

# Targeted Calibration to Adjust Stability Biases in Complex Dynamical System Models

Daniel Pals,<sup>1,‡,§</sup> Sebastian Bathiany<sup>1b,2,3,\*‡,§</sup> Joel Kuettel,<sup>2,4</sup> Richard A. Wood<sup>1b,5</sup> and Niklas Boers<sup>1b,2,3,6,†</sup>

<sup>1</sup>*Ludwig-Maximilians-University, Theresienstr. 37, 80333 Munich, Germany<sup>‡,§</sup>*

<sup>2</sup>*Technical University of Munich, Germany; Munich Climate Center;  
TUM School of Engineering and Design, Department of Aerospace and Geodesy,  
Earth System Modelling Group*

<sup>3</sup>*Potsdam Institute for Climate Impact Research, Telegrafenberg A 31, Potsdam, 14473, Germany<sup>§</sup>*

<sup>4</sup>*Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland*

<sup>5</sup>*Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, United Kingdom*

<sup>6</sup>*Department of Mathematics and Global Systems Institute, University of Exeter,  
North Park Road, Exeter, EX4 4QE, United Kingdom*



(Received 14 October 2024; accepted 16 March 2026; published 7 April 2026)

Models of complex dynamical systems like the Earth's climate often involve large numbers of uncertain parameters. Comprehensive exploration of the parameter space is typically prohibitive due to excessive computational costs, and systematic gradient-based parameter optimization is not feasible because such models are typically not differentiable. This is especially problematic in cases where the models intend to describe highly nonlinear and possibly abrupt dynamics, where sensitivity to parameter changes is high. Components of Earth's climate system, such as the North Atlantic Overturning Circulation, the polar ice sheets, or the Amazon rainforest, are at risk of undergoing critical transitions in response to anthropogenic climate change. However, estimates of the critical forcing thresholds are highly uncertain because the parameter spaces of complex climate models cannot be fully explored. Concerns have been raised that the above Earth system components are too stable in state-of-the-art models. Here, we introduce a method for efficient, systematic, and objective calibration of dynamical complex system models, targeted at adjusting system stability. Given a number of physical or observational constraints, our method moves the system in a direction where the system loses or gains stability, guided by indicators of “critical slowing down”. In contrast to a brute force approach, where the computational cost would exponentially increase with the number of parameters, our method scales polynomially and thus evades the curse of dimensionality. We successfully apply our method to a conceptual double-fold bifurcation model and a physically plausible reduced-order model of the global ocean circulation. Our method can efficiently adjust stability biases across a range of complex system models, helping to reveal hidden instabilities and resulting state transitions they induce. In particular, it can help us to explore and improve the representation of key multistable components of the Earth system in climate models.

DOI: [10.1103/624z-1p1s](https://doi.org/10.1103/624z-1p1s)

Subject Areas: Computational Physics,  
Interdisciplinary Physics,  
Nonlinear Dynamics

## I. INTRODUCTION

Numerical models of complex systems such as the Earth's climate are difficult to control because they are computationally expensive to run and depend on many free or at least highly uncertain parameters. Given the nonlinear dynamics typically represented by such models, they can respond to changes in these parameters in drastic and unforeseen ways. Prominent examples are global climate models, ranging from reduced-order box models with a few coupled equations and Earth system models of intermediate complexity (EMICs, [1]) to general circulation models (GCMs), which explicitly simulate 3D fluid dynamics in

\*Contact author: [sebastian.bathiany@tum.de](mailto:sebastian.bathiany@tum.de)

†Contact author: [n.boers@tum.de](mailto:n.boers@tum.de)

‡Also at Technical University of Munich, Germany; Munich Climate Center; TUM School of Engineering and Design, Department of Aerospace and Geodesy, Earth System Modelling Group.

§These authors contributed equally to this work.

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

atmosphere and oceans, and complex Earth system models (ESMs), which are used in the coupled model intercomparison Project CMIP [2] and typically include biogeochemical cycles [3,4]. ESM components and other geoscientific models that capture individual components of the Earth system (such as ocean-only models, terrestrial vegetation models, and ice sheet models) are often also applied in an “offline” (standalone) fashion and feature similar challenges.

Because of their complexity, and the limited availability of observations, such models involve unavoidable and often unquantifiable uncertainties of structural and parametric origin [5,6]. In particular, ESMs exhibit a large number, at the order of hundreds, of free parameters, e.g., from representing sub-grid-scale processes that cannot be explicitly resolved, such as turbulent mixing, cloud formation, or biogeochemical phenomena. Even climate models of reduced complexity [1] are still too complex for a systematic assessment of their full behavior or comprehensive exploration of parameter spaces. On the other hand, low-complexity conceptual models [7] can lack consistency with more realistic models.

The parametric uncertainty of global climate models is particularly problematic in the context of so-called tipping points. These are critical thresholds in forcing where a component of the climate system can respond abruptly and reorganize into another state, potentially in an irreversible way [8–11]. Such catastrophic phenomena can in some cases be associated with bifurcations in the underlying dynamics [12].

The most prominent elements of the Earth system that have been suspected to be able to show such tipping behavior under anthropogenic forcing are the Atlantic meridional overturning circulation (AMOC), the ice sheets on Greenland and Antarctica, and the Amazon rainforest [8,10,11]. However, the associated critical forcing levels remain unconstrained because transitions in any of the above systems in response to anthropogenic forcing would be unprecedented and because the uncertainties in the few climate model simulations targeted at quantifying these thresholds remain large. Current models do not show robust agreement on such events in future projections [13,14]. Moreover, there are reasons to believe that the above systems are in general too stable in state-of-the-art models compared to their real-world counterparts [15,16]. This implies that undesired surprises, which even the most sophisticated and comprehensive ESM projections cannot warn of, could occur in the (near) future. Methods to address existing stability biases in climate models and to identify plausible worst-case scenarios are, therefore, urgently needed.

A well-established framework for diagnosing stability changes from time series, e.g. from observations, relies on the phenomenon of critical slowing down (CSD) [11,17,18]. When the linear stability of a stable equilibrium

state is reduced and finally lost, e.g., when reaching a bifurcation, the return time to equilibrium increases, and (assuming a continuous system at a fixed point) the largest negative eigenvalue of the Jacobian for small perturbations increases toward 0 (toward 1 in the discretized representation of the system). In addition, if the system is permanently perturbed by stationary noise, the autocorrelation of its state increases toward 1 [18,19]. CSD-based indicators have been found to increase in observations of a number of suspected Earth system tipping elements such as the Greenland ice sheet [20], the AMOC [21], the Amazon rainforest [22,23], or the South American monsoon [24]; see also [11] for a recent review. CSD indicators have also been demonstrated to work in high-dimensional climate models [19,25,26], and have been successfully used to identify the perturbation pattern to which a system is least resilient [27–29]. Although CSD-based indicators can diagnose a change in stability over time, they cannot quantify stability in an absolute sense, or quantify how far in parameter space, or even in time, the system is from tipping [30].

In this study, we suggest a new CSD-based method to efficiently change the stability of a system in a given model via automatic model parameter updates, with the purpose of adjusting potential stability biases or revealing tipping risks in complex system models. Our approach resembles concepts in control theory [31,32], but with the aim of altering the stability of the system, rather than keeping it in a functioning regime. Our approach is also related to inverse problems, which seek parameter values that make the model’s output more consistent with observations. In our context, however, we aim to identify parameter values that only change the system’s stability (a property that cannot be directly observed), while keeping the observables constant.

Recently, atmospheric models that are purely based on machine learning [33–35] and a differentiable hybrid model [36] have been developed, but such models struggle to show a physically realistic response to forcing [37]. More traditional climate models are more physics-based but are not differentiable, i.e. gradients of model output with respect to parameter changes cannot be directly computed, which hinders systematic and objective parameter optimization [38,39]. Therefore, these models have typically been hand-tuned based on subjective expert judgment, with the aim of approximately matching observed features, such as the global radiative balance or a target climate sensitivity [40–42]. Although we generally strongly advocate making climate models differentiable [38] to enable efficient and transparent parameter optimization with respect to observations, differentiability will be less helpful in adjusting the stability of systems in a given model. We are interested in a targeted dynamical calibration, modifying parameters that specifically affect the linear stability of an equilibrium state while keeping observables

and emergent properties unchanged. This could not easily be achieved by differentiable programming or by emulating the numerical model in question with a differentiable machine learning model. In the adjoint-based optimization of differentiable models, where stability metrics would be differentiated through long, potentially chaotic integrations, gradients become ill-conditioned when computed over chaotic dynamics [43], and especially near bifurcations where relaxation times increase dramatically. Our approach avoids many of the problems related to ill-conditioned gradients of either natively differentiable models or (differentiable) machine learning emulators of nondifferentiable models.

Derivative-free calibration methods like simplex methods [44] or ensemble Kalman inversion [45] can also solve inverse problems by generating a perturbed-physics ensemble [46], comparing the output to desired (typically observed) values, and then iterating the ensemble of model parameters. This strategy can involve so-called emulators, which aim to capture the model’s parameter-to-output relationship, but can be evaluated (“sampled”) in a much cheaper way [47,48]. Related Bayesian approaches even aim to provide an uncertainty distribution on the parameters in question [49,50]. A different line of work is to train a machine learning emulator on a sparse number of parameters and construct the bifurcation diagram by solving the characteristic polynomial, leveraging the differentiability of the trained emulator [51]. However, running even small ensemble simulations can still be computationally costly. Moreover, emulation may be challenging for very nonlinear features like tipping events, and in situations when one needs to extrapolate into unexplored regions of parameter space. Even if such approaches were technically successful, a probabilistic interpretation of parameter values or model outcomes regarding tipping points in the real world is out of reach due to the structural uncertainty of climate models [5,6]. The problem of targeted stability tuning is complementary to the emulator approaches mentioned above. The task at hand is to efficiently identify unstable regions in the model’s phase space. Current emulation approaches such as Gaussian processes would, relying on reasonably smooth functions, not be suitable in this context. Moreover, Latin hypercube sampling (alone) would sample the relevant parameter regions too sparsely.

Here, we present a CSD-based method to systematically and objectively adjust the stability of a given state of a simulated system. The method works sequentially, without the need for costly ensemble simulations. The main types of models that we target here are relatively fast ones, allowing simulations that span several thousand times the characteristic timescale of the nonlinear system in question. Also, our method tends to work best when the model does not feature internal (chaotic) variability. This potentially includes any Earth system model component except general circulation models (GCMs), Earth system models of

intermediate complexity (EMICs), lower complexity models that serve as emulators (simplified versions) of complex ESMs (e.g., box models, conceptual models, or data-driven machine-learning models). Ideally, these models are dynamical system simulators that share properties with state-of-the-art ESMs, e.g., box models calibrated to specific target ESMs [52–54], or subcomponents taken out of more comprehensive models [55], which will allow conclusions even for comprehensive ESMs.

Our method considers dynamics on the combination of a model’s phase and parameter space. As described in detail in the following sections, we create a feedback loop between the local stability of the system’s state and its parameter values. Based on a certain initial state of the system and its parameters, our method determines the direction in parameter space where the system most effectively loses (or gains) stability under certain observational or physical constraints. Our general approach to this problem is to find a local parameter-dependent autoregressive model which effectively describes the dynamics of an observable close to the stable state of interest, and then use this model to find a new parameter combination which increases or decreases the stability of the given equilibrium state. Using the CSD phenomenon makes our approach for targeted and objective model (re-)calibration highly efficient; as we will show below, our method scales polynomially with the numbers of parameters, whereas a brute-force perturbed-parameter ensemble approach would scale exponentially (the “curse of dimensionality”).

We describe our method in its most general way in the following section, and then apply it to two example systems featuring multistability, demonstrate its efficiency, and discuss the implications of our results. Our study is accompanied by public code that can also be applied to other systems (see below).

## II. TARGETED MODEL CALIBRATION TO ADJUST SYSTEM STABILITY

We consider a general dynamical system that is discrete in time, which essentially covers all numerical models of dynamical systems, including climate and Earth system models. We denote the dynamic variables (state variables) of the system by  $\vec{x} \in \mathbb{R}^{d_x}$  and the parameters by  $\vec{p} \in \mathbb{R}^{d_p}$ . Moreover, we consider observables  $\vec{o} \in \mathbb{R}^{d_o}$ . We assume that we have access to a function  $f_o: \mathbb{R}^{d_x+d_p} \rightarrow \mathbb{R}^{d_o}$ ,  $(\vec{x}, \vec{p}) \mapsto \vec{o}$  mapping the state variables to the parameter-dependent observables ( $\vec{o}_t$ ). As an example, the system could be the global ocean circulation, and the observable could be the mass flux across a certain latitude in the North Atlantic (AMOC strength).

