Choosing observables that capture critical slowing down before tipping points: A Fokker-Planck operator approach

Johannes Lohmann[1,*] and Georg A. Gottwald[2]

[1]*Physics of Ice, Climate and Earth, Niels Bohr Institute, University of Copenhagen, Denmark*

[2]*School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia*

Abstract

Tipping points (TP) are abrupt transitions between metastable states in complex systems, most often described by a bifurcation or crisis of a multistable system induced by a slowly changing control parameter. An avenue for predicting TPs in real-world systems is critical slowing down (CSD), which is a decrease in the relaxation rate after perturbations prior to a TP that can be measured by statistical early warning signals (EWS) in the autocovariance of observational time series. In high-dimensional systems, we cannot expect a priori chosen scalar observables to show significant EWS, and some may even show an opposite signal. Thus, to avoid false negative or positive early warnings, it is desirable to monitor fluctuations only in observables that are designed to capture CSD. Here we propose that a natural observable for this purpose can be obtained by a data-driven approximation of the first non-trivial eigenfunction of the backward Fokker-Planck (or Kolmogorov) operator, using the diffusion map algorithm.

I. INTRODUCTION

We consider how critical transitions in stochastically forced complex systems may be anticipated by measuring increases in amplitude and temporal correlation of fluctuations in certain observables as early-warning signals (EWS). We consider a general class of heterogeneous systems with many interacting agents or scales, arising for instance in ecology, biology, social science, and the Earth system [1]. Such complex systems are often modeled by a first-order stochastic differential equation with a non-linear, deterministic *drift* giving rise to possibly chaotic dynamics, and a noise process that represents unresolved scales and random disturbances by the environment, as well as control parameters that represent slow changes in external boundary conditions. A critical transition occurs when upon change of a control parameter a base state, i.e., a stable invariant set of the drift, loses stability and the system undergoes an abrupt transition to an alternative state. This is usually due to a collision of the base state with an *edge* state, which is an unstable invariant set of the drift. The stable manifold of the edge state is the basin boundary separating the base state from the alternative state. The simplest case of such a transition is a noisy saddle-node bifurcation (SNB) [2], which is often considered the archetype of a tipping point (TP).

* johannes.lohmann@nbi.ku.dk

EWS arise due to so-called critical slowing down (CSD) [3–9], which is most easily understood for the SNB. Here, as a control parameter μ that modulates the drift crosses a critical value (e.g. $\mu_c = 0$) the leading eigenvalue of the Jacobian describing the linearized dynamics around the base fixed point crosses the imaginary axis. Leading up to this, for $0 < \mu \ll 1$, the dynamics along one degree of freedom (d.o.f) becomes much slower compared to all others, and after a short relaxation time the system is confined to an (extended) center manifold, or more precisely to a neighborhood thereof due to noise from random perturbations of the system's environment [10]. The drift on the manifold is one-dimensional and given by the SNB normal form after a suitable coordinate transformation. As $\mu \to 0$, CSD refers to the slowing down of the relaxation dynamics towards the equilibrium along the center manifold after an arbitrary, discrete perturbation.

In more general cases, where the base state is a limit cycle or chaotic attractor, we also expect a decrease of the relaxation rate back to steady state after a perturbation of the system. This is plausible if we assume that upon control parameter change the underlying deterministic dynamics experience a continuous change from being stable to neutrally stable in one d.o.f, before finally becoming unstable. This generic feature makes the detection of loss of resilience to perturbations the primary avenue for predicting TPs [11].

For real-world, large-scale systems controlled perturbations are not available. Instead, there is a permanent influence of random disturbances from the environment. Such noise-driven, natural fluctuations of the unperturbed system allow one to infer the system's response to perturbations if linear response theory guarantees a fluctuation-dissipation theorem [12, 13]. In particular, the size and correlation of the fluctuations are expected to grow in tandem with the system's slowing response as the critical transition is approached - which is the other side of the coin of CSD - thereby forming statistical EWS. Growing fluctuations towards the basin boundary imply a flattening of the quasipotential [14]. This happens in the direction of a particular d.o.f that is related to the location of the edge state, since the latter usually lies on the most probable path of a noise-induced escape [15].

Real-world observations have been analyzed for CSD by measuring statistical EWS of presumed critical transitions, including financial crises [16], depression [17, 18], neuron spiking [19], and climate tipping points, such as the Greenland ice sheet [20], Amazon forest [21] and Atlantic Meridional Overturning Circulation (AMOC) [22–24]. But statistical false positives and false negatives can occur. The destabilization of the system in a single (crit-

ical) d.o.f implies that the increase in noise-driven variability occurs also predominantly in a single d.o.f, which gives rise to a scalar observable where the variance is expected to diverge and the autocorrelation to tend to 1 at the bifurcation. Thus, EWS can be masked if the measurements have been taken from a dynamical observable that does not sufficiently project on the critical d.o.f [2, 15, 25–27]. In the simple case of the SNB, this means that the observable does not follow the SNB normal form to any good approximation and hence does is not subjected to significant CSD.

Consequently, the central question that will be addressed here is what observables should be used to detect CSD. This depends on how the system under the influence of noise responds to perturbations away from its steady state, and how this response changes as a control parameter slowly approaches its critical value. This can be understood in terms of the Fokker-Planck (FP) equation, which governs the temporal evolution of the probability density of the state in phase space. The density can be written as an expansion in the eigenfunctions ψ_n of the FP operator $\mathcal{L}$. For a fixed control parameter any initial density will converge to the unique stationary density $\pi(\mathbf{x})$, which is the first eigenfunction ψ_0 with eigenvalue $\lambda_0 = 0$. The system is then in statistical equilibrium, where the contributions of all other ψ_n with $\lambda_n < 0$ have decayed. The first few ψ_n (with λ_n closest to 0) signify locations in phase space where fluctuations tend to linger on a finite, but long-term time horizon.

We consider systems with a TP, i.e., the deterministic drift is (at least) bistable with a base and alternative attractor. In accordance with the paradigm of bifurcation-induced tipping [28] we assume low noise, and hence the system spends long periods of time in distinct regions around the attractors of the drift, referred to as metastable states. There are rare transitions between the states on time scales of $\mathcal{O}(1/\lambda_1)$, and the eigenfunction ψ_1 with $|\lambda_1| \ll 1$ signifies a very slow transfer of density between the metastable states. While part of the invariant density $\pi(\mathbf{x})$ occupies the alternative metastable state, we can assume that on a finite time horizon the system is in a quasi-stationary distribution concentrated entirely around the base state, where the contribution of ψ_1 has not decayed. Due to CSD, the relaxation towards the base state within the quasipotential well of the base state becomes slower along a particular mode. When close enough to the TP, this mode becomes the slowest in the system and will be expressed by the next eigenfunction ψ_2.

An observable that naturally expresses increases in fluctuations related to this critical

mode is proposed here to be given by the corresponding eigenfunction ϕ_2 of the backward (adjoint) operator $\mathcal{L}^*$, also known as the generator, which shares the same eigenvalues as $\mathcal{L}$. The latter governs the temporal evolution of expectation values of observables as a function of initial states, and the first few ϕ_n can be interpreted as patterns of initial conditions with slowest decay towards $\pi(\mathbf{x})$. This proposition is in agreement with the framework of optimal fingerprints presented in [29].

To obtain $\mathcal{L}^*$ from observational data we propose to use the diffusion map (DM) algorithm [30–32]. DM has been successfully used to define generalized collective coordinates that capture the effective dynamics of complex systems [32–34]. It gives an approximation (discretized on the set of data points) of $\mathcal{L}^*$ induced by a stochastic differential equation with drift $\nabla \ln[\pi(x)]$, i.e., a gradient system related to the quasipotential of the underlying stochastic dynamic system [14, 35]. This incurs an error when the underlying system has strong non-gradient dynamics, but it should still give useful results in our context since it preserves the flattening of the quasipotential (in the direction of a particular critical d.o.f) as a key property of CSD, which is not affected by non-gradient terms of the drift. Indeed, we show that from a DM approximation we can obtain physical observables that carry excellent EWS also for non-gradient systems, including a high-dimensional global ocean model exhibiting a TP of the AMOC.

The paper is structured as follows. Sec. II reviews some fundamentals of FP operators and its spectral properties and introduces notation, as well the DM algorithm to estimate the eigenfunctions of $\mathcal{L}^*$ from data. In Sec. III A and III B we motivate the usage of the backward FP eigenfunctions for the purposes of EWS with simple double well potential systems in one and two dimensions. In Sec. III C we show that the reconstruction of the eigenfunctions with the DM method indeed yields observables that carry strong EWS in conceptual models. Further, in Sec. III D we show that such observables are strictly necessary if one wants extrapolate from increasing fluctuations to forecast the timing of a tipping point. In Sec. III E we apply our method successfully to a high-dimensional model of the global ocean circulation, and conclude with a discussion in Sec. IV.

II. FOKKER-PLANCK EIGENFUNCTIONS AND DIFFUSION MAPS

Consider the d-dimensional system state $\mathbf{X} = (X^1, ..., X^d)$, governed by an Ito diffusion equation with time-independent coefficients written component-wise as

$$dX^\gamma = a^\gamma(\mathbf{X})dt + \sqrt{\epsilon}\sigma^\gamma_\nu(\mathbf{X})dW^\nu, \tag{1}$$

with drift $a^\gamma(\mathbf{X})$ and diffusion $\sigma^\gamma_\nu(\mathbf{X})$. The transition probability density $P(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{x_0}) \equiv P(\mathbf{x}, t|\mathbf{x_0})$ is governed by the Fokker-Planck (FP) equation

$$\partial_t P(\mathbf{x}, t|\mathbf{x_0}) = \mathcal{L}P(\mathbf{x}, t|\mathbf{x_0}) \tag{2}$$

with FP operator

$$\mathcal{L}(x) = -\frac{\partial}{\partial x^\gamma}a^\gamma(\mathbf{x}) + \frac{\epsilon}{2}\frac{\partial^2}{\partial x^\gamma \partial x^\nu}b^{\gamma\nu}(\mathbf{x}) \tag{3}$$

and diffusion tensor $b^{\gamma\nu}(\mathbf{x}) = \sigma^\gamma_\lambda(\mathbf{x})\sigma^\nu_\sigma(\mathbf{x})\delta^{\lambda\sigma}$. The stationary distribution $\pi(\mathbf{x}) \equiv \lim_{t\to\infty} P(\mathbf{x}, t|\mathbf{x_0})$ satisfies $\mathcal{L}\pi = 0$ and is an eigenfunction of $\mathcal{L}$ with eigenvalue $\lambda_0 = 0$.

We assume in the following that the system obeys detailed balance, which is the case in gradient systems with additive noise. As mentioned above, we still apply our method to non-gradient systems, observing that we will effectively reconstruct a gradient system based on the quasipotential $V_q(\mathbf{x}) \propto \ln \pi(\mathbf{x})$ of the full system, and that the presence of non-gradient terms does not change $\pi(\mathbf{x})$. The eigenfunctions of the FP operator

$$\mathcal{L}\psi_n(\mathbf{x}) = \lambda_n\psi_n(\mathbf{x}) \tag{4}$$

with eigenvalues $\lambda_0 = 0 > \lambda_1 \geq \lambda_2 \geq ... > -\infty$ then form an orthonormal basis under an inner product weighted by $\pi(\mathbf{x})$. Time-varying solutions of the FP equation can be written in the eigenfunction basis as

$$P(\mathbf{x}, t) = \sum_{n=0}^{\infty} c_n\psi_n(\mathbf{x})e^{\lambda_n t}, \tag{5}$$

with $c_n = \int \psi_n(\mathbf{x})\pi^{-1}(\mathbf{x})\pi_0(\mathbf{x})dx$, where $\pi_0(\mathbf{x}) = P(\mathbf{x}, t = 0)$. Since $\lambda_0 = 0 > \lambda_1 \geq \lambda_2 \geq ...$, eigenfunctions with small indices decay slowest. In view of studying tipping points, we consider here dynamical systems that are metastable, i.e., systems that spend a very long time in one part of phase space (a metastable set) before exhibiting a rare transition to another, and so on. In the simplest case there are two metastable sets, corresponding to neighborhoods of the attractors of the underlying deterministic dynamics, which is reflected

in the spectrum of $\mathcal{L}$ as $|\lambda_1| \ll 1$ and a spectral gap with $|\lambda_2 - \lambda_1| \gg |\lambda_1|$. To study the dynamics before the TP, it is sufficient to consider this bistable case, since a critical transition generically consists of the collision of only a single base state and a single boundary.

