A fluctuation-dissipation theorem perspective on radiative responses to temperature perturbations

Fabrizio Falasca,[a] Aurora Basinski-Ferris,[a] Laure Zanna,[a] Ming Zhao[b]

[a] *Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*
[b] *NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA*

ABSTRACT: Radiative forcing drives warming in the Earth system, leading to changes in sea surface temperatures (SSTs) and associated radiative feedbacks. The link between changes in the top-of-the-atmosphere (TOA) net radiative flux and SST patterns, is typically diagnosed by studying the response of atmosphere-only models to SST perturbations. In this work, we diagnose the pattern effect through response theory, by performing idealized warming perturbation experiments from unperturbed data alone. First, by studying the response at short time scales, where the response is dominated by atmospheric variability, we recover results that agree with the literature. Second, by extending the framework to longer time scales, we capture coupled interactions between the slow ocean component and the atmosphere, yielding a novel "sensitivity map" quantifying the response of the net radiative flux to SST perturbations in the coupled system. Here, feedbacks are captured by a spatiotemporal response operator, rather than time-independent maps as in traditional studies. Both formulations skillfully reconstruct changes in externally forced simulations and provide practical strategies for climate studies. The key distinction lies in their perspectives on climate feedbacks. The first formulation, closely aligned with prediction tasks, follows the traditional view in which slow variables, such as SSTs, exert a one-way influence on fast variables. The second formulation broadens this perspective by incorporating spatiotemporal interactions across state variables. This alternative approach explores how localized SST perturbations can alter the coupled dynamics, leading to temperature changes in remote areas and further impacting the radiative fluxes at later times.

1. Introduction

This is a post-peer-review, pre-copyedit version of an article published in Journal of Climate. The final authenticated version is available online at: https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-24-0479.1/JCLI-D-24-0479.1.xml

The response of Earth's surface temperature to changes in atmospheric CO_2 concentration is at the heart of climate change science and yet remains uncertain (e.g., Grubb et al. 2022; Zelinka et al. 2020; Meehl et al. 2020). For example, the equilibrium global mean surface temperature change to a doubling of CO_2, i.e., "equilibrium climate sensitivity", ranges between $1.8-5.6^{o}$C in the latest generation of climate models, an even larger uncertainty than the previous Coupled Model Intercomparison Project (CMIP) (Meehl et al. 2020). A common framework to understand the global mean climate response to external forcing is through the linear global energy balance, in which we decompose the net heat flux (N) at the top of the atmosphere as a radiative forcing (Q) and a radiative response of the system (H) as $N = Q - H$ (Gregory et al. 2004). The radiative response can be expressed as $H \approx \lambda \Delta T$, where λ $[W/(m^2 K)]$ is a climate feedback parameter and ΔT $[K]$, the global mean temperature change (Gregory et al. 2004). The magnitude of the climate feedback parameter λ is a large contributor to the uncertainty of climate sensitivity in models (Chenal et al. 2022; Roe and Armour 2011), primarily due to the poor representation of clouds (e.g., Zelinka et al. 2020).

The climate feedback parameter (λ) is often approximated as constant; however, it has been demonstrated in coupled climate models that λ evolves in time, even under a time-independent CO_2 forcing (e.g., Murphy 1995; Senior and Mitchell 2000; Williams et al. 2008; Andrews et al. 2015; Meyssignac et al. 2023). It is generally hypothesized that the temporal evolution of λ is due to the evolution of the spatial pattern of surface temperature, which can initiate different climate feedbacks over time (Armour et al. 2013; Zhao 2022), referred to as the "pattern effect" (Stevens et al. 2016). In other words, given the same global mean temperature change, different spatial patterns of sea surface temperature change can lead to very different radiative feedbacks. Given the importance of λ for constraining climate sensitivity, the pattern effect has been a source of continuous investigation in recent years (Zhou et al. 2017; Zhang et al. 2023; Dong et al. 2020, 2019; Alessi and Rugenstein 2023; Bloch-Johnson et al. 2020; Kang et al. 2023; Bloch-Johnson et al. 2024).

The current approach for diagnosing the pattern effect relies on Green's function experiments (Barsugli and Sardeshmukh 2002), by leveraging atmosphere-only models to diagnose the response of the top of the atmosphere (TOA) radiative fluxes to perturbations in the sea surface temperature field (Zhou et al. 2017; Zhang et al. 2023; Dong et al. 2020, 2019; Alessi and Rugenstein

Corresponding author: Fabrizio Falasca, fabri.falasca@nyu.edu

2023). Despite some variations, current studies tend to have the following features in common, as summarized in Bloch-Johnson et al. (2024): (i) the assumption that global mean TOA radiative response is linearly related to variations in the sea surface temperature (SST) spatial pattern, (ii) utilizing an atmosphere-only model to diagnose the atmosphere's sensitivity to SST boundary conditions, and (iii) using relatively large perturbations (e.g., $1-4K$) in the SST field when constructing the Green's functions. Large perturbations are beneficial for detecting a clear response with a shorter integration time but also lead to additional problems as large perturbations can result in nonlinear responses (Williams et al. 2023), invalidating assumption (i) stated above. In addition to the Green's function approach, there have been efforts to estimate a response operator from existing simulation output to avoid the computational cost of numerical perturbation experiments. In particular, Zhou et al. (2017); Bloch-Johnson et al. (2020); Kang et al. (2023) have all used forms of linear regression to estimate the pattern effect and explore sensitivities from internal variability. Regression approaches successfully eliminate the computational expenses of previous Green's function methods, but may not be optimally designed to infer spatiotemporal-dependent causal links among climate fields.

In this work, we present a general method for diagnosing radiative responses to temperature perturbations based on the Fluctuation-Dissipation Relation, also often referred to as Fluctuation-Dissipation Theorem (FDT) (see, e.g., Majda et al. 2005), and response theory (Marconi et al. 2008). The Fluctuation-Response formalism provides a strategy to compute the ensemble average response of a physical system to *external* perturbations, solely given correlation functions of the *unperturbed* system (Kraichnan 1959; Kubo 1966; Majda et al. 2005; Marconi et al. 2008). The Earth's climate is a multiscale, complex dynamical system for which the application of the FDT formalism is only possible by focusing on the proper observables (Gritsun and Lucarini 2017) and on a narrow range of spatiotemporal scales (Majda et al. 2005; Baldovin et al. 2022b). The protocol proposed here, building on the framework recently presented in Falasca et al. (2024), leverages the FDT formalism together with dimensionality reduction techniques to infer a response operator in a coarse-grained representation of the climate system. The success of linear response theory for climate data is closely tied to coarse-graining methods (Colangeli et al. 2012). In the case of spatiotemporal dynamical systems (such as climate), coarse-graining procedures refer to (i) averaging over large spatial regions (or choosing a few relevant modes), (ii) selecting a limited range of temporal scales, and (iii) considering a limited set of variables. These coarse-graining steps need to be carefully considered in applications of FDT in climate and they will be detailed throughout this paper. The response operator, together with the appropriate convolution formulas, allows us to study the causal relation between the SST and the TOA net radiative flux fields across multiple spatial and temporal scales.

The method presented in this work bridges the gap between Green's function approach (Bloch-Johnson et al. 2024) and previous statistical methods (Zhou et al. 2017; Bloch-Johnson et al. 2020). Like Green's function techniques derived from atmospheric models (Zhang et al. 2023), our framework infers causal linkages among climate fields within the paradigm of responses to external perturbations (Baldovin et al. 2020). At the same time, it remains data-driven, akin to prior statistical approaches, offering a computationally fast, causal protocol to diagnose the pattern effect directly from data.

We apply the proposed framework to the GFDL-CM4 model (Held et al. 2019), demonstrating its relevance to the literature on the pattern effect in several ways. First, by focusing on responses at the shortest time scales, where the atmospheric dynamics dominates the response to SST perturbations, we recover results that align with the existing literature. This strategy is justified as the response of TOA fluxes to changes in SST is fast (e.g., ~ 20 days in tropical large-scale convective overturning motion). We refer to this formulation as "atmospheric-only" formulation. The "atmospheric-only" formulation is tailored for the pattern effect problem where the instantaneous sea surface temperature pattern determines the radiative response. In a second step, we incorporate responses over longer time scales and refer to this as the "atmosphere-ocean coupled", or "coupled", formulation. Here, the slow ocean response to SST perturbations becomes significant, and the net flux at the TOA can be further modulated by slow changes driven by the system's coupled dynamics reacting to the initial perturbation patterns. This extended analysis yields a novel sensitivity map, characterized by negative sensitivity across the tropical Pacific, in contrast to the traditional view of a negative-positive dipole in the basin. This second perspective it is not interpretable in the traditional pattern effect framework whereby an instantaneous prescribed SST pattern maps into the instantaneous TOA budget. Both formulations of the protocol demonstrate good predictive skill in capturing changes in radiative flux at the TOA across two independent forced experiments. In Section 7, we reconcile the two approaches from both conceptual and practical angles, emphasizing that their primary distinction lies in the formulation of climate feedbacks. The first formulation adopts the traditional view, closely linked to prediction tasks, treating feedbacks as quasi-instantaneous, with one-way directionality from slow variables (e.g., SST) to fast ones (e.g., TOA radiative flux). In contrast, the second formulation introduces a novel perspective where initial perturbations at time t drive cascades of responses in the coupled climate dynamics at $t+\tau$, which can further amplify or dampen the radiative balance. Feedbacks, in this case, are encoded in a spatially and temporally dependent response operator rather than in static maps. This distinction builds on the framework proposed by Lucarini (2018), also in the context of response theory, and formalizes ideas and considerations previously discussed in the climate literature by Taylor et al. (2006). From an information theory perspective, this view of feedbacks is closely linked to the idea of "flow information" as first shown in Ay and Polani (2008). As further detailed in Sections 7, 8 and Appendix A, this "coupled" formulation also introduces new challenges, as it requires accounting for the full complexity of the climate system's response. In the coupled system, focusing on only two variables represents a more significant simplification than in the atmosphere-only system.

Importantly, in both formulations, we eliminate the need to prescribe large perturbations to the system to construct the Green's function operator. This ensures the validity of the linearity assumption when studying the response of TOA fluxes to SST perturbations. The FDT response operator is derived in the limit of infinitesimally small perturbations, where the linearity assumption holds (Majda et al. 2005; Marconi et al. 2008). Evaluation tests presented in this paper demonstrate that, in the specific context of the pattern effect and with the proposed coarse-graining procedure, the FDT formalism can also provide accurate predictions for global mean TOA response to finite-amplitude perturbations in the SST pattern.

In what follows, we describe the proposed methodology and data preprocessing in Sections 2 and 3. We present the general, data-driven protocol to perform perturbation experiments in a multivariate climate system in Section 4. The "atmospheric-only" and "coupled" formulations are presented in Sections 5 and 6 respectively. The main differences across approaches, mainly related to a different perspective on climate feedbacks are further outlined in Section 7. Conclusions and future work are given in Section 8. Useful considerations for the correct application of FDT to realistic climate data, including limitations and caveats, are further outlined in detail in Appendix A and are especially relevant for future applications of FDT.

2. Proposed theoretical framework

We present our proposed framework, which focuses on computing a response operator from the unforced fluctuations of the climate system. In practice, to meet the necessary assumptions in linear response theory, we compute the response operator for a coarse-grained representation of the relevant state variables by focusing on large-scale spatial averages and a limited range of time scales. We detail the dimensionality reduction technique in Section 2c and the data preprocessing in Section 3. Further details on the necessary assumptions and implementation choices are discussed in Appendix A.

a. Linear response theory and the Fluctuation-Dissipation Theorem

Consider a dynamical system with state vector $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]$. Here, N is the dimensionality of the system and t is a time index. We study the system from the perspective of stochastic dynamics, interpreting $\mathbf{x}(t)$ as a set of N coarse-grained variables. The integrated effects of the unresolved degrees of freedom on the resolved ones are modeled as suitable noise terms (Hasselmann 1976; Penland 1989; Majda et al. 1999, 2001; Lucarini and Chekroun 2023). We perturb the system by applying a small, time-dependent perturbation $\delta x_j(t) = \Delta x_j \delta f_j(t)$ directly to the j-th degree of freedom x_j, i.e. $x_j \to x_j + \delta x_j(t)$. The leading order of the response in the k-th variable can be written as:

$$\begin{aligned}\delta\langle x_k(t)\rangle &= \langle x_k(t)\rangle_{\mathrm{p}} - \langle x_k(t)\rangle \\ &= \sum_j \int_0^t R_{k,j}(\tau)\delta x_j(t-\tau)d\tau,\end{aligned} \tag{1}$$

where $\langle\cdot\rangle$ and $\langle\cdot\rangle_{\mathrm{p}}$ indicate ensemble averages of the system before and after applying the perturbation. The response operator $R_{k,j}(t)$ is defined as the average response of the variable $x_k(t)$ to an *impulse* perturbation $\Delta x_j \delta(t)$ applied to the variable $x_j(0)$, where $\delta(t)$ denotes the Dirac delta function (Risken 1996; Baldovin et al. 2022a). If the perturbation $\delta x_j(t)$ is instead represented as a Heaviside step function, $\delta x_j(t) = \Delta x_j \Theta(t)$, the response simplifies to:

$$\delta\langle x_k(t)\rangle = \sum_j \delta x_j \int_0^t R_{k,j}(\tau)d\tau. \tag{2}$$

The primary challenge in quantifying responses to perturbations, as described in Eqs. (1) and (2), is to infer the response operator in terms of the *unperturbed* system. The Fluctuation-Dissipation Theorem (FDT) addresses this challenge by establishing a causal link between the variability of the *unperturbed*, stationary system and its response to *external* perturbations (e.g., Marconi et al. 2008; Hairer and Majda 2010). Verifying the existence of such a link, for specific cases or variables of interest, is highly relevant for climate dynamics, whether to study the climate response to external forcing (Leith 1975; Majda et al. 2005) or to analyze the spatio-temporal interactions among different climate variables (Lucarini 2018; Falasca et al. 2024).

(i) FDT: general formulation. Given a sufficiently smooth and non-vanishing invariant probability distribution, $\rho(\mathbf{x})$, of the stochastic system $\mathbf{x}(t)$, the following result holds:

$$R_{k,j}(t) = \lim_{\delta x_j(0)\to 0} \frac{\delta\langle x_k(t)\rangle}{\delta x_j(0)} = -\left\langle x_k(t) \frac{\partial \ln\rho(\mathbf{x})}{\partial x_j}\Big|_{\mathbf{x}(0)}\right\rangle. \tag{3}$$

This formulation represents a general form of the FDT (Falcioni et al. 1990). Eq. (3) enables us to evaluate the response of a stochastic dynamical system to infinitesimal impulse perturbations using only the unperturbed dynamics of the system. We note that by focusing on stochastic systems, we circumvent the technical difficulties associated with deterministic dissipative dynamical systems, where the invariant measure is singular (see Ruelle 1998, 2009; Colangeli et al. 2012).