The dynamics of the system is given in the form of an evolution function

$$f_p: \mathbb{R}^{d_x+d_p} \rightarrow \mathbb{R}^{d_x}, \quad (\vec{x}_t, \vec{p}) \mapsto \vec{x}_{t+1},$$

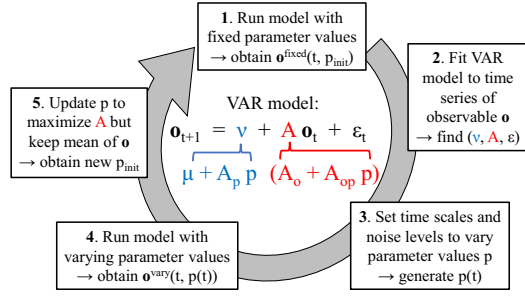


FIG. 1. Main steps of our calibration method to automatically update parameter values that change the stability of a simulated system in a targeted way.  $o$  are observables constructed from the output of the complex model,  $p$  are the model parameters to be calibrated (in general, both are vectors). A vector-autoregressive model (VAR) is fitted to the data with the aim of varying  $p$  such that it maximizes the largest eigenvalue of matrix  $A$  (red), while keeping the time mean of  $o$  constant. The nature of vector and matrix multiplications are not specified here; for more detail on each step see Sec. II.

which defines the one-step ahead propagation of the system state  $\vec{x}_t$  to  $\vec{x}_{t+1}$  and allows us to integrate it. It is not necessary to know the exact expression of  $f_p$ ; we only need the ability to run parameter-dependent simulations of the system. We assume that the system's state fluctuates around a dynamic equilibrium which has a certain local stability, defined in terms of the linear restoring rate. The target is to adjust the system's parameters in a way that changes this local stability.

Our method consists of the following iterative steps (Fig. 1), which all run fully automatically once a system has been set up. For the sake of clarity, we focus on the case of destabilizing a given system; adjustments to the case of increasing stability are straightforward.

- (1) Generate a trajectory of observables ( $\vec{o}_t$ ) of length  $T$  with parameters fixed to their initial values  $\vec{p}_{\text{init}}$ . To this end, we integrate the system to obtain a trajectory of the system variables ( $\vec{x}_t$ ), to which we apply  $f_o$ . During the integration process we force the underlying system with small additive white noise  $\vec{u}_t^{\text{fixed}} \in \mathbb{R}^{d_x}$  (“fixed” because the parameters are fixed here) in order to drive the system out of equilibrium and control the scale of the region in phase space used to compute the Jacobian at equilibrium.

$$\vec{x}_{t+1} = f_p(\vec{x}_t, \vec{p}_{\text{init}}) + \vec{u}_t^{\text{fixed}} \quad (1)$$

$$\vec{o}_t = f_o(\vec{x}_t, \vec{p}_{\text{init}}) \quad (2)$$

It generally suffices to use state noise levels that are so small that they do not affect the system's general state and its dynamics. The only purpose is to be able to sample the system's linear stability around a fixed point.

- (2) Fit a vector-autoregressive model [VAR(1) model] of the form

$$\vec{o}_{t+1} = \vec{v} + A\vec{o}_t + \vec{\epsilon}_t^{\text{fixed}} \quad (3)$$

to the data from step 1, including uncertainty ranges (see Appendix A for theoretical context on this approach), where  $\vec{v}$  is a constant baseline state, the square matrix  $A$  contains constant parameters of the autoregressive model, and  $\vec{\epsilon}_t^{\text{fixed}}$  are the residuals of the fit (resulting both from the imposed noise and the potential model error). Since the equilibrium is stable, all eigenvalues of  $A$  should be smaller than 1 and the stability of the system is characterized by the largest eigenvalue  $\lambda_{\text{max}}$ .

- (3) Vary the parameter values. To this end, we use univariate autoregressive [AR(1)] processes of the form

$$p_{i,t+1} = (1 - \lambda_p)p_{i,\text{init}} + \lambda_p p_{i,t} + \sigma_i w_{i,t} \quad (4)$$

to generate a time series for each parameter  $p_i$ , fluctuating around  $p_{i,\text{init}}$ , where  $(\vec{w}_{i,t})$  describes white noise with  $w_{i,t} \sim \mathcal{N}(0, 1)$ . Here, we choose the persistence parameter of the AR(1) process,  $\lambda_p$ , such that  $\lambda_p \leq \lambda_{\text{max}}$ . In this way, the parameters vary on a timescale that is longer than the timescale of the system, expressed as the slowest mode of the observable dynamics. This is necessary in order to record the response of the system to the parameter changes. We set the scalar  $\lambda_p$  to  $\lambda_{\text{max}}$  plus two standard deviations from  $\lambda_{\text{max}}$  to ensure this condition. In Appendix B, we discuss how to find appropriate noise amplitudes  $\sigma_i$  and in Sec. III we discuss how we implemented the method based on the model at hand.

- (4) Further integrate the system, forcing it with the time-dependent parameter values:

$$\vec{x}_{t+1}^{\text{vary}} = f_p(\vec{x}_t^{\text{vary}}, \vec{p}_t) + \vec{u}_t^{\text{vary}} \quad (5)$$

$$\vec{o}_t^{\text{vary}} = f_o(\vec{x}_t^{\text{vary}}, \vec{p}_t) \quad (6)$$

Our aim is to use these simulations to understand how parameter values influence the stability—and, consequently, the VAR parameters—of the system.

Our approach is to fit a VAR(1) model of the form

$$\vec{o}_{t+1}^{\text{vary}} = \vec{\mu} + A_o \vec{o}_t^{\text{vary}} + A_p \vec{p}_t + A_{op} (\vec{o}_t^{\text{vary}} \otimes \vec{p}_t) + \vec{\epsilon}_t^{\text{vary}} \quad (7)$$

with  $A_o \in \mathbb{R}^{d_o \times d_o}$ ,  $A_p \in \mathbb{R}^{d_o \times d_p}$ ,  $A_{op} \in \mathbb{R}^{d_o \times (d_o \cdot d_p)}$  and the Kronecker product

$$\vec{o} \otimes \vec{p} \equiv (o_1 p_1, o_1 p_2, \dots, o_1 p_{d_p}, \dots, o_{d_o} p_1, \dots, o_{d_o} p_{d_p})^T \in \mathbb{R}^{d_o d_p}, \quad (8)$$

where  $d_o$  is the number of observables and  $d_p$  the number of parameters.

In the above, the second-order terms (mixed terms involving  $\vec{o}$  and  $\vec{p}$ ) are needed to detect the  $p$  dependence of the eigenvalues [represented by matrix  $A$  in Eq. (3)]. We note that the quantification of these terms can become highly uncertain if the first order effects are much more prominent than the second order effects. In this case, it would be possible to shift contributions between  $\vec{\mu}$ ,  $A_o$ ,  $A_p$  and  $A_{op}$  in Eq. (7), i.e., we would be faced with an underdetermined problem. We solve this potential problem by exploiting our knowledge of the state noise (see Appendix C).

- (5) Based on the estimated model from Eq. (7), we choose new parameter values. Our goal here is to maximize the largest eigenvalue of  $A_o + A_{op}(\mathbb{I}_{d_o} \otimes \vec{p})$  [where we used Eq. (C10), and where  $\mathbb{I}_{d_o}$  is the identity matrix of size  $d_o \times d_o$  and  $\otimes$  is the Kronecker product is defined in Eq. (8)]. We do not allow the eigenvalue to exceed 1, as we only aim at reducing the stability of the fixed point, and do not intend to make it fully unstable. Other calibration targets are possible in different contexts.

We also demand that the updated parameter values should not differ too much from the previous ones, as the VAR model is only valid in a vicinity of these values. In special cases, more efficient solutions are possible (e.g., the double-well potential in Sec. III A). In addition, we enforce a preservation of the mean values of the observables, which turns Eq. (7) into

$$\vec{o} \doteq \vec{\mu} + A_o \vec{o} + A_p \vec{p} + A_{op}(\vec{o} \otimes \vec{p}) \quad (9)$$

with  $\vec{o}$  denoting the initial equilibrium values of the observables. This condition is motivated by the fact that the mean state (for example, of the climate system) is typically constrained by observations much better than model parameters or the stability of the state. Appendix D provides details on how we use the above conditions to determine parameter updates. Model-specific details in applications of our method are discussed in the following section.

- (6) After finding a new set of parameters, we replace  $\vec{p}_{\text{init}}$  by the new parameter values and run the system until it equilibrates. After that, the whole process is repeated, now using the new parameter values.

### III. APPLICATION TO TWO EXAMPLE SYSTEMS

In order to test our method, we apply it to a simple double-well dynamical system, as well as to a conceptual, physically plausible model of the global ocean circulation [52] (see Appendix E).

#### A. Double-well system

We consider a simple double-well system with dynamics determined by the ODE

$$\dot{x} = p_1 x^3 + p_2 x^2 + p_3 x + p_4. \quad (10)$$

For simplicity, we choose the observable  $o$  to be  $x$  itself, i.e.,  $f_o(x, \vec{p}) = x$ . We use the following initial parameter values:

$$\vec{p}_{\text{init}} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}. \quad (11)$$

With this choice the system has two stable fixed points at  $x = 1$  and  $x = -1$  and one unstable fixed point at  $x = 0$ . Each fixed point has its own basin of attraction and the two basins are separated by the unstable fixed point. We aim to destabilize the fixed point at  $x = 1$  with regard to small perturbations in the  $x$  direction while maintaining the position of the fixed point at  $x = 1$ .

As state noise  $u_t$ , we add Gaussian white noise with a noise level of  $10^{-4}$  to the right-hand side of Eq. (10). We integrate the resulting stochastic differential equation using the Euler-Maruyama scheme [56], where we choose a time step of  $\Delta t = 10^{-3}$ .

We run the destabilization process for 18 iterations since further iteration would typically lead to a stochastic escape into the alternative basin of attraction. The noise amplitudes used for the parameter variation are updated in each iteration step (see Sec. II, step 3, and Appendix B). We calibrate these by first setting all parameter noise amplitudes  $\sigma_i$  simultaneously to  $10^{-3} \sqrt{1 - \lambda_p^2}$  and using these to compute varying parameter trajectories via Eq. (4). The system, forced by the varying parameters, is then integrated for 100 time steps (this number must be chosen large enough to allow the distribution of the observable to adjust to the new parameter noise). This is repeated iteratively whilst increasing the noise amplitudes in each iteration by a factor of 2, up to the point where the standard deviation of  $x$  exceeds a value of  $10^{-3}$ . The noise amplitudes for generating the parameter series are then set to the values prior to the termination condition. To ensure that the VAR model actually shows a dependence on the parameter variations, we check whether the coefficients  $A_p$  and  $A_{op}$  of the parameter-dependent VAR model differ from zero (using

the estimated errors of the coefficients, see Appendices A and B). This turns out to always be the case. With this method we typically find noise amplitudes corresponding to a standard deviation of the parameters (given by  $\sigma_i/\sqrt{1-\lambda_p^2}$  for parameter  $p_i$ ) greater or close to 0.05.

We estimate the parameter-dependent VAR model using 1000 time steps for the fixed parameter model and 100 time steps for the parameter dependent part. The parameters are then updated at the end of each iteration step by maximizing  $A_o + A_{op}\vec{p}$ , which is a scalar in this system and hence already the eigenvalue we want to increase via the parameter change. The constraint of preserving the mean value  $\bar{x} = 1$  [see Appendix D and Eq. (9)] then reads

$$1 \stackrel{!}{=} \mu + A_o + (A_p + A_{op})\vec{p}. \quad (12)$$

The system can now be destabilized under constraints as described in more detail in Appendix D.

Since the system is relatively simple, an even more efficient variant also works here. The difference to the more general approach described in Appendix D is that the constraint can be strictly imposed here instead of using a loss function. To this end, we choose a parameter update  $\Delta p$  of the form  $\Delta p = \Delta \vec{p}_\perp + L \Delta \vec{p}_\parallel$  with

$$\Delta \vec{p}_\perp = \frac{1 - \mu - A_o - \vec{n} \cdot \vec{p}_{\text{init}}}{|\vec{n}|^2} \vec{n}, \quad (13)$$

$$\Delta \vec{p}_\parallel = \left[ A_{op} - \frac{A_{op} \cdot \vec{n}}{|\vec{n}|^2} \vec{n} \right], \quad (14)$$

where  $L \in \mathbb{R}$ .

Note that in the double-well system,  $A_o$  and  $\mu$  are scalars, and  $A_{op}$ ,  $A_p$ ,  $\vec{p}$ , and  $\vec{n}$  are all vectors of dimension 4, where  $\vec{n} = A_p + A_{op}$  is the normal vector to the hyperplane  $\mathcal{H}$  spanned by all possible  $\vec{p}$  that satisfy Eq. (12). To be consistent, we here only show arrows above vectors that are vectors in any application, not only the double-well system. Equations (13) and (14) are the solution to the optimization problem where  $A_o + A_{op} \cdot \vec{p}$  (eigenvalue) must be maximized, and  $(A_p + A_{op}) \cdot \vec{p}$  must remain constant (constraint  $\bar{x} = 1$ ).

$\Delta \vec{p}_\perp$  is the component perpendicular to the hyperplane  $\mathcal{H}$  (and parallel to  $\vec{n}$ ). By adding  $\Delta \vec{p}_\perp$  to  $\vec{p}_{\text{init}}$ , we move  $\vec{p}$  onto  $\mathcal{H}$  in an orthogonal manner, in order to always keep  $\bar{x}$  at 1.  $\Delta \vec{p}_\parallel$  is the component parallel to  $\mathcal{H}$ . Adding a multiple of  $\Delta \vec{p}_\parallel$  to our parameter vector thus destabilizes the system without leaving  $\mathcal{H}$ .