The adjoint of the FP operator is the generator

$$\mathcal{L}^*(x) = a^\gamma(\mathbf{x})\frac{\partial}{\partial x^\gamma} + \frac{\epsilon}{2}b^{\gamma\nu}(\mathbf{x})\frac{\partial^2}{\partial x^\gamma \partial x^\nu}. \tag{6}$$

It governs the backward Kolmogorov equation

$$-\partial_s u(\mathbf{x}, s) = \mathcal{L}^* u(\mathbf{x}, s), \tag{7}$$

which is defined on the time interval $s \in [0, T]$ for functions $u(\mathbf{x}, s) = \mathbb{E}_{x,s}[f(X_T)] \equiv \mathbb{E}[f(X_T)|X_s = \mathbf{x}]$, i.e., conditional expectation values of observables $f(\mathbf{x}, s)$ (initialized at $\mathbf{x}$), and with the final condition $u(x, T) = f(x)$. Employing the transformation $t = T - s$ the equation can be formulated as an initial value problem

$$\partial_t u(\mathbf{x}, t) = \mathcal{L}^* u(\mathbf{x}, t), \tag{8}$$

with initial condition $u(\mathbf{x}, 0) = f(\mathbf{x})$. Here, u is the conditional expectation $u(\mathbf{x}, t) = \mathbb{E}[f(X_t)|X_0 = \mathbf{x}]$ with initial position $\mathbf{x}$. The solution of (8) can be expressed as an eigenfunction expansion

$$u(\mathbf{x}, t) = \sum_{n=0}^{\infty} d_n \phi_n(\mathbf{x}) e^{\lambda_n t}, \tag{9}$$

with coefficients $d_n = \int \phi_n(\mathbf{x}) f(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x}$, and eigenfunctions ϕ_n satisfying $\mathcal{L}^* \phi_n = \lambda_n \phi_n$. The leading eigenfunction $\phi_0(\mathbf{x})$, corresponding to the eigenvalue $\lambda_0 = 0$, is the unique solution of $\mathcal{L}^* \phi_0 = 0$ and is given by $\phi_0 = \text{const}$. This reflects the ergodicity of the underlying system, which implies that (long-term) expectation values do not depend on the initial conditions. In the reversible case the eigenfunctions of the forward and backward operators are related by $\phi_n(\mathbf{x}) = \pi(\mathbf{x})^{-1}\psi_n(\mathbf{x})$, with the same eigenvalues λ_n. Below we exploit this in order to calculate ϕ_n from the eigenfunctions of the discretized forward FP operator in low-dimensional example systems.

Any observable $g(x)$ can be expressed by an expansion in the backward eigenfunction basis with

$$g(\mathbf{x}) = \sum_{n=0}^{\infty} g_n \phi_n(\mathbf{x}), \tag{10}$$

and possibly approximated by a truncation thereof. Here, $g_n = \int g(x)\phi_n(x)\pi(x)dx$. Thus, $\phi_n(\mathbf{x})$ are can themselves be considered as observables, and in particular the first few $\phi_n(\mathbf{x})$ are special observables with expectation values that converge only slowly because they are non-constant functions in regions of phase space where there is a slow relaxation to equilibrium. From a different point of view, the subset of leading $\phi_n(x)$ are transformations of the system from the original coordinates to reduction coordinates. The reduction is meaningful in case of time scale separation, which is expected to emerge when the deterministic drift of the system approaches a bifurcation. In particular, it can be shown that the evolution of the first k eigenfunctions is approximately Markovian [36]. In this case, the long-term evolution of the system is governed by the first k backward eigenfunctions.

For a (large) data set, a discrete approximation to the (first few) $\phi_n(x)$ can be obtained by the diffusion map (DM) algorithm. We briefly summarize the method, and for more details refer the reader to, e.g., [36]. The algorithm defines a weighted graph on the data points, and subsequently computes the first few eigenvalues and eigenvectors of a random walk on this graph. To this end we define a kernel with bandwidth $\epsilon > 0$ measuring the distance of two data points $\mathbf{x}$ and $\mathbf{y}$

$$K(\mathbf{x}, \mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}||^2/\epsilon^2). \tag{11}$$

With this, given N data points $\{\mathbf{x}_i\}_{i=1}^N$, construct the $N \times N$ matrix for all pairs of data points

$$\tilde{K}_{ij} = \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{p_\epsilon(\mathbf{x}_i)p_\epsilon(\mathbf{x}_j)}}, \tag{12}$$

with $p_\epsilon(\mathbf{x}) = \sum_{j=1}^N K(\mathbf{x}, \mathbf{x}_j)$. Finally, construct the row-stochastic Markov matrix

$$M_{ij} = \frac{\tilde{K}_{ij}}{D_i} \tag{13}$$

with $D_i = \sum_{j=1}^N \tilde{K}_{ij}$. The first few eigenvectors ν_n of M, corresponding to eigenvalues $\lambda_n^{(M)}$, define the so-called diffusion coordinates $\xi_n = \lambda_n^{(M)}\nu_n$. In the limit $N \to \infty$ and $\epsilon \to 0$, the operator $(M - I)/\epsilon$ converges to the adjoint $\mathcal{L}^*$ of the FP operator (i.e. the backward FP operator), and $\lambda_n^{(M)} \to \lambda_n$ [31, 37, 38]. The Euclidean distance between data points in the DM coordinates is called the diffusion distance. The diffusion distance measures how closely two points are connected via diffusion of the Markov chain M. Two points $\mathbf{x}$ and $\mathbf{y}$ may have small Euclidian distance, but large diffusion distance, which can reflect that the dynamics evolve on a lower dimensional manifold.

To evaluate the eigenfunctions approximately at points $\mathbf{y}$ that are not part of the given data set $\{\mathbf{x}_i\}_{i=1}^N$, we employ the so-called Nyström eigenspace interpolation

$$\xi_n(\mathbf{y}) = \lambda_n^{-1} \sum_{i=1}^{N} \frac{\tilde{K}(\mathbf{y}, \mathbf{x}_i)}{D(\mathbf{y})} \xi_n(\mathbf{x}_i), \tag{14}$$

with

$$\tilde{K}(\mathbf{y}, \mathbf{x}_i) = \frac{K(\mathbf{y}, \mathbf{x}_i)}{\sqrt{p_\epsilon(\mathbf{y}) p_\epsilon(\mathbf{x}_i)}} \tag{15}$$

and $D(\mathbf{y}) = \sum_{i=1}^{N} \tilde{K}(\mathbf{y}, \mathbf{x}_i)$. In Eq. 14, $\xi_n(\mathbf{x}_i)$ denotes the entry of the n-th eigenvector corresponding to the i-th data point.

In our implementation of the DM algorithm we normalize each data variable to have unit variance. Furthermore, we remove a small number of single and double outliers so that we can use a smaller ϵ to obtain a better approximation of the backward eigenfunctions. Single outliers are those data points that have the largest distance to its nearest data point, i.e., where the minimum of the distance to all other points is largest. We find $n = 15$ points with the largest minimum distance and remove the respective columns and rows in the distance matrix $K(\mathbf{x}_i, \mathbf{x}_j)$ in (12). Thereafter we remove $n = 10$ double outliers, which are those pairs of points where the second smallest distance to all other points is largest. The number of removed outliers has been chosen by trial and error to give the best performance across all data sets used here, which have a typical sample size of about 10,000. If this step is skipped, often a quite large ϵ is needed to prevent the first diffusion coordinates from not merely acting to cluster individual outliers against the rest of the data.

III. RESULTS

A. Interpretation of Fokker-Planck eigenfunctions in one dimension

We first study the FP eigenfunctions in a one-dimensional double-well potential (DW1) with additive noise

$$dX_t = -\left(\frac{d}{dx} V(X_t)\right) dt + \sigma \, dW_t, \tag{16}$$

with potential

$$V(x) = x^4 - x^2 + \beta x \tag{17}$$

where β is a control parameter. The potential is shown in Fig. 1e for different values of β . The deterministic drift of the system undergoes a saddle-node bifurcation at $\beta_c = \sqrt{8/27} \approx$

0.544331, where one of the potential wells disappears. From here on we will simply refer to this as bifurcation or TP, also when referring to the stochastic system. In Fig. 1 we show results where the forward eigenfunctions ψ_n have been determined numerically, by a discrete approximation of $\mathcal{L}$ using the finite difference method by Chang and Cooper [39], and then an eigendecomposition of the obtained matrix using an implicitly restarted Arnoldi method (scipy.sparse.linalg.eigs package implementation of ARPACK).

As β is increased from zero towards the bifurcation, for low noise $\pi(\mathbf{x})$ quickly becomes heavily asymmetric with a dominant peak at the deeper potential well (Fig. 1a), and a peak around the shallower well that is orders of magnitude smaller. ψ_n for small $n > 0$ correspond to distinct patterns that modify the density such that it takes longest until statistical equilibrium is reached, given that the pattern projects significantly on the initial density ρ_0.

The slowest decaying pattern ψ_1 describes the situation where the probability mass in one of the wells is initially larger as it should be according to $\pi(\mathbf{x})$ (Fig. 1b). For the equilibration of such a configuration, part of the probability mass needs to diffuse uphill and overcome the potential barrier. In metastable systems with low noise, as considered here, this is associated with a long time scale, and λ_1 is approximately equal to the escape rate out of the shallow well.

The pattern ψ_2 gives a large contribution when the initial density is concentrated more prominently in the vicinity of the saddle (as compared to $\pi(\mathbf{x})$). The function shows minima slightly outwards (larger $|x|$)) of the two stable fixed points, and a broad double maximum around the saddle point (Fig. 1c). This pattern can be interpreted as additional mass that survives outside the vicinity of the two minima for some time λ_2^{-1} due to the asymmetry of each well, i.e., the smaller curvature of the potential towards the saddle. In other words, the relaxation towards equilibrium is slower in the vicinity of the saddle and on the sides of the wells that are facing the saddle. As the bifurcation is approached, the segment of the potential within the shallow well that faces the saddle becomes smaller, and thus this is the relevant mode that carries the CSD. In the one-dimensional case, higher eigenfunctions are less important for our analysis, representing higher-order corrections (see Fig. 1d for ψ_3).

With these considerations on ψ_n one may interpret the backward eigenfunctions ϕ_n. As mentioned above, ϕ_0 is constant due to the ergodicity of the system. ϕ_1 shows a sigmoidal shape, with plateaus around the two fixed points. On time scales smaller than the mean

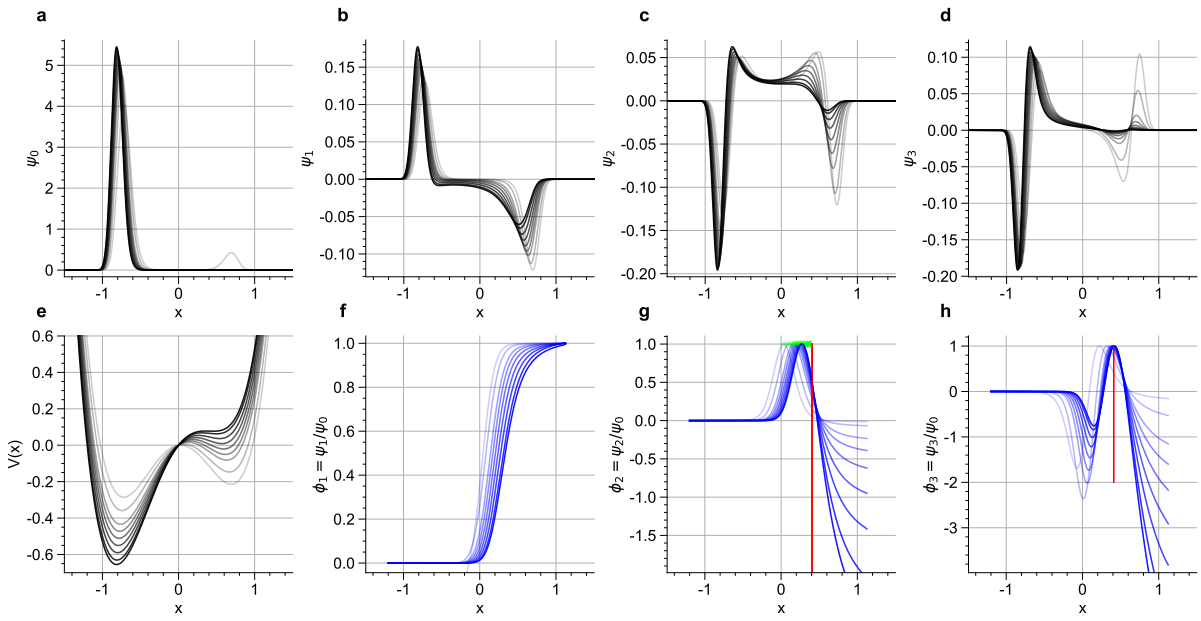

FIG. 1. Eigenfunctions of the Fokker-Planck operator for the one-dimensional double-well model (16) for a range of parameter values from $\beta = 0.05$ to $\beta = 0.53$ drawn with increasing contrast. The bifurcation occurs at $\beta \approx 0.544$. The eigenfunctions are estimated by eigendecomposition of the discrete approximation of the FP operator via the scheme of Chang and Cooper [39]. **a-d** The first four eigenfunctions of the forward FP operator for $\sigma = 0.25$. **e** Associated potential $V(x)$. **f-h** The first three non-trivial eigenfunctions of the backward operator $\phi_{1,2,3}$ (rescaled for each parameter value to have a maximum value of 1). The green dots in **g** indicate the locations of the saddle point for the respective parameter values. The vertical red line is the inflection point of the potential in the shallower well, which is independent of β.

escape time from the shallow well, observables thus have different, quasi-constant expectation values that depend on which is the starting basin. The transition zone of the sigmoid function with its midpoint at the saddle becomes narrower for decreasing noise levels.