(ii) FDT and causality. Recent literature including Aurell and Del Ferraro (2016); Lucarini (2018); Baldovin et al. (2020) has pointed out the connection between the Fluctuation-Response formalism and the role of causality in physical systems based on the notion of intervention (Pearl 2000; Ismael 2023). The main idea is that, in physical experiments, cause-effect relations are inferred by probing the system and examining its response. Specifically, the link $x_j(t) \to x_k(t+\tau)$ is inferred by studying how an *external* perturbation at variable $x_j(t)$ propagates along the system, inducing *on average* a change in variable $x_k(t+\tau)$. In the case of small perturbations, the Fluctuation-Dissipation Theorem shown in Eq. (3) allows us to do so in a straightforward way, by inferring what the response would have been if we had perturbed the system[1]. We refer the reader to Aurell and Del Ferraro (2016) and Baldovin et al. (2020) for further details and to Baldovin et al. (2022b); Cecconi et al. (2020); Falasca et al. (2024); Giorgini et al. (2024) for some examples of applications.

(iii) Quasi-Gaussian approximation. The main practical issue with the formulation given by Eq. (3) is that the functional form of $\rho(\mathbf{x})$ is not known a priori and is highly nontrivial in high-dimensional systems. While recent promising results in the estimation of $\nabla_{\mathbf{x}} \ln\rho(\mathbf{x})$ using generative modeling (Giorgini et al. 2024) have emerged, generally, the strategy has been to approximate $\rho(\mathbf{x})$ as a Gaussian distribution (Leith 1975). In the case of Gaussian distributions, Eq. (3) reduces to:

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{C}(0)^{-1}, \tag{4}$$

with the covariance function $C_{i,j}(t) = \langle x_i(\tau+t)x_j(\tau)\rangle$ (x_i is assumed to be zero mean). Eq. (4) is valid for linear systems and has been referred to as the "quasi-Gaussian approximation" (Majda et al. 2005). The form of FDT shown in Eq. (4) motivated many studies in climate, including Gritsun and Branstator (2007); Ring and Plumb (2008); Majda et al. (2010b); Hassanzadeh and Kuang (2016a,b); Christensen and Berner (2019); Lutsko et al. (2015), which focused on the implications or limitations of this formalism. It has been shown empirically that the quasi-Gaussian approximation has high skill for predicting the response in the ensemble mean in non-Gaussian regimes (Gritsun and Branstator 2007; Gershgorin and Majda 2010; Baldovin et al. 2020) and has some skill for the response in variance (Gritsun et al. 2008; Majda et al. 2010b). In fact, while the climate system is a high-dimensional, chaotic dynamical system, the probability density of many coarse-grained variables is often smooth and nearly Gaussian (Majda et al. 2010a). The quasi-Gaussian approximation is then relevant in climate studies after the appropriate spatial and temporal coarse-graining.

b. Numerics and statistics

The estimation of the response operator $\mathbf{R}(t) = \mathbf{C}(t)\mathbf{C}(0)^{-1}$ is limited by the data sample size and by the *effective* dimensionality of the dynamics (Gritsun and Branstator 2007; Martynov and Nechepurenko 2006). These limitations lead to two primary challenges: (i) the inferred covariance matrix $\mathbf{C}(0)$ is often ill-conditioned, leading to significant errors in the computation of its inverse; (ii) $\mathbf{R}(t)$ is contaminated by spurious terms due to the interplay of limited sample size and strong autocorrelations of the underlying time series. To address these issues, we employ both a regularization procedure when computing the covariance matrix and an estimation of the confidence bounds introduced in Falasca et al. (2024). We stress that the effectiveness of these tools depends on the problem at hand, including factors such as the amount of available data and the dimensionality of the system.

(i) Regularization strategy. Computing the covariance matrix, $\mathbf{C}(0)$, after reducing the dimensionality of the system, helps to lower the condition number. However, the condition number may remain high enough that a regularization step is warranted. In our case, we compute the covariance matrices in a low-dimensional space (see Section 2c) and add a Tikhonov regularization (Hansen 2010) as follows:

$$\mathbf{C}_r(0) = \mathbf{C}(0) + \lambda\mathbf{I}, \tag{5}$$

where $\mathbf{C}_r(0)$ represents the regularized matrix, λ is a parameter and $\mathbf{I}$ is the identity matrix. In practice, there is a balance between lowering the condition number of the covariance matrix and retaining important elements of the system's dynamics; if the choice of λ is too large, it can

[1] We stress the important difference between methodologies for "causal discovery", aimed at reconstructing a causal graph from time series, and the more general task of causal inference. As an example, given three variables $\{x, y, z\}$ and a causal graph such as $x \to y \to z$, the objective of a causal discovery method will be to discover the graph itself. Causal inference requires us to go one step further and study the effect of interventions on the graph; the variability of both variables x and y will cause the variability of z. See Ay and Polani (2008) and Runge et al. (2015) for details.

obscure significant correlations. Here, we regularize the data as follows: (i) we compute the maximum eigenvalue $\lambda_{\max}$ of the matrix $\mathbf{C}(0)$; (ii) we choose a regularization parameter $\lambda = 10^{-2}\lambda_{\max}$, resulting in covariance matrices $\mathbf{C}(0)$ with condition numbers ~ 100.

(ii) Confidence bounds. The theoretical tools presented in Section 2a require computing covariances through ensemble averages, which is impossible in climate applications due to the limitations of a single realization. The common way to overcome this is through the assumption of ergodicity, therefore replacing ensemble averages with temporal averages (Castiglione et al. 2008). Covariance matrices are then computed using temporal averages $C_{i,j}(t) = \overline{x_i(\tau+t)x_j(\tau)}$, i.e. $\overline{f}$ being the temporal average of f. This method will result in spurious responses $\mathbf{R}(t)$, because of (i) finite sample size (i.e., the length T of the trajectory is finite) and (ii) large autocorrelations of the time series $x_i(t)$. In order to identify spurious results of the response operator, we adopt the confidence bounds proposed in Falasca et al. (2024). Under the null hypothesis of a multivariate red noise process, a choice relevant for climate data (Allen and Smith 1996; Dijkstra 2013), it is possible to derive the null distribution of the response operator $\hat{\mathbf{R}}(t) = \hat{\mathbf{C}}(t)\hat{\mathbf{C}}(0)^{-1}$ with the following expected value and variance:

$$\begin{aligned}\mathbb{E}[\hat{R}_{k,j}(t)] &= \phi_k^t \delta_{k,j};\\ \mathbb{V}\mathrm{ar}[\hat{R}_{k,j}(t)] &= \frac{\sigma_k^2}{\sigma_j^2}\Big[\frac{\phi_k^{2t}-1}{T} + \frac{2}{T}\Big(\frac{1-\phi_k^t\phi_j^t}{1-\phi_k\phi_j}\Big)\\ &\quad - \frac{2\phi_k^t}{T}\Big(\phi_k\frac{\phi_j^t-\phi_k^t}{\phi_j-\phi_k}\Big)\Big]\,.\end{aligned} \tag{6}$$

where ϕ_i, T and σ_i^2 respectively represent the inferred autocorrelation, the length T of each time series $x_i(t)$ and their (regularized) variances $[\mathbf{C}_r(0)]_{i,i}$. Here, $\delta_{k,j}$ is the Kronecker delta. The symbol $\hat{f}$ specifies statistics f of the null model rather than of the original system. Finally, in the case $\phi_k = \phi_j$ we substitute the term $\phi_k\frac{\phi_j^t-\phi_k^t}{\phi_j-\phi_k}$ with $\phi_k^t t$.[2] We refer to Falasca et al. (2024) for details on the derivation of Eq. 6. In this paper, linear response formulas as Eq. (1) are computed after neglecting insignificant terms in the response operator. This work will focus on the ±1σ confidence level.

c. Dimensionality reduction

The formulas proposed in the previous section cannot be applied to the original high-dimensional system. First, as outlined before, the covariance matrix $\mathbf{C}(0)$ becomes ill-conditioned without dimensionality reduction due to high correlations between neighboring time series (e.g. $x_i(t)$ and $x_{i+1}(t)$) in the original data. Second, focusing on large scale averages is an important step of the coarse graining procedure, which leads to smooth and Gaussian-like probability densities (Majda et al. 2010a; Sardeshmukh and Sura 2009), thereby justifying the application of the theory from earlier Sections. Moreover, since the climate system resides on a low-dimensional attractor, we aim (at least in theory) to study the system in its *effective* dimensional space (Ghil and Lucarini 2020). This approach moves away from a grid-based representation of the system and instead emphasizes resolution-independent modes or patterns as fundamental components of the framework (Cvitanović et al. 2016; Dubrulle et al. 2022).

The method of spatial coarse-graining alone can have a large impact on results (e.g., Crommelin and Majda 2004; Lutsko et al. 2015; Hassanzadeh and Kuang 2016b). Climate studies using FDT overwhelmingly focused on Empirical Orthogonal Functions (EOFs) as the dimensionality reduction technique. However, Lutsko et al. (2015) and Hassanzadeh and Kuang (2016b) noted that the EOF step alone can be a major source of errors in FDT applications. Here, we are going to focus on a method recently proposed in Falasca et al. (2024). The method allows us to partition a climate field into a few regionally constrained patterns of highly correlated time series, leading to a large reduction in the number of degrees of freedom and, therefore, more robust and interpretable inference. More formally, consider a spatiotemporal field saved as a data matrix $\mathbf{x} \in \mathbb{R}^{N,T}$. N is the number of grid points and T is the length of each time series. For example, $\mathbf{x}$ could be the sea surface temperature field. The dimensionality reduction proposed in Falasca et al. (2024) offers a simple strategy to partition this N dimensional field in terms of n, non-overlapping patterns $c_1, c_2, c_3, ..., c_n$, with $n \ll N$. The methodology utilizes a community detection algorithm, Infomap (Rosvall et al. 2009). Each c_j represents a two-dimensional region defined as a *regionally constrained* set of time series with large average pairwise correlation, which we will refer to as a pattern or region interchangeably. Finally, to each region c_j, we associate a time series defined as the integrated anomaly inside, i.e. $X(c_j,t) = \sum_{i\in c_j} x_i(t)\cos(\theta_i)$, where θ_i represents the latitude at grid point i and $\cos(\theta_i)$ is a latitudinal scaling. To summarize, given a spatiotemporal field saved as a data matrix $\mathbf{x} \in \mathbb{R}^{N,T}$, the proposed framework allows us to define a new field $\mathbf{X} \in \mathbb{R}^{n,T}$, with $n \ll N$. See Appendix B and Falasca et al. (2024) for additional details.

3. Data and preprocessing

We focus on the state-of-the-art coupled climate model GFDL-CM4 (Held et al. 2019), used in recent studies on the pattern effect (e.g., Zhang et al. 2023). In addition,

[2] We do so as $\lim_{\phi_j\to\phi_k} \phi_k \frac{\phi_j^t-\phi_k^t}{\phi_j-\phi_k} = \phi_k^t \tau$.

it offers data from a long control run, necessary for trustworthy computations of the response operator in Eq. 4. The ocean component is the MOM6 ocean model (Adcroft et al. 2019) with a horizontal grid spacing of 0.25° and 75 vertical layers. The atmospheric component is the AM4 model (Zhao and Coauthors 2018a,b) with a horizontal grid spacing of roughly 100 km and 33 vertical layers. There is additionally a land component (LM4) and a sea-ice component (SIS2). Our protocol, as detailed in Section 4, can be applied to any model, although we focus on GFDL-CM4 here.

We consider three simulations of CM4: a pre-industrial control (piControl), and two idealized scenarios of CO_2 increase, namely $1pctCO_2$ and $4\times CO_2$. The piControl simulation is a 650-year-long, stationary run with constant CO_2 forcing at the preindustrial level. The $1pctCO_2$ and $4\times CO_2$ are idealized experiments simulating the climate system under a 1% CO_2 increase per year and an abrupt increase of 4 times CO_2 concentration respectively. Both the $1pctCO_2$ and $4\times CO_2$ experiments start from the preindustrial CO_2 concentration and are run for 150 years. The linear response operator $\mathbf{R}(t)$, shown in Eq. (4), is constructed using data from the piControl run. The forced experiments are used to test the method's performance. We consider two variables: sea surface temperature field (SST) and the global mean net radiative flux at the top of the atmosphere (TOA). We focus on the global mean TOA, i.e. one time series, rather than the full spatial field, as the spatiotemporal dynamics of TOA does not show large scale patterns of variability. In other words, it is a fast variable with respect to the slow SST variability and the main signal we are interested in is the global average. Note that this is a simplification compared to the current literature focusing on the whole TOA field. The TOA flux, hereafter referred to simply as "TOA", is computed as the (incoming shortwave) - (reflected shortwave) - (upward radiative longwave) fluxes [3], with all fluxes computed at the top of the atmosphere. The SST fields in each model experiment are remapped to 2.5° by 2° resolution; the original temporal resolution is 1 day, but our analysis will focus on monthly and 6-month averages, therefore excluding fast variability at the daily temporal scale and focusing on the slow dynamics (see also Appendix A).

a. Data preprocessing for the piControl run

We now consider the piControl run and perform four steps:

(i) Remove the first 50 years of the 650-year-long time series, given a short transient trend in the first few decades.

(ii) Compute and store the climatology of each time series $\overline{y(t)}$, calculated as the mean across all time steps $\mu = \overline{y(t)}$.

(iii) Remove the periodic signal given by the seasonal cycle, therefore focusing on the stationary internal variability of the system. Specifically, we remove the average value of each month from the data (e.g., from each January, we remove the average across all Januaries, from each February, we remove the average across all Februaries, etc.).

(iv) High-pass filter the data with a cut-off frequency of 10^{-1} years. This allows us to remove (a) low frequency (e.g., multidecadal) oscillations, which are only sampled a few times even in a 600-year long experiment and (b) any additional slow drift that is present in the piControl run (for example, the Southern Ocean SST shows a slow drift in the control CM4 simulation). This choice tacitly assumes that the response of the net radiative flux to SST perturbation experiments emerges inside a 10-year window. We stress that the highpass filtering step is performed only in the control run.

After this preprocessing, the resultant time series $y(t)$ have zero mean due to step (iii) and approximately meet the quasi-Gaussian assumption, largely due to the high-pass filtering in step (iv) and by considering monthly averaging (see Appendix C).

b. Data preprocessing for forced experiments

We now consider the 150-year long $1pctCO_2$ and $4\times CO_2$ runs. As in the section above, we describe the preprocessing steps for a time series $y^f(t)$ encoding either the variability of SST at a specific location or the global mean net flux at the TOA. Here f stands for "forced". We preprocess the forced data in the following way:

(i) Compute and store the climatology of each time series $\overline{y_i^f(t)}$, calculated as the mean across all time steps $\mu^f = \overline{y^f(t)}$.

(ii) Calculate anomalies relative to the seasonal cycle by removing the average value of each month from the data (e.g., from each January, we remove the average value across all Januaries, from each February we remove the average across all Februaries etc.).

(iii) Add the difference in means between the forced and control runs to retain the mean state difference despite removing seasonality; i.e., update $y^f(t) \longleftarrow y^f(t) + \mu^f - \mu$.

In the forced experiments, we remove the contribution to the radiative forcing coming from an increase in CO_2

[3] Each component of the TOA fluxes is referred in the CMIP6 Eyring et al. (2016) catalog as follows: incident shortwave: "rsdt"; reflected shortwave: "rsut"; the upward radiative longwave flux: "rlut".

concentration alone, allowing us to isolate and study the TOA radiative feedback to changes in SSTs. The global mean change in radiative forcing driven by changes in CO_2 alone scales approximately logarithmically with its concentration (Pierrehumbert 2010; Romps et al. 2022). As is standard, we remove a constant of $8Wm^{-2}$ from the TOA flux in the 4×CO_2 experiment (Romps et al. 2022; Zhao 2022). We then remove a time-dependent correction of the form $\alpha \log[C_t/C_0]$ in the 1pctCO_2 run, where C_0 and C_t are the concentration of CO_2 in the control run and forced experiment, respectively. We fit the α parameter by the change of $\sim 8Wm^{-2}$ in the 4×CO_2 simulation. Thus, we have an additional step:

(iv) Remove the contribution to radiative forcing coming from the CO_2 concentration alone.