By then choosing

$$L = \min \left( \frac{0.05}{|\Delta p_{\parallel,1}|}, \dots, \frac{0.05}{|\Delta p_{\parallel,4}|}, \frac{1 - A_o - A_{op}(\vec{p}_{\text{init}} + \Delta \vec{p}_\perp)}{2 \cdot A_{op} \vec{p}_\parallel} \right) \quad (15)$$

we assure by the first four arguments of the  $\min(\dots)$  function—assuming that  $|\Delta \vec{p}_\perp|$  is comparably small—that the parameter changes are not too large compared to the region explored in parameter space. The last argument of the  $\min(\dots)$  function preserves the stability of the equilibrium at  $x = 1$  by not pushing the value of  $A_o + A_{op}\vec{p}$  past 1. To this end, we compute the value that  $L$  would have to take in order for  $A_o + A_{op}(\vec{p}_{\text{init}} + \Delta \vec{p})$  to be equal to 1 and use this value, divided by a factor of 2.

The evolution of the updated parameters at the end of each iteration step is shown in Fig. 2(a) together with the evolution of the Jacobian (eigenvalue)  $\lambda$ , which indicates the stability of the equilibrium. Here,  $\lambda$  moving closer to 1 from below corresponds to a stability reduction of the system. The average value  $\bar{x}$  of  $x$  computed from a time series (obtained by integrating the stochastically forced

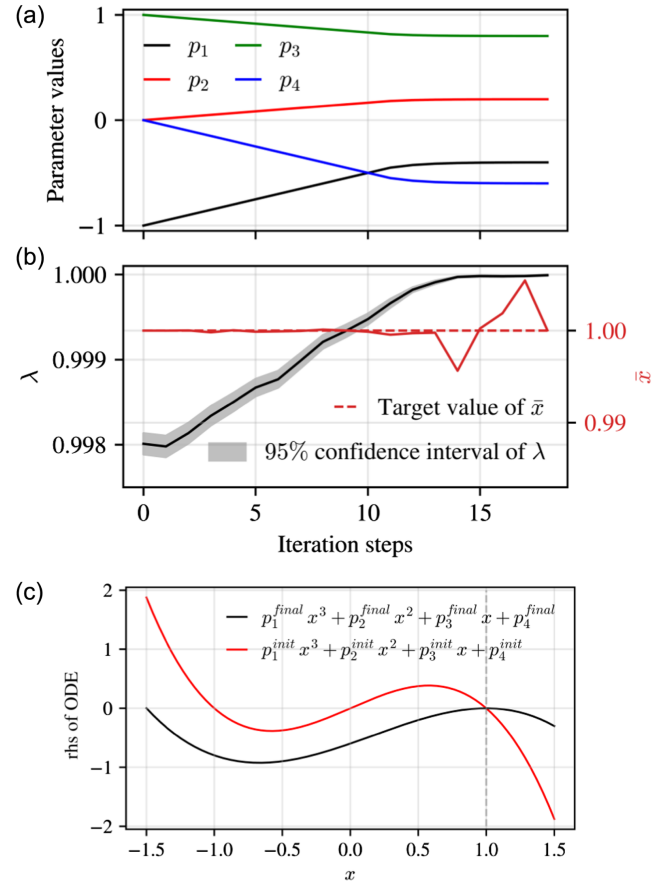


FIG. 2. Application of the destabilization method to the double-well system. (a) Evolution of the parameters  $p_1, \dots, p_4$ , and (b) evolution of the Jacobian  $\lambda$  and the mean value  $\bar{x}$  of  $x$  during the iterative destabilization of the double-well system; note the small y-axis range for the latter. (c) The right-hand side (rhs) of Eq. (10) for the parameters before (red) and after (black) the destabilization. By design, observable  $x$  is constrained to stay at  $x = 1$  (gray dotted line). The original stable state (negative slope of the red line) merges with an unstable one in a saddle-node bifurcation when the system is fully destabilized.

system) consisting of  $10^5$  data points, using the parameter setting of the respective iteration step, can be successfully stabilized by our method [Fig. 2(b)].

We notice that for the final iteration steps, both  $\lambda$  and the parameter values start to converge to prevent the equilibrium from losing its stability entirely. This is an optional feature we explicitly implemented into the parameter update scheme by adding the last argument to the  $\min(\dots)$  function in Eq. (15). Another striking aspect of Fig. 2 is that although the average value  $\bar{x}$  of  $x$  remains very close to its initial value in each iteration, there is a noticeable increase in fluctuations away from this value as  $\lambda$  approaches 1. The reason for this phenomenon likely lies in the fact that we did not change the amplitude of the noise  $u_t^{(x)}$  added to the system; so as the stability of the fixed point decreases, the standard deviation of  $x$  and therefore also of  $\bar{x}$  increases [18].

The effect of the destabilization process on the rhs of Eq. (10) is visualized in Fig. 2(c). The equilibrium at  $x = 1$  is significantly destabilized while maintaining its original position. We note that the parameter change induced by our calibration method has shifted the other stable equilibrium, initially located at  $x = -1$ , to smaller values of  $x$ , while simultaneously increasing the size of the basin of attraction of that fixed point.

### B. AMOC five-box model

In order to test our method on a more complex process-based system, we apply it to a recently proposed five-box model of the overturning circulation of the global oceans [52]. The model consists of five coupled differential equations describing the dynamics of the salinities in five boxes, where each box represents a water mass prevailing in a specific region of the Earth's oceans (Fig. 3). We made some slight modifications to the model regarding the

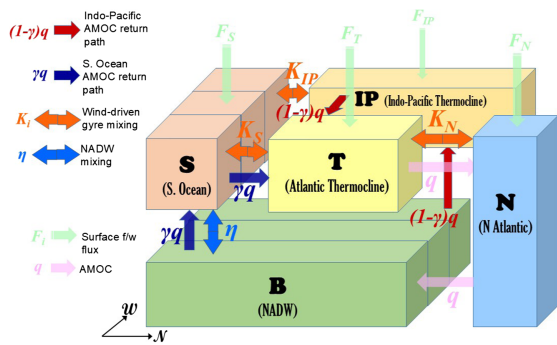


FIG. 3. Schematic illustration of the five-box ocean model (adapted from [52]). The five boxes represent major ocean basins, and arrows represent freshwater fluxes between them. The variable  $q$  (pink arrows) is the property we consider as the observable in this study. Freshwater hosing is applied by modifying the green surface fluxes  $F_i$ . The five parameters shown in strong colors ( $\gamma$ ,  $\eta$ ,  $K_N$ ,  $K_S$ , and  $K_{IP}$ ) are less well constrained by observations and are the parameters we calibrate.

conservation of total water mass (see Appendix E), which only have a negligible influence on the dynamics of the system. Here, the salinities  $\vec{S} = (S_N, S_T, S_S, S_{IP}, S_B)^T$  are the system variables called  $\vec{x}$  in the previous section.

We select the AMOC strength  $q$  as our observable  $\vec{o}$ . We vary the parameters  $\gamma$  (the relative strength of a circulation branch involving the Southern Ocean),  $\eta$  (representing mixing of North Atlantic deep water with fresher waters),  $K_N$ ,  $K_S$ , and  $K_{IP}$  (representing diffusive fluxes associated with the gyre strengths in the North Atlantic, Southern Ocean, and Indo-Pacific Ocean), as these five parameters have the highest uncertainties. The other parameters of the model are kept fixed at their original values.

We discretize the system with  $\Delta t = 0.1$  yr. The initial parameter values are chosen according to the FAMOUS<sub>A</sub> model as stated by [52]. We choose the additive noise driving the system so that the overall amount of salt in the system is conserved. To this end, we add a noise vector of the form

$$\vec{u}_t^{(x)} = \begin{pmatrix} u_t^{(1)}/V_N \\ u_t^{(2)}/V_T \\ u_t^{(3)}/V_S \\ u_t^{(4)}/V_{IP} \\ -(u_t^{(1)} + u_t^{(2)} + u_t^{(3)} + u_t^{(4)})/V_B \end{pmatrix} \quad (16)$$

to the dynamic equations updating the salinity concentrations  $\vec{S}_t$ , where each  $u_t^{(i)}$  is a different realization of Gaussian white noise with noise level  $10^{-4}$ . We then run the model for 500 iterations using  $10^6$  time steps (i.e.,  $10^5$  years) in each iteration, for estimating the VAR model with fixed parameters (method step 2 above) as well as for finding the parameter dependencies of the model (step 4) in each iteration. Also, before starting the VAR estimation, we integrate the model for a transient time of  $10^5$  time steps (10000 years) in each iteration to allow the model to equilibrate with the new parameter values. Because of their physical interpretation, the parameters are restricted to positive values and  $\gamma$  must additionally fulfill  $\gamma \leq 1$  (although it turns out that imposing this condition is not necessary since the destabilization process decreases  $\gamma$ ).

When adjusting the noise amplitudes  $\sigma_i$ , which are needed to generate the parameter series [Eq. (4)], we calibrate each parameter separately (see Appendix B). This helps determine the influence of any single parameter on the system. To this end, we fix all parameters, except one, to their initial values and iteratively increase the noise amplitude  $\sigma_i$  of the parameter of interest by powers of 2, starting from  $\sigma_i = \min(10^{-5}, p_{\text{init},i}/4 \cdot \sqrt{1 - \lambda_p^2})$ . The minimum function and factor  $p_{\text{init},i}/4$  allow the variations to also be even smaller than  $10^{-5}$  should any component of

$\vec{p}_{\text{init}}$  be close to zero. For each parameter noise amplitude, we integrate the system for  $10^5$  time steps while forcing it only by the variations of  $p_i$ . As outlined in Appendix B, the condition for stopping the iteration is when the observable's variance exceeds a threshold [here  $\text{VAR}(q) < 0.01$ ], while the VAR model parameters differ from 0. Of all noise amplitudes that meet these conditions, we select the combination of noise amplitudes (one for each parameter) for which the corresponding values of  $\text{VAR}(q)$  do not differ by more than a factor of 4 from each other. This guarantees that the variances of the observables lie on comparable scales for each separate parameter variation. At the same time, the individual noise amplitudes are maximized to maximize the region explored in parameter space. To keep computational costs minimal, we only update the noise amplitudes every 10 iteration steps. This procedure yields noise levels in the range  $\sigma_i / \sqrt{1 - \lambda_p^2} \in [5, 10]$ .

After applying steps 1 to 5 from Sec. II to find a model of the form

$$q_{t+1} = \mu + A_o q_t + A_p \vec{p} + q_t A_{op} \vec{p}, \quad (17)$$

we now systematically reduce the stability of the strong AMOC state in the five-box model under a number of physical constraints, most importantly that  $\Delta \vec{p}$  always stays on the same hyperplane. The implementation of these constraints is explained in Appendix D.

We here set the length  $l$  of the parameter update vector  $\Delta \vec{p}$  to 100 and use

$$l_i = \min \left( \frac{\sigma_i}{\sqrt{1 - \lambda_p^2}}, \frac{p_{\text{init},i}}{100} \right) \quad (18)$$

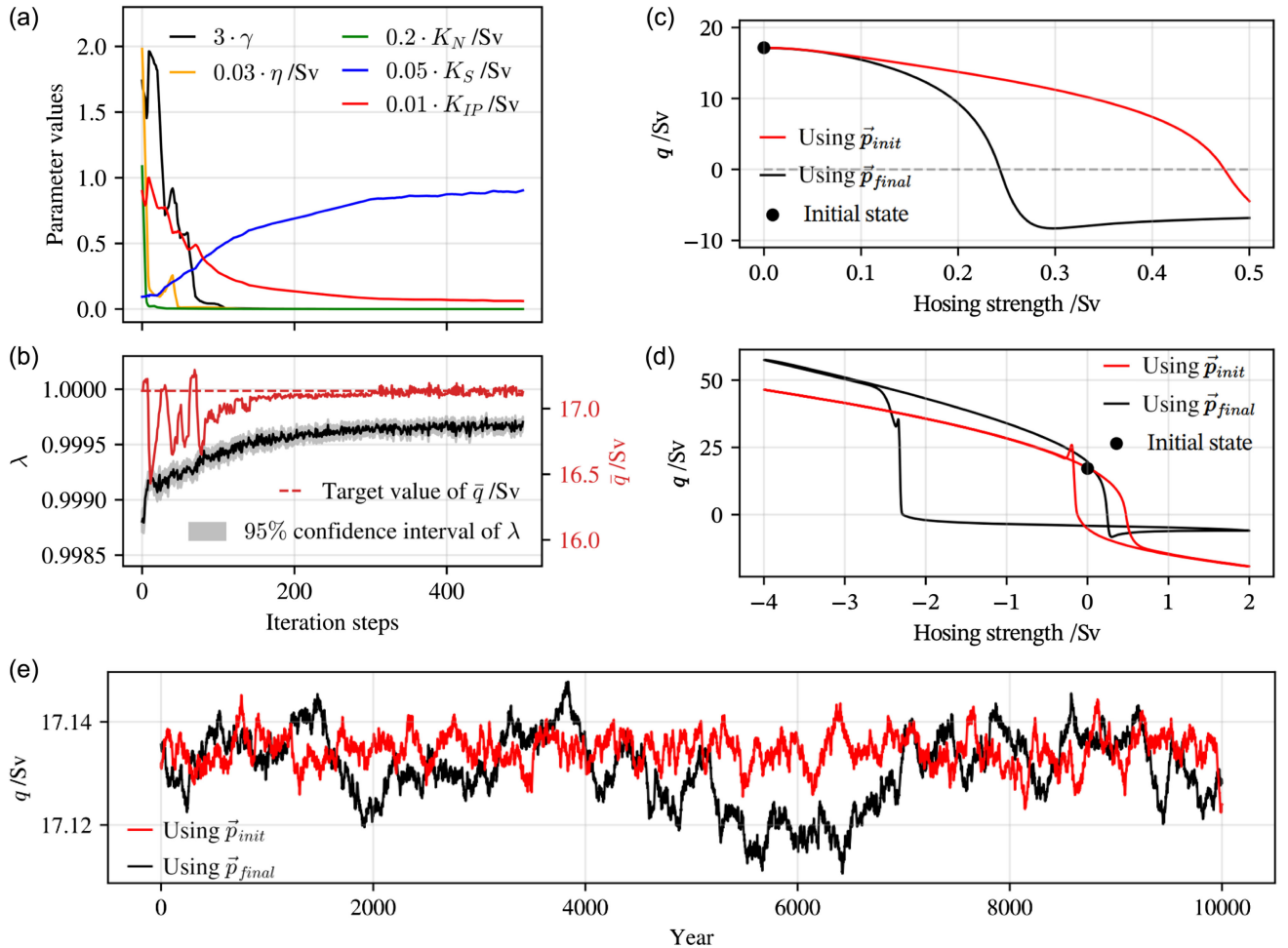


FIG. 4. Destabilization of the five-box AMOC model with (a) the evolution of the parameters  $p_1, \dots, p_5$ , as well as (b) the Jacobian  $\lambda$  and the mean value  $\bar{x}$  of  $x$  during the iterative destabilization process. Increasing  $\lambda$  toward 1 implies a successful destabilization of the system. (c) and (d) Hysteresis in the five-box AMOC model [52] before (red) and after (black) destabilization, when performing the hosing experiment as described in [52] to both systems. (c) Closeup of the hysteresis emphasizing the initial AMOC collapses due to the hosing. (d) Hysteresis curve over the full range of hosing. Note that the horizontal range of hysteresis is strongly increased by the parameter change. (e) 10 000 years of stationary time series when running the model with each parameter set, and without hosing ( $H = 0$ ).

