\mathbf{y}}[\boldsymbol{\theta}(\mathbf{x})]\pi(\mathbf{x})\,d\mathbf{x}}{\int \mathbb{1}_{d\mathbf{y}}[\boldsymbol{\theta}(\mathbf{x})]\pi(\mathbf{x})\,d\mathbf{x}}, \end{aligned} \tag{15}$$

where $d\mathbf{y}$ is a small neighborhood about $\mathbf{y}$ in CV space R^k. The subscript $\mathbf{X} \sim \pi$ means that the expectation is with respect to a random variable $\mathbf{X}$ distributed according to $\pi(\mathbf{x})$, that is, at steady state. Figure 2 uses this formula to display one-dimensional projections of the committor (first row) and lead time (second row), as well as the one standard deviation envelope incurred by projecting out the other 74 degrees of freedom. This "projection error" is defined as the square root of the conditional variance:

$$V_F(\mathbf{y}) = \mathbb{E}_{\mathbf{X}\sim\pi}[[F(\mathbf{X}) - f(\mathbf{y})]^2|\boldsymbol{\theta}(\mathbf{X}) = \mathbf{y}]. \tag{16}$$

Each quantity is projected onto two different one-dimensional CVs: $U(30\,\mathrm{km})$ (first column) and IHF (second column). In Fig. 2a, for example, we see the committor is a decreasing function of U: the weaker the wind, the more likely a vortex breakdown. Moreover, the curve provides a conversion factor between risk (as measured by probability) and a physical variable, zonal wind. An observation of $U(30\ \mathrm{km}) = 38\,\mathrm{m\,s^{-1}}$ implies a 50% chance of vortex breakdown. The variation in slope also tells us that a wind reduction from 40 to $30\,\mathrm{m\,s^{-1}}$ represents a far greater increase in risk than a reduction from 30 to $20\,\mathrm{m\,s^{-1}}$. Meanwhile, Fig. 2b shows the committor to be an increasing function of IHF, since SSW is associated with large wave amplitude and phase lag. However, IHF seems

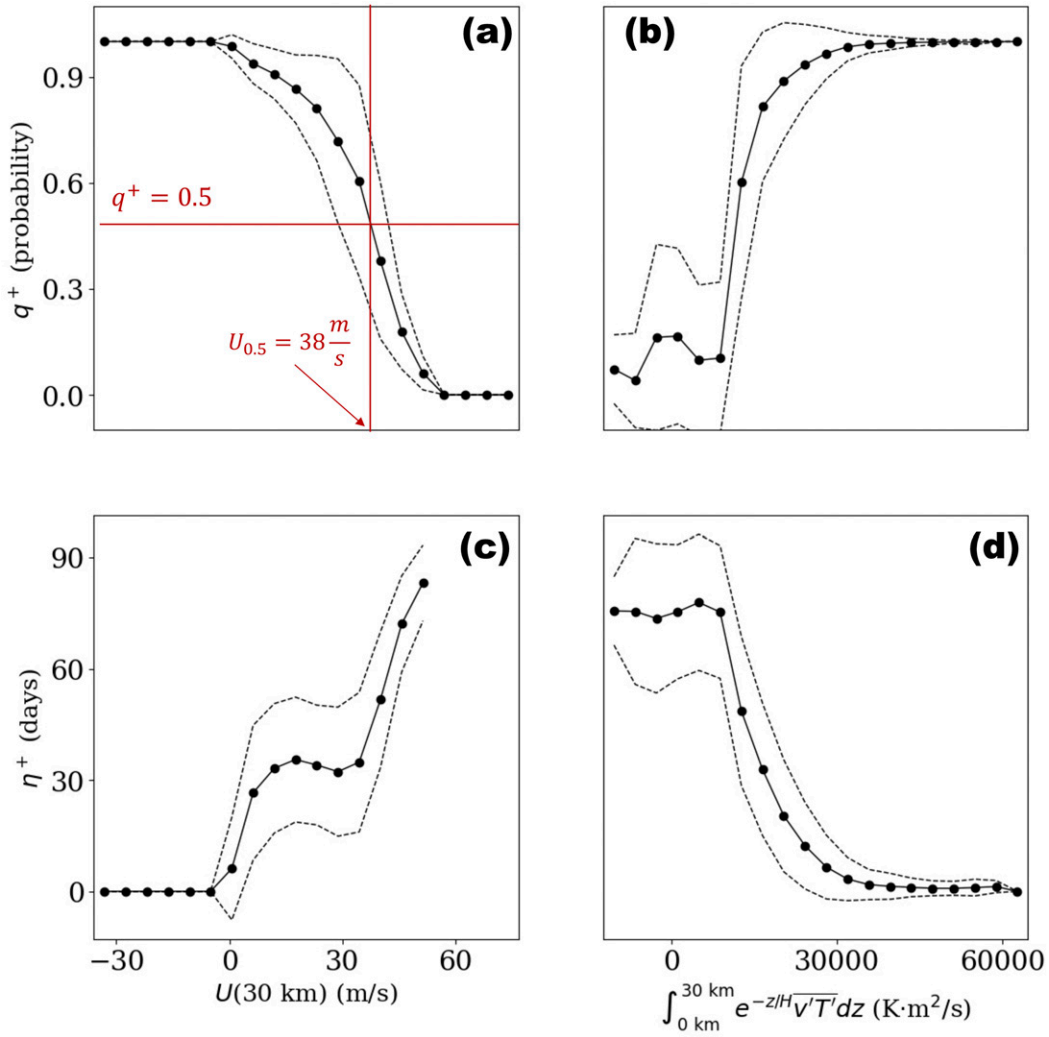

FIG. 2. (a),(b) One-dimensional projections of the forward committor and (c),(d) lead time to B. These functions depend on all $d = 75$ degrees of freedom in the model, but we have averaged across $d - 1 = 74$ dimensions to visualize them as rough functions of two single degrees of freedom: (left) U(30 km) and (right) integrated heat flux up to 30 km, IHF. Additionally (a) marks the $q^+ = 1/2$ threshold and the corresponding value of zonal wind.

inferior to zonal wind as a committor proxy, as a small change in IHF from ~15 000 to ~20 000 K m^2 s^{-1} corresponds to a sharp increase in committor from nearly zero to nearly one. In other words, knowing only IHF does not provide much useful information about the threat of SSW until it is already virtually certain. The dotted envelope is also wider in Fig. 2b than Fig. 2a, indicating that projecting the committor onto IHF removes more information than projecting onto U. While the underlying noise makes it impossible to divine the outcome with certainty from *any* observation, the projection error clearly privileges some observables over others for their predictive power.

In Figs. 2c and 2d, the lead time is seen to have the opposite overall trend as the committor: the weaker the wind, or the greater the heat flux, the closer you are on average to a vortex breakdown. The $\eta^+(\mathbf{x})$ is not defined when wind is strongest, as $\mathbf{x} \in A$ and so $q^+(\mathbf{x}) = 0$. However, an interesting exception to the trend occurs in the range $10\,\mathrm{m\,s^{-1}} \leq U \leq 40\,\mathrm{m\,s^{-1}}$: the expected lead time stays constant or slightly *decreases* as zonal wind increases, and the projection error remains large. This means that while the probability of vortex breakdown increases rapidly from 50% to 90%, the time until vortex breakdown remains highly uncertain. To resolve this seeming paradox, we will have to visualize the joint variation of q^+ and η^+.

It is of course better to consider multiple observables at once. Figure 3 shows the information gained beyond observing U(30 km) by incorporating IHF as a second observable. In the top row we project π, q^+, and η^+ onto the two-dimensional subspace, revealing structure hidden from view in the one-dimensional projections. Figure 3a is a two-dimensional extension of Fig. 1d, with density peaks visible in the neighborhoods of **a** and **b**. The white space surrounding the gray represents physically insignificant regions of state space that

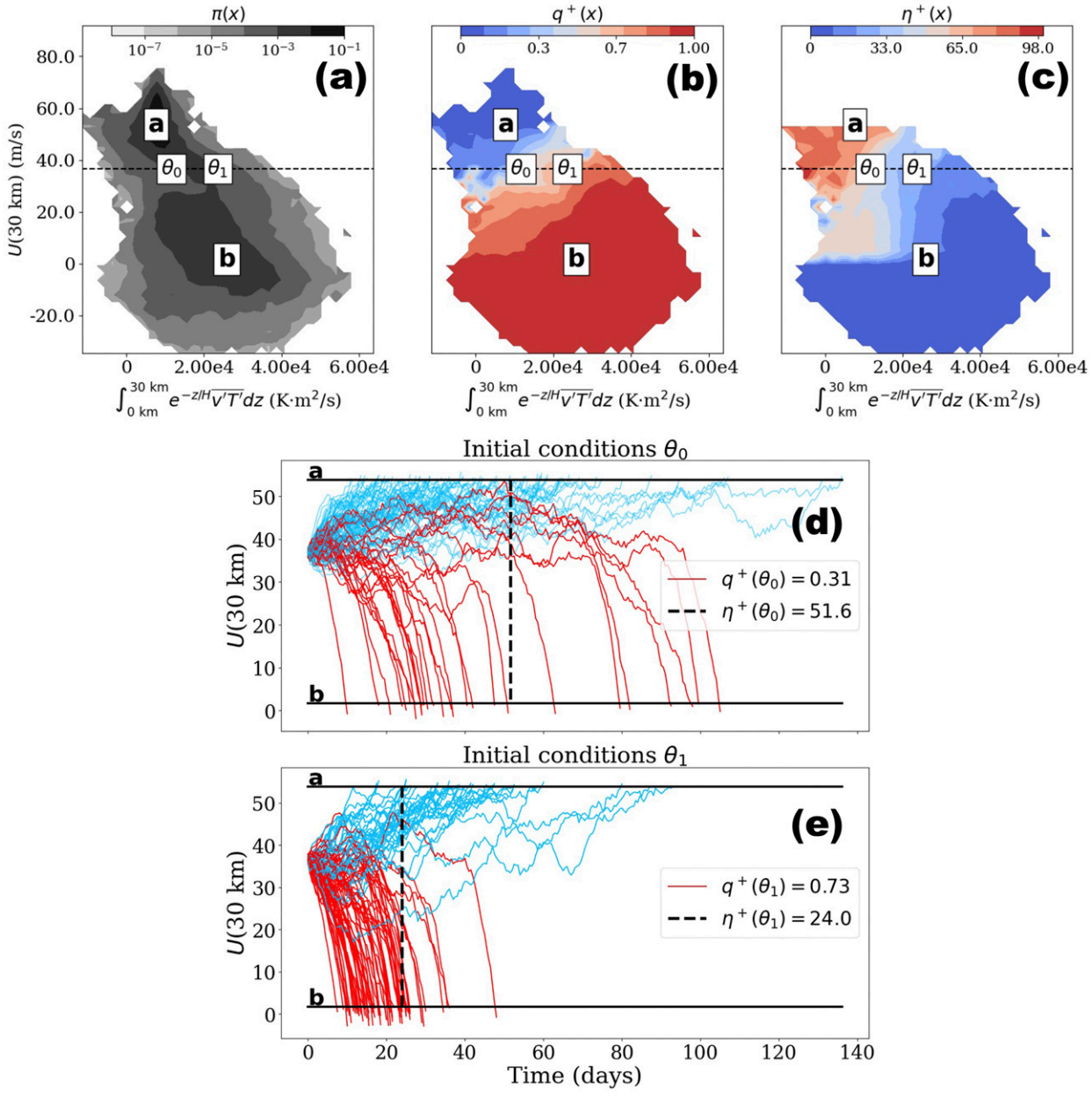

FIG. 3. The density, committor, and lead time as functions of zonal wind and integrated heat flux. (a) The steady-state distribution $\pi(\mathbf{x})$ onto the two-dimensional subspace (U, IHF) at 30 km. The white regions surrounding the gray are unphysical states with negligible probability. (b),(c) Display the committor and lead time in the same space. A horizontal transect marks the level $U(30\,\text{km}) = 38.5\,\text{m s}^{-1}$, where q^+ according to U only is 0.5. (d),(e) Ensembles initialized from two points $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ along the transect, verifying that their committor and lead time values differ from their values according to U, in a way that is predictable due to considering IHF in addition to U.

was not sampled by the long simulation. The same convention holds for the following two-dimensional figures. The committor is displayed in Fig. 3b over the same space. It changes from blue at the top (an SSW is unlikely) to red at the bottom (an SSW is likely), bearing out the negative association between U and q^+. However, there are nonnegligible horizontal gradients that show that IHF plays a role, too. Likewise, the lead time in Fig. 3c decreases from ~90 days near **a** to 0 days near **b**, when the transition is complete. Here, IHF appears even more critically important for forecasting how the event plays out, as gradients in η^+ are often completely horizontal.

A horizontal dotted line in Figs. 3a–c marks the 50% risk level $U(30\,\text{km}) = 38\,\text{m s}^{-1}$, but the committor varies along it from low risk at the left to high risk at the right: we show this concretely by selecting two points $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ along the line. According to U alone, that is, the curve in Fig. 2a, both would have the same committor of 0.5. According to both U and IHF together, that is, the two-dimensional heat map in Fig. 3b, they

have very different probabilities of $q^+(\boldsymbol{\theta}_0) = 0.31$ and $q^+(\boldsymbol{\theta}_1) = 0.73$: an SSW is more than twice as likely to occur from starting point $\boldsymbol{\theta}_1$ as $\boldsymbol{\theta}_0$.