Main assumptions and limitations. The underlying assumption behind the preprocessing procedure is that the SST field and global mean TOA variables, coarse-grained in space and time, are the "proper" variables to study the pattern effect. The integrated effect of processes active at (i) small spatial scales and (ii) fast time scales will be considered as noise (Penland 1996). These are large simplifications, leading us to consider only two variables. As we will stress in the paper, this represents a more significant simplification in the case of the "coupled" formulation of the protocol, i.e. diagnosing how perturbations propagate throughout the system across spatial and temporal scales, compared to the more traditional "atmosphere-only" formulation. The limitations and strengths of these choices will be further discussed in the conclusions and in detail in Appendix A. Future studies will focus on adding more variables and different coarse-graining methods, but we believe this study offers a valuable starting point for future perturbation experiments in the coupled system.

4. Practical implementation of the proposed framework

We now present the steps towards the practical implementation of the theoretical framework in Section 2. The main steps can be summarized by three main points: given the original, high-dimensional fields, (i) project the dynamics in a lower-dimensional representation; (ii) compute the response formulas in the low-dimensional space; (iii) project the results back to the original, high-dimensional space.

As outlined in Section 3, we are considering the SST field $\mathbf{y}^{\mathrm{SST}} \in \mathbb{R}^{N,T}$ and the global mean net radiative flux at the TOA, $y^{\mathrm{TOA}} \in \mathbb{R}^{T}$. N denotes the number of grid points and T is the length of the simulation. The linear response operator $\mathbf{R}(t)$ is inferred in the piControl run. In order to compute response formulas, we proceed as follows:

(i) We run the dimensionality reduction method presented in Section 2c for the $\mathbf{y}^{\mathrm{SST}}$ field. The identified patterns represent proxies for modes of variability. This step reduces the N dimensional field $\mathbf{y}^{\mathrm{SST}}$ into n, *regionally constrained* patterns c_j. The identified patterns are shown in Figure 1.

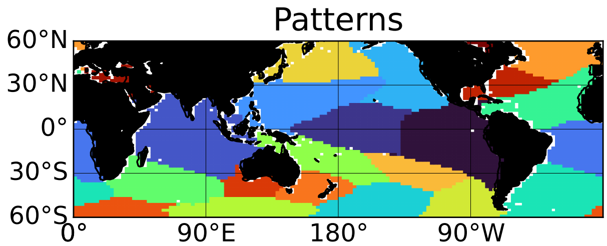

FIG. 1. Low-dimensional representation of the sea surface temperature (SST) field. Different colors are used to distinguish different patterns. The observables considered in this study are defined as the integrated SST anomalies in each pattern and by the global mean net radiative flux at the TOA.

(ii) Given each pattern c_j, we compute its SST time series, $X^{\mathrm{SST}}(c_j,t)$, by computing the integrated SST anomaly inside it as:

$$X^{\mathrm{SST}}(c_j,t) = \sum_{i \in c_j} y_i^{\mathrm{SST}}(t) \cos(\theta_i), \tag{7}$$

where θ_i is the latitude at grid point i. The inferred time series, $X^{\mathrm{SST}}(c_j,t)$, as well as the global mean TOA, y^{TOA}, are well approximated by Gaussian distributions (see Appendix C), justifying the use of the approximation presented in Eq. (4).

(iii) The new field $\mathbf{X}^{\mathrm{SST}} \in \mathbb{R}^{n,T}$ and the global mean TOA y^{TOA} are then concatenated to form the state vector $\mathbf{x} \in \mathbb{R}^{n+1,T}$. At time step t, the state of the reduced order system is then encoded in the state vector $\mathbf{x}(t)$. We then compute the regularized covariance matrix $\mathbf{C}_r(0)$ through the regularization procedure and neglect spurious responses at a $\pm 1\sigma$ level, see Section 2b.

(iv) The evaluation step requires us to reconstruct the TOA flux in idealized forced experiments as a function of temperature changes and the response operator $\mathbf{R}(t)$. In the case of the "coupled" formulation, we project the SST perturbation, $\delta\mathbf{T}(t)$, in the low-dimensional space in the same way as for $\mathbf{y}^{\mathrm{SST}}$ in point (ii). We then form the perturbation pattern by concatenating $\delta\mathbf{T}(t)$ with a zero forcing pattern in the degree of freedom related to the TOA net flux.

(v) Response formulas (Section 2) are computed for the low-dimensional system $\mathbf{x} \in \mathbb{R}^{n+1,T}$ using Eq. (1) or

(2). The resulting response is referred to as $\delta\langle \mathbf{x}(t)\rangle \in \mathbb{R}^{n+1}$. In this paper we are interested in computing global average responses of the TOA field. To do so, we consider the last entry of response $\delta\langle \mathbf{x}(t)\rangle$, denoted here as $\delta\langle x_{\text{TOA}}(t)\rangle$.

We note that the SST and TOA fields are associated with different units and magnitudes. Therefore, the numerical implementation of the protocol benefits from a standardization step, where the response operator and the responses (with Eq. (1) or (2)) are computed in a standardized space. Standardization is performed with the standard deviations of the piControl run for each respective variable.

"Sensitivity map" metric. A key metric for understanding the pattern effect is to examine the sensitivity of changes in the global mean net radiative flux at the TOA to local SST perturbations; this is generally approached through "sensitivity maps". In previous studies, (e.g., Zhang et al. 2023; Bloch-Johnson et al. 2024), sensitivity maps have been estimated by (i) applying a step function perturbation in the SST field at grid point i, e.g., a constant $1K$ for $t > 0$, (ii) computing the change in global mean TOA fluxes after equilibration, (iii) plotting this value at each grid point i. These maps show values in units of $[W/(m^2K)]$: positive sensitivity corresponds to positive global radiative feedbacks to local SST forcings. Such positive feedbacks decrease the upward radiative response, amplifying the initial global mean temperature changes. The opposite scenario is true for negative sensitivity. In the case of the coarse-grained representation of the system where our variables are not at the grid scale, we define an equivalent metric as follows: given a pattern c_j, with $j = 1, ..., n$, we prescribe a step function perturbation of 1 Kelvin $[K]$ in each grid point i belonging to c_j. The total perturbation in the j^{th} pattern is equal to its area (dimensionality of $[K]$):

$$\Delta T_j = \sum_{i \in c_j} \cos(\theta_i); \text{ for } t > 0. \tag{8}$$

Where θ_i is the latitude at grid point i. The perturbation field is defined as $\delta\mathbf{T} \in \mathbb{R}^{n+1,T}$, with the j^{th} term equal to ΔT_j and all other entries equal to zero. We then compute the linear response in global mean net flux $\delta\langle x_{\text{TOA}}(t)\rangle$. The sensitivity map $\mathbf{S}_t \in \mathbb{R}^N$ is a gridded map of the same dimensionality N as the original space. The map is defined by plotting at each grid point i belonging to pattern c_j, the same value, defined as the global mean TOA response caused by the perturbation ΔT_j:

$$S_{t,i} = \frac{\delta\langle x_{\text{TOA}}(t)\rangle}{\Delta T_j}; \ \forall i \in c_j, \tag{9}$$

The subscript t denotes the time scale up to which the responses are computed. As discussed later in Section 5, studying the "atmosphere-only" formulation requires us to focus on the shortest time scales, in this case defined by $t = 1$ month. In this case, the link SST → TOA is quasi-instantaneous in agreement with the standard view of feedbacks. In contrast, in the "atmosphere-ocean coupled" formulation, we are interested in the equilibrated response to a step function perturbation, obtained by computing the integral in Eq. (2) for $t \to \infty$. In practice, the equilibrated response can be computed by setting an upper bound of the integral in Eq. (2) that is much larger than the characteristic time of the response; in our case, we take $\tau_\infty = 10$ years, as data are high-passed filtered with a cut-off frequency of 10^{-1} years (see Section 3), so we do not expect any variability beyond some statistical noise at time scales longer than a decade.

5. Pattern effect in the "atmosphere-only" system

The traditional formulation of the pattern effect links the change in SST, $\Delta T_i(t)$ at grid point i and time t, to the global mean response in TOA flux, $\Delta\overline{\text{TOA}}(t)$, at the same time t. The "Green's function protocol" (Bloch-Johnson et al. 2024) defines a standardized protocol to infer such linkages through perturbation experiments in atmospheric-only models forced by SST boundary conditions. Therefore, the SST field is prescribed and cannot respond to perturbations from either atmospheric or oceanic processes as it would in a true coupled system. In such a setup, the atmospheric model equilibrates to an SST step function perturbation pattern very quickly, and the feedbacks can be considered as instantaneous. More formally, from a data analysis point of view we can consider the feedback as instantaneous when the temporal resolution is coarse enough to be larger than the characteristic time scale of the response. Feedbacks are encoded in $\partial\overline{\text{TOA}}/\partial \text{T}_i$, $\overline{\text{TOA}}$ representing the global mean net flux at the TOA and T_i the SST at the grid point i. We will show that the same approach is contained in our method without the need for expensive model runs.

We consider the response of the global mean TOA flux to SST perturbations at the shortest time scale $t = 1$ month. Assuming a fast response of the atmosphere compared to the ocean, the subsequent SST change at $t = 1$ month will be small and corresponds to a system without an active atmosphere-ocean coupling such that the SST can be regarded as an imposed boundary condition. This is verified by considering the sensitivity map shown in Figure 2(a) computed for $t = 1$ month, which shows strong qualitative similarities to previous approaches involving atmosphere-only models (Bloch-Johnson et al. 2024; Zhang et al. 2023). Specifically, we note the resemblances between our results and Kang et al. (2023) who used Ridge Regression (compare our Figure 2(a) to Extended Data Fig. 7(b) in Kang et al. (2023)). Our work focuses on seasonal anomalies, and it complements previous work, which focused on annual mean deviations from a global average. A distinct feature of sensitivities at short time scales is a dipole in the tropical Pacific Ocean, with marked negative values on the western side of the basin and positive values on the eastern side, see for example Dong et al. (2019). This feature is well captured by our sensitivity map in Figure 2(a). The physical mechanisms behind such feedbacks has been described, for example, in Guilyardi et al. (2009); Williams et al. (2023). The positive sensitivity in the eastern Pacific is associated with the high climatology of low-level clouds. A surface warming leads to a decrease in cloud cover, leading to positive anomalies in incoming shortwave radiation. However, negative sensitivity is commonly found in regions of deep convection, such as the western Pacific. An additional warming in these regions travels vertically through the troposphere and horizontally through gravity waves, strengthening the inversion layer in regions with high climatological low-cloud amount. This leads to larger cloud cover in regions such as the eastern Pacific, further increasing the reflected shortwave at the TOA.

The change in global mean TOA fluxes at time t ($\Delta\overline{\mathrm{TOA}}(t)$) is retrieved as a dot product of the sensitivity map $\mathbf{S}$ with changes in the temperature pattern. This is given by $\Delta\overline{\mathrm{TOA}}(t) = \mathbf{S}_1\Delta\mathbf{T}(t) = \sum_i S_{1,i}\Delta T_i(t)$ (each $\Delta T_i(t)$ weighted with the latitudinal weight $\cos\theta_i$), which mirrors Eq. 5 in Bloch-Johnson et al. (2024). As is standard, we test the framework by reconstructing the change in the global mean net radiative flux at the TOA at time t given the change in the SST field $\Delta\mathbf{T}$ in the 1pctCO2 and 4xCO2 forced experiments introduced in Section 3). Such reconstructions are computed using monthly anomalies and then shown as yearly averages in Figure 2(b,c). The analysis shows a good reconstruction of the global mean TOA flux at each time t given the change in SST at the same time t. Quantitatively, for the 1pctCO2 run, the trend in the global mean TOA net flux has been found to be $-0.042\ Wm^{-2}yr^{-1}$ while the trend in the reconstructed response is $-0.06\ Wm^{-2}yr^{-1}$. The good skills in reconstructing the TOA response even with such a simple linear method is here ascribed to the coarse-graining procedure. Specifically, given the temporal and spatial resolutions considered, i.e. 1 month and 2.5° by 2°, focusing on large areas in the ocean as the patterns in Figure 1 allows us to effectively linearize the mapping between changes in temperature $\Delta T_i(t)$ and in global mean TOA fluxes $\Delta\overline{\mathrm{TOA}}(t)$. An additional test for the methodology is to compute the correlation between the reconstruction and the simulated global mean TOA net flux. We remove a trend in both TOA time series and compute their correlation coefficient. The results are shown in Figure 3: The correlation coefficient is relatively well captured by this methodology as $r = 0.51$ and $r = 0.69$ for the reconstruction of 1pctCO2 and 4xCO2, respectively. We note that the analysis in this Section computed the sensitivity maps from the 1 month-lag response operator $\mathbf{R}(1) = \mathbf{C}(1)\mathbf{C}(0)^{-1}$ at a $\pm 1\sigma$ confidence level. Importantly, we underscore that no tuning has been used for this experiment and better predictions can be obtained by tuning the size of the regions. However, here we emphasize simplicity and interpretability (i.e. focusing on a few components of the system) rather than more complex models that lead to better evaluation procedures (see also discussion in Held (2005)).

To conclude, the method presented in this section provides a fast and practical tool for studying the pattern effect based on the paradigm of responses to perturbations, akin to approaches in the literature using atmosphere-only models (e.g., Bloch-Johnson et al. (2024); Zhang et al. (2023)), but relying solely on a long piControl run of a climate model, avoiding the need for computationally expensive integrations.

6. Pattern effect in the "atmosphere-ocean coupled" system

a. Calibration and evaluation

In this Section, we consider a new idealized experiment, where we impose step function perturbations in the SST field in the coupled climate system and study the response of the net flux at the TOA. Such a perturbation experiment has not yet been performed with a climate model, but it can be explored within our proposed framework under the assumptions listed in Section 2 and Appendix A. In this case, the SST field is not merely a boundary condition, but an active component of the coupled system: at longer time scales, local SST perturbations can feedback over remote ocean regions through coupled dynamics, e.g. via "atmospheric bridge"-type mechanisms (Chiang and Sobel 2002). Therefore, capturing the appropriate time scales of signal propagation is necessary to accurately build our reduced-order dynamics. Here, the system is viewed through the lens of coarse-grained dynamics, after averaging the SST and TOA flux variables over large regions. However, reducing the spatial dimensionality affects the effective dynamics of the low-dimensional system, which may require a different averaging time scale than the original one-month resolution. We utilize a practical solution by using the two independent forced experiments, i.e., the 1pctCO2 and 4xCO2 runs. We inferred the response operator $\mathbf{R}(t)$ in the control run after re-processing the data with a reasonable range of temporal resolutions ranging from 3 to 6 months. We tested each implementation against the 4xCO2 simulation and found good agreement for 6-month averages; we therefore chose this temporal resolution. Choosing different temporal resolution mainly affects the magnitude of the mean response, e.g. whether the TOA response after 100 years is $\sim -6W/m^2$ or $\sim -15W/m^2$, but it does not impact the year-to-year variability predictions.