with  $\sigma_i$  and  $\lambda_p$  from Eq. (4). With this choice, we guarantee that no parameter value becomes negative and that we stay sufficiently close to the region explored in parameter space. In more general cases, the limitation to positive parameters can be dropped by setting  $l_i$  to the first term in the  $\min()$  function [Eq. (D11)]. We set the penalty coefficients in the optimization [see Eq. (D9)] to  $\eta_1 = 10^{12}$ ,  $\eta_2 = 10^3$  and  $\eta_3 = 1$ . The very high penalty value  $\eta_1$  ensures that the mean state  $\bar{q}$  is well conserved.

The parameter values and the observable state converge and are close to their final values after a few hundred iterations [Figs. 4(a) and 4(b)]. From Fig. 4(a) we can also infer that the two model parameters  $K_N$  and  $\eta$  have a major influence on the stability of the system, as they both rapidly tend towards zero within the first few iteration steps whilst simultaneously causing a significant decrease in the stability of the system. In principle, this result confirms that our method is physically meaningful; for example, the gyre exchange coefficients  $K_i$  represent the sensitivity to wind stress, which is known to be an important factor for AMOC stability in ESMs [57].

These parameter values, however, may be physically implausible without further constraints. For example, fresh water exchange by the wind-driven gyre circulation, represented by  $K_N$ , is known to be a stabilizing factor, and observations and climate models show that this process is active in the North Atlantic [57,58], suggesting that the minimum allowable value of  $K_N$  should be set somewhat larger than zero. In addition, we have used only a single constraint on the destabilized equilibrium solution (the overturning strength  $q$ ) for simplicity. We find that our destabilized equilibrium solution has box-mean salinities that are inconsistent with observations. In particular, the box representing the Indo-Pacific oceans is too salty, while the boxes representing the North Atlantic and Southern Ocean are too fresh. Such additional observational constraints can, in principle, be added to our method, and would presumably result in a solution that is somewhat less unstable than the one we find here.

After destabilizing the five-box model, we verify that its sensitivity to parameter changes in the form of freshwater forcing (so-called hosing) has increased. To this end, we apply freshwater hosing as described in [52] to both the original and the destabilized system [Figs. 4(c) and 4(d)]. The effect of the parameter change induced by our method is similar to the one it had on the double-well system: When exposed to the same hosing, the destabilized model [black lines in Figs. 4(c) and 4(d)] reaches an alternative steady state at much smaller hosing than the model with the original parameter values (red lines). Moreover, the negative hosing (reversed forcing) that is required for the system to recover to its original state is much larger for the destabilized system, i.e., the regime with hysteresis has become wider [Fig. 4(d)]. Moreover, Fig. 4(e) shows that the AMOC time series in the reparametrized model (when

run without hosing), has almost the same mean as before, but much larger autocorrelation due to the successful destabilization. As expected from theory [18,59], and in similarity to the double-well system [Fig. 2(b)] the time series consequently also has larger variance even though the additive state noise was the same.

#### IV. ANALYZING THE COMPUTATIONAL COST OF THE METHOD

An important question regarding our calibration method for adjusting stability biases in complex system models is how well it performs in terms of computational cost when compared to a brute force parameter search. In particular, we are interested in how the computational cost scales with the number of parameters in our model.

Given a target stability of the dynamical system in question, the computational cost for finding a suitable parameter combination would scale exponentially with the number of adjustable parameters when applying a brute force parameter search. This manifestation of the so-called curse of dimensionality has so far prevented systematic calibration of climate or Earth system models, also because they are not differentiable. In the case of our method, the number of iterations needed to achieve the destabilization goal does not necessarily show any systematic dependency on the number of parameters, as the parameters evolve along a gradient of decreasing or increasing linear stability of the system. We therefore expect that the significant factor determining how the computational cost scales with the number of parameters will depend on how the length of the trajectory needed to estimate the parameter-dependent VAR model in each iteration step scales with the number of parameters. Since the length of such a trajectory would typically be independent of the number of parameters in a brute force setting, the main question is whether the length of the trajectory in an iteration step scales subexponentially.

We determine the scaling behavior of the trajectory length in the two example systems presented above. To this end, we consider each system with its respective initial parameter constellation. Then, in a first step, we only vary one parameter at a time, fixing all other parameter values. Using a trajectory of fixed length, we collect the errors in the coefficient  $A_{op}$  of the resulting VAR model for each separate parameter variation. This results in a vector  $\sigma_{p_i}^2 \in \mathbb{R}^{d_p}$ , where the  $i$ th entry corresponds to the variances of  $A_{op}$  in a VAR model where only the  $i$ th parameter is varied. The reason why we only focus on the error of  $A_{op}$ , as opposed to also considering further coefficients, is due to simplicity and the fact that  $A_{op}$  is the most relevant coefficient for parameter updates when purely aiming for a change in stability.

In a second step, we iterate through all possible parameter constellations, including at least two parameters. For each such parameter set, we iteratively increase the

trajectory length used to compute a VAR model in each iteration, and compare the variance of the sum of all entries of  $A_{op}$  (taking correlations into account) of the resulting VAR model to the sum of the respective entries in  $\sigma_{p_i}^2$ . As soon as the variance of the sum of the entries in  $A_{op}$  drops below the latter sum, the current trajectory length necessary to fulfill this condition is noted. The reason why we do not directly compare the variances of each coefficient in  $A_{op}$  to their counterpart in  $\sigma_{p_i}^2$  is that this can lead to extremely long trajectory lengths necessary to meet this condition, in the case that the entries in  $\sigma_{p_i}^2$  differ in magnitude. Thus, if the total number of parameters that can possibly be varied is given by  $d_p$  and the currently considered set of parameters consists of  $p$  parameters, this results in  $\binom{d_p}{p}$  trajectory lengths, corresponding to the case of  $p$  parameters being varied. By averaging all values corresponding to a given value of  $p$  and visualizing these data, we can try to infer the functional dependence of the trajectory length with respect to the number of parameters that are varied.

For each given parameter set, we always use the same noise amplitudes  $\sigma_i$  for a given parameter  $p_i$  regardless of the other parameters it is paired with. In the case of the double-well system we used  $0.05\sqrt{1-\lambda_p^2}$  (see Sec. III A) as noise amplitude for every parameter. For the five-box AMOC model we used  $(\sigma_\gamma, \sigma_\eta, \sigma_{K_N}, \sigma_{K_S}, \sigma_{K_{IP}}) = (0.020, 5.243, 0.328, 0.082, 2.621)\sqrt{1-\lambda_p^2}$ , where we determined the noise amplitudes using the procedure described in Sec. III B. As this choice of noise amplitudes for the five-box model led to quite long trajectory lengths necessary to meet the convergence condition, we also empirically modified the noise amplitudes to be

$(\tilde{\sigma}_\gamma, \tilde{\sigma}_\eta, \tilde{\sigma}_{K_N}, \tilde{\sigma}_{K_S}, \tilde{\sigma}_{K_{IP}}) = (0.020 \cdot 32, 5.243/32, 0.328/2, 0.082 \cdot 2, 2.621/8)\sqrt{1-\lambda_p^2}$ , which resulted in much shorter trajectory lengths. We will refer to the resulting model as the optimized box model system in the following.

When determining the values of  $\sigma_{p_i}^2$  we used VAR models computed from data series of length 6000 in the case of the double-well system and length 1000 for the two versions of the five-box model. In order to suppress stochastic effects, we repeated this procedure 100 times and took the final value of  $\sigma_{p_i}^2$  to be the average over all of these runs. For each parameter set including at least two parameters, we computed the trajectory length needed to fulfill the condition described above as the average over 20 such runs. To this end, we increased the trajectory length by 1000 in each iteration step, starting from 6000 for the double-well system and from 6000 for the optimized five-box model system). For the nonoptimized five-box model, we increased the trajectory length in the following fashion:  $10^3, 2 \times 10^3, 5 \times 10^3, 10^4, 2 \times 10^4, 5 \times 10^4, 10^5, \dots$  to save computation time.

The results of the analysis on the computational cost of the method are shown in Fig. 5. We notice that the data, displayed in a log-log plot, agree well with respective linear fits. One can also observe that the range of trajectory lengths, which we observe for parameter subsets with a fixed number of parameters (small data points), decreases as this number grows in all three cases.

## V. DISCUSSION

In both examples that we considered, namely, the simple double-well system and the five-box model of the global ocean circulation, our method is able to adjust the stability of the modeled system. Specifically, focusing on the task of

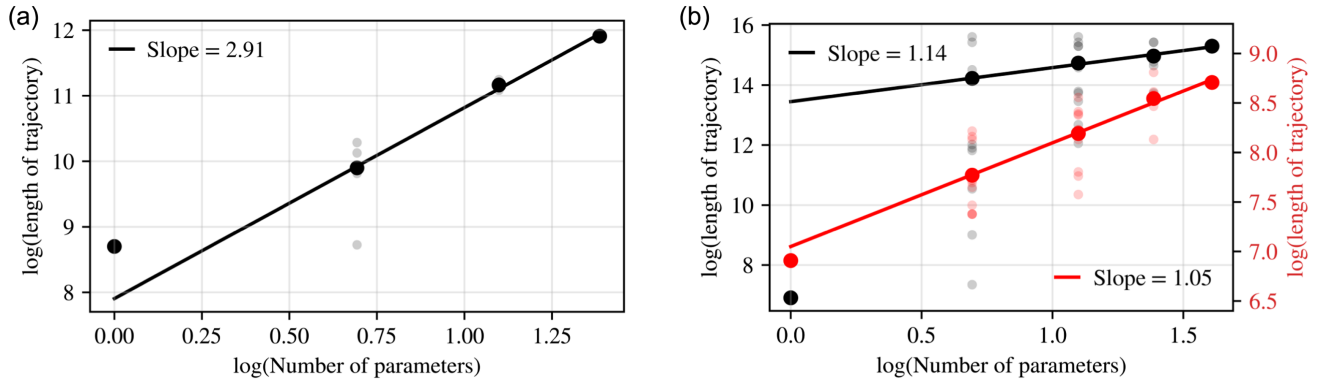


FIG. 5. Computational cost of our calibration method to adjust stability for (a) the double-well system, [(b), black] the box model system, and [(b), red] the optimized box model system. Both figures show how the length of the trajectory that is needed in a single iteration step in order to fulfill the accuracy condition (see Sec. IV), depends on the number of parameters under consideration. The small transparent data points correspond to (averaged) measurements for a given fixed parameter subset whereas the large data points display the average value taken over all parameter subsets, containing the same number of parameters. By fitting a linear function to the averaged data points in the log-log plot for each system, respectively, we infer a polynomial dependence of the trajectory length on the number of parameters. We excluded the data points corresponding to one parameter being varied from the linear fits due to its unique significance for the generation process of the shown data.

reducing stability, the largest eigenvalue  $\lambda$  is moved closer to 1. The computational cost when destabilizing the double-well system as well as the five-box model shows a polynomial dependency on the number of parameters under consideration, although the leading order power seems to differ between the two models (Fig. 5). This polynomial dependency implies great improvement compared to a brute force parameter search (e.g., the exponential scaling when populating the parameter space with always the same density). This is particularly relevant for nondifferentiable models, where gradients of model output with respect to parameter variations cannot be computed. However, brute-force approaches are prohibitive due to the enormous computational costs of running these models. Ensemble Kalman inversion (EKI) [45,60] may be a potential alternative, but several caveats would have to be overcome. For example, (i) the subspace spanned by the initial EKI ensemble must already cover the solution, which can require a very large (initial) ensemble; (ii) though EKI can scale well (i.e., approximately quadratically with the number of parameters), it does not provide any guaranty regarding the convergence time or the quality of the estimated parameters (i.e., closeness to a global or local minimum); (iii) it may be challenging to apply constraints, e.g., conserving the system's mean state during parameter updates, and (iv), since the system's stability is a property of the dynamics rather than an observable, one would need to artificially construct an "observed" property (including its uncertainty), which may pose problems. For example, the maximum value of a bounded value, like autocorrelation = 1, cannot be a valid target for EKI.