While those committor values come from the DGA method to be described in section 5, we confirm them empirically by plotting an ensemble of 100 trajectories originating from each of the two initial conditions in Figs. 3d and 3e below, coloring A-bound trajectories blue and B-bound trajectories red. Only 28% of the sampled trajectories through $\boldsymbol{\theta}_0$ exhibit an SSW, next going to state B, while 68% of the integrations from $\boldsymbol{\theta}_1$ end at B. In both cases, the heat maps and ensemble sample means roughly match. The small differences between the projected committor and the empirical "success" rate of trajectories arises both from errors in the DGA calculation (which we analyze in section 5) and the finite size of the ensemble.

The lead time prediction is improved similarly by incorporating the second observable. According to U alone, Fig. 2 predicts a lead time of 40 days for both $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Considering IHF additionally, the two-dimensional heat map in Fig. 3 predicts a lead time of 52 and 24 days for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively. Referring to the ensemble from $\boldsymbol{\theta}_1$ in Figs. 3d and 3e, the arrival times of red trajectories to B provide a discrete sampling of the lead time distributions of $\tau_B|\tau_B < \tau_A$. The sample means are 50 and 32 days respectively from $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, again roughly matching with our predictions.

These two-dimensional projections still leave out 73 remaining dimensions, which we could incorporate to make the forecasts even better. After accounting for all 75 dimensions, we would obtain the full committor function $q^+: \mathbb{R}^d \to \mathbb{R}$. This is still a probability, that is, an expectation over the unresolved turbulent processes and uncertain initial condition. Low-dimensional committor projections simply treat the projected-out dimensions as random variables sampled according to π. Whether projected to a space of 1 or 75 dimensions, the committor is the function of that space that is closest, in the mean-square sense, to the binary indicator $\mathbb{1}_B[\mathbf{X}(\tau)]$; this is the defining characteristic of conditional expectation (Durrett 2013). In the case that the system does hit B next, the lead time is closest in the mean-square sense to τ_B.

While high-dimensional systems offer many coordinates to choose from, we argue that the committor and lead time are the most important nonlinear coordinates to monitor for forecasting purposes. We will explore their relationship in the next subsection. Although both encode some version of proximity to SSW, they are independent variables that deserve separate consideration.

d. Relationship between risk and lead time

A forecast is most useful if it comes sufficiently early (to leave some buffer time before impact) and is sufficiently precise to time your response. For example, in June we can say with certainty it will snow next winter in Minnesota. To be useful, we want to know the date of the first snow as early as possible. By relating levels of risk (quantified by q^+) and lead time (quantified by η^+), we can now assess the limits of early prediction. Such a relationship would answer two questions: For an SSW transition, 1) how far in advance will we be aware of it with some prescribed confidence, say 80%? 2) Given some prescribed lead time, say 42 days, how aware or ignorant could we be of it?

The committor and lead time have an overall negative relationship, but they do not completely determine each other, as the contours in Figs. 3a and 3b do not perfectly line up. We treat them as independent variables in Fig. 4, which maps zonal wind and IHF as functions of the coordinates q^+ and η^+ in an inversion of Fig. 3. The density $\pi(\mathbf{x})$ projected on this space in Fig. 4a shows again a bimodal structure around **a** and **b**, which occupy opposite corners of this space by construction. Meanwhile, zonal wind and IHF are indicated by the shading in Fig. 4b and 4c. The bridge between **a** and **b** is not a narrow band, but rather includes a curious high-committor, high lead time branch that seems paradoxical: points at $q^+ = 0.9$ have a greater spread in η^+ than points at $q^+ = 0.5$, contrary to the intuition that closeness to B in probability means closeness in time. The color shading shows that q^+ is strongly associated with U(30 km), while η^+ is more strongly associated with IHF(30 km). In particular, the horizontal contours in Fig. 4c show that the large spread in lead time near B is due almost completely to variation in IHF. In other words, the system can be highly committed to B with a low zonal wind, but if IHF is low, it may take a long time to get there. We can also see this from the lower-left region of Figs. 3a and 3b, where committor is high and lead time is high.

There are two complementary explanations for this phenomenon. First, the low-U, low-IHF region of state space corresponds to a temporary restoration phase in a vacillation cycle, which delays the inevitable collapse of zonal wind below the threshold defining B. In fact, the ensemble of pathways starting from $\boldsymbol{\theta}_0$ in Fig. 3c has several members whose zonal wind either stagnates at medium strength, or dips low and partially restores before finally plunging all the way down. The second explanation is that many of these partial restoration events are not part of an $A \to B$ transition, but rather a $B \to B$ transition. In a highly irreversible system such as the Holton–Mass model, these two situations are quite dynamically distinct. To distinguish them using DGA, we would have to account for the *past* as well as the future, calculating backward-in-time forecasts such as the backward committor $q^-(\mathbf{x}) = \mathbb{P}_{\mathbf{x}}\{\mathbf{X}(\tau^-) \in A\}$, where $\tau^- < 0$ is the most-recent hitting time. Backward forecasts will be analyzed thoroughly in a forthcoming paper, but they are beyond the scope of the present one.

In summary, q^+ and η^+ are principled metrics to inform preparation for extreme weather. For example, a threatened community might decide in advance to start taking action when an event is very likely, $q^+ \geq 0.8$, and somewhat imminent, $\eta^+ \leq 10$ days, or rather, when an event is somewhat likely, $q^+ \geq 0.5$, and very imminent, $\eta^+ \leq 3$ days. Because of partial restoration events, the committor does not determine the lead time or vice versa, and so a good real-time disaster response strategy should take both of them into account, defining an "alarm threshold" that is not a single number, but some function of both the committor and lead time. This idea is similar in spirit to that of the Torino scale, which assigns a single risk metric to an asteroid or comet impacts based on both probability and severity (Binzel 2000). Of course, after many near-SSW events, a lot of material damage may have already occurred, which may be a reason to define a higher threshold for the definition of B, or even a continuum for different severity levels

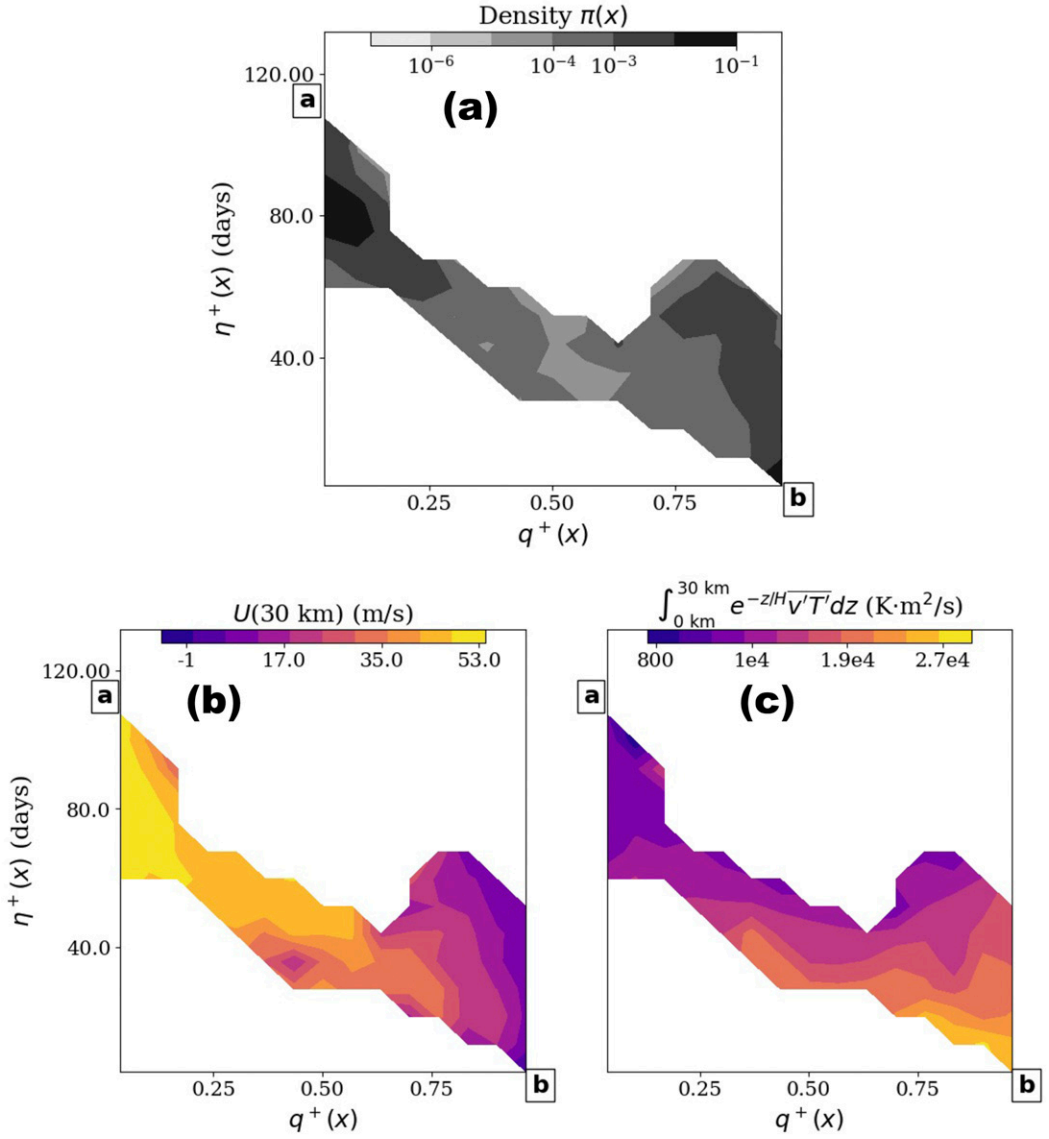

FIG. 4. Committor and lead time as independent coordinates. This figure inverts the functions in Fig. 3, considering the zonal wind and integrated heat flux as functions of committor and lead time. The two-dimensional space they span is the essential goal of forecasting. (a) The steady-state distribution on this subspace, which is peaked near **a** and **b** (darker shading), weaker in the "bridge" region between them, and completely negligible the white regions unexplored by data. (b),(c) Zonal wind and heat flux in color as functions of the committor and lead time.

of SSW. We emphasize that the choice of A, B, and alarm thresholds are more of a community and policy decision than a scientific one. The strength of our approach is that it provides a flexible numerical framework to quantify and optimize the consequences of those decisions.

4. Sparse representation of the committor

The committor projections showed give only an impression of its high-dimensional structure. While Eq. (15) says how to optimally represent the committor over a given CV subspace, optimizing $S[f;\boldsymbol{\theta}]$ over f, it does not say which subspace $\boldsymbol{\theta}$ is optimal. If the committor does admit a sparse representation, we could specifically target observations on these high-impact signals. In this section we address this much harder problem of optimizing $S[f;\boldsymbol{\theta}]$ over subspaces $\boldsymbol{\theta}$.

The set of CV spaces is infinite, as observables $\boldsymbol{\theta}$ can be arbitrarily complex nonlinear functions of the basic state variables $\mathbf{x}$. Machine learning algorithms such as artificial neural networks are designed exactly for that purpose: to represent functions nonparametrically from observed input–output pairs. However, to keep the representation interpretable, we will restrict ourselves

to physics-informed input features based on the Eliassen–Palm (EP) relation, which relates wave activity, PV fluxes and gradients, and heating source terms in a conservation equation. From Yoden (1987b), the EP relation for the Holton–Mass model takes the form

$$\partial_t\left(\frac{q'^2}{2}\right) + (\partial_y\overline{q})\rho_s^{-1}\nabla\cdot\mathbf{F} = -\frac{f_0^2}{N^2}\rho_s^{-1}\overline{q'\partial_z(\alpha\rho_s\partial_z\psi')}, \quad (17)$$

$$\text{where} \quad \mathbf{F} = (-\rho_s\overline{u'v'})\mathbf{j} + (\rho_s\overline{v'\partial_z\psi'})\mathbf{k}.$$

The EP flux divergence has two alternative expressions: $\rho_s^{-1}\nabla\cdot\mathbf{F} = \overline{v'q'} = \rho_s^{-1}(R/Hf_0)\partial_z(\rho_s\overline{v'T'})$. If there were no dissipation ($\alpha = 0$) and the background zonal state were time-independent ($\partial_t\overline{q} = 0$), dividing both sides by $\partial_y\overline{q}$ would express local conservation of wave activity $\mathcal{A} = \rho_s\overline{q'^2}/(2\partial_y\overline{q})$. Neither of these is exact in the stochastic Holton–Mass model, so we use the quantities in Eq. (17) as diagnostics: enstrophy $\overline{q'^2}$, PV gradient $\partial_y\overline{q}$, PV flux $\overline{v'q'}$, and heat flux $\overline{v'T'}$. Each field is a function of (y, z) and takes on very different profiles for the states **a** and **b**, as found by Yoden (1987b). A transition from A to B, where the vortex weakens dramatically, must entail a reduction in $\partial_y\overline{q}$ and a burst in positive $\overline{v'T'}$ (negative $\overline{v'q'}$) as a Rossby wave propagates from the tropopause vertically up through the stratosphere and breaks. This is the general physical narrative of a sudden warming event, and these same fields might be expected to be useful observables to track for qualitative understanding and prediction. For visualization, we have found $U(30\,\text{km})$ and $\text{IHF}(30\,\text{km}) = \int_{0\,\text{km}}^{30\,\text{km}} e^{-z/H}\overline{v'T'}\,dz$ to be particularly helpful. However, this does not necessarily imply they are optimal predictors of q^+, and regression is a more principled way to find them.