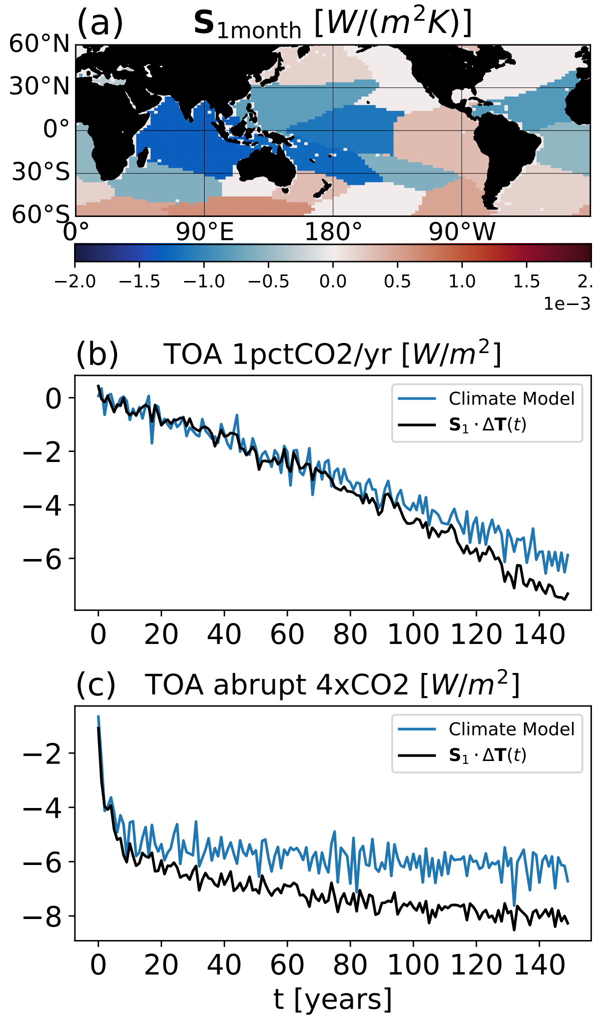

FIG. 2. Pattern effect in the "atmosphere-only" system. Panel (a): Sensitivity maps $\mathbf{S}_1$ computed only for the 1 month response (Section 4). Physically, focusing on the shortest time scales allows us to approximate the static sensitivity map in the absence of atmosphere-ocean coupling. At each grid point i, we plot the global average response of the net radiative flux at the TOA given a step function SST perturbation of 1 Kelvin imposed at that point. Positive sensitivity corresponds to a positive radiative feedback, therefore amplifying the initial global mean temperature changes; the opposite is true for negative sensitivity. Panel (b,c): Given the sensitivity maps $\mathbf{S}_1$ shown in Panel (a) and the sea surface temperature field $\Delta\mathbf{T}(t)$ at time t in the 1pctCO2 and 4xCO2 forced experiments, we compute the dot product $\Delta\overline{\text{TOA}}(t) = \sum_i S_{1,i}\Delta T_i(t)$. Results are shown as yearly averages. This analysis is akin to the one proposed in the "Green's function protocol", as in Zhang et al. (2023); Bloch-Johnson et al. (2024) and valid for atmosphere-only models where the sea surface temperature is a boundary condition. The response operator has been inferred using the statistical bounds in Section 2 at the ±1σ level.

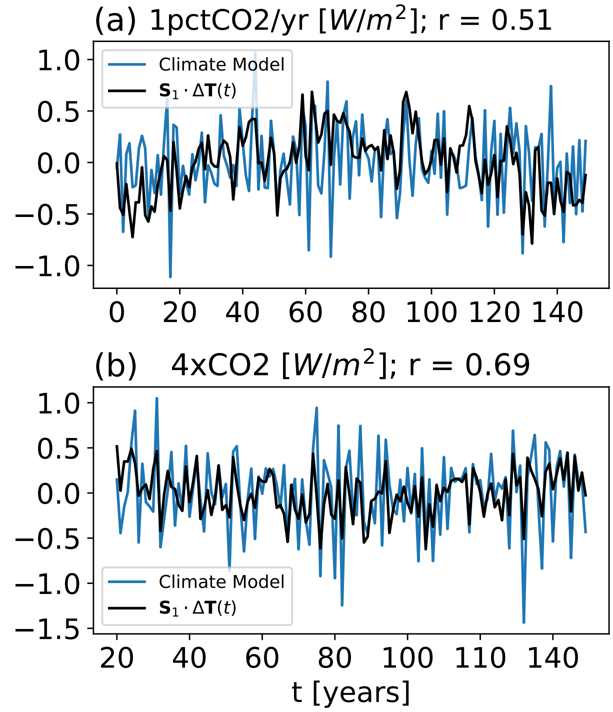

FIG. 3. Panel (a): Detrended yearly global mean changes in net radiative flux at the TOA as reconstructed by the regression strategy $\mathbf{S}_1\Delta\mathbf{T}(t)$ and simulated by the climate model in the 1pctCO2 run. Panel (b): same as panel (a) but for the the 4xCO2 run after removing the first 20 years. The correlation coefficient r between the reconstruction and the model simulation is reported.

We refer to this step as a calibration step. We refer to Section 7 for additional details on this calibration step. As shown later on, this calibration leads to good performance on the (independent) 1pctCO2 run, thus giving confidence in this data-processing step. Therefore, given the large area of the patterns in Figure 1, the effective time scales to capture time-dependent responses to external perturbations is ~ 6 months. In Figure 4 we show the reconstruction of the simulated mean radiative flux at TOA in both forced experiments. For the 1pctCO2 run, the model global mean TOA net flux trend is $-0.042\ Wm^{-2}yr^{-1}$, which is well captured by our reconstruction which is estimated at $-0.043\ Wm^{-2}yr^{-1}$. As in the previous Section, we compute the correlation of the reconstructed and simulated TOA signals after detrending (Figure 5). The correlation coefficients are $r = 0.68$ and $r = 0.73$ for the 1pctCO2 and 4xCO2 reconstruction, respectively. As opposed to the atmosphere-only formulation, the variability of the predicted TOA is increased with respect to the model's simulation in this coupled atmosphere-ocean formulation. The 6 months averages for the calibration could be a potential limitation of the framework, as it averages short lags responses. Considering shorter time scales might be possible by (i) reducing the spatial resolution of the patterns and (ii) adding new state variables relevant at higher frequency and small spatial scales. These considerations are further detailed in Appendix A.

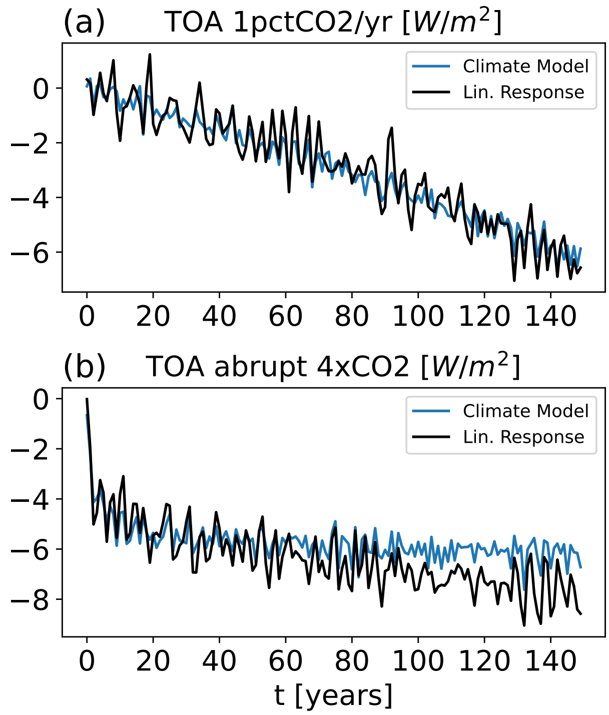

FIG. 4. Pattern effect in the "atmospheric-ocean" coupled system. Panel (a,b): Changes in global mean radiative flux at the top of the atmosphere (TOA) in the 1pctCO2 and 4xCO2 experiments. The net flux at the TOA as simulated by the fully coupled GFDL-CM4 model is shown in blue. A prediction of TOA by linear response theory solely as a function of the changes in the SST field is in black. Linear responses have been computed using the convolution in Eq. (1): the response of TOA at time t is then computed as the integrated effect of the perturbation patterns across all previous time scales $t-\tau$. Results are shown as yearly averages.

b. A sensitivity map for the coupled system

We now address the following question: What is the cumulative (in time) causal relationship between warming perturbations and changes in the global mean radiative flux at the TOA in a coupled system? In Figure 6, we show the equilibrated, cumulative (in time) response of the net TOA flux to a step function perturbation of 1 Kelvin at each grid point in the SST field. Here, $t = 5$ to 10 years is considered long enough for the atmosphere-ocean coupled system to equilibrate to SST perturbations, see Appendix D. The sensitivity map agrees with previous studies on the large negative sensitivity in the western Pacific but clearly differs in terms of sensitivity of the eastern Pacific, see (e.g., Bloch-Johnson et al. 2024; Kang et al. 2023). Positive sensitivities (responses) are found in the Indian Ocean, the North Tropical Atlantic, and interestingly, in higher latitudes regions such as the Southern Ocean, the North Pacific, and the North Atlantic. In particular, the North Atlantic carries the largest positive sensitivity, and such sensitivity is found only at longer time scales (around 5 years). However, if the statistical significance is increased to ±3σ, only the tropical domains (and the North Pacific) have non-zero sensitivity, see Appendix E. The positive sensitivity in high-latitudes regions is interesting as it differs to the traditional analysis of the pattern effect but will be left for future studies. We will concentrate mainly on the robust negative sensitivity in the eastern Pacific.

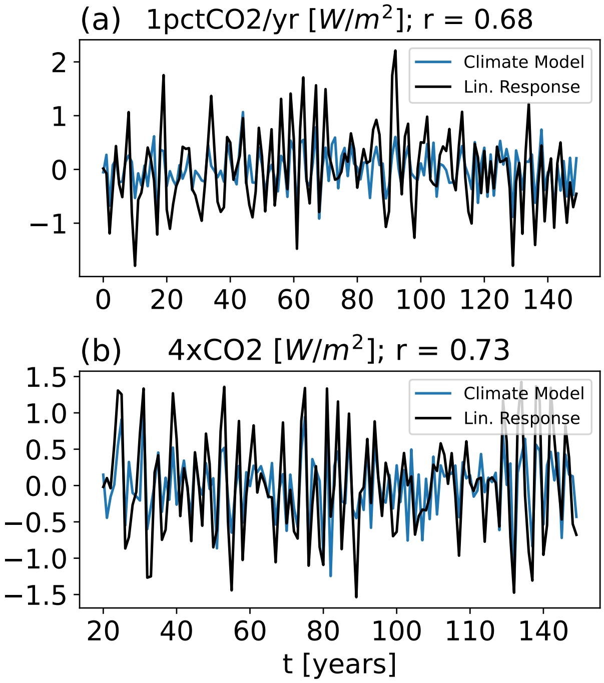

FIG. 5. Panel (a): Detrended yearly global mean changes in net radiative flux at the TOA as reconstructed by the convolution integral in Eq. (1) and simulated by the climate model in the 1pctCO2 run. Panel (b): same as panel (a) but for the the 4xCO2 run after removing the first 20 years. The correlation coefficient r between the reconstruction and the model simulation is reported.

The sensitivity map shown in Figure 6 differs from the one in 2(a) by considering longer time scales in the estimation of the response operator and therefore captures the coupled dynamics. A temperature perturbation pattern prescribed at the ocean surface impacts the radiative balance at both the TOA and surface on very short time scales. However, over longer periods, an initial warming localized to a specific region can alter dominant modes of variability, redistributing heat globally through coupled dynamics and leading to additional warming and cooling patterns across a broader area than the original perturbation region.

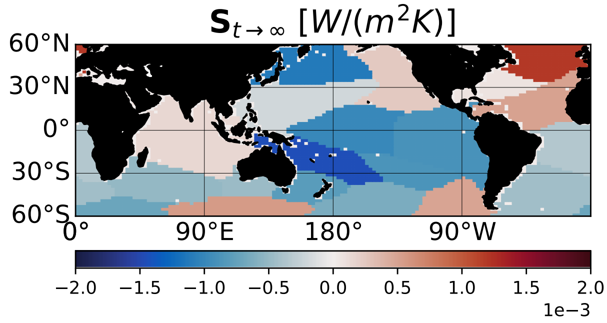

FIG. 6. Sensitivity map such that at each grid point i, we plot the equilibrated, global average response of the net radiative flux at the TOA given a constant perturbation of 1 Kelvin in SST imposed at that point. Positive sensitivity corresponds to a cumulative (in time) positive radiative feedback, therefore amplifying the initial global mean temperature changes; the opposite is true for negative sensitivity. The time $t \to \infty$ is here approximated as $t = 10$ years, as all data have been high-pass filtered with a 10^{-1} years cut-off frequency.

In other words, the warming caused by an external perturbation will evolve over time as a result of the coupled dynamics. These cumulative changes can lead to a larger outgoing longwave emission, or short wave feedbacks, than would occur if the warming remained confined to its initial location. Although building a mechanistic understanding of such a cascade of feedbacks is challenging, the response theory formalism integrates these effects across temporal and spatial scales when diagnosing radiative feedbacks (Ghil and Lucarini 2020).

A few insights on the response to eastern Pacific warming. We further examine changes in radiative fluxes to eastern Pacific warming through a lag-covariances analysis. We stress that covariance does not imply causality, but combined with the causal analysis in the previous section, it can provide additional insights into the time-dependent feedbacks associated with eastern Pacific warming. A constant, external warming applied to the eastern Pacific region drives the system into an El-Niño-dominated state, weakening the temperature gradient across the equatorial Pacific and enhancing deep convection towards the central-to-east Pacific. Additionally, at longer time scales, the persistent local heating over the region forces atmospheric wave responses (Alexander et al. 2002; Chiang and Sobel 2002) or, at much longer time scales, an oceanic response (Bracco et al. 2005; Wang 2019), therefore leading to a heat redistribution over remote ocean basins, changes in cloud cover far from the tropical Pacific Ocean and potentially feedback onto the net radiative flux. To further analyze such changes in net radiative fluxes at the TOA we proceed as follows: (i) we consider the signal (Eq. (7)) of the eastern Pacific pattern in Figure 1; (ii) we compute the lag-covariance between the signal and time series of the radiative fluxes at the grid level. We focus on (i) the net flux at the TOA, here referred to as TOA and positive downward (i.e., towards the Earth's surface); (ii) the reflected shortwave radiation, here referred to as SW_{up} and (iii) the outgoing longwave radiation, here referred to as LW_{up}. Both SW_{up} and LW_{up} are positive upward (i.e. leaving the planet). As in Section 3, we are dealing with anomalies, such that the incoming shortwave radiation is zero, and therefore, we have $TOA = -SW_{up} - LW_{up}$. We refer to the covariance between the eastern Pacific signal and radiative fluxes at a lag τ respectively as $C(EP,TOA,\tau)$, $C(EP,SW_{up},\tau)$ and $C(EP,LW_{up},\tau)$. The following decomposition holds: $C(EP,TOA,\tau) = -C(EP,SW_{up},\tau) - C(EP,LW_{up},\tau)$. We show the results in Figure 7 for three lags, where the lag $\tau = 0$ means synchronous correlations and $\tau = 1,2$ indicates a lead of eastern Pacific towards the radiative fluxes. We remind the reader that the data are saved as 6-months averages. Let us first focus on the $\tau = 0$ analysis. The shift in deep convection from the warm pool region towards the central and east Pacific during an El Niño is captured by the negative covariances with LW_{up} in Figure 7(a). The resulting change in the Walker circulation leads to anomalous sinking in the western Pacific, decreased cloud cover, and, therefore, positive anomalies in LW_{up}. As the eastern Pacific warms, there is a decrease in the stratocumulus deck west of South America, leading to positive anomalies in LW_{up}. As the spatial distribution of outgoing longwave radiation changes, so thus the reflected shortwave radiation: an increase in deep convection leads to a decrease in outgoing longwave radiation but also to an increase in reflected shortwave radiation, as shown in Figure 7(d). In general, the changes in reflected shortwave radiation during an El Niño episode are of the opposite sign to the ones in outgoing longwave radiation. The balance between changes in SW_{up} and LW_{up} defines the net radiation change at the TOA and is shown in Figure 7(g). The net change in radiation shows a positive anomaly in incoming radiation in the eastern Pacific and the equatorial Pacific region. Negative anomalies are seen in the north and south of the equatorial Pacific region, as well as in the west Pacific and Indian and tropical Atlantic basins. At lag $\tau = 1$, the covariance between the eastern Pacific signal and the net TOA radiative fluxes are mostly negative, as shown in Figure 7(h). Comparison between Figures 7(b) and 7(e) implies that the negative anomalies in net radiation at longer time scales result from a larger outgoing longwave radiation compared to the incoming shortwave radiation. At longer time lag $\tau = 2$, the main change in net radiation remains negative (Figure 7(i)) for the same reason. Such negative changes intensify up to $\tau = 4$ (not shown), before fading away after. Therefore, the integrated negative changes in net incoming radiative fluxes following a warming of the eastern Pacific can be, in part, explained as a result of a larger cumulative (in time) emitted longwave radiation across the tropical oceans compared to the shortwave radiation feedback.