We note that our method is not designed to find the most unstable version of the system. Since the result depends on the specific initial conditions of the system state and the parameters, its purpose is to efficiently approach regional minima of the stability landscape. Because there may, in principle, be multiple unstable regions, the search algorithm may need to be run many times from different initial parameter sets, and the computational efficiency of our method is therefore crucial. In order to scan larger regions in parameter space, global methods have to be used. For example, one can perform perturbed-parameter ensemble simulations [46], where the combinations of model parameters are determined by a Latin Hypercube sampling [61]. Such global approaches and our local optimization method are complementary approaches that can be combined in a novel way to illuminate a model's dependence on parameters. In cases where more physical or observational constraints are applied to the parameter ranges and the resulting equilibrium solutions than we used here, the resulting target region of parameter space may already be rather constrained, and our local search method may indeed already deliver a global optimum by itself.

A practical challenge associated with applying our method is that it can be difficult and computationally

expensive to find suitable hyperparameters, involving the calibration of the amplitudes of the noise imposed on the system state or the parameters, or the choice of suitable observables. In general, the final parameter configuration can potentially depend on the choice of the hyperparameters. Depending on how we choose the observables and the state noise, our method offers flexibility as to which aspects the system is stabilized or destabilized (through the choice of observables) and also what types of perturbation are relevant (by choosing suitable noise).

The general recipe provided above has been explicitly demonstrated for the two example systems in Sec. III, but it will also work for many other systems. There is no guarantee, however, that the implementation we chose is also the most optimal approach in a specific case with a considerably different model. Depending on the model at hand, users will need to reconsider how to introduce state noise, how to calibrate parameter noise levels, and how to constrain certain parameters when updating them. The small additions and variations applied in our two example systems in this regard showcase how the method may be adapted. Moreover, we also propose a number of additional methods to evaluate the quality of the VAR model fit (Appendix F), to test the whiteness of the residuals (Appendix G), and an approach to improve the VAR estimation by denoising (Appendix H). We implemented and applied the latter method to our two example systems, which yielded similar results to those discussed above (not shown).

The fact that the result depends on the constraints on the parameter values is a desired property. For example, parameters in complex climate models are often constrained by observations or by their physical meaning. Considering the five-box model, we notice that even after 500 iteration steps we were not able to push  $\lambda$  as close to 1 [(Fig. 4(b))] as in the case of the double-well system. This is likely due to the fact that some parameters hit the boundaries of their restricted ranges [they become zero, see Fig. 4(a)], which might prevent the system from reaching an arbitrary degree of destabilization. In other words, one cannot always expect to be able to change the stability of a system's fixed point, given a set of changeable parameters. Such a result is informative, since it points to structural limitations of a model (or, potentially, the real world) which cannot be overcome and which prevent tipping behavior. By integrating domain-specific constraints and real-world observables, our method enables the exploration of climate change trajectories that are consistent with observations and theory, yet expose possible worst-case tipping-point scenarios.

An indicator of potential improvements is the convergence of the tuned parameter values. In general, each parameter has a clear tendency for the system to lose stability, despite the stochastic elements in our approach. We see an exception to this within the first  $\sim 50$  iteration

steps for the five-box model, where some parameters seem to evolve contrary to their overall tendency [Fig. 4(a)]. This effect seems to be correlated with the mean value  $\bar{q}$  of  $q$  that significantly deviates from its target value. A possible explanation could be that at this point the VAR model does not extrapolate well, causing parameter changes to affect the system in an unforeseen way. On the one hand, this issue could be fixed by temporarily and adaptively decreasing the step size of the parameter updates between iteration steps. On the other hand, this is not required as long as  $q$  recovers to its original value after a temporary anomaly. Moreover, the fixed step size reduces the computational cost of our method.

Our method is flexible in the requirements that one imposes on the updated parameter values, which, in general, could

- (i) Preserve the equilibrium values of the observables by fulfilling Eq. (9).
- (ii) Only take on allowed values in case the parameters are restricted.
- (iii) Not exceed a certain step change, which would typically depend on the noise amplitudes  $\sigma_i$  used for the parameter variation, as these determine the explored region in parameter space.
- (iv) Minimize the errors of the new value of the maximum eigenvalue or the error on the accuracy on how well the condition for preserving the observable values is fulfilled.

All of these conditions (and more) can be included in the parameter update by either implementing these as constraints or as penalty terms in the optimization problem that needs to be solved in order to determine the new parameter values. For example, one could constrain the variance of the observable to some target range (e.g., matching an ESM) by adding an additional penalty to Eq. (D9).

Further research on this topic could be dedicated to applying our method to more complex models and finding more efficient ways for determining appropriate hyperparameter values. Also, a more detailed investigation on why the computational efficiency scales polynomially with the number of parameters, compared to exponential scaling for brute force methods, or more sophisticated methods like EKI, could be of great interest. In addition, machine learning methods [32,62,63] could be used, instead of VAR fitting, for targeted calibration to control system stability.

Our implementation of the method is fully automated, i.e., all steps outlined above, including the noise calibration and iterative parameter updates, work without intervention (see Sec. VI and code documentation online). In principle, it is straightforward to apply our method to more complex models, though the parameter updates and the additive state noise would require one to incorporate the compilation of the model code into the procedure. In Earth system models of intermediate complexity (EMICs), internal dynamics are

often missing or already parametrized as noise [64], rendering them particularly suited for our method (which can exploit knowledge on the realization of the imposed state noise). The same applies to complex components of ESM with little internal variability, such as (parts of) dynamic global vegetation models [55] or ice sheet models [65]. Relevant problems our method could tackle in EMICs are, for example, (i) the “Green Sahara” problem [66], where parameter combinations controlling precipitation and vegetation dynamics are required that allow reproducing abrupt shifts as seen in paleo reconstructions, and (ii) the potential for Amazon forest dieback via parameters affecting moisture recycling and/or vegetation fire dynamics [67,68].

Regarding comprehensive ocean or atmosphere general circulation models, a more fundamental caveat can be that the “noise level” of the observables cannot be directly controlled since their variability emerges from chaotic internal dynamics on short timescales. However, such internal variability does not exclude using our method in principle. For example, depending on the observable, the internal variability can still be sufficiently small, which is, in fact, the case when averaging salinities from ocean GCMs to the scale of the five-box model [53]. Even in the case of larger variance, the method may still yield useful results, although it will generally become less accurate since second-order terms cannot be estimated well anymore due to lack of knowledge of the “noise” realization (see Appendix C). A possible solution could be to fit a model to the slow internal modes, allowing one to separate its recovery dynamics from the noise term (fast chaotic geophysical fluid dynamics). In this situation, the presence of internal variability would make it unnecessary to add any artificial noise, which would improve the usability of the calibration method.

In the case of analyzing slow Earth system components (coupled to an atmosphere and/or ocean GCM), the most important limitation of our method is probably that the large number of uncertain parameters, and the long timescales of the processes of interest, make direct application of our method to calibrate parameters computationally difficult for some problems (including AMOC stability in ESMs). Although the iterative parameter updates work only sequentially, the efficiency of the method may be improved by using an initial condition ensemble to generate the time series used in step 1 in each iteration.

We see our method as an efficient way to identify the key targets for ESM tuning, to access model versions with realistic stability properties. The right choice of such metrics is not obvious *a priori*. Our method is probably best suited to exploring the parameter space of physically based emulators such as the AMOC box model used in our study, to identify unstable regions of its parameter space that are consistent with available observational constraints. The five-box model can be calibrated to represent different

comprehensive ESMs [52–54] and is a good example of how applying our method to a reduced-order model (where technical and theoretical demands for users are still tractable), can be informative for the stability of complex models. Because the parameters of the AMOC box model themselves correspond to emergent (observable, large-scale) properties of the climate system [52], identifying the unstable parameter regions provides a framework for understanding why different ESMs have different stability properties. A pipeline outlining how our approach can support parameter tuning in a given comprehensive ESM proceeds as follows: (i) Calibrate a physically based emulator to the ESM. (ii) Run a targeted stability calibration in the emulator parameter space. (iii) Insert the resulting “unstable but still observationally consistent” parameters back into the ESM.

In case the ESM features larger internal variability than the emulator, the identified optimal parameters may already cause a transition to an alternative regime (“*N*-tipping,” see Ref. [69]), but since our method has tracked the parameter updates, one can step back from the brink as far as required. The optimal parameter regions serving as initial parameter conditions for our method become a set of target metrics for ESM tuning, which can be performed either through conventional tuning approaches, or using systematic methods that are now being developed at several major modeling centers (e.g., [70]).

Because of its flexibility regarding the purpose of the model, the nature of the constraints, and the target property to be optimized, our method is not restricted to applications with the goal of varying the stability of climate models, but can potentially be applied to a wide class of optimization problems within the context of complex dynamical systems.

## VI. CONCLUSION

We have introduced a targeted method for systematic and objective parameter calibration to adjust system stability in nondifferentiable complex system models, under given physical and observational constraints. Our method considers dynamics on the combination of a given model’s phase and parameter space and exploits the phenomenon of critical slowing down to identify the optimal direction in parameter space to adjust the stability of modeled systems in a desired way. This makes the method computationally highly efficient, breaking the curse of dimensionality by scaling only polynomially in the number of parameters. Our results are particularly promising given the persisting concerns that major Earth system components are too stable in state-of-the-art models, which are very challenging to calibrate objectively. The Julia code implementing our method, as well as the two numerical models used in this study, can be found at [71].

## ACKNOWLEDGMENTS

This is ClimTip contribution No. 4; the ClimTip project has received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No. 101137601. N. B. and S. B. also acknowledge funding by the Volkswagen foundation. We are grateful to Brian Groenke for helpful discussions, to Max Gelbrecht for education and support concerning Julia programming, to Mycroft Holmes for vital clues given from his armchair, and to two anonymous reviewers who helped us to improve the manuscript by their constructive concerns and suggestions.

N. B. and S. B. defined the research goal, D. P. designed the implementation and execution of the method under supervision of N. B. and S. B., D. P., and R. W. implemented the five-box model; D. P. wrote a first description of the methods and results, S. B. wrote the paper with contributions from all authors.

## APPENDIX A: VAR(1) MODEL ESTIMATION (STEP 2)

We assume a general setting in which we are given two time series  $(\vec{z}_t)$  and  $(\vec{y}_t)$  with  $t \in \{0, \dots, T\}$ , which can potentially be identical. Our goal is to find a model of the form

$$\vec{z}_{t+1} = \vec{\nu} + A\vec{y}_t + \vec{\epsilon}_t, \quad (\text{A1})$$

which fits the data best in the sense of a least-squares estimate concerning the error vectors  $\vec{\epsilon}_t$ .

In order to provide compact formulas, we introduce the notation

$$Z = (\vec{z}_1, \dots, \vec{z}_T) \quad (d_z \times T) \quad (\text{A2})$$

$$Y_t = \begin{pmatrix} 1 \\ \vec{y}_t \end{pmatrix} \quad [(d_y + 1) \times 1] \quad (\text{A3})$$

$$Y = (Y_0, \dots, Y_{T-1}) \quad [(d_y + 1) \times T] \quad (\text{A4})$$

$$E = (\vec{\epsilon}_1, \dots, \vec{\epsilon}_T) \quad (d_z \times T) \quad (\text{A5})$$

$$B = (\vec{\nu}, A) \quad [d_z \times (1 + d_y)], \quad (\text{A6})$$

where  $d_z$  and  $d_y$  are the dimensions of  $\vec{z}_t$  and  $\vec{y}_t$ , respectively. We then find [also see Eq. (3.2.10) in [72]] that the unbiased least square estimate for  $B$  is given by

$$B = ZY^T(YY^T)^{-1}. \quad (\text{A7})$$

Further, we also obtain an unbiased estimate for the covariance matrix of the error terms  $\vec{\epsilon}_t$  given by

$$\Sigma_\epsilon = \frac{T}{T - d_y - 1} Z(\mathbb{I}_T - Y'(YY')^{-1}Y)Z' \quad (\text{A8})$$

as described in Eq. (3.2.19) in [72].

In order to give an expression of the errors and correlations of the coefficients in matrix  $B$  we introduce the following:

$$\vec{\beta} = (B_{1,1}, B_{2,1}, \dots, B_{d_z,1}, \dots, B_{1,1+d_y}, \dots, B_{d_z,1+d_y})^T \quad (\text{A9})$$

The covariance matrix of  $\vec{\beta}$  (for asymptotically large  $T$ ) can then be estimated as

$$\Sigma_\beta = \frac{\left(\frac{YY'}{T}\right)^{-1} \otimes \Sigma_\epsilon}{T} \quad (\text{A10})$$

which can be found in section 3.2.2. of [72].

We also estimate the errors of the coefficients by using  $(\vec{z}_t) = (\vec{o}_t)$  and  $(\vec{y}_t) = [(\vec{o}_t^T, \vec{p}_t^T, \vec{o}_t^T \otimes \vec{p}_t^T)^T]$ . We make use of these errors when calibrating the noise amplitudes to test the dependence of the observables on the varying parameters (see Appendix B).