While ϕ_0 can be called the trivial eigenfunction because it is constant, and ϕ_1 the dominant eigenfunction since $|\lambda_2 - \lambda_1| \gg |\lambda_1|$, we refer to ϕ_2 as the first subdominant eigenfunction. ϕ_2 peaks close to the saddle point, and converges to 0 at the deep well while reaching lower values in the shallow well. On time scales of order λ_2^{-1}, expectation values are thus altered when starting close to the saddle. ϕ_3 (and similarly higher eigenfunctions) is noteworthy in the sense that it is not a monotonic function within the shallow well. It

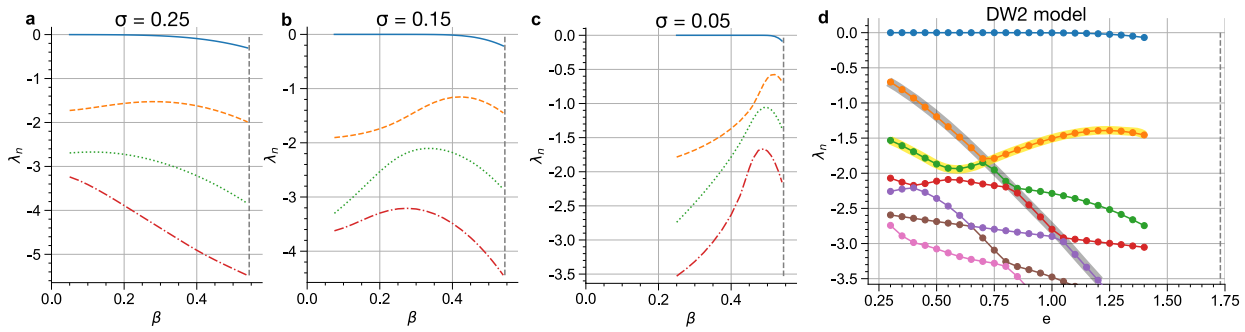

FIG. 2. **a-c** Eigenvalues $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ of the FP operator for the one-dimensional double-well potential as function of the control parameter β, for different noise levels σ. The critical value corresponding to the bifurcation is marked by the vertical dashed line. **d** Eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_7\}$ of the FP operator of the two-dimensional DW2 model (19) as a function of the control parameter e, using the noise level $\sigma = 0.3$. The bifurcation point is marked with the vertical dashed line.

first increases towards the inflection point in the shallow well (Fig. 1h), and then decreases again towards the saddle. In contrast, for the purpose of EWS, we are interested in observables that are monotonic from the base attractor towards the edge state, because otherwise fluctuations of increasing length along the critical d.o.f towards the edge state due to CSD are suppressed in the measured observable. ¨

The CSD as β is changed towards the bifurcation is reflected in the eigenvalues, though this requires a low noise level to be observed. Since ψ_2 captures the slowing of relaxation towards equilibrium as the curvature in the shallow well decreases, the relaxation rate λ_2 should go towards zero as a manifestation of CSD. This can be seen for low noise levels in Fig. 2b,c. For finite noise levels the relaxation time remains finite (Fig. 2a), and even slightly decreases very close to the bifurcation. It is bounded by the noise-induced escape rate λ_1, which decreases drastically and becomes O(1) at the bifurcation, at which point the potential is so flat that the relevant time scale for the decay of ψ_2 is not the deterministic relaxation, but pure diffusion dynamics. Additionally, the decrease in distance of the saddle and the fixed point in the shallow well may play a role.

B. Eigenfunctions in two dimensions

In the previous one-dimensional example, there were no other slow d.o.f that compete with the critical d.o.f to be in the position of mode ϕ_2. But generally the situation is different, especially in high-dimensional systems with multi-scale behaviour, where the correct physical mode first needs to slow down enough so that it emerges as ϕ_2. We illustrate this with a system of two variables x and y in the double-well potential (DW2)

$$V(x, y) = x^2(x^2 + y^2 - a) + y\frac{cy + d}{x^2 + b} + ex. \tag{18}$$

3 Adding Gaussian white noise independently to both variables yields the system of stochastic differential equations

$$\begin{pmatrix} dx_t \\ dy_t \end{pmatrix} = \begin{pmatrix} -\frac{\partial V}{\partial x} \\ -\frac{\partial V}{\partial y} \end{pmatrix} dt + \begin{pmatrix} \sigma_x \, dW_{x,t} \\ \sigma_y \, dW_{y,t} \end{pmatrix}, \tag{19}$$

where $W_{x,t}$ and $W_{y,t}$ are independent, standard Wiener processes. Fixed values $a = 2.5$, $b = 0.5$, $c = 0.2$, $d = 0.5$ are used, and e is the control parameter. For small e, there are two stable fixed points and one saddle point in the deterministic system. There is a saddle-node bifurcation of the deterministic drift at $e_c \approx 1.73$, where the potential well with $x > 0$ disappears. Figure 3f shows isolines of the potential, as well as the fixed points and basin boundary at $e = 0.5$.

The first non-trivial eigenfunction ψ_1 is again related to the slow transport of density from one well to the other (Fig. 3b), and accordingly the backward function ϕ_1 is approximately constant in the two basins, with a steep transition layer along the basin boundary (Fig. 3g). Next, compared to DW1 there is an additional eigenmode because of the slow time scale from the generally slower deterministic dynamics in the y direction. Far from the bifurcation, ψ_2 represents probability mass that is more slowly contracted in the y-direction and for some time (λ_2^{-1}) has the tendency to linger at strongly negative y values (Fig. 3c) instead of converging to either fixed point. Hence, ϕ_2 identifies initial conditions that take longest to converge to either of the two wells along the y-direction (Fig. 3h). The next mode ψ_3 corresponds (at this value of the control parameter e) to ψ_2 of the one-dimensional case, i.e., a result of slow convergence to the fixed points in the more flat parts of the asymmetric potential wells towards the saddle point (Fig. 3d). Correspondingly, ϕ_3 peaks near the saddle, and it shows that in particular initial conditions starting near the stable manifold of the

saddle will lead to a slow relaxation of conditional expectation values. ψ_4 is a higher-order, antisymmetric pattern, analogous to ψ_3 of the DW1 model.

As e approaches the bifurcation, and given the noise is low enough, the eigenvalues of the abovementioned patterns cross (Fig. 2d), leading to a different ordering of the modes (Fig. 3k-t). The eigenvalue of the pattern associated with the d.o.f. in the y-direction decreases (gray band in Fig. 2d), and the pattern drops to higher n. Instead, the pattern related to the low potential curvature towards the saddle (yellow band in Fig. 2d) becomes ϕ_2 (Fig. 3r), i.e., the pattern subdominant only to the pattern ϕ_1 that reflects noise-induced escape.

C. Eigenfunction reconstruction from diffusion maps and observables for early-warning signals

We now reconstruct the backward eigenfunctions $\phi_n(x)$ via the DM approach from data of the DW2 model (19) obtained by simulation with an Euler-Maruyama scheme with time step $dt = 0.005$. We simulate an ensemble of 100 uniformly distributed initial conditions covering both wells for a fixed simulation time $T = 100$, allowing the ensemble to converge to $\pi(\mathbf{x})$. We only use data after $t = 75$, i.e., we discard any transient dynamics. Finally, the simulated data is subsampled and all ensemble members combined to yield a set of 15,000 data points.

The scaled eigenvectors ξ_n (i.e. the diffusion coordinates) obtained from a spectral decomposition of the Markov matrix (13) define functions that can be evaluated approximately at any point in the original phase space via Eq. 14-15. Evaluation on an evenly spaced grid for DW2 shows that the first few non-trivial diffusion coordinates $\xi_{1,2,3,4}$ are indeed in good qualitative agreement with the (scaled) eigenfunctions $\phi_{1,2,3,4}$ obtained from the discretized FP operator (compare Fig. 4a-d and Fig. 3g-j).

We now restrict our attention to the scenario of bifurcation-induced tipping, where we assume the system resides in one of the metastable sets, and where we consider time scales much shorter than λ_1^{-1} associated with noise-induced escape. We consider the well that contains the base state with $x > 0$ (cf. Fig. 4 or Fig. 3). Assuming a slowly varying control parameter, the system is close to a quasi-stationary distribution $p_{qs}(x) \approx \psi_0(x) + c_1\psi_1(x)$ at any given instantaneous control parameter value. Here, $c_1\psi_1$ compensates ψ_0 such that all

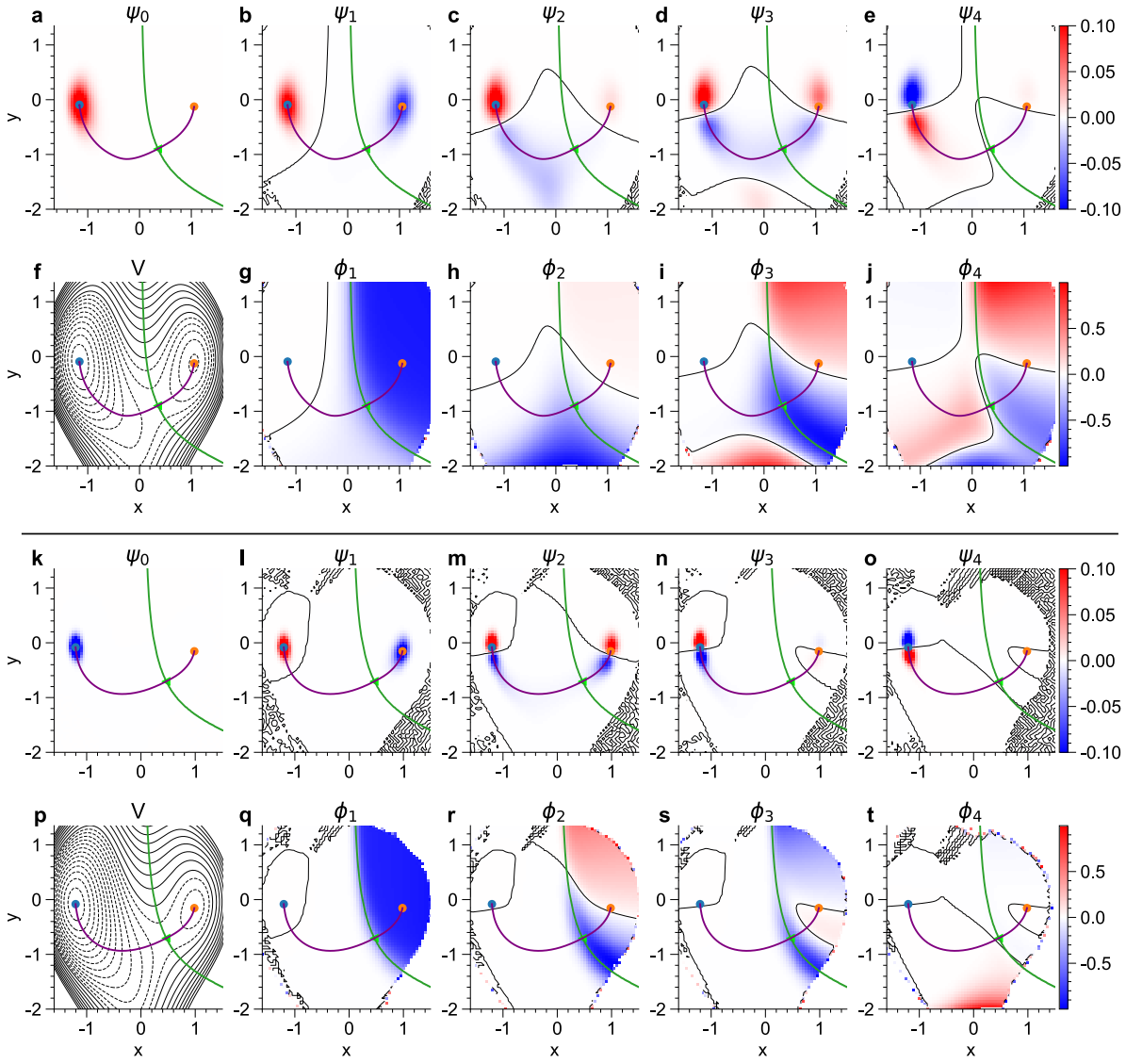

FIG. 3. **a-j** Forward (**a-e**) and backward (**g-j**) FP eigenfunctions of the two-dimensional double well model (19) with $\sigma = 0.6$ and control parameter $e = 0.5$, computed using the method by Chang and Cooper [39]. The black contour depicts the level where $\psi_n = \phi_n = 0$. The potential $V(x, y)$ of the system is shown as level sets in (**f**). The instanton (computed by the method in [40]) is drawn in purple, and the basin boundary in green. **k-t** Same but for the model with $\sigma = 0.3$ and control parameter $e = 1.0$, which is closer to the bifurcation at $e \approx 1.73$ compared to the case in panels **a-j**. Note there is numerical noise due to the very low probabilities that occur at the steepest parts of the potential around the boundaries of the domain. This produces numerical artefacts in the zero contour-line of the eigenfunctions, where erroneously the values in the computed eigenfunctions rapidly alternate in sign.