As explained throughout the paper the framework focuses on time-dependent feedbacks that are difficult to address with a covariance analysis only. Consequently, the interpretation above, while useful, remains incomplete. A growing literature has recently investigated the complex, time-dependent linkages among modes of variability in the tropical Pacific, Indian, and Atlantic Oceans and it could serve as future guidance for building mechanistic understandings of the analysis discussed in our work; see, for example, Cai et al. (2019); Wang (2019); Zhao and Capotondi (2024). To give a practical example, let us consider the observed positive sensitivity in the Indian Ocean (IO) in Figure 6. It has been recently recognized that the IO variability can influence both the tropical Pacific and Atlantic through atmospheric and oceanic pathways. For example, a warming projected towards the Indian Ocean Basin mode (Klein et al. 1999) results in enhanced convection over the region. This instability generates eastward-propagating Kelvin waves, which, in turn, can further intensify easterly wind anomalies in the western Pacific, driving the variability in the basin into a La Niña state (Cai et al. 2019; Wang 2019). As shown, a warming in the eastern Pacific is associated with a negative sensitivity. It follows that the cooling of the tropical Pacific, as driven by an Indian Ocean warming can lead to a net positive response of the global mean radiative flux at the TOA.

7. Atmospheric-only and coupled system formulations: conceptual and practical differences

We have introduced a comprehensive protocol for studying the effects of idealized warming perturbation experiments from data alone. By examining responses at the shortest time scales or across multiple time scales, these experiments can be framed as atmospheric-only or coupled climate models formulations. Our proposed framework encompasses both approaches, with the primary distinction arising from differences in the formulation of climate feedbacks.

a. Conceptual differences across approaches: climate feedbacks

In the "Green's function protocol" (Bloch-Johnson et al. 2024; Zhang et al. 2023; Dong et al. 2019), the data-driven sensitivity studies (Bloch-Johnson et al. 2020; Kang et al. 2023) and the "atmosphere-only" formulation of our protocol (Section 5) radiative feedbacks are encoded in a sensitivity map. In the literature, sensitivity maps are commonly referred to as the radiative feedback and are defined at each grid point i by $\frac{\partial \overline{\mathrm{TOA}}}{\partial \mathrm{T}_i}$, where $\overline{\mathrm{TOA}}$ is the globally averaged net radiative flux at the TOA and T_i is the SST at point i. This perspective is theoretically motivated in the framework of energy balance models by Taylor expanding the radiative response over the steady state temperature (see e.g. Section 2 of Meyssignac et al. (2023)). The dot product of a sensitivity map with an imposed SST pattern at time t will return an estimate of the globally averaged change in net flux $\overline{\mathrm{TOA}}$ at the same time t. This means that an imposed change in SST patterns at time t cannot drive changes in average TOA fluxes at later times $t+\tau$. Naturally, neither our approach (see Section 5) nor the "Green's function protocol" (Bloch-Johnson et al. 2024) involves a truly instantaneous feedback (i.e., equal to zero at lag $\tau = 0$). Instead, the term "instantaneous" refers to the radiative response to imposed SST perturbations occurring on time scales much shorter than the evolution of the SST itself. We note that this is standard in the analysis of climate feedbacks, and is not restricted to the pattern effect (Taylor et al. 2006).

The main object of our proposed protocol is the linear response operator, defined as $R_{k,j}(\tau) = \frac{\delta\langle x_k(t+\tau)\rangle}{\delta x_j(t)}$ in the limit of $\delta x_j(t) \to 0$. In the general formulation of our framework, where responses are not limited to the shortest time scales, x_k in $\delta\langle x_k(t+\tau)\rangle$ can be a component of the SST or the TOA flux field, responding to a perturbation $\delta x_j(t)$ imposed to the SST or the TOA flux field. Therefore every variable can respond to perturbations in any other variable. Additionally, the operator $R_{k,j}(\tau)$, i.e., the impulse Green's function defined in our framework, is time-dependent as in a physical system, a perturbation in one variable will propagate through the system and impact another variable at later times. The response to external perturbations at time t is then computed as the integrated effect of the perturbation patterns across all previous times $t-\tau$ (Christensen and Berner 2019). In this case, feedbacks are spatially and temporally dependent and encoded in the response operator. This framework allows us to diagnose responses from imposed external perturbations and from the internal time-dependent responses within the coupled system, which further influence radiative feedbacks at later times. The sensitivity map in Figure 6 then represents the cumulative in time feedback (response) to a perturbation. These different perspectives on feedbacks have been mentioned in the climate literature, see for example both the Introduction and Section 2a in Taylor et al. (2006). However, response theory provides both a rigorous mathematical foundation and a practical computational tool for their analysis.

b. Practical differences across approaches: next time step prediction vs cumulative effect of perturbations

The response, using the coupled formulation in its discrete form, is given by $\delta\langle x_k(t)\rangle$ to time-dependent perturbations by $\sum_j \sum_{\tau=0}^{t} R_{k,j}(\tau)\delta x_j(t-\tau) = \sum_j \Big(R_{k,j}(0)\delta x_j(t) + R_{k,j}(1)\delta x_j(t-1) + R_{k,j}(2)\delta x_j(t-$

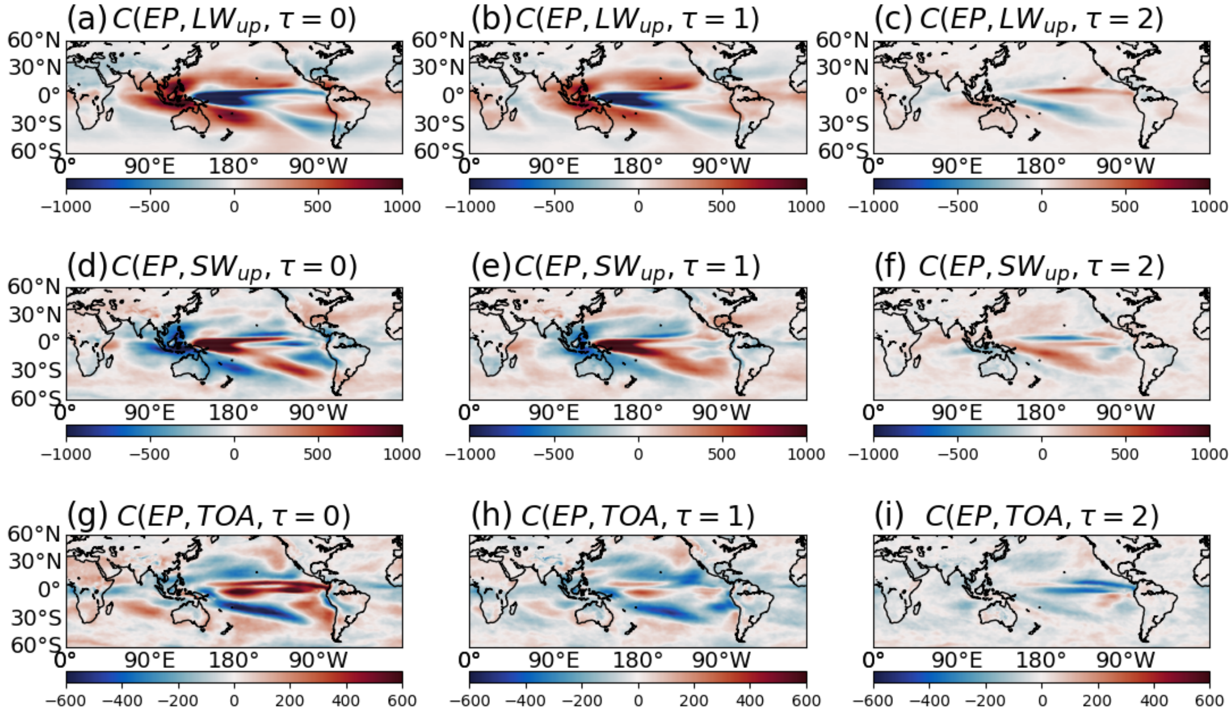
(a) $C(EP, LW_{up}, \tau = 0)$
(b) $C(EP, LW_{up}, \tau = 1)$
(c) $C(EP, LW_{up}, \tau = 2)$
(d) $C(EP, SW_{up}, \tau = 0)$
(e) $C(EP, SW_{up}, \tau = 1)$
(f) $C(EP, SW_{up}, \tau = 2)$
(g) $C(EP, TOA, \tau = 0)$
(h) $C(EP, TOA, \tau = 1)$
(i) $C(EP, TOA, \tau = 2)$
60°N
30°N
0°
30°S
60°S
0°
90°E
180°
90°W
−1000
−500
0
500
1000
−600
−400
−200
0
200
400
600

Fig. 7. In the first row: lag-covariance in $[KW/m^2]$ between the eastern Pacific (EP) signal, computed via Eq (7) and the outgoing longwave radiation, LW_{up}, for $\tau = 0, 1, 2$. Second and third rows: same as the first row but for the reflected shortwave radiation SW_{up} and the net radiative flux at the TOA, TOA. As all quantities are anomalies, the incoming short wave radiation at the TOA is zero and we have $TOA = -SW_{up} - LW_{up}$. The exact following decomposition holds: $C(EP, TOA, \tau) = -C(EP, SW_{up}, \tau) - C(EP, LW_{up}, \tau)$. Note that the fields are saved at 6-months averages. Negative anomalies of net TOA are seen in the tropical until $\tau = 4$ and fade away after then. Note that the EP signal is computed as the integrated temperature anomaly, not the average, inside the EP region in Figure 1, leading to the large absolute values of covariances seen in the colorbars.

$2) + ...\Big)$. Instead, the standard formulation of feedbacks is encoded in the sensitivity map at the shortest time-scale (of 1 month), namely $\mathbf{S}_1$. To evaluate the framework, we focused on reproducing changes in the global mean radiative flux at the TOA, expressed as $\Delta\overline{\text{TOA}}(t) \sim \mathbf{S}_1\Delta\mathbf{T}(t) = \sum_i S_{1,i}\Delta T_i(t)$, where $\Delta\mathbf{T}(t)$ are gridded maps of temperature changes. Strictly speaking, the relationship should be $\Delta\overline{\text{TOA}}(t+1) \sim \mathbf{S}_1\Delta\mathbf{T}(t)$, but (a) at monthly resolution, we empirically observe that $\Delta\overline{\text{TOA}}(t+1) \sim \Delta\overline{\text{TOA}}(t)$, and (b) this distinction becomes less important when $\mathbf{S}_1$ is applied to annual temperature changes, as is common in many current studies. This approach can be reformulated in the low-dimensional space by studying the change in the net flux at the TOA at time $t+1$ as $\sum_j R_{k,j}(1)\delta x_j(t)$, which, at monthly resolution, is nearly equal to the first term in the corresponding discrete convolution above $\sum_j R_{k,j}(1)\delta x_j(t-1)$. In this case the variable x_k represents the global mean TOA.

Thus, there is an equivalence between the traditional evaluation step $\sum_i S_{1,i}\Delta T_i(t)$ and a 1-step prediction $\sum_j R_{\text{TOA},j}(1)\delta x_j(t-1)$. In other words, in our formulation of an "atmospheric-only" protocol, there is an exact equivalence between the traditional feedback analysis and a Linear Inverse Model prediction (Penland 1989). This further underscores how the conceptual distinctions in the formulation of feedbacks influence the practical implementation of the protocol.

Finally, the distinctions in feedback formulations can be well illustrated through the framework of causal inference. Consider a three-dimensional climate system represented as $x \longrightarrow y \longrightarrow z$. Here, x and y represent the variability in sea surface temperature in two distinct regions, and z the variability of net radiative fluxes at the TOA. The arrows indicate a causal association: $y \longrightarrow z$ implies that a change in temperature in y directly impacts the TOA flux z. Following the traditional view of feedbacks, the main radiative feedback will be linked to the direct causal link $y \longrightarrow z$. This standard approach is closely aligned with prediction tasks, as demonstrated earlier in this section: information on the variability of y at time t is sufficient to predict z at the next time step. In contrast, the perspective advanced here and in Lucarini (2018) emphasizes the study of how external perturbations propagate along the causal graph. In our simple example, changes in the radiative flux z are directly influenced by SST variability y. However, a change in temperature in region x also impacts y, subsequently affecting z. This second *indirect* link between x and z is captured at longer time scales. This view links feedbacks to the time-dependent "flow of information" within a physical system (Ay and Polani 2008), i.e. both direct and indirect causal links, rather than focusing solely on predictions.

If the goal is to reconstruct TOA fluxes, then the traditional atmosphere-only formulation may suffice: TOA fluxes respond almost instantaneously (at monthly resolution) to SST perturbations, supporting a direct causal link between the two variables. The alternative "coupled" formulation offers a novel way to explore the complex cascade of coupled feedbacks across spatial and temporal scales, with great relevance for building understanding of causal mechanisms in coupled climate dynamics. Crucially, this highlights that the goals of (i) reconstructing radiative fluxes and (ii) diagnosing radiative feedbacks need not align.

Finally, a note on reconstruction tasks in the coupled setting is warranted. While in the response operator $\mathbf{R}(t)$ in the "coupled" perspective can be used to reconstruct changes in radiative flux from temperature changes, this is not its primary purpose. In atmosphere-only settings, the reconstruction is straightforward. In the coupled case, applying the convolution formula (Eq. 2) to temperature changes in forced climate runs, requires care: the "temperature forcing" is itself a product of coupled dynamics, not an externally prescribed input. In principle, it could be inferred via an inverse problem, but in practice, we find that using smoothed temperature changes (e.g., 6-month averages) as a proxy yields negligible error. Therefore, the convolved input is approximated using the observed temperature change from from 4xCO2 and 1pctCO2 runs. This is reasonable so long as the forced signal dominates internal variability, as we empirically observed after a 6-months average.