We start by projecting the committor onto each observable at each altitude separately, in hopes of finding particularly salient altitude levels that clarify the role of vertical interactions. The first five rows of Fig. 5 display, for five fields (U, $|\Psi|$, $\overline{q'^2}$, $\partial_y\overline{q}$, and $\overline{v'q'}$) and for a range of altitude levels, the mean and standard deviation of the committor projected onto that field at that altitude. Each altitude has a different range of the CV; for example, because U has a Dirichlet condition at the bottom and a Neumann condition at the top, the lower levels have a much smaller range of variability than the high levels. We also plot the integrated variance, or L^2 projection error, at each level in the right-hand column. A low projected committor variance over U at altitude z_0 means that the committor is mostly determined by the single observable $U(z_0)$, while a high projected variance indicates significant dependence of q^+ on variables other than $U(z_0)$. To compare different altitudes and fields as directly as possible, the L^2 projection error at each altitude is an average over discrete bins of the observable.

In selecting good CVs, we generally look for a simple, hopefully monotonic, and sensitive relationship with the committor. Of all the candidate fields, U and $\partial_y\overline{q}$ stand out the most in this respect, being clearly negatively correlated with the forward committor at all altitudes. The associated projection error tends to be greatest in the region $q^+ \approx 0.5$, as observed before, but interestingly there is a small altitude band around 15–25 km where its magnitude is minimized. This suggests an optimal altitude for monitoring the committor through zonal wind, giving the most reliable estimate possible for a single state variable. In contrast, the projection of q^+ onto $|\Psi|$, displays a large variance across all altitudes. The eddy enstrophy and potential vorticity flux are also rather unhelpful as early warning signs, despite their central role in SSW evolution. For example, the large, positive spikes in heat flux across all altitudes generally occur after the committor ≈ 0.5 threshold has already been crossed. Furthermore, the relationship of $\overline{v'q'}$ with the committor is not smooth. The $q^+ < 0.5$ region at each altitude is a thin band near zero.

The exhaustive CV search in Fig. 5 is visually compelling in favor of some fields and some altitudes over others, but it is not satisfactory as a rigorous comparison. Differences between units and ranges make it difficult to objectively compare the L^2 projection error. Furthermore, restricting to one variable at a time is limiting. Accordingly, we also perform a more automated approach to identify salient variables in the form of a generalized linear model for the forward committor, using sparsity-promoting least absolute shrinkage and selection operator (LASSO) regression due to Tibshirani (1996), as implemented in the scikit-learn Python package (Pedregosa et al. 2011). As input features, we use all state variables $\text{Re}\{\Psi\}$, $\text{Im}\{\Psi\}$, U, the integrated heat flux $\int_0^z e^{-z/H}\overline{v'T'}\,dz$, the eddy PV flux $\overline{v'q'}$, and the background PV gradient $\partial_y\overline{q}$, at all altitudes z simultaneously. The advantage of a sparsity-promoting regression is that it isolates a small number of observables that can accurately approximate the committor in linear combination. Considering that regions close to A and B have low committor uncertainty, we regress only on data points with $q^+ \in (0.2, 0.8)$, and of those only a subset weighted by $\pi(\mathbf{x})q^+(\mathbf{x})[1 - q^+(\mathbf{x})]$ to further emphasize the transition region $q^+ \approx 0.5$. To constrain committor predictions to the range $(0, 1)$, we regress on the committor after an inverse-sigmoid transformation, $\ln[q^+/(1 - q^+)]$. First we do this at each altitude separately, and in Fig. 6a we plot the coefficients of each component as a function of altitude. The bottom row of Fig. 5 also displays the committor projected on the height-dependent LASSO predictor.

The height-dependent regression in Fig. 6a shows each component is salient for some altitude range. In general, U and $\text{Im}\{\Psi\}$ dominate as causal variables at low altitudes, while $\text{Re}\{\Psi\}$ dominates at high altitudes. The overall prediction quality, as measured by R^2 and plotted in Fig. 6b, is greatest around 21.5 km, consistent with our qualitative observations of Fig. 5. Note that not all single-altitude slices are sufficient for approximating the committor, even with LASSO regression; in the altitude band 50–60 km, the LASSO predictor is not monotonic and has a large projected variance, as seen in the bottom row of Fig. 5. The specific altitude can matter a great deal. But by using all altitudes at once, the committor approximation may be improved further. We thus repeat the LASSO with all altitudes simultaneously and find the sparse coefficient structure shown in Fig. 6c, with a few variables contributing the most, namely the state variables Ψ and U in

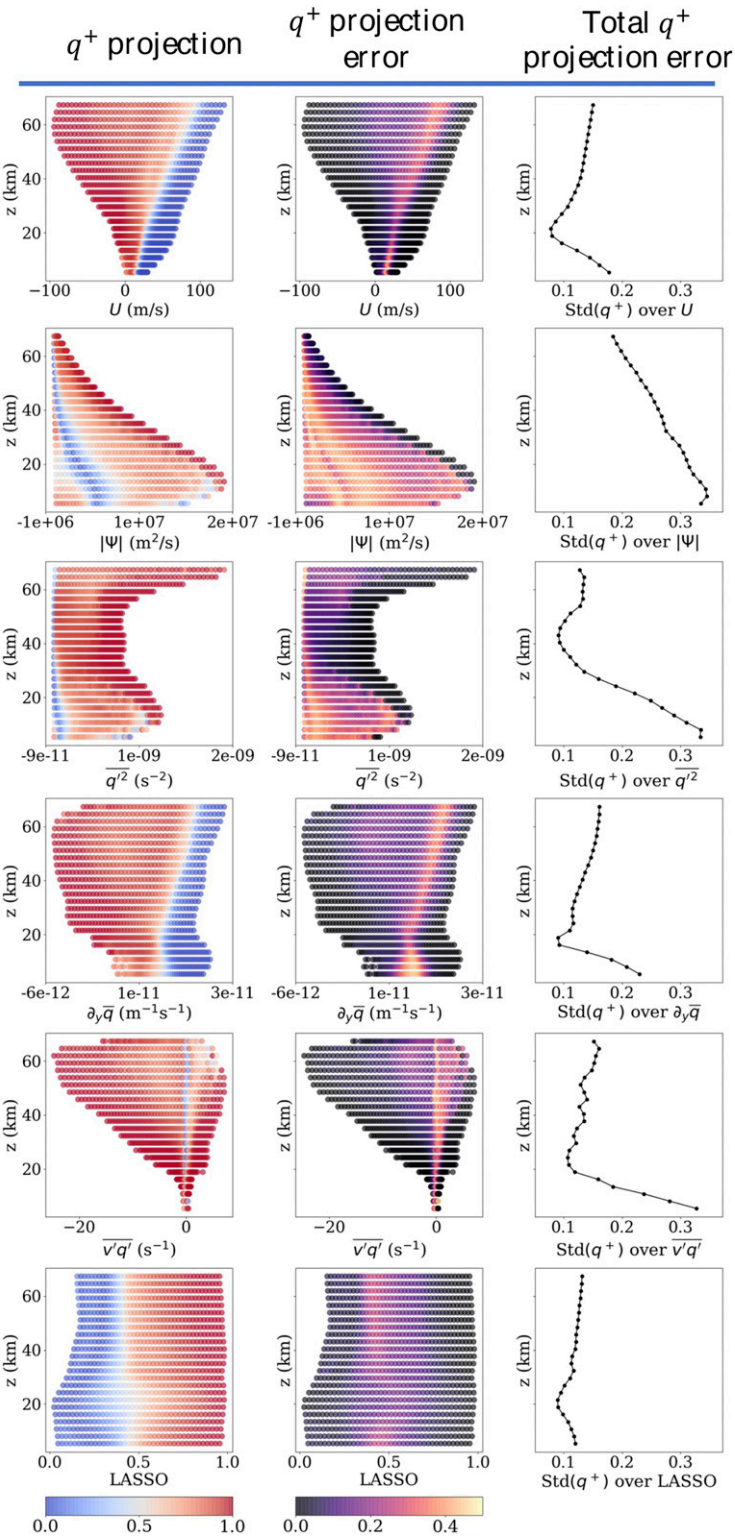

FIG. 5. Projection of the forward committor onto a large collection of altitude-dependent physical variables. (top left) Heat maps of q^+ as a function of U and z; white regions denote where $U(z)$ is negligibly observed. (top center) The standard deviation in q^+ as a function of U and z; this uncertainty stems from the remaining 74 model dimensions. (top right) The total mean-squared error due to the projection for each altitude, i.e., $\sqrt{S[f;\boldsymbol{\theta}]}$ from Eq. (14). A low value indicates that this level is ideal for prediction. The remaining rows show the same quantities as in the top row for other physical variables: streamfunction magnitude, eddy enstrophy, background PV gradient, eddy PV flux, and LASSO.

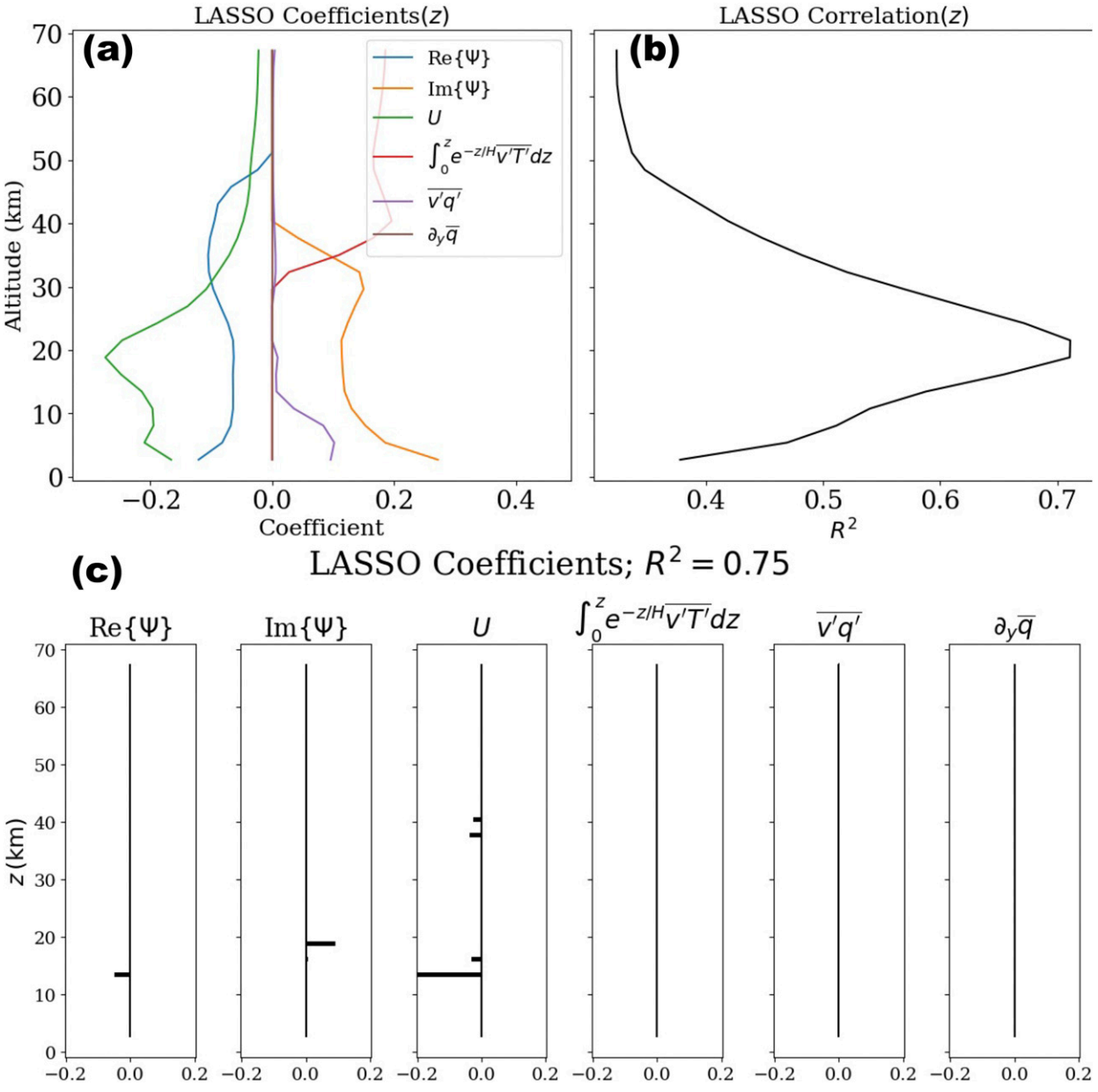

FIG. 6. Results of LASSO regression of the forward committor with linear and nonlinear input features. (a) The coefficients when q^+ is regressed as a function of only the variables at a given altitude, and (b) the corresponding correlation score. 21.5 km seems the most predictive (where $z \equiv 0$ at the tropopause, not the surface). (c) The coefficient structure when all altitudes are considered simultaneously. Most of the nonzero coefficients appear between 15 and 22 km, distinguishing that range as highly relevant for prediction.

the altitude range 15–22 km. The nonlinear CVs failed to make any nonzero contribution to LASSO, and this remained stubbornly true for other nonlinear combinations not shown, such as $\overline{v'T'}$. With multiple lines of evidence indicating 21.5 km as an altitude with high predictive value for the forward committor, we can make a strong recommendation for targeting observations here. This conclusion applies only to the Holton–Mass model under these parameters, but the methodology explained above can be applied similarly to models of arbitrary complexity.