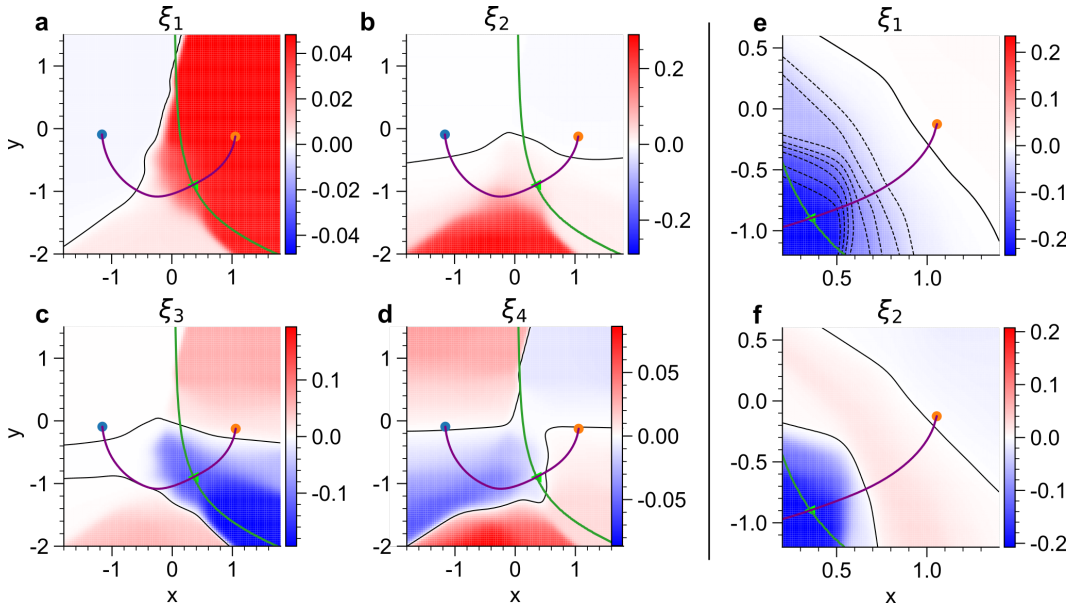

FIG. 4. **a-d** First four diffusion coordinates $\xi_{1,2,3,4}$, which approximate the eigenfunctions $\phi_{1,2,3,4}$ of $\mathcal{L}^\star$, of the DW2 model (19) with control parameter $e = 0.5$ and noise strength $\sigma = 0.6$, estimated by DM on simulated data sampling the whole phase space. The diffusion coordinates are evaluated at evenly spaced grid points using Eq. 14-15. **e,f** First two non-trivial diffusion coordinates of the DW2 model (19) from simulated data restricted to dynamics that remains in the shallow well, with control parameter $e = 0.5$ and noise strength $\sigma = 0.3$.

mass is concentrated in the shallow well. The dominant eigenfunction ϕ_1 is approximately constant in the shallow well, and thus variations in expectation values are determined by ϕ_2 and onward. When sufficiently close to the bifurcation, the first backward eigenfunctions are almost constant $\phi_2 \approx \phi_3 \approx \phi_4 \approx 0$ in the basin of the alternative state (Fig. 3q-t). Hence, it should be possible to approximate them from data restricted to the basin of the base state.

We compute DMs from simulation data with initial conditions restricted to a square domain around the base state that lies entirely within its basin. The initial conditions quickly converge to the quasi-stationary distribution $p_{qs}(x)$ in the shallow well and the transient during equilibration is discarded. A small noise strength $\sigma = 0.3$ is used, ensuring that noise-induced transitions to the other well during the simulation time are extremely rare. Realization leading to transitions are discarded. The features of the first two non-trivial eigenfunctions (Fig. 4e,f) are consistent with the corresponding (higher) eigenfunctions obtained from the full state space (Fig. 3r,s or Fig. 3i,j). Specifically, the level sets show that

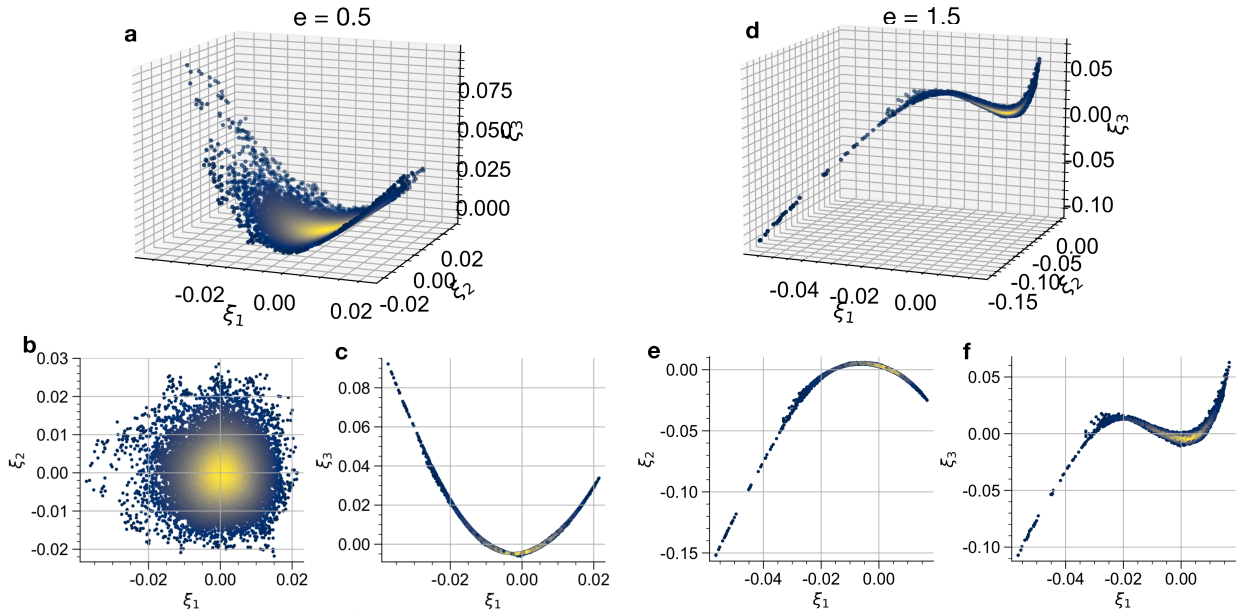

FIG. 5. **a, d** Simulated data points of the DW2 model (19) in the space spanned by the first three diffusion coordinates, for parameter value $e = 0.5$ far from the bifurcation (left) and for $e = 1.5$ (right), which is closer to the bifurcation at $e \approx 1.73$. The lower panels **b,c,e,f** show the same data in two-dimensional projections onto the diffusion coordinates (ξ_1,ξ_2) and (ξ_1,ξ_3). We use a lower noise level of $\sigma = 0.09$ in order to obtain simulation data restricted to the shallow well when very close to the bifurcation, and thus the relation of ξ_1 and ξ_2 is different compared to Fig. 4e-f, where $\sigma = 0.3$.

ξ_1 is increasing monotonically and non-linearly towards the saddle. Level sets of ξ_2 have similar shape, but do not exhibit a monotonic increase toward the saddle. In fact, there is a quadratic relationship between ξ_2 and ξ_1.

The leading ξ_n are not necessarily all independent d.o.f's. In the system restricted to the shallow well, the critical d.o.f becomes the slowest upon approaching the bifurcation and the time scale separation with respect to all other d.o.f's becomes larger, at which point the first few backward eigenfunctions (and hence the first few diffusion coordinates) are all expected to parameterize the slowest d.o.f. This is particular for single-well systems. For instance, in a multi-dimensional parabolic potential with a slow variable x and a spectral gap, ξ_1 is a function of x and the next k eigenfunctions ξ_k (with k dependent on the magnitude of the spectral gap) are polynomially related to ξ_1 [32, 36]. In this case, ξ_1 is sufficient as a reduction coordinate and is the only diffusion coordinate that indicates monotonically how

far fluctuations evolve towards the saddle.

Accordingly, in the DW2 example restricted to the shallow well, the dynamics in the space of the diffusion coordinates $\xi_{1,2,3}$ evolves from a two-dimensional to a one-dimensional manifold as the bifurcation is approached (Fig. 5).

Further away from the bifurcation, ξ_2 represents the slow mode in the y-direction, and the two other diffusion coordinates are quadratically related and parameterize the d.o.f related to the asymmetry of the potential well towards the edge state (Fig. 5a-c). Close to the bifurcation, the dynamics becomes constrained to an approximately one-dimensional curve where ξ_2 (ξ_3) is a quadratic (cubic) function of ξ_1 (Fig. 5d-f). The observed pattern formerly associated with ξ_2 at $e = 0.5$ drops to higher eigenfunctions.

The observations above suggest three main ways to leverage information contained in the diffusion coordinates ξ_n for early-warning of TPs. First, one may observe the qualitative change of the functional dependencies of the first ξ_n as a result of the emerging time scale separation, as just discussed. Second, one can directly evaluate and compare ξ_1 for data sets obtained at different observational time slices (i.e. for different values of the control parameter) via the Nyström interpolation (14)-(15). In particular, we can estimate ξ_1 from a data set believed to be closest to a TP, for example from climate observations closest to present-day, and then evaluate the observable ξ_1 on data sets sampled further away from the TP, for example using climate observations of the past. If the variability and correlation of the values of ξ_1 is increased significantly in the former data set, this is an indication of decrease in resilience in the critical d.o.f and an impending TP. Note that the same normalization of the variables that is applied before the DM algorithm to the data set where ξ_1 is estimated has to be applied to data at parameter values further away from the bifurcation. How ξ_1 can be used in this manner is illustrated in Fig. 6a, where a clear change in variability is seen when evaluating ξ_1 estimated at $e = 1.5$ on data simulated farther away from the bifurcation at $e = 0.5$. Note that here and in most of the following we focus our presentation on the variance as EWS, but similar plots could be shown for the autocorrelation.

Third, a physical observable can be constructed by expressing ξ_1 as function of (possibly a subset of) the state variables. ξ_1 is estimated for an observational time slice closest to the TP and the values of ξ_1 at the data points (i.e. the entries of the eigenvectors of the Markov matrix M) are fit to a suitable function, e.g., a polynomial. The fitted function can then be

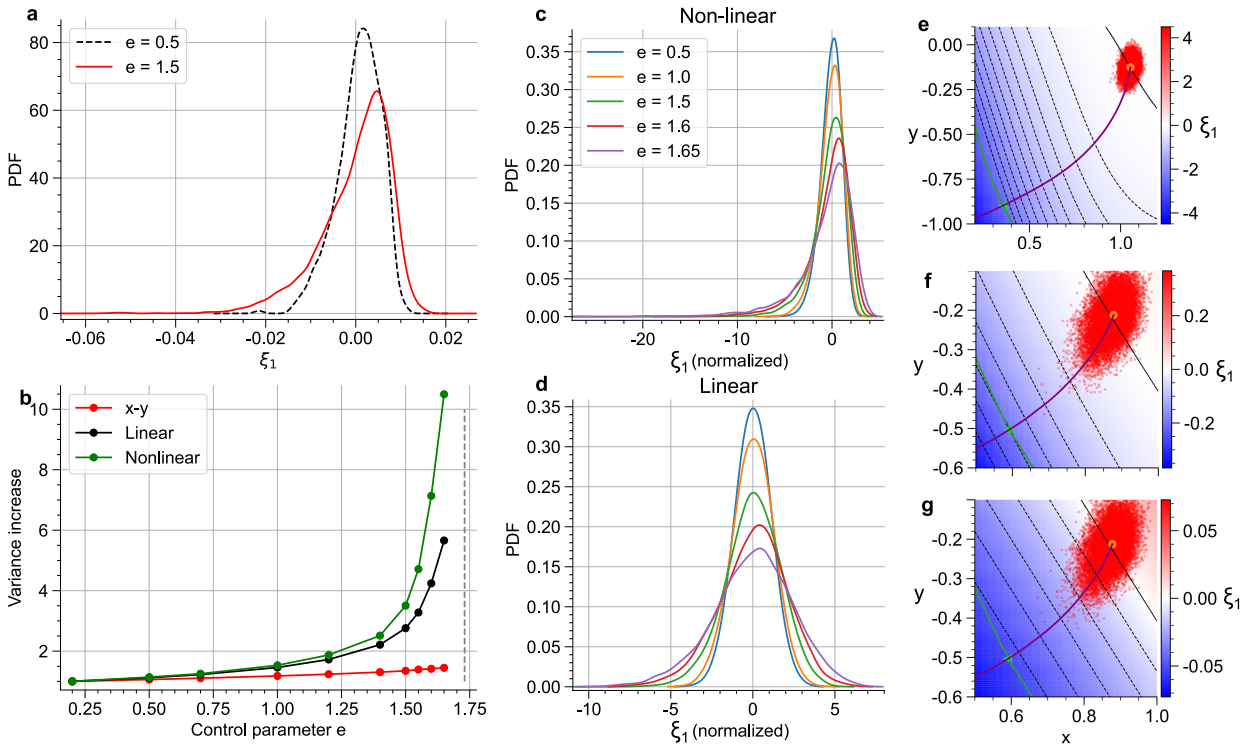

FIG. 6. **a** Distribution of the ξ_1-values evaluated at the data points of a DW2 simulation (19) at $e = 1.5$ and $\sigma = 0.09$ (red line). The black dashed line is a distribution of the same ξ_1 (i.e. obtained from the DM at $e = 1.5$), but evaluated on data points of a simulation at $e = 0.5$ via the Nyström interpolation (14)-(15). **b-d** The diffusion coordinate ξ_1, estimated for DW2 simulation data close to the tipping point ($e = 1.65$), is fit to a polynomial of the state variables (x,y), and used as observable to detect CSD by evaluating it on residual data of simulations at lower values of the control parameter. All simulations are initialized in the shallow well. Shown are distributions of the values of a linear (**d**) and cubic (**c**) polynomial, normalized to the fluctuations at $e = 0.2$. Panel **b** shows the variance increase of the linear and non-linear observable, as well as the observable $x - y$, compared to the variance of the fluctuations at $e = 0.25$. **e-g** Polynomial fits of the first non-trivial diffusion coordinate ξ_1 estimated from simulations of the DW2 system (19) (dots) with $\sigma = 0.09$. Panel **e** is a cubic fit to data at the parameter value $e = 0.5$ far from the bifurcation, and (**f,g**) is a cubic and linear fit for $e = 1.5$, which is closer to the bifurcation.