If one's interest is to reconstruct the TOA fluxes then the traditional, atmosphere-only formulation is sufficient, as TOA fluxes respond almost instantaneously (at monthly resolution) to perturbations in SST, leading to the existence of a direct causal link between the two variables. The skill in reconstruction is expected to converge across approaches with infinite data and all variables. Our alternative formulation can both perform well for the prediction task and offers novel insights into the complex cascade of coupled feedbacks across spatial and temporal scales, with great relevance for building understanding of causal mechanisms in coupled climate dynamics. Importantly, this distinction highlights how the goals of (i) reconstructing radiative fluxes based on temperature changes and (ii) diagnosing radiative feedbacks do not necessarily have to align.

Finally, a note on reconstruction tasks. While in the response operator $\mathbf{R}(t)$ in the "coupled" perspective can be used to reconstruct changes in radiative flux from temperature changes, this is not its primary purpose. In the "atmosphere-only" settings, this reconstruction is straightforward. However, in the coupled case, applying the convolution formula (Eq.2) requires care: the "temperature forcing" used in the convolution is itself the result of coupled dynamics and is not externally prescribed. In principle, this forcing could be recovered through an inverse problem, but in practice, we find that using smoothed temperature changes (e.g., 6-month averages) as a proxy leads to negligible error. Thus, in our implementation, when trying to reconstruct the radiative flux in the forced runs the convolved temperature input is approximated using the observed temperature change from 4xCO2 and 1pctCO2 runs. This is reasonable so long as the forced signal dominates internal variability, as we empirically observed after a 6-months average.

8. Conclusion and discussion

In this work, we developed a protocol based on coarse-graining and the Fluctuation-Response formalism to diagnose and understand the relation between patterns of SST warming and the radiative feedbacks from a long control simulation of a climate model. At least two main classes of methods have been utilized previously to investigate this problem: (i) a Green's function approach, estimating the response of atmospheric fields to local perturbation patches in the ocean using an atmosphere-only model (Zhou et al. 2017; Zhang et al. 2023; Dong et al. 2020, 2019; Alessi and Rugenstein 2023; Bloch-Johnson et al. 2024) and (ii) statistical regression approaches (Zhou et al. 2017; Bloch-Johnson et al. 2020; Kang et al. 2023) with coupled climate models. The theory-driven approach presented here, building on the framework recently proposed in Falasca et al. (2024), balances the strengths of each of the previous two methods. Namely, as in previous Green's function methods, it causally studies the atmospheric response to SST perturbations, and it can be applied using only a model's control run, as in previous statistical approaches. Fluctuation-response theory allows us to infer what the response of a dynamical system to small perturbations would have been without actually perturbing the system (Marconi et al. 2008). The response at time t is then computed by convolving the operator $\mathbf{R}(t)$ with perturbations across all previous time scales $t-\tau$ (Christensen and Berner 2019).

Our proposed approach is general and simple to apply to study idealized perturbation experiments from data. The method requires two main ingredients: (i) coarse-graining the system (spatially and temporally) in terms of physically relevant, projected dynamics and (ii) computing a response operator utilizing the variability of a long and stationary control simulation. The protocol focused on the dimensionality reduction proposed in Falasca et al. (2024), reducing the system's dimensionality into only a few (here order 10) components. This focus on a small set of patterns facilitates the practical application of FDT in data-scarce scenarios, where only limited variables and samples are available. Comparing with previous studies, Gritsun and Branstator (2007) demonstrated remarkable skills in computing responses with FDT but utilized a dataset comprising four million days at half-day resolution. This extensive dataset included multiple variables and a substantial number of EOFs (1800), allowing perturbations to be projected near the grid scale. Our approach emphasizes working with significantly fewer samples and only two variables, reflecting the constraints of real-world scenarios with limited data availability. This constraint necessitates a coarse-grained, stochastic representation of the system's variability, achieved by averaging over large spatial regions and time scales, with variables active at finer spatial and temporal scales modeled as noise. Importantly, such choices are constrained by various assumptions about the system's dynamics, and we further discuss their limitations and strengths in Appendix A. In general, no single dimensionality reduction tool is expected to be optimal for all applications, and a key area for future research will involve exploring various methods to decompose the original datasets, tailored to the available sample size and the specific type of dynamics investigated.

Traditional experiments that diagnose the response of atmosphere-only models to SST boundary conditions can be reframed within the proposed framework by focusing on the shortest time scales, under the assumption that the response at short time scales is dominated by the atmospheric dynamics. When longer time scales are considered, the protocol enables idealized climate experiments that capture the joint evolution of the atmosphere and the slow oceanic dynamics. The long term memory of the sea surface temperature field (Fraedrich and Blender 2003; Galfi and Lucarini 2021) is the key physical factor enabling differentiation across the two frameworks. In the context of the pattern effect, our framework reproduces results consistent with existing literature in the atmosphere-only formulation. However, when the coupled dynamics is included – accounting for cumulative responses over extended time scales – the framework generates a novel sensitivity map, offering a qualitative prediction of TOA radiative flux responses to SST perturbations in coupled climate model experiments. The sensitivity map in Figure 6, reveals a negative response to SST warming throughout the tropical Pacific. This finding contrasts with instantaneous feedbacks reported in the literature (and in our "atmosphere-only" implementation), which emphasize a dipole sensitivity pattern in the tropical Pacific, with positive and negative sensitivities on the east and west sides of the basin, respectively. In both cases, i.e. "atmosphere-only" and "coupled" formulation, the framework allows us to perform well at the "prediction" task and reproduce changes in the net radiative flux in forced simulations. The key difference between the two formulations is an additional perspective on the theory of climate feedbacks. The traditional approach defines climate feedbacks in terms of instantaneous relationships and it is centered around the problem of prediction/reconstruction. Instead, the "coupled' formulation of our framework extends this concept to account for spatially and temporally dependent linkages across climate fields. This broader perspective aligns with the theoretical proposals of Lucarini (2018) and focuses on understanding how perturbations propagate through the coupled climate system.

The "coupled" formulation presents an intrinsically more challenging problem than the traditional analysis of pattern effect. In this case – diagnosing how perturbations propagate throughout the system across spatial and temporal scales – focusing on only two variables represents a more significant simplification than in the "atmosphere-only" formulation. Future studies should aim to explore the "coupled" formulation across a broader set of variables and finer spatial and temporal scales. Nonetheless, we consider this study a valuable starting point for such future exploration and a preliminary blueprint for future perturbation experiments with coupled climate models.

A clear limitation of this work is that we focused exclusively on the GFDL climate model. As is well-known, climate models are affected by model errors and they are limited in their ability to represent the "real" climate (Smith 2002). Future work should aim to move beyond model land (Thompson and Smith 2019) and explore the applicability of the proposed framework to real-world observations. A significant constraint in this direction is the extremely limited time span of reliable satellite observations on radiative fluxes, which began only in 1998 (Wielicki et al. 1996). A potential solution could involve integrating the proposed framework with Bayesian analysis, where "prior" response functions are inferred from a catalog of climate models, under the assumption of a reduced model error through multi-model ensemble analysis. This prior response functions could then be updated using a substantial portion of observational data and tested on the remainder. Furthermore, future research could investigate the pattern effect from the theoretical perspective introduced by Ruelle (1998, 2009) and its computational implementation in Lucarini et al. (2017); Lembo et al. (2020). This would allow for a deeper quantification of the limitations/strengths of the assumptions and preprocessing procedures proposed in this work.

Finally, the results in this work serve as additional evidence for the relevance of FDT in climate studies, even in its simple quasi-Gaussian approximation, after carefully choosing the relevant observables. Given appropriate consideration of the coarse-graining steps, the FDT can be utilized throughout climate science beyond what is examined here, with great relevance for inferring causal linkages and building understanding of climate dynamics.

APPENDIX A

A few considerations on the application of linear response theory for spatiotemporal climate data

We briefly review current applications of linear response theory in climate science, including the limitations of the fluctuation-dissipation Theorem (FDT). We then highlight important considerations for applying the fluctuation-response formalism in the climate context. These considerations are important steps for the effective application of FDT to climate data.

a. A summary of linear response theory in climate science

The Earth's climate is a complex, spatiotemporal dynamical system with variability across a large range of spatial and temporal scales (Ghil and Lucarini 2020). It has been recently argued that linear response theory serves as a comprehensive framework to understand and quantify (i) large scale climate dynamics and (ii) its response to external forcings. As outlined in the main text, there has been two main versions of linear response theory applied to climate problems: the Fluctuation-Dissipation Theorem/Relation (Leith 1975; Marconi et al. 2008; Majda et al. 2005) and Ruelle response theory (Ruelle 1998; Lucarini et al. 2017; Lucarini and Chekroun 2023). In the last two decades there has been active research on strengths and limitations of both approaches (see e.g., Lucarini et al. 2014; Gritsun and Lucarini 2017; Christensen and Berner 2019; Lucarini and Chekroun 2023).

Ruelle (1998) proposed a new perspective on linear response grounded in dynamical system theory rather than near-equilibrium statistical mechanics as the original formulation of FDT (Kubo et al. 2008; Sarracino and Vulpiani 2019). This different perspective is recently emerging as a general, rigorous tool to study and attribute changes in the climate system to external forcing with impressive results at both global and regional levels (e.g., Lucarini et al. 2017; Lembo et al. 2020). In practice, the general strategy is to define a Green's function through a few simulations of a climate model, for example, by using a control and a step-function run. It is then possible to convolve the Green's function with, for example, new CO_2 forcing and investigate different possible climate change scenarios (e.g., Lembo et al. 2020). Recently, Gutiérrez and Lucarini (2022) showed how to link the forcing to free modes of variability in the context of Ruelle linear response, therefore adding to interpretability and understanding of the system's response. As with FDT (Aurell and Del Ferraro 2016; Baldovin et al. 2020), Ruelle response theory is causal, in the sense of interventional causality (Ismael 2023; Pearl 2008). We highlight the contributions by Lucarini and Colangeli (2012); Lucarini (2018); Tomasini and Lucarini (2021); Lembo et al. (2020); Basinski-Ferris and Zanna (2024); Lucarini and Chekroun (2023) within the broader framework of Ruelle response theory. In particular, our work shares strong similarities with the proposal of Lucarini (2018). Such approaches are highly relevant for future studies on the pattern effect and, more generally, as rigorous methods for studying feedbacks and performing causal attribution of climate change (Lucarini and Chekroun 2024).

The FDT formalism, as considered in this study, is less general than the strategy outlined above (Ghil and Lucarini 2020), and it has been argued that results can be affected by dimensionality reduction procedures (Hassanzadeh and Kuang 2016b), by the variables of choice (Gritsun and Lucarini 2017), by the length of the dataset analyzed (Lucarini et al. 2014) and on whether the forced response in question projects strongly onto the internal variability (Gritsun and Lucarini 2017).

b. Practical application of FDT: assumptions, limitations and strengths

Despite drawbacks, the FDT has proven to be relevant and useful in climate studies (e.g., Majda et al. 2005; Gritsun and Branstator 2007; Lacorata and Vulpiani 2007; Cooper and Haynes 2011) and, in general, in dynamical systems with many degrees of freedom (Colangeli et al. 2012; Sarracino and Vulpiani 2019). In its domain of applicability, the FDT approach is extremely powerful as it eliminates the need to perform new simulations to construct Green's functions and focus only on long stationary simulations or, ideally, on observational data.

While the theoretical formalism behind the FDT framework is general, its practical implementation is often non-trivial and influenced by coarse-graining procedures, data preprocessing, and, most critically, the underlying assumptions about the system. We highlight several important limitations and caveats - often overlooked in the literature - that may affect interpretations in future studies. The FDT formalism enables us to study how external perturbations propagate in a dynamical system, assuming access to the full state vector $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]$. However, in practice, this is rarely the case, necessitating coarse-graining procedures and a stochastic reformulation of the system. In our analysis, this led to specific choices: (i) averaging sea surface temperature (SST) variability over large spatial regions, (ii) using globally averaged net radiative flux at the TOA, and (iii) adopting a temporal resolution of six-month averages. These decisions were guided by a calibration process (see Section 6a) and prior studies. For instance, earlier work demonstrated that coarse-grained sea surface temperature (SST) fields, particularly in the tropics, can be modeled as a Markov process (Penland 1989; Penland and Sardeshmukh 1995), while more recent research has identified SST and top-of-atmosphere (TOA) net radiative flux as key variables for studying the pattern effect (Bloch-Johnson et al. 2024; Dong et al. 2019). However, the climate system encompasses far more variables than just SST and TOA radiative flux. In this context, coarse-graining becomes essential, treating processes at smaller spatial and faster temporal scales as noise. Crucially, such methodological choices – common to all applications of the FDT – are not dictated by rigorous mathematical criteria but are based on assumptions made by researchers. These assumptions should be taken into account as they significantly shape the results and their interpretation. We also briefly note that dealing with incomplete observations is a common problem across many fields. In order to reconstruct the full system, there have been proposals to consider applications of the Takens' embedding theorem (Takens 1981). The practical implementation of such strategy is, however, severely constrained by the dimensionality of the system and it is not valid for stochastic systems. Therefore, Takens' theorem cannot be a valid option in this study and we refer to the discussions in Baldovin et al. (2020); Lucente et al. (2022) for more details on this point.

The FDT application in this work focused on a simple form of FDT presented in Eq. (4), referred to as "quasi-Gaussian approximation" by Majda et al. (2005). This form of FDT is the one used in many previous applications (e.g., Hassanzadeh and Kuang 2016b, and references therein) and it is valid for linear systems. The climate system is nonlinear and it is therefore not obvious why Eq. (4) should work. Again, the coarse-graining procedures play a central role in enabling the use of this simplified form. As noted, in several previous works, (Sardeshmukh and Sura 2009; Majda et al. 2010a), the probability distribution of coarse-grained climate variables is often smooth and Gaussian. This holds true in our analysis as well, as shown in Appendix C. Thus, the use of Eq. (4) is justified.

Finally, we note that the name "quasi-Gaussian" rather than "Gaussian" approximation, stresses the fact that while $\rho(\mathbf{x})$ is approximated as a Gaussian, lag-covariances are computed at each time t by averaging over the data (whether with linear or nonlinear dynamics), rather than automatically assuming linear dynamics with Gaussian invariant density. This is an important difference from linear regression modeling strategies, leading for example to skill in capturing changes in variance in contrast to linear inverse models (see Majda et al. (2010b) for details and in-depth comparisons).