We have presented the committor and lead time as "ideal" forecasts, especially the committor, which we have devoted considerable effort to approximating in this section. We want to emphasize that q^+ and η^+ are not competitors to ensemble forecasting; rather, they are two of its most important end results. So far, we have simply advocated including q^+ and η^+ as quantities of interest. Going forward, however, we do propose an alternative to ensemble forecasting aimed specifically at the committor, lead time, and a wider class of forecasting functions, as they are important enough in their own right to

warrant dedicated computation methods. Our approach uses only short simulations, making it highly parallelizable, and shifts the numerical burden from online to offline. Figures 2–6 were all generated using the short-simulation algorithm. While the method is not yet optimized and in some cases not competitive with ensemble forecasting, we anticipate such methods will be increasingly favorable with modern trends in computing.

5. The computational method

In this section we describe the methodology, which involves some technical results from stochastic processes and measure theory. After describing the theoretical motivation and the numerical pipeline in turn, we demonstrate the method's accuracy and discuss its efficiency compared to straightforward ensemble forecasting.

a. Feynman–Kac formulas

The forecast functions described above—committors and passage times—can all be derived from general conditional expectations of the form

$$F(\mathbf{x};\lambda) = \mathbb{E}_{\mathbf{x}}\left[G[\mathbf{X}(\tau)]\exp\left\{\lambda\int_0^\tau \Gamma[\mathbf{X}(s)]\,ds\right\}\right], \tag{18}$$

where again the subscript $\mathbf{x}$ denotes conditioning on $\mathbf{X}(0) = \mathbf{x}$; G, Γ are arbitrary known functions over $\mathbb{R}^d$; and τ is a stopping time, specifically a first-exit time like Eq. (10) but possibly with D replaced by another set. The term λ is a variable parameter that turns F into a moment-generating function. To see that the forward committor takes on this form, set $G(\mathbf{x}) = \mathbb{1}_B(\mathbf{x})$, $\lambda = 0$ (Γ can be anything), and $\tau = \tau_{A\cup B}$. Then $F(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}_B[\mathbf{X}(\tau)]] = \mathbb{P}_{\mathbf{x}}\{\mathbf{X}(\tau_{D^c}) \in B\} = q^+(\mathbf{x})$. For the η^+, set $\tau = \tau_B$, $G = \mathbb{1}_B$, and $\Gamma = 1$. Then

$$F(\mathbf{x};\lambda) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}_B[\mathbf{X}(\tau)]\exp(\lambda\tau)] \tag{19}$$

$$\frac{1}{q^+(\mathbf{x})}\frac{\partial}{\partial\lambda}F(\mathbf{x};0) = \frac{\mathbb{E}_{\mathbf{x}}[\tau\mathbb{1}_B[\mathbf{X}(\tau)]]}{\mathbb{E}_{\mathbf{x}}[\mathbb{1}_B[\mathbf{X}(\tau)]]} \tag{20}$$

$$= \eta^+(\mathbf{x}). \tag{21}$$

So we must also be able to differentiate F with respect to λ.

More generally, the function G is chosen by the user to quantify risk at the terminal time τ; in the case of the forward committor, that risk is binary, with an SSW representing a positive risk and a radiative vortex no risk at all. The function Γ is chosen to quantify the risk accumulated up until time τ, which might be simply an event's duration, but other integrated risks may be of more interest for the application. For example, one could express the total poleward heat flux by setting $\Gamma = \overline{v'T'}$, or the momentum lost by the vortex by setting $\Gamma(\mathbf{x}) = U(\mathbf{a}) - U(\mathbf{x})$. Extending (20), one can compute not only means but higher moments of such integrals by expressing the risk with Γ. Repeated differentiation of $F(\mathbf{x};\lambda)$ gives

$$\partial_\lambda^k F(\mathbf{x};0) = \mathbb{E}_{\mathbf{x}}\left[G[\mathbf{X}(\tau)]\left\{\int_0^\tau \Gamma[\mathbf{X}(s)]\,ds\right\}^k\right] \tag{22}$$

We choose to focus on expectations of the form (18) in order to take advantage of the Feynman–Kac formula, which represents $F(\mathbf{x};\lambda)$ as the solution to a PDE boundary value problem over state space. As PDEs involve local operators, this form is more amenable to solution with short trajectories that do not stray far from their source. The boundary value problem associated with (18) is

$$\begin{cases} (\mathcal{L} + \lambda\Gamma)F(\mathbf{x};\lambda) = 0 & \mathbf{x} \in D \\ F(\mathbf{x};\lambda) = G(\mathbf{x}) & \mathbf{x} \in D^c \end{cases}. \tag{23}$$

The domain D here is some combination of A^c and B^c. The operator $\mathcal{L}$ is known as the *infinitesimal generator* of the stochastic process, which acts on functions by pushing expectations forward in time along trajectories:

$$\mathcal{L}f(\mathbf{x}) := \lim_{\Delta t\to 0}\frac{\mathbb{E}_{\mathbf{x}}[f[\mathbf{X}(\Delta t)]] - f(\mathbf{x})}{\Delta t}. \tag{24}$$

In a diffusion process like the stochastic Holton–Mass model, $\mathcal{L}$ is an advection–diffusion partial differential operator that is analogous to a material derivative in fluid mechanics. The generator encapsulates the properties of the stochastic process. In addition to solving boundary value problems (18), its adjoint $\mathcal{L}^*$ provides the Fokker–Planck equation for the stationary density $\pi(\mathbf{x})$:

$$\mathcal{L}^*\pi(\mathbf{x}) = 0. \tag{25}$$

We can also write equations for moments of F, as in (22), by differentiating (23) repeatedly and setting $\lambda = 0$:

$$\mathcal{L}[\partial_\lambda^k F](\mathbf{x};0) = -k\Gamma\partial_\lambda^{k-1}F. \tag{26}$$

This is an application of the Kac moment method (Fitzsimmons and Pitman 1999). Note that we never actually have to solve (23) with nonzero λ. Instead we implement the recursion above. Note that the base case, $k = 0$, with $G = \mathbb{1}_B$ gives $F^+ = q^+$, no matter what the risk function Γ. In this paper we compute only up to the first moment, $k = 1$. Further background regarding stochastic processes and Feynman–Kac formulas can be found in Karatzas and Shreve (1998), Oksendal (2003), E et al. (2019).

b. Dynamical Galerkin approximation

To solve the boundary value problem (23) with $\lambda = 0$, we start by following the standard finite element recipe, converting to a variational form, and projecting onto a finite basis. First, we homogenize boundary conditions by writing $F(\mathbf{x}) = \hat{F}(\mathbf{x}) + f(\mathbf{x})$, where $\hat{F}$ is a guess function that obeys the boundary condition $\hat{F}|_{D^c} = G$, and $f|_{D^c} = 0$. Next, we integrate the equation against any test function ϕ, weighting the integrand by a density μ (which is arbitrary for now, but will be specified later):

$$\begin{aligned}\int_{\mathbb{R}^d}\phi(\mathbf{x})\mathcal{L}f(\mathbf{x})\mu(\mathbf{x})\,d\mathbf{x} &= \int\phi(\mathbf{x})(G - \mathcal{L}\hat{F})(\mathbf{x})\mu(\mathbf{x})\,d\mathbf{x} \\ \langle\phi, \mathcal{L}f\rangle_\mu &= \langle\phi, G - \mathcal{L}\hat{F}\rangle_\mu.\end{aligned} \tag{27}$$

The test function ϕ should live in the same space as f, that is, with homogeneous boundary conditions $\phi(\mathbf{x}) = 0$ for $\mathbf{x} \in A \cup B$. We refer to the inner products in (27) as being "with respect to" the measure (with density) μ. We approximate f by expanding in a finite basis $f(\mathbf{x}) = \sum_{j=1}^M \xi_j\phi_j(\mathbf{x})$ with unknown coefficients ξ_j, and enforce that (27) hold for each ϕ_i. This reduces the problem to a system of linear equations,

$$\sum_{j=1}^{M} \langle \phi_i, \mathcal{L}\phi_j \rangle_\mu \xi_j = \langle \phi_i, G - \mathcal{L}\hat{F} \rangle_\mu \quad i = 1, \ldots, M, \tag{28}$$

which can be solved with standard numerical linear algebra packages.

This procedure consists of three crucial subroutines. First, we must construct a set of basis functions ϕ_j. Second, we have to evaluate the generator's action on them, $\mathcal{L}\phi_j$. Third, we have to compute inner products. With standard PDE methods, the basis size would grow exponentially with dimension, quickly rendering the first and third steps intractable. Successful approaches will involve a representation of the solution F, suitable for the high dimensional setting, that is, representations of the type commonly employed for machine learning tasks. DGA is one such method, whose special twist is to construct a "data informed" basis of reasonable size, evaluate the generator by implementing Eq. (24) with the same dataset, and finally evaluate the inner products (27) with a Monte Carlo integral. The data consist of short trajectories launched from all over state space, which the system of linear equations stitches together into a global function estimate. We sketch the procedure here, but for the implementation details we refer to the appendix and to Thiede et al. (2019) and Strahan et al. (2021), where DGA has already been developed for molecular dynamics.

Step 1: Generate the data, in the format of N initial conditions $\{\mathbf{X}_n : 1 \le n \le N\}$. Evolve each initial condition forward for a "lag time" Δt to obtain a set of short trajectories $\{\mathbf{X}_n(t) : 0 \le t \le \Delta t,\ n = 1, \ldots, N\} \subset \mathbb{R}^d$. (Lag time is an algorithmic parameter for DGA. It is not to be confused with the forecast time horizon between the prediction and the event of interest in meteorology.) Here and going forward, $\mathbf{X}_n$ will mean $\mathbf{X}_n(0)$. The choice of starting points is flexible, but crucial for the efficiency and accuracy of DGA. Because our goal here is to demonstrate interpretable results, we prioritize simplicity and accuracy over efficiency, and defer optimization to later work. We simply draw initial conditions at random from the long control simulation of 5×10^5 days, and then generate new short trajectories from those points. We do not sample the points with equal probability, but instead reweight to get a uniform distribution over the space $[U(30\,\mathrm{km}), |\Psi|(30\,\mathrm{km})]$, within the bounds realized by the control simulation, which are approximately $-30\,\mathrm{m\,s^{-1}} \le U(30\,\mathrm{km}) \le 70\,\mathrm{m\,s^{-1}}$ and $0\,\mathrm{m^2\,s^{-1}} \le |\Psi|(30\,\mathrm{km}) \le 2 \times 10^7\,\mathrm{m^2\,s^{-1}}$. This sampling procedure, and any other version, implicitly defines a *sampling measure* μ on state space, where $\mu(\mathbf{x})d\mathbf{x}$ is the expected fraction of starting points in the neighborhood $d\mathbf{x}$ about **x.** Sampling points with equal weight from the control run would induce $\mu = \pi$, a very inefficient choice because probability concentrates around the metastable states **a** and **b.** The reweighting procedure ensures data coverage of intermediate-wind regions between A and B, as well as the large bursts of wave amplitude that characterize the transition pathways. Our main results use $N = 5 \times 10^5$ short trajectories with a lag time of $\Delta t = 20$ days, sampled at a frequency of twice per day. This dataset is more than needed to get a reasonable committor estimate, but we have sampled generously in order to visualize the functions in high detail. The final section will show the method is robust, capable of reasonably approximating the committor even with an order-of-magnitude reduction in data.

Step 2: Define the basis. The Galerkin method works for any class of basis functions that becomes increasingly expressive as the library grows and becomes capable of estimating any function of interest. However, with a finite truncation, choosing basis functions is a crucial ingredient of DGA, greatly impacting the efficiency and accuracy of the results. In our current study, we restrict to the simplest kind of basis, which consists of indicator functions $\phi_i(x) = \mathbb{1}_{S_i}(x)$, where $\{S_1, \ldots, S_M\}$ is a disjoint partition of state space. In practice we will construct these sets by clustering the initial data points as described in more detail in the appendix. This is a common practice in the computational statistical mechanics community for building a Markov state model (MSM) (Chodera et al. 2006; Frank and Fischer 2008; Pande et al. 2010; Bowman et al. 2013; Chodera and Noé 2014). MSMs are a dimensionality reduction technique that has also been used in conjunction with analysis of metastable transitions, primarily in protein folding dynamics (Noé et al. 2009). MSMs have also been used recently to study garbage patch dynamics in the ocean (Miron et al. 2021) as well as complex social dynamics (Helfmann et al. 2021). In Maiocchi et al. (2020), the authors take an interesting approach to MSMs by clustering points based on proximity to unstable periodic orbits, a potentially useful paradigm for general chaotic weather phenomena (Lucarini and Gritsun 2020). DGA can be viewed as an extension of MSMs, though, rather than producing any reduced complexity model, the explicit goal in DGA is estimating specific functions as in Eq. (18).