evaluated as an observable for any data sets further away from the TP. In Fig. 6e-g we show polynomial fits to ξ_1 estimated from simulation data of the DW2 model. The directionality of the level sets of the observable functions is consistent with the direction of the edge state,

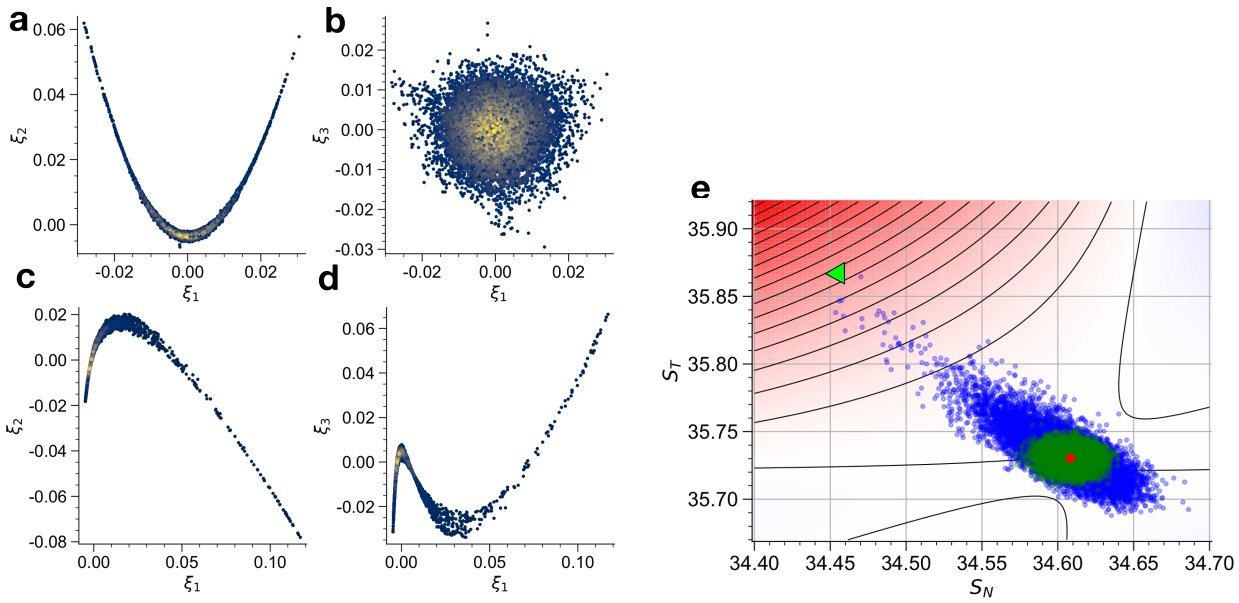

FIG. 7. **a-d** Scatterplots of simulated data from the five-box AMOC model in projections onto the first three diffusion coordinates, for two values of the control parameter $H = 0.25$ (**a,b**) and $H = 0.315$ (**c,d**). **e** Physical observable associated with the leading diffusion coordinate ξ_1, using as cubic polynomial of the variables S_N and S_T, at $H = 0.315$. The position of the edge state is marked by the green triangle, and the base attractor is the red dot. The blue point cloud is the data at $H = 0.315$ used for the construction of the DM. The green point cloud is corresponding simulation data further from the bifurcation ($H = 0.25$), shifted such that the mean is at the base attractor at $H = 0.315$.

as expected. Far from the bifurcation (assuming low noise) the dynamics samples only the relatively flat part of ξ_1 far from the saddle point, provided the noise is sufficiently small, as shown in Fig. 6e. Still, the fitted function shows a more rapid decrease towards the saddle point, thus indicating that it already carries the crucial information for detecting CSD. When close to the TP, non-linear functions tend to be required for an adequate fit of the ξ_1 data (Fig. 6f). However, linear fits preserve the directionality of the edge state well (Fig. 6g). Using the fits estimated from data sampled close to the TP (Fig. 6f,g), we again see that the variability of the values of ξ_1 decreases when evaluated for data sample further away from the TP (Fig. 6c,d). The non-linear observable shows a significantly stronger change in variance compared to the computed linear one (Fig. 6b).

Similar results are obtained when applying the method to a slightly more complex non-

gradient system, which is summarized in the following for a four-dimensional conceptual model of the AMOC [41]. The variables ($\{S_N, S_T, S_S, S_I\}$) are the average salinities in four boxes of the global ocean (see Sec. A for the equations). The large difference in the box volumes gives a time scale separation, with S_S and S_I being the slowest variables, and S_N being the fastest. The model is bistable for a range of the control parameter H from $H \approx 0.04$ until the bifurcation at $H \approx 0.3214$, where the stable fixed point corresponding to a present-day AMOC disappears.

Using simulation data restricted to the present-day AMOC state, far from the bifurcation ξ_1 represents a correlated relaxation mode in the slow variables S_S and S_I. ξ_2 is a quadratic function of ξ_1, and ξ_3 is independent and strongly correlated with S_N only (Fig. 7a,b). When approaching the TP, the slow mode corresponding to relaxation along the direction of the edge state emerges due to CSD, and eventually rises to the position of ξ_1, with ξ_2 (ξ_3) being a quadratic (cubic) function thereof (Fig. 7c,d). ξ_1 exhibits a strong non-linear anti-correlation with S_N and a non-linear positive correlation with S_T (see also [15]). A good physical observable representing ξ_1 is found by considering polynomial functions in a projected space of a subset of (e.g. two) variables. The best cubic polynomial fit of ξ_1 (estimated at $H = 0.315$) is a function of S_N and S_T. The resulting function shows a monotonic, non-linear increase from the base state towards the edge state, while being essentially flat in other directions (Fig. 7e). This renders the observable very sensitive to excursions towards the edge state, and thus ideal for EWS.

D. Observables and extrapolation of tipping times

So far we discussed the construction of observables that show a large increase in variance (and also autocorrelation) as a result of CSD, which is a qualitative indication that the system moves towards a TP. One may go further and attempt a quantitative prediction of the expected time of tipping by extrapolating the CSD signal in observational time series, as was done in the context of real-world climate observations in [20, 24]. This works by assuming that an observed time series samples the critical dynamics by obeying the SNB normal form, implying that such a prediction is sensitive to the choice of observable.

Consider the general multi-dimensional system described by the coupled stochastic differential equations (1), where we now assume that the drift $a^\gamma(\mathbf{X}, \mu)$ depends on a control

parameter μ and the noise $\sigma_\nu^\gamma(\mathbf{X}) = \sigma_\nu^\gamma$ is additive. If the system undergoes SNB the noise-driven dynamics is expected to become restricted to the vicinity of a one-dimensional center manifold, and is described by the normal form for a SNB

$$dx = (x^2 - \mu)dt + \sigma dW_t, \tag{20}$$

where $\mu = 0$ demarcates the bifurcation. Close to the fixed point, the system can be linearized and approximated by the Ornstein-Uhlenbeck process $d\tilde{x}_t = -\lambda\tilde{x}dt + \sigma dW_t$. Crucially, the linear restoring rate λ is related to the control parameter with $\lambda = 2\sqrt{\mu}$. Data that are sampled at small time intervals Δ_t can be approximated by an AR(1) process

$$X_{k+1} = e^{-\lambda\Delta t}X_k + \epsilon_k, \tag{21}$$

where ϵ_k are Gaussian random variables with variance $\sigma^2(2\lambda)^{-1}(1 - e^{-2\lambda\Delta t})$. For an AR(1) process of this form the autocorrelation at lag 1 is given by $\rho_1 = e^{-\lambda\Delta t}$. This means that from the above relation of λ and the control parameter μ the autocorrelation tends to 1 at the SNB, and we can reconstruct μ from data by

$$\mu = \left(\frac{\ln\rho_1}{2\Delta t}\right)^2. \tag{22}$$

Thus, in a sliding window one can estimate ρ_1 as a function of time, and, assuming a linear trend in μ, estimate with a linear fit to the function on the righthand-side at what time the control parameter will cross zero.

If only given data from an arbitrary scalar observable of the system, this extrapolation to the time of tipping can fail since the observable need not obey the SNB normal form, or may only approximately do so when arbitrarily close to the bifurcation. In fact, even for bi-stable systems with only one variable (where the question of observable is obsolete), the scaling $\mu \propto [\ln(\rho_1(X))]^2$ according to the saddle-node normal form only applies when close to the bifurcation. For the DW1 model (16) under a linear change in time of the control parameter μ the function $[\ln(\rho_1(X))]^2$ is a convex function of time (Fig. 8). Thus, the TP, which occurs at $t \approx 420$, would be predicted too early at $t \approx 390$ (red line in Fig. 8b) when the linear extrapolation is performed based on data not close enough to the bifurcation. An estimation of the autocorrelation in a moving window introduces a further bias towards a later estimated tipping, because the autocorrelation is underestimated by removing some correlation during the necessary step of detrending within each window.

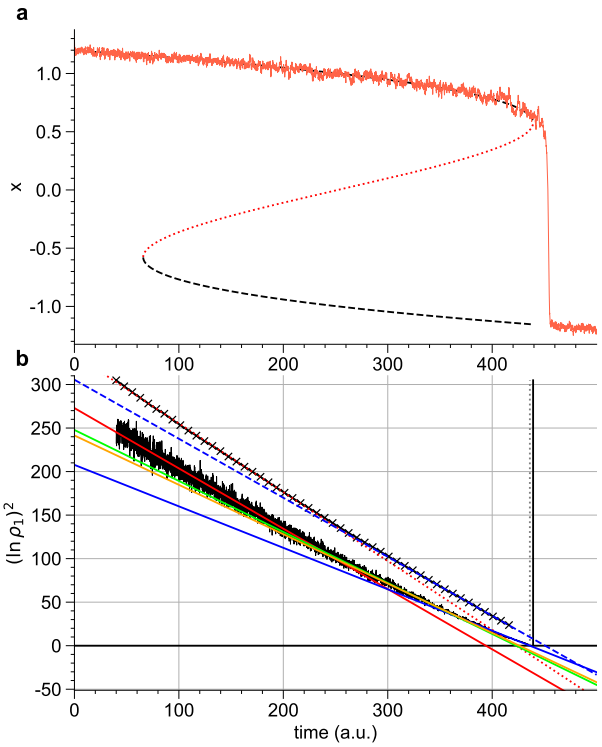

FIG. 8. Simulations of the DW1 model (16) with $\sigma = 0.05$ and a linear increase of the control parameter within 500 time units from $\beta = -0.7$ past the bifurcation point $\beta_c = \sqrt{8/27}$ to $\beta = 0.7$. **a** Time series of one simulation overlaid on the bifurcation diagram. **b** Reconstructed control parameter (Eq. 22) obtained by the lag-1 autocorrelation ρ_1, estimated at each time step (sample spacing of $\Delta t = 0.05$ time units) for an ensemble of 15,000 simulations (black trajectory) until a cutoff time where 98% of the ensemble members have not tipped yet (evaluated by crossing a threshold of $x = 0.2$). At the noise level $\sigma = 0.05$, this is only very shortly before the bifurcation is reached. The 2% of realizations that tipped are removed. The solid lines are linear fits using different segments of the data, with the red (blue) line using the first half (last sixth) of the data. The crosses as well as dotted and dashed lines are for ρ_1 estimated in a moving window of length 40.