APPENDIX B

Dimensionality reduction through community detection

For completeness we report here the main steps of the dimensionality reduction step proposed in Falasca et al. (2024), and refer to that paper for further details. Consider a spatiotemporal field saved as a data matrix $\mathbf{x} \in \mathbb{R}^{N,T}$. N is the number of grid points and T is the length of each time series. For example, $\mathbf{x}$ could be the sea surface temperature field. The dimensionality reduction proposed in Falasca et al. (2024) works in a few simple steps:

- Compute the correlation matrix $\mathbf{C}$, defined as $C_{i,j} = \overline{x_i(t)x_j(t)}$, where the overline stands for temporal averages, and each x_i has been scaled to zero mean.
- Define an Adjacency matrix $\mathbf{A}$ from the correlation matrix $\mathbf{C}$ by setting $A_{i,j} = 1$ if (i) $C_{i,j}$ exceeds a threshold k and (ii) the distance between grid points i and j is smaller than a threshold η. If (i) and (ii) are not satisfied, then $A_{i,j} = 0$. $\mathbf{A}$ is the matrix representation of a graph where nodes i and j are connected if sharing large covariability and if they are close on a longitude-latitude grid. Regionally constrained patterns of variability can then be identified by finding "communities" in the graph (Barabási 2016; Newman 2010; Lancichinetti and Fortunato 2009). With communities of a graph, we refer to group of nodes that are much more connected to each other than to the rest of the graph. Importantly, parameters k and η are automatically defined by two simple heuristics. The two heuristics depend on two parameters $q_k = 0.95$ and $q_\eta = 0.1$ and we refer the reader to Falasca et al. (2024) for details. We chose a value of $q_\eta = 0.1$ rather than 0.15 as in Falasca et al. (2024) in order to split the ENSO region (Figure 2 in Falasca et al. (2024)) into an eastern and central Pacific region.
- Each node i, correspondent to a grid point i on the map, will then be associated to a community/pattern.

In other words, we partitioned a spatiotemporal climate field of spatial dimension N, in a series of n regions c_j, with $j = 1, ..., n$. We identify communities through the Infomap community detection algorithm (Rosvall and Bergstrom 2007, 2008; Rosvall et al. 2009; Smiljanić et al. 2023) as shown in Falasca et al. (2024). The size and number of the identified patterns will roughly depend on the q_k and q_η parameters presented above.

- Finally, to each community c_j, we are going to associate a time series defined as the integrated anomaly inside, i.e. $X(c_j, t) = \sum_{i \in c_j} x_i(t) \cos(\theta_i)$. Where θ_i represents the latitude at grid point i and $\cos(\theta_i)$ a latitudinal scaling.

To summarize, given a spatiotemporal field saved as a data matrix $\mathbf{x} \in \mathbb{R}^{N,T}$, the proposed framework allows us to define a new field $\mathbf{X} \in \mathbb{R}^{n,T}$, with $n \ll N$.

APPENDIX C

Probability distributions

In Figure C1, we show the histogram of the integrated SST anomalies in each one of the patterns in Figure 1 and of the global mean net radiative flux at the TOA. Each time series have been scaled to zero mean and unit variance. A standard normal distribution is also shown in black for comparison. This analysis demonstrates that the quasi-Gaussian approximation shown in Eq. (4) is indeed relevant for the system studied. The Gaussianity of the process is a direct consequence of our preprocessing by coarse-graining in both the temporal and spatial directions, further confirming the ideas and findings of previous papers such as Sardeshmukh and Sura (2009).

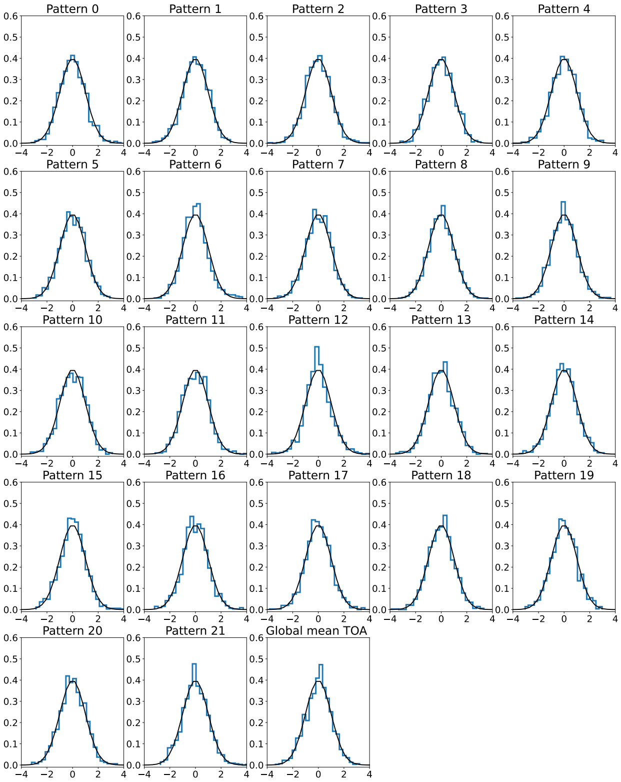

FIG. C1. Probability distributions of the cumulative time series of sea surface temperature in each pattern in Figure 1 and of the global mean net radiative flux at the TOA. Each signal had the mean removed and has been standardized to unit variance. A Gaussian fit with zero mean and unit variance is shown in black on top of each histogram.

APPENDIX D

Diagnosing the characteristic time scale of the response

The sensitivity map in Figure 6 integrates all responses up to $t = 10$ years. Such threshold has been considered mainly because data have been high-pass filtered with a cut-off frequency of 10^{-1} years. Therefore variability present for $t \geq 10$ years is here considered as noise. However, the characteristic time scale of the response in TOA can appear sooner than 10 years. In fact, the response operator is defined as the response to a small impulse perturbation. Therefore, at long time scales ($t \longrightarrow \infty$), the response $R_{k,j}(t)$ between any variable x_j and x_k should (i) go to zero, see for example Figure 1 in (Baldovin et al. 2020), or (ii) become statistically insignificant, see for example Figure 1 in (Falasca et al. 2024). In practice, then, it is common not to compute the response operator $\mathbf{R}(t)$ for all t and set $R_{k,j}(t) = 0$ after a long characteristic time scale τ, see for example Hassanzadeh and Kuang (2016b). The value of τ will depend on the system itself. To explore this time scale (and the robustness of our choice of 10 years) we then evaluate the framework as done in the main text (i.e. by reconstructing the change in TOA mean flux given changes in the Temperature field) while setting to zero every $R_{k,j}(t)$ for $t > \tau$, with $\tau = 6$ months, 5 and 10 years. Results are shown in Figure D1. Independent on the tested data, i.e. 1pctCO2 ot 4xCO2 experiment, considering the impact of longer time scales, helps reducing the bias in the reconstruction (i.e. $\tau = 5$ and 10 lead to better reconstruction of $t = 1$). Small differences are observed when going from $\tau = 5$ to 10 years in the 1pctCO2, with slightly better values observed for $\tau = 10$ years. In the case of 4xCO2, both $\tau = 5$ and $\tau = 10$ years result in relatively good reconstructions. For $\tau = 5$ years, better reconstructions are observed after approximately 70 years compared to $\tau = 10$ years, but the reconstruction is worse in the first 60 to 70 years. We conclude that a time scale of $\tau \geq 5$ is long enough to observe the cumulative response to perturbations. Note that the confidence bounds cosidered here can further influence this analysis, and we would expect to have clearer results in the case of much longer datasets.

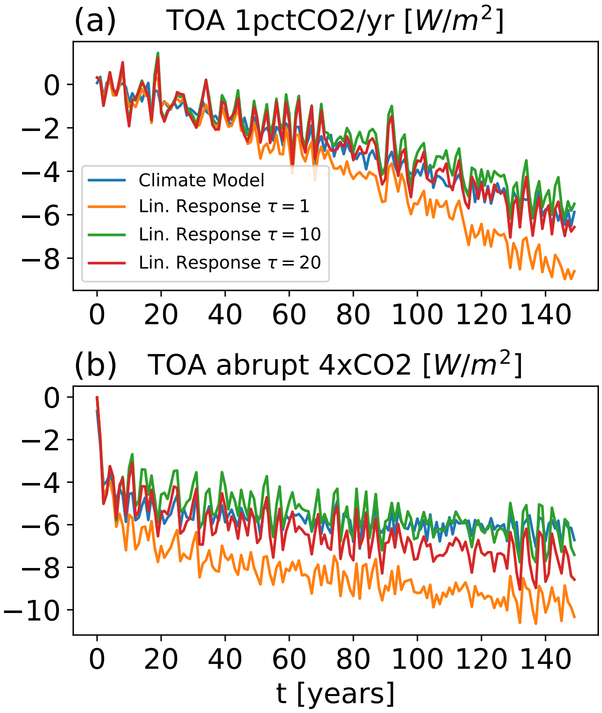

Fig. D1. Pedicting the change in the net radiative flux at the top of the atmosphere (TOA) as a function of a time-scale parameter τ. We compute responses of the global mean TOA to changes in SST in the forced simulation using after setting $R_{k,j}(t) = 0$ for $t \geq \tau$, with $\tau = 1, 10, 20$, correspondent to + 6 months, + 5 years and + 10 years.

APPENDIX E

A sensitivity map in the coupled system using strict confidence bounds

We plot the sensitivity map as in Section 6 but focusing on a very strict confidence bound of $\pm 3\sigma$. In this case only tropical domains becomes relevant to describe the pattern effect. However, such strict confidence bounds almost surely mask also “true” responses and future work should focus on analysis on much larger datasets or on multi-model ensembles.

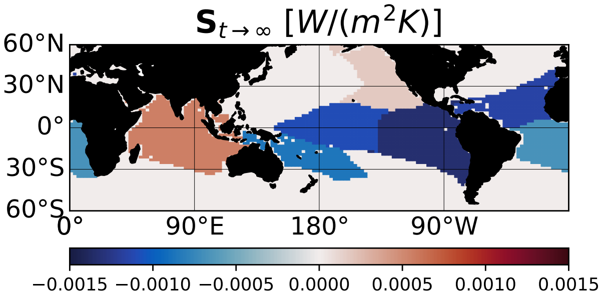

Fig. E1. Sensitivity map such that at each grid point i, we plot the equilibrated, global average response of the net radiative flux at the TOA given a constant perturbation of 1 Kelvin in SST imposed at that point. Positive sensitivity corresponds to a cumulative (in time) positive radiative feedback, therefore amplifying the initial global mean temperature changes; the opposite is true for negative sensitivity. The time $t \longrightarrow \infty$ is here approximated as $t = 10$ years, as all data have been high-pass filtered with a 10^{-1} years cut-off frequency.

Acknowledgments. The authors would like to thank two anonymous reviewers and Valerio Lucarini for their comments and suggestions which greatly improved the quality of the manuscript. F.F. acknowledges numerous discussions with Pavel Perezhogin. This work was supported by the NOAA Grant NA20OAR4310411, National Science Foundation Grant OCE-2048576, and by the KITP Program "Machine Learning and the Physics of Climate" supported by the National Science Foundation under Grant No. NSF PHY-1748958.

Data availability statement. Codes for the community detection and implementation of Fluctuation-Dissipation formulas, can be found in https://github.com/FabriFalasca/Linear-Response-and-Causal-Inference.

References

Adcroft, A., W. Anderson, V. Balaji, C. Blanton, M. Bushuk, C. O. Dufour, and Coauthors, 2019: The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, **11**, 3167–3211, https://doi.org/doi.org/10.1029/2019MS001726.

Alessi, M. J., and M. A. Rugenstein, 2023: Surface temperature pattern scenarios suggest higher warming rates than current projections. *Geophysical Research Letters*, **50 (23)**, e2023GL105 795.

Alexander, M., I. Bladé, M. Newman, J. Lanzante, N.-C. Lau, and J. Scott, 2002: The Atmospheric Bridge: The Influence of ENSO Teleconnections on Air–Sea Interaction over the Global Oceans. *Journal of Climate*, 2205–2231.

Allen, M., and L. Smith, 1996: Monte Carlo SSA: Detecting irregular oscillations in the Presence of Colored Noise . *Journal of Climate*, **9**, 3373–3404, https://doi.org/https://doi.org/10.1175/1520-0442(1996)009⟨3373:MCSDIO⟩2.0.CO;2.

Andrews, T., J. M. Gregory, and M. J. Webb, 2015: The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *Journal of Climate*, **28 (4)**, 1630–1648.

Armour, K. C., C. M. Bitz, and G. H. Roe, 2013: Time-varying climate sensitivity from regional feedbacks. *Journal of Climate*, **26 (13)**, 4518–4534.

Aurell, E., and G. Del Ferraro, 2016: Causal analysis, correlation-response, and dynamic cavity. *J. of Phys.: Conference Series*, **699**, 012 002, https://doi.org/doi:10.1088/1742-6596/699/1/012002.

Ay, N., and D. Polani, 2008: Information flows in causal networks. *Adv. Complex Syst.*, **11 (1)**, 17–41.

Baldovin, M., L. Caprini, A. Puglisi, A. Sarracino, and A. Vulpiani, 2022a: *The Many Faces of Fluctuation-Dissipation Relations Out of Equilibrium*, 29–57. Springer International Publishing, Cham, https://doi.org/10.1007/978-3-031-04458-8_3, URL https://doi.org/10.1007/978-3-031-04458-8_3.

Baldovin, M., F. Cecconi, A. Provenzale, and A. Vulpiani, 2022b: Extracting causation from millennial-scale climate fluctuations in the last 800 kyr. *Scientific Reports*, **12**, 15 320.

Baldovin, M., F. Cecconi, and A. Vulpiani, 2020: Understanding causation via correlations and linear response theory. *Physical Review Research*, **2**, 043 436.

Barabási, A. L., 2016: Network science. *Cambridge, UK: Cambridge University Press*.

Barsugli, J. J., and P. D. Sardeshmukh, 2002: Global atmospheric sensitivity to tropical sst anomalies throughout the indo-pacific basin. *Journal of Climate*, **15 (23)**, 3427–3442.

Basinski-Ferris, A., and L. Zanna, 2024: Estimating freshwater flux amplification with ocean tracers via linear response theory. *Earth System Dynamics*, **15 (2)**, 323–339.

Bloch-Johnson, J., M. Rugenstein, and D. S. Abbot, 2020: Spatial radiative feedbacks from internal variability using multiple regression. *Journal of Climate*, **33 (10)**, 4121–4140.

Bloch-Johnson, J., and Coauthors, 2024: The green's function model intercomparison project (gfmip) protocol. *Journal of Advances in Modeling Earth Systems*, **16 (2)**, e2023MS003 700, https://doi.org/https://doi.org/10.1029/2023MS003700, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023MS003700.

Bracco, A., F. Kucharski, F. Molteni, and et al., 2005: Internal and forced modes of variability in the Indian Ocean. *Geophysical Research Letters*, **32**, L12 707.

Cai, W., and Coauthors, 2019: Pantropical climate interactions. *Science*, **363**, eaav4236, https://doi.org/DOI:10.1126/science.aav4236.

Castiglione, P., M. Falcioni, A. Lesne, and A. Vulpiani, 2008: *Chaos and coarse-graining in statistical mechanics*. Cambridge University Press.

Cecconi, F., G. Costantini, C. Guardiani, M. Baldoving, and A. Vulpiani, 2020: Correlation, response and entropy approaches to allosteric behaviors: a critical comparison on the ubiquitin case. *Physical Biology*, **5**, 056 002.

Chenal, J., B. Meyssignac, A. Ribes, and R. Guillaume-Castel, 2022: Observational constraint on the climate sensitivity to atmospheric co 2 concentrations changes derived from the 1971–2017 global energy budget. *Journal of Climate*, **35 (14)**, 4469–4483.

Chiang, J., and A. Sobel, 2002: Tropical tropospheric temperature variations caused by enso and their influence on the remote tropical climate. *Journal Of Climate*, **15**, 2616–2631, https://doi.org/https://doi.org/10.1175/1520-0442(2002)015⟨2616:TTTVCB⟩2.0.CO;2.