Step 3: Apply the generator. The forward difference formula

$$\widehat{\mathcal{L}\phi}(\mathbf{X}_n) = \frac{\phi[\mathbf{X}_n(\Delta t)] - \phi(\mathbf{X}_n)}{\Delta t} \tag{29}$$

suggested by the definition of the generator (24), results in a systematic bias when Δt is finite. On the other hand, small values of Δt lead to large variances in our Monte Carlo estimates of the inner products in (28). To resolve these issues, we use an integrated form of the Feynman–Kac equations that involves stopping trajectories when they enter A or B. Details are provided in the appendix.

Step 4: Compute the inner products. The inner products in Eq. (28) are integrals over high-dimensional state space that are intractable with standard quadrature, but can be approximated using Monte Carlo integration. If $\mathbf{X}$ is an $\mathbb{R}^d$-valued random variable distributed according to μ, and we have access to random samples $\{X_1, \ldots, X_N\}$ (which we do), the law of large numbers gives, for any function g with finite expectation,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(\mathbf{X}_n) = \int_{\mathbb{R}^d} g(\mathbf{x})\mu(\mathbf{x})\, d\mathbf{x}. \tag{30}$$

Setting $g(\mathbf{x}) = \phi_i(\mathbf{x})\mathcal{L}\phi_j(\mathbf{x})$, the sample average on the left-hand side of (30) therefore provides an estimator of $\langle \phi_i, \mathcal{L}\phi_j \rangle_\mu$. Of course, our approximation uses finite N and nonzero Δt. A similar sample average approximation can

be used to estimate the inner product on the right-hand side of (28).

These same steps apply to both q^+ and $\mathbb{E}[\tau_B]$, as well as the recursion in (26) for η^+. For the Fokker–Planck Eq. (25), one extra step is needed to convert an equation with $\mathcal{L}^*$ into an equation with $\mathcal{L}$. Our procedure for estimating π is described in appendix A.

Step 5: Solve the Eq. (28). With a reasonable basis size $M \lesssim 1000$, a lower–upper (LU) solver such as in Linear Algebra Package (LAPACK) via Numpy can handle Eq. (28). In the case of the homogeneous system for $w(\mathbf{x})$, a quantile regression (QR) decomposition can identify the null vector.

c. DGA fidelity and sensitivity analysis

To illustrate the effect of parameter choices on performance, we present here a simple sensitivity analysis. Figure 7 verifies the numerical accuracy and convergence of DGA by plotting the committor as a function of $U(30\,\text{km})$, estimated both with DNS and DGA, for various DGA parameters. The red curves $q^+_{\text{DGA}}[U(30\,\text{km})]$ are calculated by projecting the committor as in Fig. 2a, while the black curve $q^+_{\text{DNS}}[U(30\,\text{km})]$ is an empirical committor estimate equal to the fraction of control simulation points seen at a particular value of $U(30\,\text{km})$ that next hit B.

In Figs. 7a, 7b, and 7d, the lag time Δt increases from 5 to 10 to 20 days while the number of short trajectories stays fixed at $N = 5 \times 10^5$. Figure 7c has a long lag of 20 days, but a small dataset of $N = 5 \times 10^4$, allowing us to see the trade-off between N and Δt. The basis size M is chosen heuristically as large as possible within reason for the clustering algorithm (see the appendix). While DGA tends to systematically overestimate q^+ relative to q^+_{DNS} in the midrange of U, it seems to approach the empirical estimate as the data size and lag time increase. Each plot also displays the root-mean-square deviation between the two estimators over this subspace, $\varepsilon = \sqrt{\langle (q^+_{\text{DGA}} - q^+_{\text{DNS}})^2 \rangle_\pi}$. Within this regime, it seems that increasing the lag time has a greater impact on the deviation than increasing the number of data points. Figures 7b and 7c have approximately the same deviation ε, but Fig. 7c uses only one fifth the data, measured by total simulation time. On the other hand, more short trajectories can be parallelized more readily than fewer long trajectories, and the optimal choice will depend on computing resources.

It is natural to ask whether our short trajectory based approach is more efficient than DNS in which many independent "long" trajectories are launched from a single initial condition $\mathbf{x}$ and the committor probability $q^+(\mathbf{x})$ (or another forecast) is estimated directly. For a single value of $\mathbf{x}$ for which $q^+(\mathbf{x})$ is not very small (so that a nonnegligible fraction of trajectories reach B before A) and for which the lead time $\eta^+(\mathbf{x})$ is not too large (so that trajectories reaching B do so without requiring long integration times), DNS will undoubtedly be more efficient. This is often the situation in real-time weather forecasting. However, a key feature of our approach is that it simultaneously estimates forecasts at all values of $\mathbf{x}$, allowing the subsequent analysis of those functions that has been the focus of much of this article. Global knowledge of the committor and lead time is more pertinent for oft-repeated forecasts, for long-term risk assessment of extreme event climatology, and for targeting observations optimally. Building accurate estimators in all of state space by DNS would be extremely costly even for the reduced complexity model studied here.

6. Conclusions

Forecasting rare events is, by the very nature of rare events, an extremely difficult computational task, and one of science's most pressing challenges. We have described a computational framework, a dynamical Galerkin approximation to the Feynman–Kac equations, that combines the minimalistic philosophy of dimensionality reduction with the fidelity of high-resolution models. We identify a set of reduced coordinates, the committor probability and expected lead time, that provide the essential information that large ensemble forecasts hope to compute. DGA uses relatively short simulations of the full model to estimate these quantities of interest, allowing for prediction on much longer time scales than that of the simulation. In its focus on directly estimating statistics of interest, DGA differs from previous reduced-order modeling methods that attempt to capture general qualities of the system, including both physics-based models (Lorenz 1963; Charney and DeVore 1979; Legras and Ghil 1985; Crommelin 2003; Timmermann et al. 2003; Ruzmaikin et al. 2003) and more recent data-driven models making use of machine learning (Giannakis and Majda 2012; Giannakis et al. 2018; Berry et al. 2015; Sabeerali et al. 2017; Majda and Qi 2018; Wan et al. 2018; Bolton and Zanna 2019; Chattopadhyay et al. 2020; Chen and Majda 2020; Kashinath et al. 2021; Chattopadhyay et al. 2021).

We have shown numerical results in the context of a stochastically forced Holton–Mass model with 75 degrees of freedom, which points to the method's promise for forecasting. By systematically evaluating many model variables for their utility in predicting the fate of the vortex, we have identified some salient physical descriptions of early warning signs. We have furthermore examined the relationship between probability and lead time for a given rare event, a powerful pairing for assessing predictability and preparing for extreme weather. Our results suggest that the slow evolution of vortex preconditioning is an important source of predictability. In particular, the zonal wind and streamfunction in the range of 10–20 km above the tropopause seems to be optimal among a large class of dynamically motivated observables.

Beyond the problem of real-time weather forecasting, it is also important to assess the climatology, that is, long-term frequency, intensity, and other characteristics of rare events. For this goal as well, our methodology offers advantages over large ensemble simulations, which are currently the most detailed source of data (e.g., Schaller et al. 2018). The committor and lead time are ingredients in a larger framework called transition path theory (TPT) for describing rare transition events *at steady state*, meaning average properties over long time scales. TPT describes not only the future evolution from an initial condition ($\mathbf{x} \to B$), but the ensemble of full vortex breakdown events ($A \to B$), and how they differ from restoration events ($B \to A$). In principle, interrogating the ensemble of transition paths requires direct

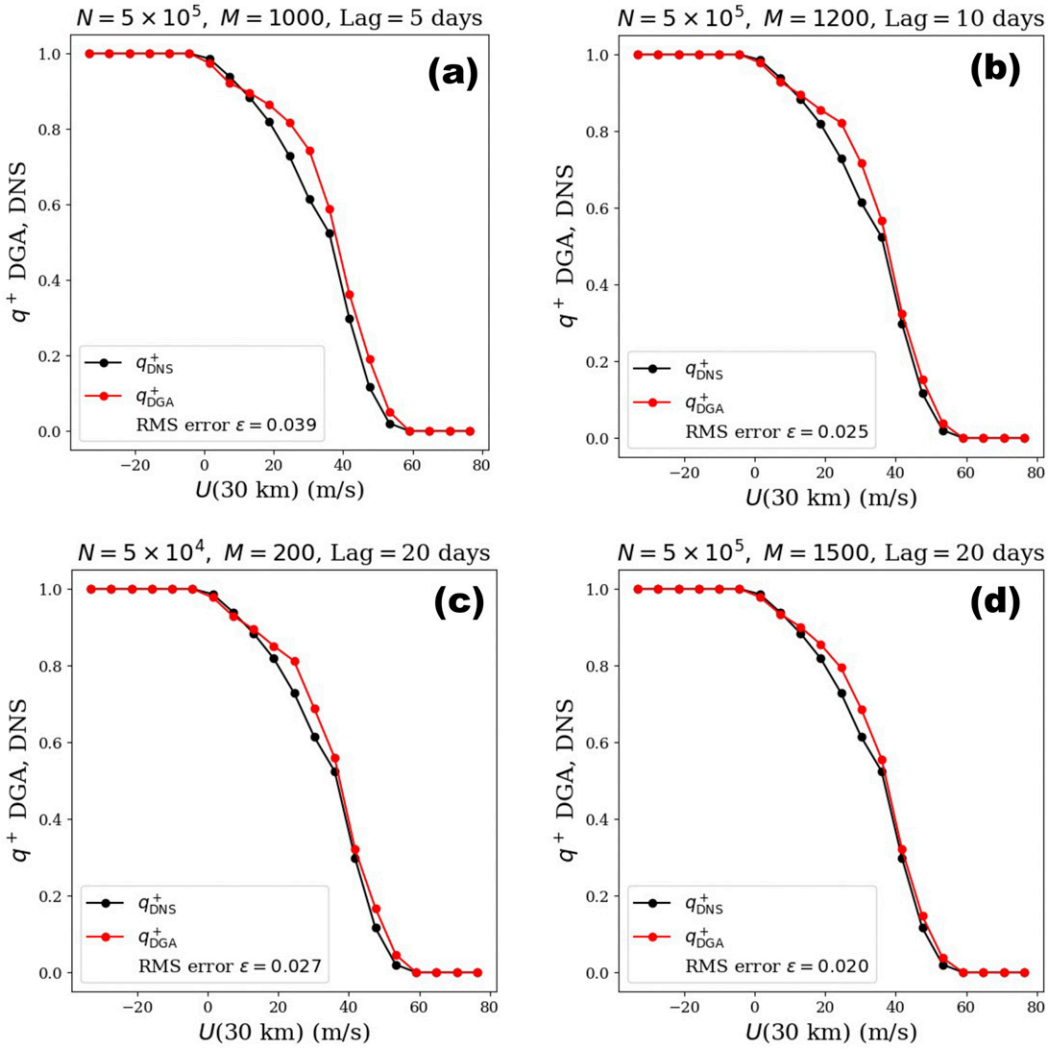

FIG. 7. Fidelity of DGA. For several DGA parameter values of N (the number of data points), M (the number of basis functions), and lag time, we plot the committor calculated from DGA and DNS (from the long control simulation), both as a function of U(30 km). The mean-square difference ε in the legend is used as a global error estimate for DGA.

simulation of the system long enough to observe many transition events. However, using TPT, quantities computable by our framework can be combined to yield key statistics describing the ensemble of transition paths (Metzner et al. 2006, 2009; E and Vanden-Eijnden 2010, 2006; Finkel et al. 2020). In a following paper, we will apply the same short-trajectory forecasting approach together with TPT to compute transition path statistics such as return times and extract insight about physical mechanisms of the transition process.

Scaling our approach up to state-of-the-art weather and climate models will require significant further development. In particular, a completely new procedure for generating trajectory initial conditions will need to be introduced. Generation of a trajectory long enough to thoroughly sample transitions will not be practical for more complicated models. One promising alternative is launching many trajectories in parallel and selectively replicating those that explore new regions of state space, especially transition regions. Such an approach could build on exciting progress over the last decade in targeted rare event simulation schemes (Hoffman et al. 2006; Weare 2009; Bouchet et al. 2011, 2014; Vanden-Eijnden and Weare 2013; Chen et al. 2014; Yasuda et al. 2017; Farazmand and Sapsis 2017; Dematteis et al. 2018; Mohamad and Sapsis 2018; Dematteis et al. 2019; Webber et al. 2019; Bouchet et al. 2019a,b; Plotkin et al. 2019; Simonnet et al. 2021; Ragone and Bouchet 2020; Sapsis 2021). A potential challenge here is that GCMs may not be set up for short simulations that start and stop frequently. For this reason, it may be sensible to use longer lag times and a sliding window to define short trajectories. Furthermore, the communication overhead required for adaptive sampling with GCMs would impose additional costs. We have deferred the sampling problem to future work, acknowledging that this step is crucial to make DGA competitive. The utility of committor and lead time, however, is independent of the method for computing them.

Defining the source of stochasticity is also an important step that varies between models. Explicitly stochastic parameterization (e.g., Berner et al. 2009; Porta Mana and Zanna 2014) will automatically lead to a spread in the short-trajectory ensemble, but in deterministic models, uncertainty will arise from perturbing the initial conditions. This may require special care depending on the model.