## APPENDIX B: FINDING APPROPRIATE NOISE AMPLITUDES FOR THE PARAMETER VARIATION (STEP 3)

The optimal choice of the noise amplitudes  $\sigma_i$  when generating the parameter series as presented in Eq. (4) can be problem-specific, when finding a satisfactory trade-off between the importance of well chosen noise amplitudes and the time it takes to compute these. Quality features of well chosen noise amplitudes are that the respective coefficients of  $A_p$  and  $A_{op}$  from Eq. (7) significantly differ from zero while having small errors. In principle, we aim to maximize each noise level in order to maximize the volume in parameter space which informs us on the effect of the parameters. However, the noise amplitude should not be too large to assure that the parameter variations do not overshadow the additive state noise allowing us to also estimate the parameter independent parts of the model properly. Moreover, no potential restrictions on the parameter ranges should be violated during the parameter variation process, and the overall structure of the dynamical system (close to the considered equilibrium) should remain preserved. We consequently implement a method that tries different noise levels iteratively. In general, the starting value should be sufficiently small; in case of prior knowledge about the system, larger initial values make the method more efficient (we hence picked different initial values for both systems in Sec. III).

Our general approach is to vary the noise levels one by one, i.e., we perform the iteration for each parameter separately. Varying the parameters all together is, of course, more efficient. However, this only works if all parameters

have a similar influence at similar noise amplitudes, which applies to the special case of the double-well system but not to the five-box model (see Sec. III). We also note that for most systems it might not be necessary to update the noise amplitudes in each iteration, which can save computational costs (as is the case in the five-box model).

The iteration ends when the variance of an observable substantially exceeds the variance we obtain without any parameter noise (by setting a fixed threshold value somewhat larger than this natural variability). Thereby, the parameter variation does not overshadow the additive noise, allowing us to also estimate the parameter independent parts of the model properly. In addition, we check whether the observables show a relationship to the varying parameter by testing if the VAR model coefficients differ from zero.

As a more sophisticated approach, one may also test if the respective coefficients of  $A_p$  and  $A_{op}$  significantly differ from zero when considering a given parameter  $p_i$  as shown in section III. 6 in [72]: Let  $C$  be a  $[n \times d_z(d_y + 1)]$  matrix such that  $C\vec{\beta}$  only consists of the  $n$  coefficients in  $\vec{\beta}$  that are relevant for the coupling of the parameter under consideration with  $\vec{o}$ . Then we can compute the following statistic

$$\lambda_F = \frac{1}{n} (C\vec{\beta})' [C((YY')^{-1} \otimes \hat{\Sigma}_\epsilon)C']^{-1} C\vec{\beta}, \quad (\text{B1})$$

which we expect to follow an  $F(n, T - d_y - 1)$  distribution in the case where there is no causal relationship from  $\vec{p}$  to  $\vec{o}$ .

In order to find suitable noise amplitudes, possible approaches could also include grid searches or more sophisticated methods where the noise amplitudes are simultaneously varied until certain conditions are met (in similarity to the double-well system, Sec. III A).

## APPENDIX C: ESTIMATION OF VAR MODEL SECOND ORDER TERMS (STEP 4)

If the magnitude of the variations of the observables  $\vec{o}$  and the parameters  $\vec{p}$  are significantly smaller than the equilibrium values, the second-order term,  $A_{op}$ , is difficult to estimate even from a long time series. This, however, inevitably leads to wrong predictions also of the coefficients describing higher order terms. This can be understood by expressing

$$\vec{p}_t = \vec{p}_{\text{init}} + \Delta\vec{p}_t, \quad (\text{C1})$$

$$\vec{o}_t^{\text{vary}} = \vec{o}_{eq} + \Delta\vec{o}_t^{\text{vary}}, \quad (\text{C2})$$

where  $\vec{o}_{eq}$  is the current equilibrium value of the observable for  $\vec{p} = \vec{p}_{\text{init}}$ . Rewriting Eq. (7) yields

$$\begin{aligned}
\vec{o}_{t+1}^{\text{vary}} = & (\vec{\mu} + A_o \vec{o}_{eq} + A_p \vec{p}_{\text{init}} + A_{op}(\vec{o}_{eq} \otimes \vec{p}_{\text{init}})) \\
& + (A_o \Delta \vec{o}_t^{\text{vary}} + A_{op}(\Delta \vec{o}_t^{\text{vary}} \otimes \vec{p}_{\text{init}})) \\
& + (A_p \Delta \vec{p}_t + A_{op}(\vec{o}_{eq} \otimes \Delta \vec{p}_t)) \\
& + (A_{op}(\Delta \vec{o}_t^{\text{vary}} \otimes \Delta \vec{p}_t)) + \vec{\epsilon}_t^{\text{vary}}. \quad (\text{C3})
\end{aligned}$$

$A_{op}$  can only be estimated from the second order term  $A_{op}(\Delta \vec{o}_t^{\text{vary}} \otimes \Delta \vec{p}_t)$  and, if incorrectly predicted, will distort also the approximation of the higher order coefficients, which leads to wrong conclusions concerning the parameter dependence of the linearized model.

In order to solve this problem, we compare each predicted future time step from the full VAR model to a prediction using a model without parameter dependencies, where we use our knowledge of the state noise trajectory ( $\vec{u}_t^{(x)}$ ) used to force the dynamical system. The trajectory of the observable for the system with fixed parameters, ( $\vec{o}_t^{\text{fixed}}$ ), is computed by

$$\vec{o}_{t+1}^{\text{fixed}} = f_o(f_p(\vec{x}_t^{\text{vary}}, \vec{p}_{\text{init}}) + \vec{u}_t^{(x)}, \vec{p}_{\text{init}}). \quad (\text{C4})$$

Assuming that the time evolution of the observables can indeed be described as suggested in Eq. (7), we obtain

$$\begin{aligned}
\vec{o}_{t+1}^{\text{vary}} = & \vec{\mu} + A_o \vec{o}_t^{\text{vary}} + A_p \vec{p}_t \\
& + A_{op}(\vec{o}_t^{\text{vary}} \otimes \vec{p}_t) + \vec{\epsilon}_t^{\text{vary}}, \quad (\text{C5})
\end{aligned}$$

$$\begin{aligned}
\vec{o}_{t+1}^{\text{fixed}} = & \vec{\mu} + A_o \vec{o}_t^{\text{vary}} + A_p \vec{p}_{\text{init}} \\
& + A_{op}(\vec{o}_t^{\text{vary}} \otimes \vec{p}_{\text{init}}) + \vec{\epsilon}_t^{\text{fixed}}, \quad (\text{C6})
\end{aligned}$$

where Eq. (C5) is Eq. (7) from above, and is repeated here for practical purposes.

Defining the time series  $(\hat{o}_t) \equiv (\vec{o}_t^{\text{vary}} - \vec{o}_t^{\text{fixed}})$ , and taking the difference Eqs. (C5) and (C6), we find

$$\hat{o}_{t+1} = A_p \Delta \vec{p}_t + A_{op}(\vec{o}_t^{\text{vary}} \otimes \Delta \vec{p}_t) + \Delta \vec{\epsilon}_t \quad (\text{C7})$$

with  $\Delta \vec{p}_t \equiv \vec{p}_t - \vec{p}_{\text{init}}$  and  $\Delta \vec{\epsilon}_t \equiv \vec{\epsilon}_t^{\text{vary}} - \vec{\epsilon}_t^{\text{fixed}}$ . If we now fit a VAR(1) model to the data  $\vec{z}_t = \hat{o}_t$  and  $Y_t = [\Delta \vec{p}_t^T, (\vec{o}_t^{\text{vary}} \otimes \Delta \vec{p}_t)^T]^T$  (see Appendix A for details of the notation) we can estimate  $A_p$  and  $A_{op}$ . Note that by using  $Y_t$  instead of  $y_t$  we assume  $\vec{v} = 0$  in our estimation. By eliminating the zeroth order terms as well as the term of first order in  $\Delta \vec{o}_t^{\text{vary}}$  the model becomes much simpler and it is therefore easier to estimate the remaining model parameters. In particular, there is no longer an ambiguity whether variations in the observable  $\vec{o}_t$  must be captured by the coefficient  $A_o$  or by  $A_{op}$  making it easier to correctly estimate  $A_{op}$ , despite the fact that it nevertheless describes a second order contribution.

In order to determine all coefficients from Eq. (7) [Eq. (C5)], we equate the VAR model from Eq. (3) to the model from Eq. (C6), as both models describe the

evolution for parameter values fixed to  $\vec{p}_{\text{init}}$ . By comparing coefficients, we obtain

$$\vec{\mu} = \vec{v} - A_p \vec{p}_{\text{init}} \quad (\text{C8})$$

$$A_o = A - A_{op}(\mathbb{I}_{d_o} \otimes \vec{p}_{\text{init}}). \quad (\text{C9})$$

We here used the Kronecker product properties:

$$\begin{aligned}
A_{op}(\vec{o} \otimes \vec{p}) &= (A_{op}(\vec{o} \otimes \mathbb{I}_{d_p}))\vec{p} \\
&= (A_{op}(\mathbb{I}_{d_o} \otimes \vec{p}))\vec{o}. \quad (\text{C10})
\end{aligned}$$

$A_p$  and  $A_{op}$  are given as estimated in step 4 in Sec. II. The errors of the coefficients, as well as their correlations, can be computed using error propagation techniques. We consider the errors of both VAR estimations (i.e., the estimation for fixed and for varying parameters) as independent, since they arise from independent data.

#### APPENDIX D: PARAMETER UPDATING UNDER CONSTRAINTS (STEP 5)

Our aim is to maximize the largest eigenvalue of the VAR model's autoregressive term  $A_o + A_{op}(\mathbb{I}_{d_o} \otimes \vec{p})$  via parameter changes (step 5 of our recipe in Sec. II). We also wish to preserve the mean state [Eq. (9), duplicated here]:

$$\vec{o} \stackrel{!}{=} \vec{\mu} + A_o \vec{o} + A_p \vec{p} + A_{op}(\vec{o} \otimes \vec{p}). \quad (\text{D1})$$

Rearranging this equation gives

$$(\mathbb{I}_{d_o} - A_o)\vec{o} - \vec{\mu} = [A_p + A_{op}(\vec{o} \otimes \mathbb{I}_{d_p})]\vec{p}, \quad (\text{D2})$$

where we again used expression (C10), this time in the form where  $\vec{p}$  stands on the very right, after the brackets.

We abbreviate the left-hand side of Eq. (D2) as  $\vec{s}$ , and the term in the brackets on the right-hand side as a matrix  $n$  (in  $\mathbb{R}^{d_o \times d_p}$ ):

$$n \equiv A_p + A_{op}(\vec{o} \otimes \mathbb{I}_{d_p}). \quad (\text{D3})$$

The constraint (D1) implies that any parameter vector  $\vec{p}$  must point on a hyperplane defined by

$$\mathcal{H} = \{\vec{p} \in \mathbb{R}^{d_p} | n\vec{p} = \vec{s}\}. \quad (\text{D4})$$

The hyperplane consists of all possible parameter combinations that comply with the constraint of keeping the time mean observable fixed. Each row of matrix  $n$  is a vector that is perpendicular to this hyperplane, and takes care of a specific component of the observable  $\vec{o}$ .

In the following, we discuss the case of a one-dimensional observable ( $d_o = 1$ ), which also applies to the systems we use in Sec. III. In this situation,  $n$  is a vector of length  $d_p$ , and  $s$  is a scalar, and the eigenvalue to be

maximized is the scalar  $A_o + A_{op}\vec{p}$ .  $A_o$  is constant under noisy parameter fluctuations, and only changes in the next iteration after updating the parameters and fitting a new VAR model. Moreover, the new parameters we seek can be written as

$$\vec{p} = \vec{p}_{\text{init}} + \Delta\vec{p}, \quad (\text{D5})$$

where  $\vec{p}_{\text{init}}$  is already given.

Hence, the property we optimize is

$$A_{op} \cdot \Delta\vec{p}. \quad (\text{D6})$$

Here,  $A_{op}$  is a vector, and we abbreviate it as  $\vec{v}$ .

The condition of keeping the equilibrium value of  $q$  at  $\bar{q}$  can now be written as

$$u \stackrel{!}{=} \vec{n} \cdot \Delta\vec{p} \quad (\text{D7})$$

with

$$u = \bar{o}(1 - A_o) - \mu - \vec{n} \cdot \vec{p}_{\text{init}}. \quad (\text{D8})$$

Overall, we hence compute  $\Delta\vec{p}$  by maximizing

$$\begin{aligned} M \equiv & \vec{v} \cdot \Delta\vec{p} - \eta_1 (\vec{n} \cdot \Delta\vec{p} - u)^2 \\ & - \eta_2 (\Delta\vec{p}^T, -1, 0, \dots, 0) \Sigma_{nuv} \begin{pmatrix} \Delta\vec{p} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ & - \eta_3 (0, \dots, 0, \Delta\vec{p}^T) \Sigma_{nuv} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \Delta\vec{p} \end{pmatrix}. \end{aligned} \quad (\text{D9})$$

Here,  $\Sigma_{nuv}$  denotes the correlation matrix of  $(\vec{n}^T, u, \vec{v}^T)^T$  which can be computed from  $\Sigma_\beta$  [Eq. (A10)]: Since  $\Sigma_\beta$  contains all errors and their correlations of the VAR model parameters, and since  $n$ ,  $v$  and  $u$  are functions of these

parameters, we can compute their errors and correlations via standard error propagation techniques.