Moreover, apart from these biases, as more dimensions are involved the prediction depends on the choice of observable. For the DW2 model (19), the variable x can give an accurate prediction when data are available close enough to the TP (Fig. 9b). In contrast, y initially shows a quasi-linear relation of μ and $[\ln(\rho_1(X))]^2$, but then a much steeper relationship closer to the bifurcation (Fig. 9c). Extrapolating based on the initial slope would

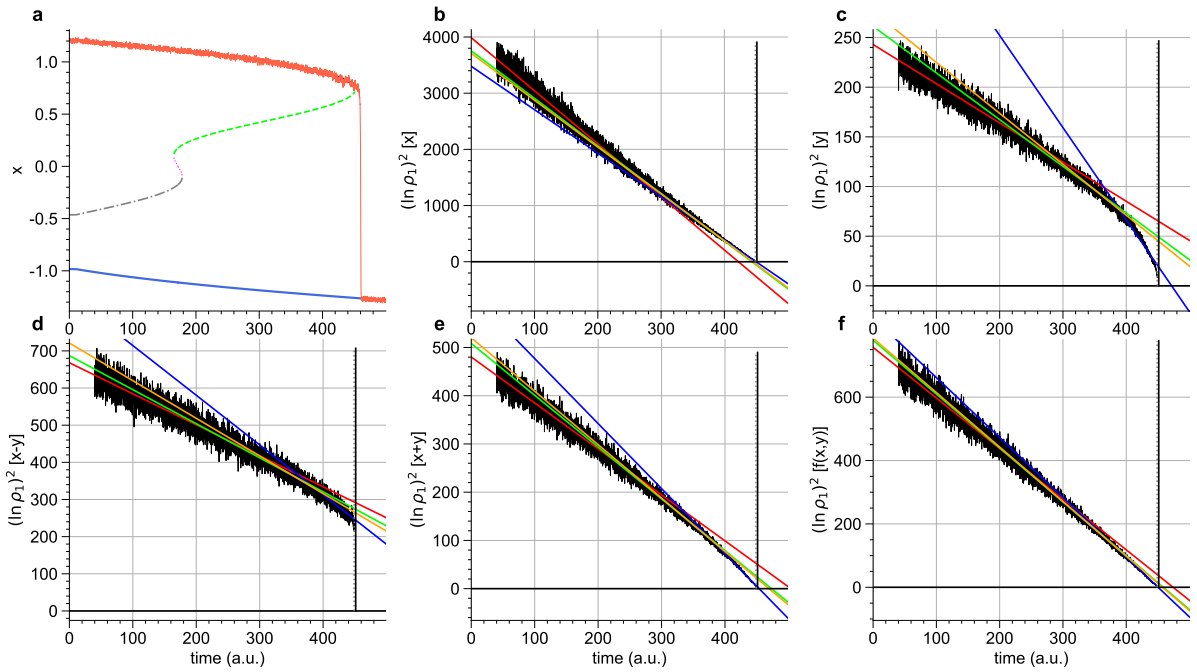

FIG. 9. Same as Fig. 8, but for simulations with the DW2 model (19) with $\sigma_x = \sigma_y = 0.05$, where the parameter e is increased linearly past the bifurcation. Shown in **a** is a single example trajectory overlaid on the bifurcation diagram. Panels **b-f** show the evolution of the quantity $[\ln(\rho_1(O))]^2(2\Delta t)^{-2}$ for different observables $O(x, y)$, along with linear fits to different parts of the time series.

lead to an estimated tipping time that is far too late. Especially unsuited observables exist, such as $O(x, y) = x - y$, where no tipping can be predicted before a noise-induced transitions would occur (Fig. 9d). In contrast, $O(x, y) = x + y$ is very closely aligned with the direction of the edge state [15] and permits to predict the time of tipping accurately for data sufficiently close to the bifurcation (Fig. 9e). Finally, the non-linear observable obtained from the DM approximation to the first subdominant backward eigenfunction (Fig. 6f) is most accurate, even when evaluated at data far from the bifurcation (Fig. 9f).

E. Application to tipping points in a global ocean model

As the final result, we show our method is capable of successfully detecting TPs in a high-dimensional system. We consider the global ocean model Veros [42], which shows a TP of the AMOC from its present-day state to a collapsed state as a result of increasing

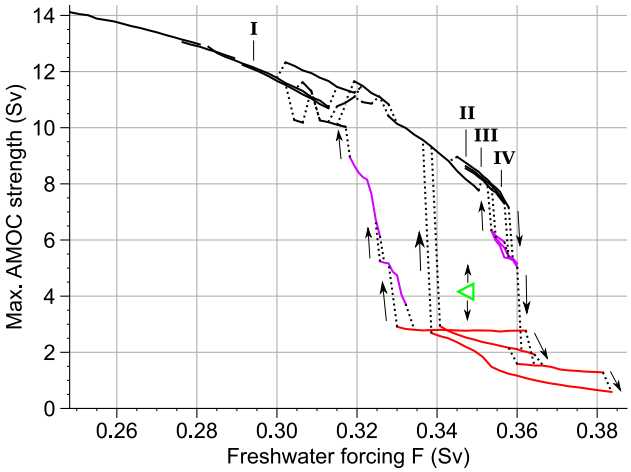

FIG. 10. Bifurcation diagram of the Veros ocean model without noise forcing (obtained in [44]), with the maximum AMOC strength as order parameter, and the freshwater forcing F as control parameter. All individual solid lines correspond to different branches of attractors, and the dotted lines as well as the arrows indicate the transition path of the system as a given attractor loses stability. The edge state at $F = 0.3472$ (computed in [45]) is marked by the green triangle.

meltwater input to the North Atlantic. Veros is a primitive-equation finite-difference ocean model forced with a fixed atmospheric climatology, and discretized on a grid of 40 latitudinal and 90 longitudinal grid points, as well as 40 depth levels. This is a coarse-resolution setup, but it enables long steady-state simulations to get good statistics beneficial for our feasibility study. As a dynamical system, the model possesses almost one million degrees of freedom. For more details on the model, see [42–44]. The meltwater input F is the control parameter, and the stability landscape with respect to F (computed in [44]) is shown in Fig. 10. There are several branches of attractors with an AMOC similar to present-day, but these collapse at a high freshwater forcing of $F \approx 0.36$. After this TP, there remain only attractor branches with a collapsed AMOC.

We use four 33,000-year long simulations (after removal of a transient for equilibration) performed at four, fixed values of F leading up to the TP, and sampled as 5-year averages of the state variables. These are referred to as simulations I to IV, see Fig. 10. Surface temperature and salinity noise forcing drives fluctuations of the system around its deterministic attractors (for more details see [15]), which otherwise feature relatively small-amplitude chaotic oscillations [44]. The system without noise forcing has been investigated previously to determine an edge state on the separatrix of the present-day and collapsed AMOC regimes

[45]. By analyzing its mean climatology it was found that the edge state distinguishes itself most strongly from the mean states on the attractors in terms of its fresh and cold deep Atlantic. This is the "fingerprint" of the edge state. Subsequently, it was argued that increased fluctuations towards the edge state as a result of CSD should be most prominent in variables quantifying this fingerprint [15]. Indeed, only a very small subset of all d.o.f, coinciding exactly with the variables describing the deep ocean fingerprint, shows significant variance increase prior to the AMOC collapse [15]. The variable that was found to exhibit the largest increase in variance is shown in Fig. 11a-d across the four data sets. In the following, we show that similar (if not better) results can be obtained with our the DM method which does not require prior knowledge of the edge state or a brute-force search across all d.o.f (risking false positives). Instead, only observational data close to the TP is required.

It would be feasible to compute the DM distance Kernel in the full space of the three-dimensional fields, perhaps after a weighting of the different physical units (temperature, salinity, density and velocity). But for simplicity we perform an initial dimensionality reduction, by averaging the salinity, temperature and density fields over boxes covering the entire ocean at different depths, and by summarizing the strength of the main ocean currents in terms of the spatial maxima of the meridional and barotropic stream functions (see [45]). This yields time series of 83 variables covering most important aspects of the model state. Fig. 11a-h shows time series of two of these variables at four different values of F, with the mean removed. Before applying the DM algorithm, we normalize all variables to have unit variance. A bandwidth of $\epsilon = 35$ for the similarity Kernel (11) was found to be optimal to resolve the data manifold at all parameter values F without being influenced by outliers, the largest of which are removed beforehand (see Sec. II).

The physical meaning of the first two inferred diffusion coordinates is summarized in Tab. I, where the five physical variables with the highest correlation to ξ_1 and ξ_2 are listed. As the control parameter is changed towards the TP, we can see that the expected critical mode emerges. When far from the TP, ξ_1 is best correlated with deep ocean density in the Indo-Pacific, South Atlantic and Southern Ocean. The next mode, represented by ξ_2, is best explained by variability in the tropical subsurface ocean. When increasing the control parameter to $F = 0.3515$, it is replaced by a mode correlated with temperature and salinity in the deep northern and tropical Atlantic. Increasing F further to $F = 0.3557$ shortly

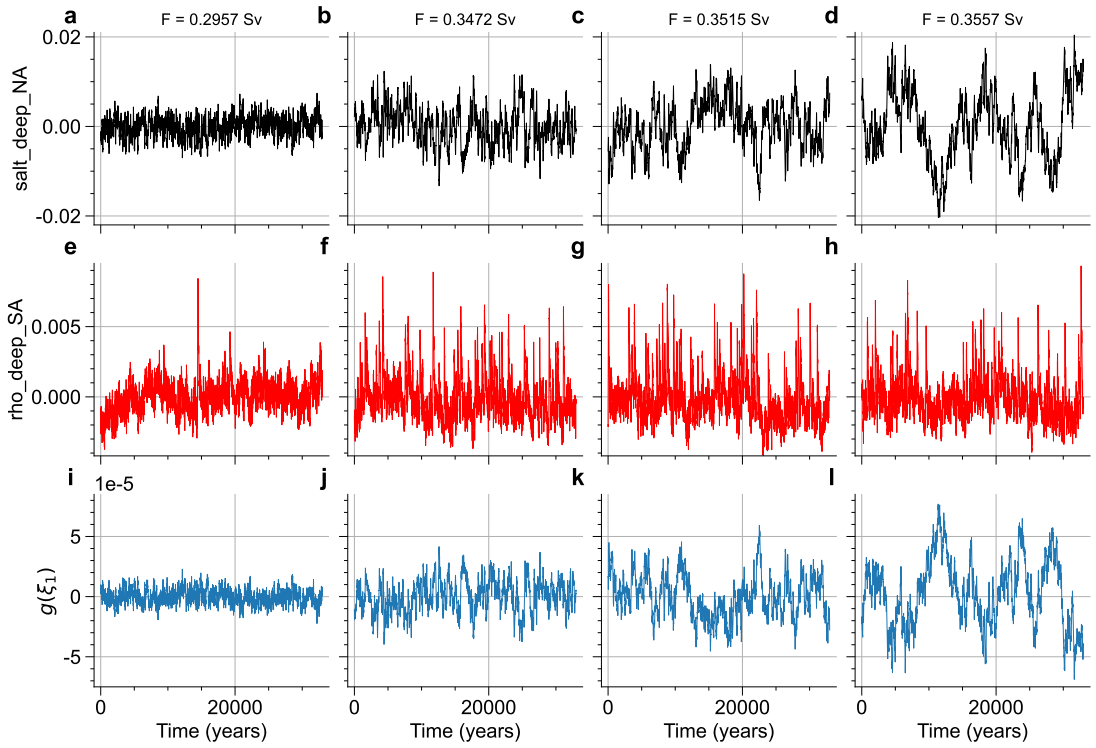

FIG. 11. **a-h** Example time series of two variables in the four Veros data sets used here. Shown is the north Atlantic deep ocean salinity (**a-d**) at the four different values of the control parameter F, as well as the south Atlantic deep ocean density (**e-h**). The time series are shown as anomalies with the mean removed. **i-l** Time series of an observable constructed by projection onto the leading diffusion coordinate ξ_1 in the subspace of deep ocean salinity. The eigenmode was estimated with the DM algorithm from data sampled at $F = 0.3557$ Sv.

before the TP, this becomes the leading mode ξ_1. The most important variables in this mode (upper right column in Tab. I) are exactly those that make up the fingerprint of the edge state [45] and feature the largest increase in variance [15].

A high-dimensional feature set gives much freedom in designing physical observables for EWS from ξ_1. The first option is again to use the Nyström extension based on all variables and interpolate the function ξ_1 to observations further back in time, in order to find evidence for increased fluctuations in the critical mode. Next, one may fit ξ_1 as a linear or non-linear function of the variables. In high dimensions it is sensible to only consider a subset of variables to find a parsimonious observable with the best signal-to-noise ratio when applied for EWS. We leave a treatment of this statistical optimization problem for

TABLE I. Spearman correlation of the top five Veros features with the first two diffusion coordinates ξ_1 and ξ_2, where the ξ_n were estimated at the four different values of the control parameter F. In bold text are key variables ('rho' refers to density) related to the mode of cold and fresh excursions in the deep north and tropical Atlantic directed towards the edge state. In italic are key variables related to the mode of fast cold excursions in the Southern ocean, which strongly increase the density in large parts of the deep ocean.

	F=0.2957		**F=0.3472**		**F=0.3515**		**F=0.3557**	
	Variable	r_S	Variable	r_S	Variable	r_S	Variable	r_S
ξ_1	rho deep TP	0.854	*rho deep SA*	-0.874	*rho deep SA*	-0.886	**salt deep NA**	-0.873
	rho deep TA	0.848	rho deep TA	-0.870	rho deep TA	-0.868	**salt subs NA**	-0.834
	rho deep SP	0.835	*rho deep IO*	-0.842	*rho deep IO*	-0.860	**temp deep NA**	-0.789
	rho deep IO	0.822	*rho deep SP*	-0.803	*rho deep SP*	-0.827	**salt deep TA**	-0.735
	rho deep SA	0.813	*rho deep SO*	-0.794	*rho deep SO*	-0.821	temp deep IO	0.730
ξ_2	temp subs TP	0.627	temp subs TA	0.624	**salt deep NA**	-0.775	*rho deep SO*	-0.782
	rho subs TP	-0.607	rho subs TA	-0.615	**salt subs NA**	-0.765	*rho deep SA*	-0.763
	rho subs TA	-0.597	rho subs TP	-0.567	**temp deep NA**	-0.690	*rho deep SP*	-0.723
	temp subs TA	0.582	salt subs NA	-0.563	**temp subs TA**	0.667	salt deep SO	-0.701
	temp deep TP	0.518	temp subs TP	0.554	rho subs TA	-0.643	rho deep TP	-0.681

future research, and consider here three simple examples. First, simply take the variable with highest correlation to ξ_1 as observable. This is NA deep ocean salinity (Fig. 11a-d), which indeed has the highest increase in variance of all individual features [15], increasing by a factor of 12.23 when going from $F = 0.2957$ to $F = 0.3557$. Second, fit ξ_1 to a linear combination of the two best features in Tab. I (at $F = 0.3557$). This yields a variance increase by a factor of 14.21. Third, fit ξ_1 to a cubic polynomial of two variables, which are chosen as the two features among the best four (Tab. I) that represent two different ocean sectors, thus giving some degree of spatial independence. This observable captures the directionality of the edge state well (Fig. 12), and shows an increase in variance by a factor of 12.60.