Another area of algorithmic improvement is selecting a basis expansion of the forecast functions. In upcoming work we will explore more flexible representations using kernel methods and neural networks. The solution of high-dimensional PDEs is an active research area that is making innovative use of machine learning, particularly in the fields of computational chemistry, quantum mechanics, and fluid dynamics (e.g., Carleo and Troyer 2017; Han et al. 2018; Khoo et al. 2018; Li et al. 2020; Mardt et al. 2018; Li et al. 2019; Raissi et al. 2019; Lorpaiboon et al. 2020). Similar approaches may hold great potential for understanding predictability in atmospheric science.

Acknowledgments. J.F. is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award DE-SC0019323.[1] R.J.W. was supported during this project by New York University's Dean's Dissertation Fellowship and by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via Grant RTG/DMS-1646339. E.P.G. acknowledges support from the U. S. National Science Foundation through Grant AGS-1852727. This work was partially supported by the NASA Astrobiology Program, Grant 80NSSC18K0829 and benefited from participation in the NASA Nexus for Exoplanet Systems Science research coordination network. J.W. acknowledges support from the Advanced Scientific Computing Research Program within the DOE Office of Science through award DE-SC0020427. The computations in the paper were done on the high-performance computing clusters at New York University and the Research Computing Center at the University of Chicago. We extend special thanks to Thomas Birner, Pedram Hassanzadeh, and an anonymous reviewer from *Monthly Weather Review*, who provided invaluable feedback on both the technical and high-level aspects of the manuscript. Their insight has helped us to sharpen and clarify the message. Freddy Bouchet also offered constructive feedback. We thank John Strahan, Aaron Dinner, and Chatipat Lorpaiboon for helpful methodological advice. Mary Silber, Noboru Nakamura, and Richard Kleeman offered invaluable scientific insight. J.F. benefitted from many helpful discussions with Anya Katsevich.

Data availability statement. The code for simulating the model, performing DGA, and producing plots is publicly available in the SHORT Github repository, "Solving for Harbingers of Rare Transitions," at https://github.com/justinfocus12/SHORT. J.F. is happy to provide further guidance upon request.

APPENDIX

Feynman–Kac Formula and DGA

In this section we spell out the DGA procedure in more detail than the main text, explaining the variants that get us to the more intricate conditional expectations. The theoretical background can be found in, for example, Karatzas and Shreve (1998), Oksendal (2003), E et al. (2019). Let $\mathbf{X}(t)$ be a time-homogeneous stochastic process with continuous sample paths in $\mathbb{R}^d$. Associated to this process is the infinitesimal generator $\mathcal{L}$, which acts on functions of state space (also called "observable" functions) by evolving their expectation forward in time:

$$\mathcal{L}f(\mathbf{x}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}_{\mathbf{x}}[f[\mathbf{X}(\Delta t)]] - f(\mathbf{x})}{\Delta t}, \tag{A1}$$

where $\mathbb{E}_{\mathbf{x}}[\cdot] := \mathbb{E}[\,\cdot\,|\mathbf{X}(0) = \mathbf{x}]$. It can be shown that under the above assumptions on $\mathbf{X}$, the Itô chain rule gives

$$df[\mathbf{X}(t)] = \mathcal{L}f[\mathbf{X}(t)]dt + d\mathbf{M}(t), \tag{A2}$$

where $\mathbf{M}(t)$ is a martingale. More concretely, in this paper, $\mathbf{X}(t)$ is an Itô diffusion obeying the stochastic differential equation

$$\mathbf{X}(t) = \mathbf{X}(0) + \int_0^t b[\mathbf{X}(s)]\,ds + \int_0^t \sigma[\mathbf{X}(s)]\,d\mathbf{W}(s), \tag{A3}$$

with infinitesimal generator and martingale terms

$$\mathcal{L}f(\mathbf{x}) = \sum_{i=1}^{d} b_i(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{2}[\sigma(\mathbf{x})\sigma(\mathbf{x})^{\mathrm{T}}]_{ij} \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial \mathbf{x}_j}, \quad \text{and} \tag{A4}$$

$$d\mathbf{M}(t) = \sum_{i=1}^{d} \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \sigma_{ij}(\mathbf{x}) d\mathbf{W}_j(t). \tag{A5}$$

The key forecasting quantities in this paper are of the form (18) and can be solved with (23), a linear equation involving the generator. We now lay out a brief derivation of the Feynman–Kac formula and our numerical discretization, roughly following E et al. (2019).

[1] This report was prepared as an account of work sponsored by an agency of the U.S. government. Neither the U.S. government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. government or any agency thereof.

a. Feynman–Kac formula

Let D be a domain in $\mathbb{R}^d$ [e.g., $(A \cup B)^c$] and $\tau_{D^c} = \min\{t \geq 0 : \mathbf{X}(t) \notin D\}$ be the first exit time from this domain starting at time zero. This is a random variable that depends on the starting condition $\mathbf{x} \in D$. Let $G : \partial D \to \mathbb{R}$ be a boundary condition, $\Gamma : D \to \mathbb{R}$ a source term, and $\Gamma : D \to \mathbb{R}$ a term to represent accumulated risk. We seek a PDE for the conditional expectation from (18):

$$F(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\left[G[\mathbf{X}(\tau)]\exp\left\{\lambda\int_0^{\tau}\Gamma[\mathbf{X}(s)]\,ds\right\}\right], \tag{A6}$$

where $\mathbb{E}_{\mathbf{x}}[\cdot] = \mathbb{E}[\,\cdot\,|\mathbf{X}(0) = \mathbf{x}]$. To derive the PDE (23), consider the following stochastic process:

$$Z(t) = F[\mathbf{X}(t)]Y(t), \tag{A7}$$

where $Y(t) := \exp\{\lambda\int_0^t \Gamma[\mathbf{X}(s)]\,ds\}$. Itô's lemma gives us that $dY(t) = \lambda\Gamma[\mathbf{X}(t)]\mathbf{Y}(t)dt$. Hence, applying the product rule to $Z(t)$,

$$dZ(t) = dF[\mathbf{X}(t)]Y(t) + F[\mathbf{X}(t)]\,dY(t), \tag{A8}$$

$$= \mathcal{L}F[\mathbf{X}(t)]Y(t)\,dt + d\mathbf{M}(t)Y(t) + \lambda F[\mathbf{X}(t)]\Gamma[\mathbf{X}(t)]Y(t)\,dt, \quad \text{and} \tag{A9}$$

$$= [\mathcal{L}F + \lambda\Gamma F][\mathbf{X}(t)]Y(t)\,dt + Y(t)d\mathbf{M}(t), \tag{A10}$$

where in (A8) we have left out the quadratic cross-variation of $F[\mathbf{X}(t)]$ and $Y(t)$ because Y has finite variation. If the bracketed term $[\mathcal{L} + \lambda\Gamma(\mathbf{x})]F(\mathbf{x}) = 0$ for all $\mathbf{x}$, then $Z(t)$ is a martingale and it follows that

$$Z(0) = \mathbb{E}_{\mathbf{x}}[Z(t)], \quad \text{and} \tag{A11}$$

$$F(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\left[F[\mathbf{X}(t)]\exp\left\{\lambda\int_0^t\Gamma[\mathbf{X}(s)]\,ds\right\}\right]. \tag{A12}$$

Finally, the formula still holds if we substitute a stopping time for t. By choosing τ, the first exit time from D, the $F[\mathbf{X}(t)]$ inside the brackets becomes its boundary value $G[\mathbf{X}(\tau)]$. Thus $F(\mathbf{x})$ as defined in (A6) also solves the PDE boundary value problem (23):

$$\begin{cases} [\mathcal{L} + \lambda\Gamma(\mathbf{x})]F(\mathbf{x};\lambda) = 0 & \mathbf{x} \in D \\ F(\mathbf{x};\lambda) = G(\mathbf{x}) & \mathbf{x} \in D^c \end{cases}, \tag{A13}$$

where we have inserted the additional dependence of F on λ in order to lead directly to the recursive formulas (20) and (26).

b. Dynkin's formula and finite lag time

We have presented (29) as a mathematically concise approximation to the generator. In practice, we achieve better numerical stability integrating the generator (A1) to a finite lag time Δt, following Strahan et al. (2021). The theorem that allows this is called Dynkin's formula (e.g., Oksendal 2003), which states that for any suitable function $f : \mathbb{R}^d \to \mathbb{R}$ and a stopping time θ (not to be confused with CV coordinates),

$$\mathbb{E}_{\mathbf{x}}[f[\mathbf{X}(\theta)]] = f(\mathbf{x}) + \mathbb{E}_{\mathbf{x}}\left[\int_0^{\theta}\mathcal{L}f[\mathbf{X}(t)]\,dt\right]. \tag{A14}$$

The left-hand side, $\mathbb{E}_{\mathbf{x}}[f[\mathbf{X}(\theta)]]$, is known as the *transition operator* $\mathcal{T}^{\theta}f(\mathbf{x})$, a finite-time version of the generator. Note that this is a deterministic operator despite θ being a random variable, because by definition $\mathcal{T}^{\theta}$ only has θ inside of expectations. We can apply Dynkin's formula to (A13) *before* numerical approximation, setting $\theta = \min(\Delta t, \tau)$. That is, the short trajectory $\{\mathbf{X}(t) : 0 \leq t \leq \Delta t = 20 \text{ days}\}$ is stopped early if it exits the domain D before Δt. Applying Dynkin's formula to $F(\mathbf{x};\lambda)$, we find

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}[F[\mathbf{X}(\theta)]] &= F(\mathbf{x}) + \mathbb{E}_{\mathbf{x}}\left[\int_0^{\theta}\mathcal{L}F[\mathbf{X}(t)]\,dt\right] \\ &= F(\mathbf{x}) - \lambda\mathbb{E}_{\mathbf{x}}\left[\int_0^{\theta}\Gamma[\mathbf{X}(t)]F[\mathbf{X}(t)]\,dt\right] \\ \mathcal{T}^{\theta}F(\mathbf{x}) &= F(\mathbf{x}) - \lambda\mathcal{K}^{\theta}[\Gamma F](\mathbf{x}),\end{aligned} \tag{A15}$$

where $\mathcal{K}^{\theta}$ is shorthand notation for the integral operator on the right. Equation (A15), along with the boundary conditions $F|_{D^c} = G|_{D^c}$, gives us a linear equation for $F(\mathbf{x})$ that can be solved by DGA. As outlined in section 5, we write $F = \hat{F} + f$, where $\hat{F}$ obeys the boundary conditions and f obeys

$$\begin{aligned}&(\mathcal{T}^{\theta} - 1)f(\mathbf{x}) + \lambda\mathcal{K}^{\theta}[\Gamma f](\mathbf{x}) \\ &\quad = -(\mathcal{T}^{\theta} - 1)\hat{F}(\mathbf{x}) - \lambda\mathcal{K}^{\theta}[\Gamma\hat{F}](\mathbf{x}).\end{aligned} \tag{A16}$$

We then expand $f = \sum_{j=1}^{M}\xi_j\phi_j(\mathbf{x})$ with basis functions $\{\phi_j\}$ that are zero on D^c, and take μ-weighted inner products with ϕ_i on both sides to obtain

$$\begin{aligned}\sum_{j=1}^{M}\xi_j(&\langle\phi_i, (\mathcal{T}^{\theta} - 1)\phi_j\rangle_{\mu} + \lambda\langle\phi_i, \mathcal{K}^{\theta}[\Gamma\phi_j]\rangle_{\mu}) \\ &- \langle\phi_i, (\mathcal{T}^{\theta} - 1)\hat{F}\rangle_{\mu} - \lambda\langle\phi_i, \mathcal{K}^{\theta}[\Gamma\hat{F}]\rangle_{\mu}.\end{aligned} \tag{A17}$$

Finally, the inner products can be estimated with short trajectories using (30). For two functions ϕ and ψ, the first left-hand side inner product is approximately

$$\langle\phi, (\mathcal{T}^{\theta} - 1)\psi\rangle_{\mu} \approx \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{X}_n)\{\psi[\mathbf{X}_n(\theta_n)] - \psi(\mathbf{X}_n)\}, \tag{A18}$$

where θ_n is the sampled first-exit time of the nth trajectory, or Δt if it never exits. The second left-hand side inner product is approximately

$$\begin{aligned}&\langle\phi, \mathcal{K}^{\theta}[\Gamma\psi]\rangle_{\mu} \\ &\quad \approx \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{X}_n)\int_0^{\theta_n}\Gamma[\mathbf{X}_n(t)]\psi[\mathbf{X}_n(t)]\,dt,\end{aligned} \tag{A19}$$

where the time integral on the right is computed with the trapezoid rule on the trajectory, which is sampled every 0.5 days. The error from numerical quadrature is likely small compared to the error from basis set construction, but higher-order integration methods do merit further investigation.