The first penalty term hence penalizes the distance of  $\Delta\vec{p}$  to the hyperplane determined by  $\vec{n}$  and  $u$  (to keep the mean of the observable constant).

The terms involving  $\eta_2$  and  $\eta_3$  in Eq. (D9) penalize the uncertainties on how  $\vec{v} \cdot \Delta\vec{p}$  (destabilization of the system) and the first penalty term ( $\eta_1$  term) change for a parameter update by  $\Delta\vec{p}$ . The second term penalizes the error on  $\vec{n} \cdot \Delta\vec{p} - u$ , and the third term penalizes the error on  $\vec{v} \cdot \Delta\vec{p}$ . Thus, these last two terms in Eq. (D9) assure that, with a high probability, our parameter update indeed has the desired effect of stabilizing or destabilizing the system and preserving the AMOC strength. The trade-off for this is that the destabilization process could potentially become less efficient.

Apart from the direction of vector  $\Delta\vec{p}$  as discussed above, we also need to control its length. To this end, we rescale all parameters by their respective exploration range in order to make them comparable, and then demand that the Euclidean norm of this rescaled parameter vector must equal a prescribed value  $l$  with

$$l^2 \stackrel{!}{=} \sum_i \frac{\Delta p_i^2}{l_i^2}, \quad (\text{D10})$$

where each  $l_i$  measures how large the explored region is for parameter  $p_i$ . We here use step sizes that match the region in parameter space that we explored via the parameter noise

$$l_i = \frac{\sigma_i}{\sqrt{1 - \lambda_p^2}} \quad (\text{D11})$$

with  $\sigma_i$  and  $\lambda_p$  from Eq. (4). Additional constraints on certain parameters can easily be implemented (e.g., see Sec. III B).

## APPENDIX E: THE FIVE-BOX AMOC MODEL

The model represents the meridional global ocean circulation (“global conveyor belt”) in the form of five boxes. The model has been designed to represent the Atlantic meridional overturning circulation, and its connected major circulation features on the globe [52,73]. The set of equations we use is

$$q = \frac{\lambda}{1 + \lambda\alpha\mu} [\alpha(T_S - T_0) + \beta(S_N - S_S)]. \quad (\text{E1})$$

$$\text{For } q \geq 0 \quad (\text{E2})$$

$$V_N \frac{dS_N}{dt} = q(S_T - S_N) + K_N(S_T - S_N) + F_N S_0 - F_N S_N \quad (\text{E3})$$

$$V_T \frac{dS_T}{dt} = q[\gamma S_S + (1 - \gamma)S_{IP} - S_T] + K_S(S_S - S_T) + K_N(S_N - S_T) + F_T S_0 + F_N(\gamma S_S + (1 - \gamma)S_{IP}) + F_S S_S + F_{IP} S_{IP} \quad (\text{E4})$$

$$V_S \frac{dS_S}{dt} = q\gamma(S_B - S_S) + K_{IP}(S_{IP} - S_S) + K_S(S_T - S_S) + \eta(S_B - S_S) + F_S S_0 + \gamma F_N(S_B - S_S) - F_S S_S \quad (\text{E5})$$

$$V_{IP} \frac{dS_{IP}}{dt} = q(1 - \gamma)(S_B - S_{IP}) + K_{IP}(S_S - S_{IP}) + F_{IP} S_0 + (1 - \gamma)F_N(S_B - S_{IP}) - F_{IP} S_{IP} \quad (\text{E6})$$

$$V_B \frac{dS_B}{dt} = q(S_N - S_B) + \eta(S_S - S_B) + F_N(S_N - S_B), \quad (\text{E7})$$

$$\text{and for } q < 0 \quad (\text{E8})$$

$$V_N \frac{dS_N}{dt} = |q|(S_B - S_N) + K_N(S_T - S_N) + F_N S_0 + (F_T + F_S + F_{IP})S_B \quad (\text{E9})$$

$$V_T \frac{dS_T}{dt} = |q|(S_N - S_T) + K_S(S_S - S_T) + K_N(S_N - S_T) + F_T S_0 - F_T S_T \quad (\text{E10})$$

$$V_S \frac{dS_S}{dt} = |q|\gamma(S_T - S_S) + K_{IP}(S_{IP} - S_S) + K_S(S_T - S_S) + \eta(S_B - S_S) + F_S S_0 + \gamma F_T(S_T - S_S) - F_S S_S \quad (\text{E11})$$

$$V_{IP} \frac{dS_{IP}}{dt} = |q|(1 - \gamma)(S_T - S_{IP}) + K_{IP}(S_S - S_{IP}) + F_{IP} S_0 + (1 - \gamma)F_T(S_T - S_{IP}) - F_{IP} S_{IP} \quad (\text{E12})$$

$$V_B \frac{dS_B}{dt} = |q|\gamma S_S + (1 - \gamma)|q|S_{IP} - |q|S_B + \eta(S_S - S_B) + F_S(S_S - S_B) + F_T[\gamma S_S + (1 - \gamma)S_{IP} - S_B] + F_{IP}(S_{IP} - S_B). \quad (\text{E13})$$

In contrast to the original model, we do not only demand conservation of salt, but also a conservation of total water volume in each box. This results in two slight differences to the original model:

- (1) We removed a factor of  $\gamma$  from the second-to-last term of Eq. (11) in [52].
- (2)  $F_i$  describes the flux of freshwater that the ocean surface of box  $i$  exchanges with the atmosphere. In the original model, there is a water flux into the boxes with index  $i \in \{N, S, T, IP\}$  if  $F_i < 0$ , which is not removed from the box, while boxes with  $F_i > 0$  lose water volume over time.

We therefore added additional fluxes between the boxes. The flux between the boxes  $N$  and  $T$  remains unchanged as this flux is supposed to describe the strength of the AMOC. Furthermore, we split up the additional flux going from box  $B$  into the boxes  $S$  and  $IP$  by the same factor of  $\gamma$  as used for the AMOC.

We use the same parameter values as given in [52] for the FAMOUS<sub>A</sub> simulation, with the exception that in the modified version we use  $S_0 = 0.035 \text{ psu}$  instead of  $S_0 = 35 \text{ psu}$ . The reason is that  $S_0$  in the modified model can be

interpreted as fresh water salinity (the salinity of the rain) whereas in the original model it represents the average salt water salinity. The original equations are then an approximation of our modified equations. As the modification is very small numerically, all results in this study are very likely to be independent of the model version applied.

## APPENDIX F: EVALUATION OF THE VAR(1) MODEL

As an additional suggestion, we propose to evaluate the accuracy of the VAR approximation by comparing forecasts from the original model and the approximated VAR model, defining some expected error bounds. Assume that we have an estimated VAR model and  $(\vec{z}_t^{\text{new}})$  and  $(\vec{y}_t^{\text{new}})$  represent some additional data not used to estimate the model. Then the one-step ahead prediction  $\hat{z}_t^{\text{new}}(1)$  of  $\vec{z}_{t+1}^{\text{new}}$  is given by  $\vec{v} + A\vec{y}_t^{\text{new}}$ . As similarly presented in Sec. III. 5 in [72], the covariance matrix of this one-step prediction is given by

$$\Sigma_z(1) = \frac{T + d_y + 1}{T} \Sigma_e. \quad (\text{F1})$$

Here,  $\Sigma_e$  is the covariance matrix of the offsets  $\vec{e}$ , which can be estimated as given in Eq. (17), and  $T$  is the length of the data series used to estimate the model.

A possible procedure to test the model using one-step forecasts could be the following: Assume we have  $n$  validation samples available. Then for each  $i \in \{1, \dots, n\}$  the one step forecast  $\hat{z}_i^{\text{new}}(1)$  is computed as

$$\hat{z}_i^{\text{new}}(1) = BY_i^{\text{new}} = \vec{v} + A\vec{y}_i^{\text{new}}. \quad (\text{F2})$$

We can then use this for each  $i$

$$[z_i^{\text{new}} - \hat{z}_{i-1}^{\text{new}}(1)]' \Sigma_z(1)^{-1} [z_i^{\text{new}} - \hat{z}_{i-1}^{\text{new}}(1)] \sim \chi^2(d_z). \quad (\text{F3})$$

If we now pick a random subset  $A \subsetneq \{1, \dots, n\}$  of cardinality  $|A| \ll n$  and for  $n$  sufficiently large, we can assume that the corresponding one-step forecasts are uncorrelated. This leads to

$$\sum_{i \in A} [z_i^{\text{new}} - \hat{z}_{i-1}^{\text{new}}(1)]' \Sigma_z(1)^{-1} [z_i^{\text{new}} - \hat{z}_{i-1}^{\text{new}}(1)] \sim \chi^2(d_z | A), \quad (\text{F4})$$

which is a criterion that can be tested.

## APPENDIX G: TESTING FOR WHITENESS OF THE VAR RESIDUALS

Here we suggest a hypothesis test for quantifying the whiteness of the residuals (see Sec. IV.4 in [72]). This means testing whether the autocorrelation of the series ( $u_t$ ) vanishes. We set up a model

$$\varepsilon_t = D_1 \varepsilon_{t-1} + \dots + D_h \varepsilon_{t-h} + e_t \quad (\text{G1})$$

with  $e_t$  as error terms and test the hypothesis  $H_0: D_1 = \dots = D_h = 0$  against  $H_1: D_j \neq 0$  for at least one  $j \in \{1, \dots, h\}$ .

Using a Lagrange multiplier test, we first define

$$\hat{E} = Z - BY \quad (\text{G2})$$

$$F_i = \begin{bmatrix} 0_{(i \times T-i)} & 0_{(i \times i)} \\ \mathbb{I}_{T-i} & 0_{(T-i \times i)} \end{bmatrix} \quad (\text{G3})$$

$$F = (F_1, \dots, F_h) \quad (\text{G4})$$

$$\hat{E} = (\mathbb{I}_h \otimes \hat{E}) F' \quad (\text{G5})$$

$$e = (e_1, \dots, e_T) \quad (\text{G6})$$

$$D = (D_1, \dots, D_h), \quad (\text{G7})$$

using the same notation as introduced in Appendix A. As described in Appendix C.7 and Sec. IV.4.4 in [72], it can be shown that the hypothesis test as described above is equivalent to testing whether

$$\lambda_{LM}(h) = \text{vec}(\hat{E}\hat{E}')'([\hat{E}\hat{E}' - \hat{E}Y'(YY')^{-1}Y\hat{E}']^{-1} \otimes \hat{\Sigma}_e^{-1}) \times \text{vec}(\hat{E}\hat{E}') \quad (\text{G8})$$

is compatible with a  $\chi^2(hd_x^2)$  distribution. To this end, we use that we know the mean value and variance of the  $\chi^2$  distribution, which allows us to define a confidence interval where we can test whether or not  $\lambda_{LM}$  lies inside it or not.

## APPENDIX H: IMPROVEMENT OF VAR ESTIMATION BY DENOISING

In order to potentially increase the accuracy of the VAR estimation with fixed parameters, which is also needed for fixed parameter values (step 2 of our recipe above), one can again exploit the fact that we know the noise values  $\vec{u}_t^{(x)}$  used as offsets for computing the trajectories of the system variables  $\vec{x}_t$ . To this end, we first record the trajectory (including noise) of system variables ( $\vec{x}_t$ ), which can then be used to compute the corresponding observable trajectory ( $\vec{o}_t$ ) using  $f_o$ . We then compute a second trajectory ( $\vec{\delta}_t$ ) defined by

$$\vec{\delta}_t = f_o[f_p(\vec{x}_{t-1}, \vec{p}_{\text{init}}), \vec{p}_{\text{init}}], \quad (\text{H1})$$

i.e., we use the same trajectory but remove the noise in each step. Note that we are considering the case of fixed parameters, which is the reason why we use  $\vec{p} = \vec{p}_{\text{init}}$  as second argument for  $f_o$  and  $f_p$ . If we now estimate the VAR model with  $\vec{z}_t = \vec{o}_t$  and  $\vec{y}_t = \vec{\delta}_t$  using the procedure and the notation from Appendix A, we maintain the benefits of forcing our system with additive white noise—i.e. forcing the system out of equilibrium and “coarse graining” the Jacobian at the equilibrium to a desired scale—but remove the noise from the VAR model estimation. We note that this does not always increase the quality of the VAR model estimation even in the case of simple linear dynamical systems, since  $f_o(\cdot, \vec{p})$  is not injective in general. This means that multiple configurations of the system variables could lead to the same value for the observable, but the values of the observable in the next time step might differ. In cases where this effect plays a dominant role, the benefit of cutting out the noise forcing might be negligible. In the case of the five-box model, we found that including the noise correction for the fixed-parameter VAR model only has a negligible effect on the quality of the VAR model estimation, so we did not apply it for simplicity and for the sake of cutting down the computational costs.