Finally, ξ_1 can be mapped back into the physical space of full dimension, whereafter spatio-temporal anomalies from different time periods can be projected onto the mode. In particular, we propose to obtain the physical representation of this mode by averaging over the time points where ξ_1 is extremized. For instance, one can choose the data points with the

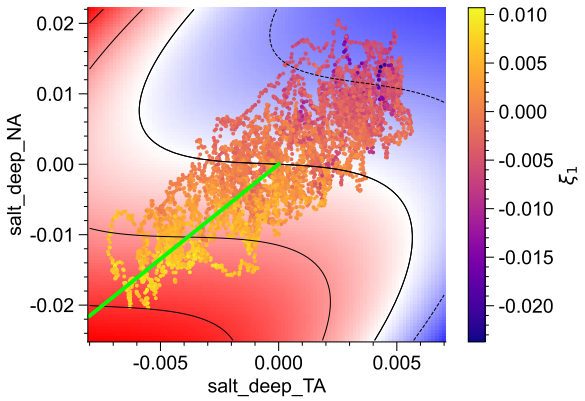

FIG. 12. Two-dimensional observable (color map and contours) created from a cubic polynomial fit of two variables of the Veros data (north and tropical Atlantic deep ocean salinity at parameter value $F = 0.3557$) to the values of ξ_1 obtained from the diffusion map. The point cloud depicts the data with color coding according to corresponding the value of the diffusion coordinate ξ_1. Shown are anomalies with respect to the mean state of the model. The green line is a vector pointing from this mean state to the edge state [15].

top 5% largest and 5% smallest values of ξ_1. Averaging independently over these two sets of data points defines a positive and a negative phase of the mode. By taking the difference of the positive and negative patterns (or vice versa) we obtain a pattern that describes the mode as a whole and that we can project data onto. This may be viewed as linear approximation that interpolates the physical mode linearly as a function of the value of ξ_1. Fig. 13 gives a comparison of the first subdominant modes extracted in this way by extremizing ξ_1 far from (Fig. 13a,d) and close to the TP (Fig. 13b,e). The modes are projected down to the two-dimensional physical space of vertically averaged deep ocean of temperature and salinity. Far from the TP, the mode is characterized by a global cooling of the deep ocean initiated by abrupt (decadal-scale) cooling events in the Southern Ocean (seen as spikes of density increase in Fig. 11e-h), which are excited by noise in the multistability regime, as discussed in [15]. The resulting dense deep ocean water spreads throughout the deep ocean, before the signal decays on a multi-centennial time scale. There is no significant salinity signal.

Close to the TP, the new physical mode that has emerged as subdominant eigenfunction is characterized by a cold and fresh anomaly of the deep northern and tropical Atlantic (Fig. 13b,e), with a spatial pattern that very closely resembles the anomaly of the edge state

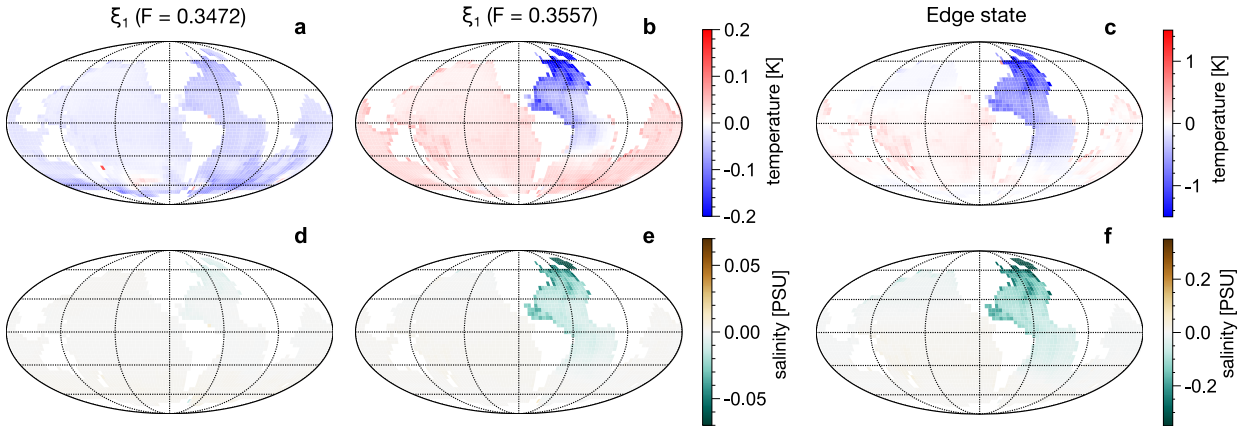

FIG. 13. **a,b,d,e** Anomaly maps of deep ocean (vertical average below 1000m) temperature (**a,b**) and salinity (**d,e**) describing the first subdominant spatiotemporal mode ξ_1 of the Veros model obtained by the DM algorithm at the control parameter values $F = 0.3472$ (**a,d**) and $F = 0.3557$ (**b,e**). **c,f** Corresponding anomaly map of the edge state with respect to the mean state on an attractor with active AMOC (state II in Fig. 10), as estimated from the deterministic version of the model in [45].

(Fig. 13c,f). A scalar observable is created by projecting the snapshots of the data fields (at each time step, and as anomalies with respect to a mean state) onto this pattern via the scalar product of the two fields. Comparing different observational slices (such as the simulations I to IV) then reveals the changes in variability of this critical mode, which serves as EWS. The variance increases by a factor of 21.05 when going from state I to state IV, and the time series of this observable are shown in Fig. 11i-l.

While these are encouraging results, in practice there is a risk of false positives and negatives, since here the critical physical mode only appears in its correct place at ξ_1 when already quite close to the TP. One needs to verify that the leading mode captured by ξ_1 is likely a critical mode. The guiding principle should be to look how strongly fluctuations in the leading diffusion coordinate, estimated at the current time slice, have increased compared to data slices back in time, and then set a level of statistical significance based on a reference period. In the Veros data one can in this way rule out the initially leading mode (Fig. 13a,d), because its variability does not keep increasing towards the TP (Fig. 11f-h), and the associated excursions do not last longer. But in principle there remains a chance for false positives of a new time scale separation would arise upon change of a control parameter

for reasons not related to a TP.

IV. DISCUSSION

Here we propose a method to obtain observables for detecting CSD before TPs in multi-stable systems driven by low noise via a data-driven approximation of the FP operator. While its first k eigenfunctions are very slowly decaying modes related to rare noise-induced escape between metastable states [46], the subsequent eigenfunctions describe probability density patterns in phase space that relax slowest towards the quasi-stationary distribution centered around the metastable states. In the context of TP, where a base state loses stability and the system transitions to an alternative state, we can consider without loss of generality the bistable situation with $k = 1$. Further, we consider bifurcation-induced tipping [28], where the system prior to the TP is only observed in the basin of the base state and the contribution of the eigenfunction ψ_1 remains quasi-constant (until infinitesimally close to the TP). In this case, the first non-trivial eigenfunction that can be observed in data is ψ_2, which describes the slowest relaxation mode towards the quasi-stationary distribution around the base state within its basin of attraction. As the TP is approached, due to CSD this mode will eventually represent the slowing relaxation along the critical d.o.f.

We suggest to approximate the corresponding eigenfunction of the backward FP operator by the first diffusion coordinate ξ_1, obtained as scaled eigenvector of the DM Markov matrix. From ξ_1, one can obtain a physical observable - e.g. by projecting onto patterns obtained as average over data points that extremize ξ_1 - that shows a monotonic increase from the base state along the critical d.o.f towards the edge state. With several examples of low-dimensional bi-stable models we demonstrated that such an observable shows increases in noise-driven fluctuations that provide robust statistical EWS of the CSD associated with the impending TP. We also showed that measuring CSD in the correct observable is crucial when attempting to predict the time of tipping by extrapolating the scaling of variance or autocorrelation of a scalar time series based on the SNB normal form. We furthermore showed that the method can be applied successfully to simulation data from a high-dimensional global ocean model that features a TP of the AMOC. A critical mode was extracted that is in excellent agreement with the mode that would be expected from knowledge of the edge state [15, 45], and a scalar observable was derived from ξ_1 that shows highly significant

increases in variance that are useful as statistical EWS.

A general caveat for this and other methods aiming to measure CSD from high-dimensional systems is that the critical mode may only emerge as the first subdominant eigenmode when already very close to the TP. If there are competing non-critical modes that are very slow, it may not be possible to identify the critical mode, since one would need a very long observational time horizon while being close enough to the TP where the critical mode is finally the slowest. Indeed, it was seen for the Veros model that the correct physical mode takes on the role of ξ_1, but only as the system was arguably quite close to the TP. Unless one knows beforehand which d.o.f should be measured, for instance by knowledge of the edge state [15] or robust physical considerations, it may only be possible to issue a reliable warning when very close to the TP, and potentially only after the probability of noise-induced transitions has become substantial.

There are other previously proposed dimensionality reduction methods aiming to extract a scalar observable that can be used to detect CSD. These include variance-based techniques [8, 47–49], where the first principle component is identified by empirical orthogonal functions (EOF) - as originally proposed to obtain the critical mode for EWS [8] - or principal oscillation patterns (POP). An autocorrelation-based method has also been proposed, where the directions of maximum variance of the first differences of multivariate time series are found, which gives components of high autocorrelation that should indicate directions of lowest resilience [50]. Other methods look for a SNB in the full set of (observed) variables via the eigenvalues of a reconstructed Jacobian, which is determined by fitting a multivariate autoregressive model [51] or a multidimensional Langevin equation [52].

Our approach is distinguished by combining several attributes. It yields an observable derived from the first subdominant backward FP eigenfunction that is designed to represent the critical mode displaying CSD, based on the flattening of the quasipotential expected for a broader class of TPs. This mirrors the reasoning for a natural tipping observable recently proposed in [29], and it is also supported by other operator-theoretic work on the topic [53–57]. The specific DM algorithm that we propose to use is a non-linear dimensionality reduction method and thus allows for observables to be non-linear functions of the state variables. It can be deployed for relatively high dimensional systems, since the quality of its approximation for a given sample size is not dependent on the full state space dimension, but on the intrinsic dimension of the underlying data manifold, which may be much lower.

An equal time spacing of data points, or any time ordering at all, is not required. The method furthermore allows for a qualitative assessment of changes in the dominant physical modes, by observing the functional relation between eigenfunctions, and the relation of eigenfunctions and physical variables. This is useful for determining an emerging time scale separation before the TP, and it may help to prevent false positives. It would be interesting in future work to compare our approach to the abovementioned methods.

Future work should address two shortcomings. First, a modification of the method that is viable in higher dimensions and able to reconstruct the correct backward FP eigendecomposition also including non-gradient terms is desirable. This may broaden its applicability and further improve the significance of obtained observables for EWS. Second, in many cases of real-world relevance, such as tipping of the polar ice sheets, the change in the control parameter is fast compared to the critical relaxation mode ψ_2 and perhaps many other modes. This means that the system is not in a quasi-stationary state, as was assumed here, and it is likely that the critical mode is not displayed before crossing the TP. An extended framework based on non-autonomous dynamical systems theory can hopefully yield useful insights on whether EWS before the de-facto TP exist in this case.

---

[1] J. Ladyman, J. Lambert, and K. Wiesner, What is a complex system?, Eur. J. Phil. Sci. **3**, 33 (2013).

[2] C. Kuehn, A mathematical framework for critical transitions: Bifurcations, fast–slow systems and stochastic dynamics, Physica D **240**, 1020 (2011).

[3] H. Haken, *Synergetics, 2nd ed.* (Springer, Berlin Heidelberg New York, 1980).

[4] G. R. North, R. F. Cahalan, and J. J. A. Coakley, Energy Balance Climate Models, Rev. Geo. Space Phys. **19**, 91 (1981).

[5] C. Wissel, A universal law of the characteristic return time near thresholds, Oecologia **65**, 101 (1984).

[6] T. J. Crowley and G. R. North, Abrupt Climate Change and Extinction Events in Earth History, Science **240**, 996 (1988).

[7] T. Kleinen, H. Held, and G. Petschel-Held, The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation, Ocean Dynamics

53, 53 (2003).

[8] H. Held and T. Kleinen, Detection of climate system bifurcations by degenerate fingerprinting, Geophys. Res. Lett. **31**, L23207 (2004).

[9] A. Morr and N. Boers, Detection of Approaching Critical Transitions in Natural Systems Driven by Red Noise, Phys. Rev. X **14**, 021037 (2024).

[10] E. Knobloch and K. A. Wiesenfeld, Bifurcations in Fluctuating Systems: The Center-Manifold Approach, J. Stat. Phys. **33**, 611 (1983).

[11] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara, Early-warning signals for critical transitions, Nature **461**, 53 (2009).

[12] R. Kubo, The fluctuation-dissipation theorem, Rep. Prog. Phys. **29**, 255 (1966).

[13] M. Hairer and A. J. Majda, A simple framework to justify linear response theory, Nonlinearity **23**, 909 (2010).

[14] R. Graham, A. Hamm, and T. Tél, Nonequilibrium potentials for dynamical systems with fractal attractors or repellers, Phys. Rev. Lett. **66**, 3089 (1991).

[15] J. Lohmann, A. B. Hansen, A. Lovo, R. Chapman, F. Bouchet, and V. Lucarini, The role of edge states for early-warning of tipping points, Proc. Roy. Soc A **481**, 20240753 (2025).