Given a fixed Γ and G, and with the inner products in hand, we now have (A17) as a family of matrix equations with λ a continuous parameter:

$$(P+\lambda Q)\xi(\lambda)=\mathbf{v}+\lambda\mathbf{r}. \tag{A20}$$

We can then differentiate in λ and evaluate at $\lambda=0$ to obtain a ready-to-solve discretization of the recursion (26):

$$P\xi(0)=\mathbf{v}, \tag{A21}$$

$$P\xi'(0)=\mathbf{r}-Q\xi(0), \quad \text{and} \tag{A22}$$

$$P\xi^{(k)}(0)=-kQ\xi^{(k-1)}(0) \quad \text{for} \quad k\geq 2, \tag{A23}$$

where the kth derivative $\boldsymbol{\xi}^{(k)}(0)$ is the coefficient expansion in the basis $\{\phi_j\}$ of the kth moment from (22):

$$\partial_\lambda^k F(\mathbf{x};0)=\mathbb{E}_{\mathbf{x}}\left[G[\mathbf{X}(\tau)]\left\{\lambda\int_0^\tau \Gamma[\mathbf{X}(s)]ds\right\}^k\right]. \tag{A24}$$

c. Change of measure

We now specify how to compute the change of measure from μ (the sampling distribution) to π (the steady-state distribution), using an adjoint version of the Feynman–Kac formula. Each of the basis functions ϕ_i has an expectation at time zero with respect to the steady-state distribution: $\mathbb{E}_{\mathbf{X}(0)\sim\pi}[\phi_i[\mathbf{X}(0)]]=\int\phi_i(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}$. Evolving the dynamics from 0 to Δt induces another expectation: $\mathbb{E}_{\mathbf{X}(0)\sim\pi}[\phi_i[\mathbf{X}(\Delta t)]]=\int\mathcal{T}^{\Delta t}\phi_i(\mathbf{x})\pi(d\mathbf{x})$. π is the *invariant* distribution, which means that these two integrals are equal:

$$\int(\mathcal{T}^{\Delta t}-1)\phi_i(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x}=0. \tag{A25}$$

Furthermore, with a change of measure they can be rewritten with respect to the sampling measure μ instead of π, so

$$\int(\mathcal{T}^{\Delta t}-1)\phi_i(\mathbf{x})\frac{d\pi}{d\mu}(\mathbf{x})\mu(\mathbf{x})\,d\mathbf{x}=0. \tag{A26}$$

The change of measure $(d\pi/d\mu)(\mathbf{x})$, which we abbreviate $w(\mathbf{x})$, is yet another unknown function that we expand in the basis as $w(\mathbf{x})=\sum_j\xi_j\phi_j(\mathbf{x})$. Putting this into the integral and using Monte Carlo, we cast the coefficients ξ_j as the solution to a null eigenvector problem:

$$0=\int(\mathcal{T}^{\Delta t}-1)\phi_i(\mathbf{x})\sum_{j=1}^M\xi_j\phi_j(\mathbf{x})\mu\,(d\mathbf{x}), \quad \text{and} \tag{A27}$$

$$\approx\sum_{j=1}^M\xi_j\sum_{n=1}^N\{\phi_i[\mathbf{X}_n(\Delta t)]-\phi_i(\mathbf{X}_n)\}\phi_j(\mathbf{X}_n). \tag{A28}$$

This last equation is simply the Fokker–Planck equation, $\mathcal{L}^*\pi=0$, in weak form and integrated in time using Dynkin's formula. Note that the matrix elements in (A28) are the transpose of those in (A18).

d. DGA details

We will provide more details here on our particular construction of basis functions. The partition $\{S_1,\ldots,S_M\}$ to build the basis function library $\phi_j(\mathbf{x})=\mathbb{1}_{S_j}(\mathbf{x})$, $n=1,\ldots,N$ should be chosen with a number of considerations in mind. The partition elements should be small enough to accurately represent the functions they are used to approximate, but large enough to contain sufficient data to robustly estimate transition probabilities. We form these sets by a hierarchical modification of k-means clustering on the initial points $\{\mathbf{X}_n\}_{n=1}^N$. The K-means method is a robust method that can incorporate new samples by simply identifying the closest centroid, and is commonly used in molecular dynamics (Pande et al. 2010). However, straightforward application of k-means, as implemented in the scikit-learn software (Pedregosa et al. 2011), can produce a very imbalanced cluster size distribution, even with empty clusters. This leads to unwanted singularities in the constructed Markov matrix. To avoid this problem we cluster hierarchically, starting with a coarse clustering of all points and iteratively refining the larger clusters, at every stage enforcing a minimum cluster size of five points, until we have the desired number of clusters (M). After clustering on the initial points $\{\mathbf{X}_n\}$, the other points $\{\mathbf{X}_n(t), 0<t\leq\Delta t\}$ are placed into clusters using an address tree produced by the k-means cluster hierarchy. For boundary value problems with a domain D and boundary D^c, we need only cluster points in D since the basis should be homogeneous. The total number of clusters should scale with dataset. In our main results with $N=5\times10^5$, we found $M=1500$ to be enough basis functions to resolve some of the finer details in the structure of the forecast functions, but not so many as to require an unmanageably deep address tree, which manifests in dramatic slowdown past a certain threshold. At this point, the cluster number is still a manually tuned hyperparameter.

Because the committor and lead time obey Dirichlet boundary conditions on $A\cup B$, the basis functions used to construct them should be zero on $A\cup B$, meaning only data points $\mathbf{X}_n\notin A\cup B$ should be used to produce the clusters. On the other hand, the steady-state distribution has no boundary condition to satisfy, only a global normalization condition. Hence, the basis for the change of measure w must be different from the basis for q^+ and η^+, with its clusters including all data points in $A\cup B$. Furthermore, the basis must be chosen so that the matrix $\langle(\mathcal{T}^{\Delta t}-1)\phi_i,\phi_j\rangle$ has a nontrivial null space; this is guaranteed by the indicator basis set we use but can otherwise be guaranteed by including a constant function in the basis.

The use of an indicator basis follows the Markov state modeling literature (e.g., Chodera et al. 2006; Pande et al. 2010), which has the advantage of simplicity and robustness. In particular, the discretization of $\mathcal{T}^\theta-1$ is a properly normalized stochastic matrix (with nonnegative entries and rows summing to 1), which guarantees the maximum principle $0\leq q^+(\mathbf{x})\leq1$ and $0\leq w(\mathbf{x})$ for all data points $\mathbf{x}$. However, alternative basis sets have been shown to be promising, perhaps with much less data. Thiede et al. (2019) used diffusion maps, while Strahan et al. (2021) used a PCA-like procedure to construct the basis. More generally, there is no requirement to use a linear Galerkin method to solve the Feynman–Kac formulas. More flexible functional forms may have an important role to play as well. In the low-data regime, some preliminary experiments have suggested that Gaussian process regression (GPR) is a useful way to constrain the committor estimate with a prior, following the framework in Bilionis (2016) to solve PDEs with Gaussian processes. As mentioned in the conclusion, there is rapidly growing interest in the use of artificial neural networks to solve PDEs. As with many novel methods, however, DGA is

likely to work best on new applications when its simplest form is applied first. This will be our approach in coming experiments on more complex models.

REFERENCES

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626, https://doi.org/10.1175/2008JAS2677.1.

Berry, T., J. R. Cressman, Z. Gregurić-Ferenček, and T. Sauer, 2013: Time-scale separation from diffusion-mapped delay coordinates. *SIAM J. Appl. Dyn. Syst.*, **12**, 618–649, https://doi.org/10.1137/12088183X.

——, D. Giannakis, and J. Harlim, 2015: Nonparametric forecasting of low-dimensional dynamical systems. *Phys. Rev.*, **91E**, 032915, https://doi.org/10.1103/PhysRevE.91.032915.

Bilionis, I., 2016: Probabilistic solvers for partial differential equations. arXiv:1607.03526, 9 pp., https://arxiv.org/pdf/1607.03526.pdf.

Binzel, R. P., 2000: The Torino impact hazard scale. *Planet. Space Sci.*, **48**, 297–303, https://doi.org/10.1016/S0032-0633(00)00006-4.

Birner, T., and P. D. Williams, 2008: Sudden stratospheric warmings as noise-induced transitions. *J. Atmos. Sci.*, **65**, 3337–3343, https://doi.org/10.1175/2008JAS2770.1.

Bolton, T., and L. Zanna, 2019: Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.*, **11**, 376–399, https://doi.org/10.1029/2018MS001472.

Bouchet, F., J. Laurie, and O. Zaboronski, 2011: Control and instanton trajectories for random transitions in turbulent flows. *J. Phys. Conf. Ser.*, **318**, 022041, https://doi.org/10.1088/1742-6596/318/2/022041.

——, ——, and ——, 2014: Langevin dynamics, large deviations and instantons for the quasi-geostrophic model and two-dimensional Euler equations. *J. Stat. Phys.*, **156**, 1066–1092, https://doi.org/10.1007/s10955-014-1052-5.

——, J. Rolland, and E. Simonnet, 2019a: Rare event algorithm links transitions in turbulent flows with activated nucleations. *Phys. Rev. Lett.*, **122**, 074502, https://doi.org/10.1103/PhysRevLett.122.074502.

——, ——, and J. Wouters, 2019b: Rare event sampling methods. *Chaos*, **29**, 080402, https://doi.org/10.1063/1.5120509.

Bowman, G. R., V. S. Pande, and F. Noé, 2013: *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Advances in Experimental Medicine and Biology, Vol. 797, Springer Science & Business Media, 139 pp.

Carleo, G., and M. Troyer, 2017: Solving the quantum many-body problem with artificial neural networks. *Science*, **355**, 602–606, https://doi.org/10.1126/science.aag2302.

Charlton, A. J., and L. M. Polvani, 2007: A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Climate*, **20**, 449–469, https://doi.org/10.1175/JCLI3996.1.

Charney, J. G., and J. G. DeVore, 1979: Multiple flow equilibria in the atmosphere and blocking. *J. Atmos. Sci.*, **36**, 1205–1216, https://doi.org/10.1175/1520-0469(1979)036<1205:MFEITA>2.0.CO;2.

Chattopadhyay, A., E. Nabizadeh, and P. Hassanzadeh, 2020: Analog forecasting of extreme-causing weather patterns using deep learning. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001958, https://doi.org/10.1029/2019MS001958.

——, M. Mustafa, P. Hassanzadeh, E. Bach, and K. Kashinath, 2021: Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving spatial transformers in a case study with ERA5. *Geosci. Model Dev. Discuss.*, https://doi.org/10.5194/gmd-2021-71.

Chen, N., and A. J. Majda, 2020: Predicting observed and hidden extreme events in complex nonlinear dynamical systems with partial observations and short training time series. *Chaos*, **30**, 033101, https://doi.org/10.1063/1.5122199.

——, D. Giannakis, R. Herbei, and A. J. Majda, 2014: An MCMC algorithm for parameter estimation in signals with hidden intermittent instability. *SIAM/ASA J. Uncertainty Quantif.*, **2**, 647–669, https://doi.org/10.1137/130944977.

Chodera, J. D., and F. Noé, 2014: Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, **25**, 135–144, https://doi.org/10.1016/j.sbi.2014.04.002.

——, W. C. Swope, J. W. Pitera, and K. A. Dill, 2006: Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, **5**, 1214–1226, https://doi.org/10.1137/06065146X.

Christiansen, B., 2000: Chaos, quasiperiodicity, and interannual variability: Studies of a stratospheric vacillation model. *J. Atmos. Sci.*, **57**, 3161–3173, https://doi.org/10.1175/1520-0469(2000)057<3161:CQAIVS>2.0.CO;2.

Crommelin, D. T., 2003: Regime transitions and heteroclinic connections in a barotropic atmosphere. *J. Atmos. Sci.*, **60**, 229–246, https://doi.org/10.1175/1520-0469(2003)060<0229:RTAHCI>2.0.CO;2.

DelSole, T., and B. F. Farrell, 1995: A stochastically excited linear system as a model for quasigeostrophic turbulence: Analytic results for one- and two-layer fluids. *J. Atmos. Sci.*, **52**, 2531–2547, https://doi.org/10.1175/1520-0469(1995)052<2531:ASELSA>2.0.CO;2.

Dematteis, G., T. Grafke, and E. Vanden-Eijnden, 2018: Rogue waves and large deviations in deep sea. *Proc. Natl. Acad. Sci. USA*, **115**, 855–860, https://doi.org/10.1073/pnas.1710670115.

——, ——, M. Onorato, and E. Vanden-Eijnden, 2019: Experimental evidence of hydrodynamic instantons: The universal route to rogue waves. *Phys. Rev. X*, **9**, 041057, https://doi.org/10.1103/PhysRevX.9.041057.

Durrett, R., 2013: *Probability: Theory and Examples.* Cambridge University Press, 428 pp.

E, W., and E. Vanden-Eijnden, 2006: Towards a theory of transition paths. *J. Stat. Phys.*, **123**, 503, https://doi.org/10.1007/s10955-005-9003-9.

——, and ——, 2010: Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, **61**, 391–420, https://doi.org/10.1146/annurev.physchem.040808.090412.

——, T. Li, and E. Vanden-Eijnden, 2019: *Applied Stochastic Analysis*. Graduate Studies in Mathematics, Vol. 199, American Mathematical Society, 305 pp.

Esler, J. G., and M. Mester, 2019: Noise-induced vortex-splitting stratospheric sudden warmings. *Quart. J. Roy. Meteor. Soc.*, **145**, 476–494, https://doi.org/10.1002/qj.3443.

Farazmand, M., and T. P. Sapsis, 2017: A variational approach to probing extreme events in turbulent dynamical systems. *Sci. Adv.*, **3**, e1701533, https://doi.org/10.1126/sciadv.1701533.