- [1] M. Claussen, L. A. Mysak, A. J. Weaver, M. Crucifix, T. Fichefet, M.-F. Loutre, S. L. Weber, J. Alcamo, V. A. Alexeev, A. Berger, R. Calov, A. Ganopolski, H. Goosse, G. Lohmann, F. Lunkeit, I. I. Mokhov, V. Petoukhov, P. Stone, and Z. Wang, *Earth system models of intermediate complexity: Closing the gap in the spectrum of climate system models*, *Clim. Dyn.* **18**, 579 (2002).
- [2] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, *Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization*, *Geosci. Model Dev.* **9**, 1937 (2016).
- [3] J. M. Gutierrez and A.-M. Treguier, *Annex II: Models. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, United Kingdom and New York, USA, 2021), pp. 2087–2138.
- [4] M. Scholze, J. I. Allen, W. J. Collins, S. E. Cornell, C. Huntingford, M. M. Joshi, J. A. Lowe, R. S. Smith, and O. Wild, *Earth system models: A tool to understand changes in the earth system*, in *Understanding the Earth System: Global Change Science for Application*, edited by S. E. Cornell, I. C. Prentice, J. I. House, and C. J. Downy (Cambridge University Press, Cambridge, England, 2012) Chap. 5, pp. 129–159.
- [5] J. Curry and P. Webster, *Climate science and the uncertainty monster*, *Bull. Am. Meteorol. Soc. (BAMS)* **92**, 1667 (2011).
- [6] D. Stainforth, M. Allen, E. Tredger, and L. Smith, *Confidence, uncertainty and decision-support relevance in climate predictions*, *Phil. Trans. R. Soc. A* **365**, 2145 (2007).
- [7] H. A. Dijkstra, *The role of conceptual models in climate research*, *Physica (Amsterdam)* **457D**, 133984 (2024).
- [8] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber, *Tipping elements in the Earth's climate system*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1786 (2008).
- [9] D. Chen, M. Rojas, B. Samset, K. Cobb, A. Diongue Niang, P. Edwards, S. Emori, S. Faria, E. Hawkins, P. Hope, P. Huybrechts, M. Meinshausen, S. Mustafa, G.-K. Plattner, and A.-M. Treguier, *Framing, Context, and Methods. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, United Kingdom and New York, USA, 2021), pp. 147–286.
- [10] S. Wang, A. Foster, E. A. Lenz, J. D. Kessler, J. C. Stroeve, L. O. Anderson, M. Turetsky, R. Betts, S. Zou, W. Liu, W. R. Boos, and Z. Hausfather, *Mechanisms and impacts of earth system tipping elements*, *Rev. Geophys.* **61**, 1 (2023).
- [11] N. Boers, T. Liu, S. Bathiany, M. Ben-Yami, L. L. Blaschke, N. Bochow, C. A. Boulton, T. M. Lenton, A. Morr, D. Nian, M. Rypdal, and T. Smith, *Destabilization of earth system tipping elements*, *Nat. Geosci.* **18**, 949 (2025).
- [12] N. Boers, M. Ghil, and T. Stocker, *Theoretical and paleoclimatic evidence for abrupt transitions in the earth system*, *Environ. Res. Lett.* **17**, 093006 (2022).
- [13] S. Drijfhout, S. Bathiany, C. Beaulieu, V. Brovkin, M. Claussen, C. Huntingford, M. Scheffer, G. Sgubin, and D. Swingedouw, *Catalogue of abrupt shifts in intergovernmental panel on climate change climate models*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5777 (2015).
- [14] S. Terpstra, S. K. J. Falkena, R. Bastiaansen, S. Bathiany, H. A. Dijkstra, and A. S. von der Heydt, *Assessment of abrupt shifts in cmip6 models using edge detection*, *AGU Adv.* **6**, e2025AV001698 (2025).
- [15] P. Valdes, *Built for stability*, *Nat. Geosci.* **4**, 414 (2011).
- [16] J. V. Mecking, S. S. Drijfhout, L. C. Jackson, and M. B. Andrews, *The effect of model bias on Atlantic freshwater transport and implications for AMOC bi-stability*, *Tellus A* **69**, 1299910 (2017).
- [17] H. Haken, *Synergetics—An Introduction* (Springer, Berlin, Heidelberg, New York, Tokyo, 1983).
- [18] 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 (London)* **461**, 53 (2009).
- [19] H. Held and T. Kleinen, *Detection of climate system bifurcations by degenerate fingerprinting*, *Geophys. Res. Lett.* **31** (2004).
- [20] N. Boers and M. Rypdal, *Critical slowing down suggests that the western greenland ice sheet is close to a tipping point*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024192118 (2021).
- [21] N. Boers, *Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation*, *Nat. Clim. Change* **11**, 680 (2021).
- [22] C. A. Boulton, T. M. Lenton, and N. Boers, *Pronounced loss of Amazon rainforest resilience since the early 2000s*, *Nat. Clim. Change* **12**, 271 (2022).
- [23] T. Smith, D. Traxl, and N. Boers, *Empirical evidence for recent global shifts in vegetation resilience*, *Nat. Clim. Change* **12**, 477 (2022).
- [24] N. Bochow and N. Boers, *The South American monsoon approaches a critical transition in response to deforestation*, *Sci. Adv.* **9**, eadd9973 (2023).
- [25] 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 Dyn.* **53**, 53 (2003).
- [26] C. Boulton, L. Allison, and T. Lenton, *Early warning signals of Atlantic meridional overturning circulation collapse in a fully coupled climate model*, *Nat. Commun.* **5**, 5752 (2014).
- [27] S. Bathiany, M. Claussen, and K. Fraedrich, *Detecting hotspots of atmosphere–vegetation interaction via slowing down—part 1: A stochastic approach*, *Earth Syst. Dyn.* **4**, 63 (2013).
- [28] S. Bathiany, M. Claussen, and K. Fraedrich, *Detecting hotspots of atmosphere–vegetation interaction via slowing down—part 2: Application to a global climate model*, *Earth Syst. Dyn.* **4**, 79 (2013).
- [29] E. Weinans, J. Lever, S. Bathiany, R. Quax, J. Bascompte, E. van Nes, M. Scheffer, and I. van de Leemput, *Finding the direction of lowest resilience in multivariate complex systems*, *J. R. Soc. Interface* **16**, 20190629 (2019).

- [30] M. Ben-Yami, A. Morr, S. Bathiany, and N. Boers, *Uncertainties too large to predict tipping times of major earth system components from historical data*, *Sci. Adv.* **10**, eadl4841 (2024).
- [31] K. J. Astroem and B. Wittenmark, *On self tuning regulators*, *Automatica* **9**, 185 (1973).
- [32] T. Baumeister, S. Brunton, and J. Kutz, *Deep learning and model predictive control for self-tuning mode-locked lasers*, *J. Opt. Soc. Am. B* **35**, 617 (2018).
- [33] I. e. a. Price, *Probabilistic weather forecasting with machine learning*, *Nature (London)* **637**, 84 (2025).
- [34] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, *Accurate medium-range global weather forecasting with 3D neural networks*, *Nature (London)* **619**, 533 (2023).
- [35] O. Watt-Meyer, B. Henn, J. McGibbon, S. K. Clark, A. Kwa, W. A. Perkins, E. Wu, L. Harris, and C. S. Bretherton, *ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses*, *npj Clim. Atmos. Sci.* **8**, 205 (2025).
- [36] D. e. a. Kochkov, *Neural general circulation models for weather and climate*, *Nature (London)* **632**, 1060 (2024).
- [37] T. e. a. Rackow, *Robustness of AI-based weather forecasts in a changing climate*, *arXiv:2409.18529*.
- [38] M. Gelbrecht, A. White, S. Bathiany, and N. Boers, *Differentiable programming for earth system modeling*, *Geosci. Model Dev.* **16**, 3123 (2023).
- [39] C. Shen, A. Appling, P. Gentile *et al.*, *Differentiable modelling to unify machine learning and physical models for geosciences*, *Nat. Rev. Earth Environ.* **4**, 552 (2023).
- [40] T. Mauritsen, B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, U. Mikolajewicz, D. Notz, R. Pincus, H. Schmidt, and L. Tomassini, *Tuning the climate of a global model*, *J. Adv. Model. Earth Syst.* **4**, M00A01 (2012).
- [41] F. Hourdin, T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, F. Rauser, C. Rio, L. Tomassini, M. Watanabe, and D. Williamson, *The art and science of climate model tuning*, *Bull. Am. Meteorol. Soc.* **98**, 589 (2017).
- [42] T. Mauritsen and E. Roeckner, *Tuning the MPI-ESM1.2 global climate model to improve the match with instrumental record warming by lowering its climate sensitivity*, *J. Adv. Model. Earth Syst.* **12**, e2019 MS002037 (2020).
- [43] D. J. Lea, M. R. Allen, and T. W. N. Haine, *Sensitivity analysis of the climate of a chaotic system*, *Tellus A* **52**, 523 (2000).
- [44] J. A. Nelder and R. Mead, *A simplex method for function minimization*, *Comput. J.* **7**, 308 (1965).
- [45] N. B. Kovachi and A. M. Stuart, *Ensemble Kalman inversion: A derivativefree technique for machine learning tasks*, *Inverse Probl.* **35**, 095005 (2019).
- [46] M. Collins, B. Booth, B. Bhaskaran, G. Harris, J. Murphy, D. Sexton, and M. Webb, *Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles*, *Clim. Dyn.* **36**, 1737 (2011).
- [47] E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, and A. M. Stuart, *Calibrate, emulate, sample*, *J. Comput. Phys.* **424**, 109716 (2021).
- [48] Q. Yang, G. S. Elsaesser, M. van Lier-Walqui, and T. Eidhammer, *A simple emulator that enables interpretation of parameter-output relationships, applied to two climate model PPEs*, *J. Adv. Model. Earth Syst.* **17**, e2024 MS004766 (2025).
- [49] K. Cranmer, J. Brehmer, and G. Louppe, *The frontier of simulation-based inference*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
- [50] K. Lux, P. Ashwin, R. A. Wood, and C. Kuehn, *Assessing the impact of parametric uncertainty on tipping points of the Atlantic meridional overturning circulation*, *Environ. Res. Lett.* **17**, 075002 (2022).
- [51] K. Kato, Y. Itoh, and T. Kousaka, *Predicting bifurcation points via extreme learning machine methods trained on time-series datasets and parameters in discrete dynamical systems*, *Phys. Rev. E* **112**, 014208 (2025).
- [52] R. A. Wood, J. M. Rodríguez, R. S. Smith, L. C. Jackson, and E. Hawkins, *Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation*, *Clim. Dyn.* **53**, 6815 (2019).
- [53] R. R. Chapman, P. Ashwin, J. Baker, and R. A. Wood, *Quantifying risk of a noise-induced AMOC collapse from northern and tropical Atlantic Ocean variability*, *Environ. Res. Commun.* **6**, 111003 (2024).
- [54] R. Chapman, S. Sinet, and P. D. L. Ritchie, *Tipping mechanisms in a conceptual model of the Atlantic meridional overturning circulation*, *Weather* **79**, 316 (2024).
- [55] S. Bathiany, D. Nian, M. Druke, and N. Boers, *Resilience indicators for tropical rainforests in a dynamic vegetation model*, *Global Change Biol.* **30**, e17613 (2024).
- [56] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, Berlin, 1992).
- [57] L. C. Jackson, R. S. Smith, and R. A. Wood, *Ocean and atmosphere feedbacks affecting AMOC hysteresis in a GCM*, *Clim. Dyn.* **49**, 173 (2017).
- [58] E. L. McDonagh, B. A. King, H. L. Bryden, P. Courtois, Z. Szuts, M. Baringer, S. A. Cunningham, C. Atkinson, and G. Gerard McCarthy, *Continuous estimate of Atlantic Oceanic freshwater flux at 26.5 °n*, *J. Clim.* **28**, 8888 (2015).
- [59] K. Wiesenfeld and B. McNamara, *Small-signal amplification in bifurcating dynamical systems*, *Phys. Rev. A* **33**, 629 (1986).
- [60] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, *Inverse Probl.* **29**, 045001 (2013).
- [61] N. Urban and T. Fricker, *A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model*, *Comput. Geosci.* **36**, 746 (2010).
- [62] F. Berkenkamp, A. Krause, and A. P. Schoellig, *Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics*, *Mach. Learn.* **112**, 3713 (2023).
- [63] F. Hourdin, B. Ferster, J. Deshayes, J. Mignot, I. Musat, and D. Williamson, *Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections*, *Sci. Adv.* **9**, eadf2758 (2023).
- [64] M. Willeit, A. Ganopolski, A. Robinson, and N. R. Edwards, *The earth system model CLIMBER-X v1.0—part 1: Climate model description and validation*, *Geosci. Model Dev.* **15**, 5905 (2022).
- [65] A. Robinson, J. Alvarez-Solas, M. Montoya, H. Goelzer, R. Greve, and C. Ritz, *Description and validation of the ice-sheet model Yelmo (version 1.0)*, *Geosci. Model Dev.* **13**, 2805 (2020).

- [66] P. Hopcroft and P. Valdes, *Paleoclimate-conditioning reveals a North Africa land–atmosphere tipping point*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2108783118 (2021).
- [67] P. M. Cox, R. Betts, M. Collins, P. P. Harris, C. Huntingford, and C. Jones, *Amazonian forest dieback under climate-carbon cycle projections for the 21st century*, *Theor. Appl. Climatol.* **78**, 137 (2004).
- [68] T. E. Lovejoy and C. Nobre, *Amazon tipping point*, *Sci. Adv.* **4**, eaat2340 (2018).
- [69] 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).
- [70] S. Peatier, B. M. Sanderson, L. Terray, and R. Roehrig, *Investigating parametric dependence of climate feedbacks in the atmospheric component of CNRM-CM6-1*, *Geophys. Res. Lett.* **49**, e2021GL095084 (2022).
- [71] Code is available at <https://github.com/TUM-PIK-ESM/targeted-calibration>, published under the MIT license
- [72] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis* (Springer, Berlin, 2005).
- [73] H. Alkhayuon, P. Ashwin, L. Jackson, C. Quinn, and R. Wood, *Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model*, *Proc. R. Soc. A* **475**, 20190051 (2019).