[16] C. Diks, C. Hommes, and J. Wang, Critical slowing down as an early warning signal for financial crises?, Empirical Economics **57**, 1201 (2019).

[17] I. A. van de Leemput, M. Wichers, *et al.*, Critical slowing down as early warning for the onset and termination of depression, PNAS **111**, 87 (2014).

[18] M. Wichers, P. C. Groot, *et al.*, Critical Slowing Down as a Personalized Early Warning Signal for Depression, Psychoterap Psychosom **85**, 114 (2016).

[19] C. Meisel, A. Klaus, C. Kuehn, and D. Plenz, Critical Slowing Down Governs the Transition to Neuron Spiking, PLoS Comput Biol **11**, e1004097 (2015).

[20] N. Boers and M. Rypdal, Critical slowing down suggests that the western Greenland Ice Sheet is close to a tipping point, PNAS **21**, e2024192118 (2021).

[21] C. Boulton, T. M. Lenton, and N. Boers, Pronounced loss of Amazon rainforest resilience since the early 2000s, Nature Clim. Change **12**, 271 (2022).

[22] N. Boers, Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation, Nature Clim. Change **11**, 680 (2021).

[23] S. L. L. Michel, D. Swingedouw, P. Ortega, G. Gastineau, J. Mignot, G. McCarthy, and M. Khodri, Early warning signal for a tipping point suggested by a millennial Atlantic Multidecadal Variability reconstruction, Nature Comm. **13**, 5176 (2022).

[24] P. Ditlevsen and S. Ditlevsen, Warning of a forthcoming collapse of the Atlantic meridional overturning circulation, Nature Comm. **14**, 4254 (2023).

[25] M. C. Boerlijst, T. Oudman, and A. M. de Roos, Catastrophic collapse can occur without early warning: Examples of silent catastrophes in structured ecological models, PLOS One **8**, e62033 (2013).

[26] C. Kuehn, A Mathematical Framework for Critical Transitions: Normal Forms, Variance and Applications, J Nonlinear Sci **23**, 457 (2013).

[27] A. Morr, N. Boers, and P. Ashwin, Internal Noise Interference to Warnings of Tipping Points in Generic Multidimensional Dynamical Systems, SIAM J. Appl. Dyn. Syst. **23**, 2793 (2024).

[28] P. Ashwin, S. Wieczorek, R. Vitolo, and P. Cox, Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system, Phil. Trans. R. Soc. A **370**, 1166 (2012).

[29] V. Lucarini and M. D. Chekroun, Detecting and Attributing Change in Climate and Complex Systems: Foundations, Green's Functions, and Nonlinear Fingerprints, Phys. Rev. Lett. **133**, 244201 (2024).

[30] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, PNAS **102**, 7426 (2005).

[31] R. R. Coifman and S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. **21**, 5 (2006).

[32] B. Nadler, R. R. Coifman, S. Lafon, and I. G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. **21**, 113 (2006).

[33] A. Bittracher, P. Koltai, S. Klus, R. Banisch, M. Dellnitz, and C. Schütte, Transition manifolds of complex metastable systems, Journal of Nonlinear Science **28**, 471 (2018).

[34] M. Lücke, S. Winkelmann, J. Heitzig, N. Molkenthin, and P. Koltai, Learning interpretable collective variables for spreading processes on networks, Phys. Rev. E **109**, L022301 (2024).

[35] G. Margazoglou, T. Grafke, A. Laio, and V. Lucarini, Dynamical landscape and multistability of a climate model, Proc. R. Soc. A **477**, 20210019 (2021).

[36] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Diffusion maps, reduc-

tion coordinates, and low dimensional representations of stochastic systems, SIAM Multiscale Model. Simul. **7**, 842 (2008).

[37] A. Singer, From graph to manifold Laplacian: The convergence rate, Applied and Computational Harmonic Analysis **21**, 128 (2006), special Issue: Diffusion Maps and Wavelets.

[38] T. Berry and J. Harlim, Variable bandwidth diffusion kernels, Applied and Computational Harmonic Analysis **40**, 68 (2016).

[39] J. S. Chang and G. Cooper, A Practical Difference Scheme for Fokker-Planck Equation, J. Comp. Phys. **6**, 1 (1970).

[40] L. Kikuchi, R. Singh, M. E. Cates, and R. Adhikari, Ritz method for transition paths and quasipotentials of rare diffusive events, Phys. Rev. Research **2**, 033208 (2020).

[41] R. Wood *et al.*, Observable, low-order dynamical controls of thresholds of the atlantic meridional overturning circulation, Climate Dynamics **53**, 6815 (2019).

[42] D. Häfner, R. L. Jacobsen, C. Eden, M. R. B. Kristensen, M. Jochum, R. Nuterman, and B. Vinter, Veros v0.1 – a fast and versatile ocean simulator in pure Python, Geosci. Model Dev. **11**, 3299 (2018).

[43] J. Lohmann and P. D. Ditlevsen, Risk of tipping the overturning circulation due to increasing rates of ice melt, PNAS **118**, e2017989118 (2021).

[44] J. Lohmann, H. A. Dijkstra, M. Jochum, V. Lucarini, and P. D. Ditlevsen, Multistability and intermediate tipping of the Atlantic Ocean circulation, Sci. Adv. **10**, eadi4253 (2024).

[45] J. Lohmann and V. Lucarini, Melancholia States of the Atlantic Meridional Overturning Circulation, Phys. Rev. Fluids **9**, 123801 (2024).

[46] W. Huisinga, S. Meyn, and C. Schütte, Phase transitions and metastability in Markovian and molecular systems, The Annals of Applied Probability **14**, 419 (2004).

[47] S. Bathiany, M. Claussen, and K. Fraedrich, Detecting hotspots of atmosphere–vegetation interaction via slowing down – Part 1: A stochastic approach, Earth Sys. Dyn. **4**, 63 (2013).

[48] F. Kwasniok, Detecting, anticipating, and predicting critical transitions in spatially extended systems, Chaos **28**, 033614 (2018).

[49] J. Prettyman, T. Kuna, and V. Livina, Generalized early warning signals in multivariate and gridded data with an application to tropical cyclones, Chaos **29**, 073105 (2019).

[50] W. Weijer, W. Cheng, S. Drijfhout, A. V. Fedorov, A. Hu, L. C. Jackson, W. Liu, E. L. McDonagh, J. V. Mecking, and J. Zhang, Stability of the Atlantic Meridional Overturning

Circulation: A Review and Synthesis, J. Geoph. Research **124**, 5336 (2019).

[51] M. S. Williamson and T. M. Lenton, Detection of bifurcations in noisy coupled systems from multiple time series, Chaos **25**, 036407 (2015).

[52] A. Morr, K. Riechers, L. R. Gorjão, and N. Boers, Anticipating critical transitions in multidimensional systems driven by time- and state-dependent noise, Phys. Rev. Res. **6**, 033251 (2024).

[53] A. Tantet, V. Lucarini, F. Lunkeit, and H. A. Dijkstra, Crisis of the chaotic attractor of a climate model: a transfer operator approach, Nonlinearity **226**, 2221 (2018).

[54] M. Chekroun, A. Tantet, H. A. Dijkstra, and J. D. Neelin, Ruelle-Pollicott Resonances of Stochastic Systems in Reduced State Space. Part I: Theory, J. Stat. Phys. **179**, 1366 (2020).

[55] A. Tantet, M. Chekroun, H. A. Dijkstra, and J. D. Neelin, Ruelle-Pollicott Resonances of Stochastic Systems in Reduced State Space. Part II: Stochastic Hopf Bifurcation, J. Stat. Phys. **179**, 1403 (2020).

[56] M. S. Gutiérrez and V. Lucarini, On some aspects of the response to stochastic and deterministic forcings, J. Phys. A: Math. Theor. **55**, 425002 (2022).

[57] N. Zagli, V. Lucarini, and G. A. Pavliotis, Response theory identifies reaction coordinates and explains critical phenomena in noisy interacting systems, J. Phys. A: Math. Theor. **57**, 325004 (2024).

ACKNOWLEDGMENTS

We thank R. Nuterman and the Danish Center for Climate Computing for supporting the simulations with the Veros ocean model. J.L. has received support from Danmarks Frie Forskningsfond (grant no. 2032-00346B) and research grant no. VIL59164 from VILLUM FONDEN.

Appendix A: Equations for the five-box ocean model

In this appendix, the equations and parameter values of the five-box ocean model, originally published in [41], are described. The boxes, labelled by $X = N, S, T, IP, B$, are coupled unidirectionally by the thermohaline overturning circulation q, and bi-directionally by the wind-driven circulation. The dynamical equation for box B can be eliminated by salt

conservation. The remaining boxes forced by an atmospheric freshwater flux F_X multiplied by the reference salinity $S_0 = 0.035$, which is then modulated by HA_X to emulate the effect of climate change, where H is the control parameter. The varying strength of the overturning q is proportional to the density difference in the northern and southern boxes, and the temperatures T_X are fixed everywhere except in the northern box, where it is assumed that $T_N = \mu q + T_0$, with a global reference temperature T_0. This yields

$$q = \lambda \frac{\alpha(T_\mathrm{S} - T_0) + \beta(S_\mathrm{N} - S_\mathrm{S})}{1 + \lambda\alpha\mu}. \tag{A1}$$

In the model, $q > 0$ corresponds to an AMOC 'ON' state, and it is assumed that in case of a reversed circulation $q < 0$ the unidirectional coupling by the overturning flow is reversed. This yields different dynamics for positive and negative q, and a non-smooth system of four ODEs, using the Heaviside function $\Theta(\cdot)$:

$$V_N \frac{dS_N}{dt} = |q|\left[\Theta(q)(S_T - S_B) + S_B - S_N\right] + K_N(S_T - S_N) - (F_N + HA_N)S_0 \tag{A2a}$$

$$V_T \frac{dS_T}{dt} = |q|\left[\Theta(q)(\gamma S_S + (1-\gamma)S_{IP} - S_N) + S_N - S_T\right] + K_S(S_S - S_T) + K_N(S_N - S_T) - (F_T + HA_T)S_0 \tag{A2b}$$

$$V_S \frac{dS_S}{dt} = \gamma|q|\left[\Theta(q)(S_B - S_T) + S_T - S_S\right] + K_{IP}(S_{IP} - S_S) + K_S(S_T - S_S) + \eta(S_B - S_S) - (F_S + HA_S)S_0 \tag{A2c}$$

$$V_{IP} \frac{dS_{IP}}{dt} = (1-\gamma)|q|\left[\Theta(q)(S_B - S_T) + S_T - S_{IP}\right] + K_{IP}(S_S - S_{IP}) - (F_{IP} + HA_{IP})S_0. \tag{A2d}$$

Time is re-scaled by $\tau_Y = 3.15 \times 10^7$ to go from seconds to years, and the remaining parameter values are listed in Tab. A1. Additive noise is included to yield stochastic differential equations of the form

$$dS_X = f_X(S_X, H)dt + \sigma_X dW_X, \tag{A3}$$

with $X \in \{N, T, S, IP\}$, $\sigma_X = 10^{-6}$. The drift f_X represents the deterministic model (A2) and W_X are standard independent Wiener processes.

TABLE A1. Parameter values used for the five-box model, adapted from the `FAMOUSA1xCO2` calibration in [41]. $\alpha = 0.12$ (thermal coefficient) and $\beta = 790$ (haline coefficient) define a linear equation of state for the density of sea water. V_i is box volume, F_i the freshwater fluxes, T are temperatures, K_i are wind fluxes and A_i determine the distribution of freshwater forcing. η is a mixing parameter between the S and B boxes, γ determines the proportion of water which takes the cold-water path, λ and μ are constants. Subscripts indicate box labels, $i \in \{N, T, S, IP, B\}$, and '0' indicates a global reference value.

Parameter	Value	Parameter	Value
$V_{\mathrm{N}}(m^3)$	3.683 $\times 10^{16}$	$F_{\mathrm{N}}(m^3 s^{-1})$	0.375 $\times 10^{6}$
$V_{\mathrm{T}}(m^3)$	5.151 $\times 10^{16}$	$F_{\mathrm{T}}(m^3 s^{-1})$	-0.723 $\times 10^{6}$
$V_{\mathrm{S}}(m^3)$	10.28 $\times 10^{16}$	$F_{\mathrm{S}}(m^3 s^{-1})$	1.014 $\times 10^{6}$
$V_{\mathrm{IP}}(m^3)$	21.29 $\times 10^{16}$	$F_{\mathrm{IP}}(m^3 s^{-1})$	-0.666 $\times 10^{6}$
$V_{\mathrm{B}}(m^3)$	88.12 $\times 10^{16}$	$F_{\mathrm{B}}(m^3 s^{-1})$	0
A_{N}	0.194	$\eta(m^3 s^{-1})$	66.061 $\times 10^{6}$
A_{T}	0.597	γ	0.1
A_{S}	-0.226	$\lambda(m^6 kg^{-1} s^{-1})$	2.66 $\times 10^{7}$
A_{IP}	-0.565	$\mu(^o C m^{-3} s)$	7.0 $\times 10^{-8}$
$K_{\mathrm{N}}(m^3 s^{-1})$	5.439 $\times 10^{6}$	$T_{\mathrm{S}}(^o C)$	5.571
$K_{\mathrm{S}}(m^3 s^{-1})$	3.760 $\times 10^{6}$	$T_0(^o C)$	3.26
$K_{\mathrm{IP}}(m^3 s^{-1})$	89.778 $\times 10^{6}$		