Finkel, J., D. S. Abbot, and J. Weare, 2020: Path properties of atmospheric transitions: Illustration with a low-order sudden stratospheric warming model. *J. Atmos. Sci.*, **77**, 2327–2347, https://doi.org/10.1175/JAS-D-19-0278.1.

Fitzsimmons, P., and J. Pitman, 1999: Kac's moment formula and the Feynman–Kac formula for additive functionals of a Markov

process. *Stochastic Process. Appl.*, **79**, 117–134, https://doi.org/10.1016/S0304-4149(98)00081-7.

Franzke, C., and A. J. Majda, 2006: Low-order stochastic mode reduction for a prototype atmospheric GCM. *J. Atmos. Sci.*, **63**, 457–479, https://doi.org/10.1175/JAS3633.1.

Giannakis, D., and A. J. Majda, 2012: Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl. Acad. Sci. USA*, **109**, 2222–2227, https://doi.org/10.1073/pnas.1118984109.

——, A. Kolchinskaya, D. Krasnov, and J. Schumacher, 2018: Koopman analysis of the long-term evolution in a turbulent convection cell. *J. Fluid Mech.*, **847**, 735–767, https://doi.org/10.1017/jfm.2018.297.

Gottwald, G. A., D. T. Crommelin, and C. L. E. Franzke, 2016: Stochastic climate theory. arXiv:1612.07474, 29 pp., https://arxiv.org/pdf/1612.07474.pdf.

Han, J., A. Jentzen, and W. E, 2018: Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA*, **115**, 8505–8510, https://doi.org/10.1073/pnas.1718942115.

Hasselmann, K., 1976: Stochastic climate models Part I. Theory. *Tellus*, **28**, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316.

Helfmann, L., J. Heitzig, P. Koltai, J. Kurths, and C. Schütte, 2021: Statistical analysis of tipping pathways in agent-based models. arXiv:2103.02883, 28 pp., https://arxiv.org/pdf/2103.02883.pdf.

Hoffman, R. N., J. M. Henderson, S. M. Leidner, C. Grassotti, and T. Nehrkorn, 2006: The response of damaging winds of a simulated tropical cyclone to finite-amplitude perturbations of different variables. *J. Atmos. Sci.*, **63**, 1924–1937, https://doi.org/10.1175/JAS3720.1.

Holton, J. R., and C. Mass, 1976: Stratospheric vacillation cycles. *J. Atmos. Sci.*, **33**, 2218–2225, https://doi.org/10.1175/1520-0469(1976)033<2218:SVC>2.0.CO;2.

Karatzas, I., and S. E. Shreve, 1998: *Brownian Motion and Stochastic Calculus.* Springer, 470 pp.

Kashinath, K., and Coauthors, 2021: Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. Roy. Soc.*, **A379**, 20200093, https://doi.org/10.1098/rsta.2020.0093.

Khoo, Y., J. Lu, and L. Ying, 2018: Solving for high-dimensional committor functions using artificial neural networks. *Res. Math. Sci.*, **6**, 1, https://doi.org/10.1007/s40687-018-0160-2.

Legras, B., and M. Ghil, 1985: Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.*, **42**, 433–471, https://doi.org/10.1175/1520-0469(1985)042<0433:PABAVI>2.0.CO;2.

Li, H., Y. Khoo, Y. Ren, and L. Ying, 2020: A semigroup method for high dimensional committor functions based on neural network. arXiv:2012.06727, 21 pp., https://arxiv.org/pdf/2012.06727.pdf.

Li, Q., B. Lin, and W. Ren, 2019: Computing committor functions for the study of rare events using deep learning. *J. Chem. Phys.*, **151**, 054112, https://doi.org/10.1063/1.5110439.

Lin, K. K., and F. Lu, 2021: Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism. *J. Comput. Phys.*, **424**, 109864, https://doi.org/10.1016/j.jcp.2020.109864.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lorpaiboon, C., E. H. Thiede, R. J. Webber, J. Weare, and A. R. Dinner, 2020: Integrated variational approach to conformational dynamics: A robust strategy for identifying eigenfunctions of dynamical operators. *J. Phys. Chem.*, **B124**, 9354–9364, https://doi.org/10.1021/acs.jpcb.0c06477.

Lucarini, V., and A. Gritsun, 2020: A new mathematical framework for atmospheric blocking events. *Climate Dyn.*, **54**, 575–598, https://doi.org/10.1007/s00382-019-05018-2.

Lucente, D., S. Duffner, C. Herbert, J. Rolland, and F. Bouchet, 2019: Machine learning of committor functions for predicting high impact climate events. *Ninth Int. Workshop on Climate Informatics*, Paris, France, École Normale Supérieure, https://hal.archives-ouvertes.fr/hal-02322370/document.

Maiocchi, C. C., V. Lucarini, A. Gritsun, and G. Pavliotis, 2020: Unstable periodic orbits sampling in climate models. *EGU General Assembly*, EGU, Abstract 18823, https://doi.org/10.5194/egusphere-egu2020-18823.

Majda, A. J., and D. Qi, 2018: Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Rev.*, **60**, 491–549, https://doi.org/10.1137/16M1104664.

——, I. Timofeyev, and E. Vanden Eijnden, 2001: A mathematical framework for stochastic climate models. *Commun. Pure Appl. Math.*, **54**, 891–974, https://doi.org/10.1002/cpa.1014.

Mardt, A., L. Pasuali, H. Wu, and F. Noé, 2018: Vampnets for deep learning of molecular kinetics. *Nat. Commun.*, **9**, 5, https://doi.org/10.1038/s41467-017-02388-1.

Matsuno, T., 1971: A dynamical model of the stratospheric sudden warming. *J. Atmos. Sci.*, **28**, 1479–1494, https://doi.org/10.1175/1520-0469(1971)028<1479:ADMOTS>2.0.CO;2.

Metzner, P., C. Schutte, and E. Vanden-Eijnden, 2006: Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.*, **125**, 084110, https://doi.org/10.1063/1.2335447.

——, ——, and ——, 2009: Transition path theory for Markov jump processes. *Multiscale Model. Simul.*, **7**, 1192–1219, https://doi.org/10.1137/070699500.

Miron, P., F. Beron-Vera, L. Helfmann, and P. Koltai, 2021: Transition paths of marine debris and the stability of the garbage patches. *Chaos*, **31**, 033101, https://doi.org/10.1063/5.0030535.

Mohamad, M. A., and T. P. Sapsis, 2018: Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, **115**, 11 138–11 143, https://doi.org/10.1073/pnas.1813263115.

Ngwira, C. M., and Coauthors, 2013: Simulation of the 23 July 2012 extreme space weather event: What if this extremely rare CME was Earth directed? *Space Wea.*, **11**, 671–679, https://doi.org/10.1002/2013SW000990.

Noé, F., and S. Fischer, 2008: Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, **18**, 154–162, https://doi.org/10.1016/j.sbi.2008.01.008.

——, and C. Clementi, 2017: Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods. *Curr. Opin. Struct. Biol.*, **43**, 141–147, https://doi.org/10.1016/j.sbi.2017.02.006.

——, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, 2009: Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, **106**, 19 011–19 016, https://doi.org/10.1073/pnas.0905466106.

Oksendal, B., 2003: *Stochastic Differential Equations: An Introduction with Applications.* Springer, 360 pp.

Pande, V. S., K. Beauchamp, and G. R. Bowman, 2010: Everything you wanted to know about Markov state models but were

afraid to ask. *Methods*, **52**, 99–105, https://doi.org/10.1016/j.ymeth.2010.06.002.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Plotkin, D. A., R. J. Webber, M. E. O'Neill, J. Weare, and D. S. Abbot, 2019: Maximizing simulated tropical cyclone intensity with action minimization. *J. Adv. Model. Earth Syst.*, **11**, 863–891, https://doi.org/10.1029/2018MS001419.

Porta Mana, P., and L. Zanna, 2014: Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modell.*, **79**, 1–20, https://doi.org/10.1016/j.ocemod.2014.04.002.

Ragone, F., and F. Bouchet, 2020: Computation of extreme values of time averaged observables in climate models with large deviation techniques. *J. Stat. Phys.*, **179**, 1637–1665, https://doi.org/10.1007/s10955-019-02429-7.

——, J. Wouters, and F. Bouchet, 2018: Computation of extreme heat waves in climate models using a large deviation algorithm. *Proc. Natl. Acad. Sci. USA*, **115**, 24–29, https://doi.org/10.1073/pnas.1712645115.

Raissi, M., P. Perdikaris, and G. Karniadakis, 2019: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, **378**, 686–707, https://doi.org/10.1016/j.jcp.2018.10.045.

Ruzmaikin, A., J. Lawrence, and C. Cadavid, 2003: A simple model of stratospheric dynamics including solar variability. *J. Climate*, **16**, 1593–1600, https://doi.org/10.1175/1520-0442-16.10.1593.

Sabeerali, C. T., R. S. Ajayamohan, D. Giannakis, and A. J. Majda, 2017: Extraction and prediction of indices for monsoon intraseasonal oscillations: An approach based on nonlinear Laplacian spectral analysis. *Climate Dyn.*, **49**, 3031–3050, https://doi.org/10.1007/s00382-016-3491-y.

Sapsis, T. P., 2021: Statistics of extreme events in fluid flows and waves. *Annu. Rev. Fluid Mech.*, **53**, 85–111, https://doi.org/10.1146/annurev-fluid-030420-032810.

Schaller, N., J. Sillmann, J. Anstey, E. M. Fischer, C. M. Grams, and S. Russo, 2018: Influence of blocking on northern European and western Russian heatwaves in large climate model ensembles. *Environ. Res. Lett.*, **13**, 054015, https://doi.org/10.1088/1748-9326/aaba55.

Simonnet, E., J. Rolland, and F. Bouchet, 2021: Multistability and rare spontaneous transitions in barotropic β-plane turbulence. arXiv:2009.09913, 38 pp., https://arxiv.org/pdf/2009.09913.pdf.

Sjoberg, J. P., and T. Birner, 2014: Stratospheric wave–mean flow feedbacks and sudden stratospheric warmings in a simple model forced by upward wave activity flux. *J. Atmos. Sci.*, **71**, 4055–4071, https://doi.org/10.1175/JAS-D-14-0113.1.

Strahan, J., A. Antoszewski, C. Lorpaiboon, B. P. Vani, J. Weare, and A. R. Dinner, 2021: Long-time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein. *J. Chem. Theory Comput.*, **17**, 2948–2963, https://doi.org/10.1021/acs.jctc.0c00933.

Tantet, A., F. R. van der Burgt, and H. A. Dijkstra, 2015: An early warning indicator for atmospheric blocking events using transfer operators. *Chaos*, **25**, 036406, https://doi.org/10.1063/1.4908174.

Thiede, E., D. Giannakis, A. R. Dinner, and J. Weare, 2019: Galerkin approximation of dynamical quantities using trajectory data. arXiv:1810.01841, 24 pp., https://arxiv.org/pdf/1810.01841.pdf.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc.*, **B58**, 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Timmermann, A., F.-F. Jin, and J. Abshagen, 2003: A nonlinear theory for El Niño bursting. *J. Atmos. Sci.*, **60**, 152–165, https://doi.org/10.1175/1520-0469(2003)060<0152:ANTFEN>2.0.CO;2.

Vanden-Eijnden, E., and J. Weare, 2013: Data assimilation in the low noise regime with application to the Kuroshio. *Mon. Wea. Rev.*, **141**, 1822–1841, https://doi.org/10.1175/MWR-D-12-00060.1.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

Wan, Z. Y., P. Vlachas, P. Koumoutsakos, and T. Sapsis, 2018: Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PLOS ONE*, **13**, e0197704, https://doi.org/10.1371/journal.pone.0197704.

Weare, J., 2009: Particle filtering with path sampling and an application to a bimodal ocean current model. *J. Comput. Phys.*, **228**, 4312–4331, https://doi.org/10.1016/j.jcp.2009.02.033.

Webber, R. J., D. A. Plotkin, M. E. O'Neill, D. S. Abbot, and J. Weare, 2019: Practical rare event sampling for extreme mesoscale weather. *Chaos*, **29**, 053109, https://doi.org/10.1063/1.5081461.

Yasuda, Y., F. Bouchet, and A. Venaille, 2017: A new interpretation of vortex-split sudden stratospheric warmings in terms of equilibrium statistical mechanics. *J. Atmos. Sci.*, **74**, 3915–3936, https://doi.org/10.1175/JAS-D-17-0045.1.

Yoden, S., 1987a: Bifurcation properties of a stratospheric vacillation model. *J. Atmos. Sci.*, **44**, 1723–1733, https://doi.org/10.1175/1520-0469(1987)044<1723:BPOASV>2.0.CO;2.

——, 1987b: Dynamical aspects of stratospheric vacillations in a highly truncated model. *J. Atmos. Sci.*, **44**, 3683–3695, https://doi.org/10.1175/1520-0469(1987)044<3683:DAOSVI>2.0.CO;2.

Zhang, F., and J. A. Sippel, 2009: Effects of moist convection on hurricane predictability. *J. Atmos. Sci.*, **66**, 1944–1961, https://doi.org/10.1175/2009JAS2824.1.

Zwanzig, R., 2001: *Nonequilibrium Statistical Mechanics*. Oxford University Press, 240 pp.