Christensen, H. M., and J. Berner, 2019: From reliable weather forecasts to skilful climate response: A dynamical systems approach. *Q. J. R. Meteorological Soc.*, **145**, 1052–1069.

Colangeli, M., L. Rondoni, and A. Vulpiani, 2012: Fluctuation-dissipation relation for chaotic non-hamiltonian systems. *Journal of Statistical Mechanics: Theory and Experiment*, **2012 (04)**, L04 002, https://doi.org/10.1088/1742-5468/2012/04/L04002.

Cooper, F., and P. Haynes, 2011: Climate Sensitivity via a Non-parametric Fluctuation–Dissipation Theorem. *Journal of the atmospheric sciences*, **68**, 937–953, https://doi.org/https://doi.org/10.1175/2010JAS3633.1.

Crommelin, D., and A. Majda, 2004: Strategies for Model Reduction: Comparing Different Optimal Bases. *Journal of the Atmospheric Sciences*, **61**, 2206–2217, https://doi.org/https://doi.org/10.1175/1520-0469(2004)061⟨2206:SFMRCD⟩2.0.CO;2.

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

Dijkstra, H. A., 2013: *Nonlinear Climate Dynamics*. Cambridge University Press.

Dong, Y., K. C. Armour, M. D. Zelinka, C. Proistosescu, D. S. Battisti, C. Zhou, and T. Andrews, 2020: Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models. *Journal of Climate*, **33 (18)**, 7755–7775, https://doi.org/10.1175/JCLI-D-19-1011.1.

Dong, Y., C. Proistosescu, K. C. Armour, and D. S. Battisti, 2019: Attributing Historical and Future Evolution of Radiative Feedbacks to Regional Warming Patterns using a Green's Function Approach: The Preeminence of the Western Pacific. *Journal of Climate*, **32 (17)**, 5471–5491, https://doi.org/10.1175/JCLI-D-18-0843.1.

Dubrulle, B., F. Daviaud, D. Faranda, L. Marié, and B. Saint-Michel, 2022: How many modes are needed to predict climate bifurcations? Lessons from an experiment. *Nonlin. Processes Geophys.*, **29**, 17–35, https://doi.org/doi.org/10.5194/npg-29-17-2022.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, **9 (5)**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Falasca, F., P. Perezhogin, and L. Zanna, 2024: Data-driven dimensionality reduction and causal inference for spatiotemporal climate fields. *Phys. Rev. E*, **109**, 044 202, https://doi.org/10.1103/PhysRevE.109.044202.

Falcioni, M., S. Isola, and A. Vulpiani, 1990: Correlation functions and relaxation properties in chaotic dynamics and statistical mechanics. *Physics Letters A*, **144 (6)**, 341–346, https://doi.org/https://doi.org/10.1016/0375-9601(90)90137-D.

Fraedrich, K., and R. Blender, 2003: Scaling of atmosphere and ocean temperature correlations in observations and climate models. *Phys. Rev. Lett.*, **90**, 108 501, https://doi.org/10.1103/PhysRevLett.90.108501.

Galfi, V., and V. Lucarini, 2021: Fingerprinting heatwaves and cold spells and assessing their response to climate change using large deviation theory. *Phys. Rev. Lett.*, **127**, 058 701, https://doi.org/10.1103/PhysRevLett.127.058701.

Gershgorin, B., and A. J. Majda, 2010: A test model for fluctuation-dissipation theorems with time-periodic statistics. *Physica D: Nonlinear Phenomena*, **239 (17)**, 1741–1757, https://doi.org/https://doi.org/10.1016/j.physd.2010.05.009.

Ghil, M., and V. Lucarini, 2020: The physics of climate variability and climate change. *Rev. Mod. Phys.*, **92**, 035 002, https://doi.org/10.1103/RevModPhys.92.035002.

Giorgini, L. T., K. Deck, T. Bischoff, and A. Souza, 2024: Response theory via generative score modeling. *Phys. Rev. Lett.*, **133**, 267 302, https://doi.org/10.1103/PhysRevLett.133.267302.

Gregory, J. M., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, **31 (3)**, 3205, https://doi.org/10.1029/2003GL018747.

Gritsun, A., and G. Branstator, 2007: Climate response using a three-dimensional operator based on the fluctuation–dissipation theorem. *Journal of The Atmospheric Science*, 2558–2575, https://doi.org/https://doi.org/10.1175/JAS3943.1.

Gritsun, A., G. Branstator, and A. Majda, 2008: Climate response of linear and quadratic functionals using the fluctuation–dissipation theorem. *Journal of The Atmospheric Science*, 2824–2841, https://doi.org/https://doi.org/10.1175/2007JAS2496.1.

Gritsun, A., and V. Lucarini, 2017: Fluctuations, response, and resonances in a simple atmospheric model. *Physica D: Nonlinear Phenomena*, **349**, 62–76, https://doi.org/https://doi.org/10.1016/j.physd.2017.02.015.

Grubb, M., and Coauthors, 2022: Introduction and framing. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley, Eds., Cambridge University Press, Cambridge, UK and New York, NY, USA, book section 1, https://doi.org/10.1017/9781009157926.003, URL https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_Chapter01.pdf.

Guilyardi, E., P. Braconnot, F.-F. Jin, S. T. Kim, M. Kolasinski, L. T., and I. Musat, 2009: Atmosphere Feedbacks during ENSO in a Coupled GCM with a Modified Atmospheric Convection Scheme. *Journal of Climate*, **2**, 5698–5718.

Gutiérrez, M. S., and V. Lucarini, 2022: On some aspects of the response to stochastic and deterministic forcings. **55 (42)**, 425 002, https://doi.org/10.1088/1751-8121/ac90fd.

Hairer, M., and A. J. Majda, 2010: A simple framework to justify linear response theory. *Nonlinearity*, **23 (4)**, 909, https://doi.org/10.1088/0951-7715/23/4/008.

Hansen, P., 2010: *Discrete Inverse Problems: Insight and Algorithms.* SIAM, https://doi.org/https://doi.org/10.1137/1.9780898718836.

Hassanzadeh, P., and Z. Kuang, 2016a: The Linear Response Function of an Idealized Atmosphere. Part I: Construction Using Green's Functions and Applications. *Journal of The Atmospheric Science*, 3423–3439, https://doi.org/https://doi.10.1175/JAS-D-15-0338.1.

Hassanzadeh, P., and Z. Kuang, 2016b: The Linear Response Function of an Idealized Atmosphere. Part II: Implications for the Practical Use of the Fluctuation–Dissipation Theorem and the Role of Operator's Nonnormality. *Journal of The Atmospheric Science*, 3441–3452, https://doi.org/https://doi.org/10.1175/JAS-D-16-0099.1.

Hasselmann, K., 1976: Stochastic climate models part i. theory. *Tellus*, **28**, 473–485, https://doi.org/https://doi.org/10.1111/j.2153-3490.1976.tb00696.x.

Held, I., 2005: The Gap between Simulation and Understanding in Climate Modeling. *Bulletin of the American Meteorological Society*, 1609–1614, https://doi.org/https://doi.org/10.1175/BAMS-86-11-1609.

Held, I. M., H. Guo, A. Adcroft, J. P. Dunne, L. W. Horowitz, J. Krasting, and Coauthors, 2019: Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, **11**, 3691–3727, https://doi.org/doi.org/10.1029/2019MS001829.

Ismael, J., 2023: Reflections on the asymmetry of causation. *Interface Focus*, **12**, 20220 081.

Kang, S. M., P. Ceppi, Y. Yu, and I.-S. Kang, 2023: Recent global climate feedback controlled by southern ocean cooling. *Nature Geoscience*, **16 (9)**, 775–780.

Klein, S., B. Soden, and N. Lau, 1999: Remote sea surface temperature variations during ENSO: evidence for a tropical atmospheric bridge. *J Clim*, **12**, 917–932.

Kraichnan, R. H., 1959: Classical fluctuation-relaxation theorem. *Phys. Rev.*, **113**, 1181–1182, https://doi.org/10.1103/PhysRev.113.1181.

Kubo, R., 1966: The fluctuation-dissipation theorem. *Reports on Progress in Physics*, **29 (1)**, 255, https://doi.org/10.1088/0034-4885/29/1/306.

Kubo, R., M. Toda, and N. Hashitsume, 2008: Statistical mechanics of linear response. *Phys. Rep.*, **461 (111)**.

Lacorata, G., and A. Vulpiani, 2007: Fluctuation-response relation and modeling in systems with fast and slow dynamics. *Nonlin. Processes Geophys.*, **14**, 681–694.

Lancichinetti, A., and S. Fortunato, 2009: Community detection algorithms: A comparative analysis. *Phys. Rev. E*, **80**, 1 –11, https://doi.org/doi:10.1103/PhysRevE.80.056117.

Leith, C. E., 1975: Climate response and fluctuation dissipation. *Journal of The Atmospheric Science*, **32**, 2022–2026.

Lembo, V., V. Lucarini, and F. Ragone, 2020: Beyond forcing Scenarios: predicting climate change through Response operators in a coupled General circulation Model. *Scientific Reports*, **10**, 8668.

Lucarini, V., 2018: Revising and Extending the Linear Response Theory for Statistical Mechanical Systems: Evaluating Observables as Predictors and Predictands. *J Stat Phys*, **173**, 1698–1721.

Lucarini, V., R. Blender, C. Herbert, F. Ragone, S. Pascale, and J. Wouters, 2014: Mathematical and physical ideas for climate science. *Rev. Geophys.*, **52**, 809–859, https://doi.org/https://doi:10.1002/2013RG00044.

Lucarini, V., and M. Chekroun, 2023: Theoretical tools for understanding the climate crisis from Hasselmann's programme and beyond. *Nat Rev Phys (2023)*.

Lucarini, V., and M. D. Chekroun, 2024: Detecting and attributing change in climate and complex systems: Foundations, green's functions, and nonlinear fingerprints. *Phys. Rev. Lett.*, **133**, 244 201, https://doi.org/10.1103/PhysRevLett.133.244201.

Lucarini, V., and M. Colangeli, 2012: Beyond the linear fluctuation-dissipation theorem: the role of causality. *J. Stat. Mech.*, **05**, P05 013.

Lucarini, V., F. Ragone, and F. Lunkeit, 2017: Predicting Climate Change Using Response Theory: Global Averages and Spatial Patterns. *J Stat Phys*, **166**, 1036–1064.

Lucente, D., A. Baldassarri, A. Puglisi, A. Vulpiani, and M. Viale, 2022: Inference of time irreversibility from incomplete information: Linear systems and its pitfalls. *Phys. Rev. Res.*, **4**, 043 103, https://doi.org/10.1103/PhysRevResearch.4.043103.

Lutsko, N., I. Held, and P. Zurita-Gotor, 2015: Applying the Fluctuation–Dissipation Theorem to a Two-Layer Model of Quasigeostrophic Turbulence . *Journal of the Atmospheric Sciences*, **72**, 3161–3177, https://doi.org/https://doi.org/10.1175/JAS-D-14-0356.1.

Majda, A., R. Abramov, and B. Gershgorin, 2010a: High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability. *Proc. Natl. Acad. Sci.*, **107 (2)**, 581–586, https://doi.org/https://doi.org/10.1073/pnas.0912997107.

Majda, A., B. Gershgorin, and Y. Yuan, 2010b: Low-Frequency Climate Response and Fluctuation–Dissipation Theorems: Theory and Practice . *Journal of the Atmospheric Sciences*, **67 (4)**, 1186–1201, https://doi.org/https://doi.org/10.1175/2009JAS3264.1.

Majda, A. J., R. V. Abramov, and M. J. Grote, 2005: *Information Theory and Stochastics for Multiscale Nonlinear Systems*. CRM Monograph Series, American Mathematical Society.

Majda, A. J., I. Timofeyev, and Vanden-Eijnden, 1999: Models for stochastic climate prediction. *Proc. Natl. Acad. Sci. USA*, **96**, 14 687–14 691.

Majda, A. J., I. Timofeyev, and Vanden-Eijnden, 2001: A mathematical framework for stochastic climate models. *Proc. Natl. Acad. Sci. USA*, **54**, 891–974.

Marconi, U., A. Puglisi, L. Rondoni, and A. Vulpiani, 2008: Fluctuation–dissipation: Response theory in statistical physics. *Physics Reports*, **461**, 111–195, https://doi.org/doi:10.1016/j.physrep.2008.02.002.

Martynov, R., and Y. Nechepurenko, 2006: Finding the response matrix to the external action from a subspace for a discrete linear stochastic dynamical system. *Comput. Math. and Math. Phys.*, **46**, 1155–1167, https://doi.org/doi.org/10.1134/S0965542506070062.

Meehl, G. A., C. A. Senior, V. Eyring, G. Flato, J.-F. Lamarque, R. J. Stouffer, K. E. Taylor, and M. Schlund, 2020: Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models. *Science Advances*, **6 (26)**, eaba1981, https://doi.org/10.1126/sciadv.aba1981, https://www.science.org/doi/pdf/10.1126/sciadv.aba1981.

Meyssignac, B., R. Guillaume-Castel, and R. Roca, 2023: Revisiting the Global Energy Budget Dynamics with a Multivariate Earth Energy Balance Model to Account for the Warming Pattern Effect . *Journal of Climate*, **36**, 8113–8126, https://doi.org/https://doi.org/10.1175/JCLI-D-22-0765.1.

Murphy, J., 1995: Transient response of the hadley centre coupled ocean-atmosphere model to increasing carbon dioxide. part iii: analysis of global-mean response using simple models. *Journal of Climate*, **8 (3)**, 496–514.

Newman, M., 2010: Networks: An introduction. *Oxford, UK: Oxford University Press*.

Pearl, J., 2000: Cambridge: Cambridge University Press.

Pearl, J., 2008: Causal inference. *JMLR Workshop and Conference Proceedings*, **6**, 39–58.

Penland, C., 1989: Random Forcing and Forecasting Using Principal Oscillation Pattern Analysis. *Monthly Weather Review*, **117**, 2165–2185.

Penland, C., 1996: A stochastic model of indopacific sea surface temperature anomalies. *Physica D: Nonlinear Phenomena*, **98 (2)**, 534–558, https://doi.org/https://doi.org/10.1016/0167-2789(96)00124-8.

Penland, C., and P. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, **8 (8)**, 1999–2024.

Pierrehumbert, R. T., 2010: *Principles of Planetary Climate*. Cambridge University Press.

Ring, M. J., and R. A. Plumb, 2008: The response of a simplified gcm to axisymmetric forcings: Applicability of the fluctuation– dissipation theorem. *Journal of The Atmospheric Sciences*, **65**, 3880–3898, https://doi.org/doi:10.1175/2008JAS2773.1.

Risken, H., 1996: *The Fokker-Planck Equation – Methods of Solution and Applications*. Springer-Verlag, https://doi.org/doi.org/10.1007/978-3-642-04898-2_150.

Roe, G., and K. Armour, 2011: How sensitive is climate sensitivity? *Geophysical Research Letters*, **38 (14)**.

Romps, D. M., J. T. Seeley, and J. P. Edman, 2022: Why the forcing from carbon dioxide scales as the logarithm of its concentration. *Journal of Climate*, **35 (13)**, 4027–4047, https://doi.org/https://doi.org/10.1175/jcli-d-21-0275.1.

Rosvall, M., D. Axelsson, and C. Bergstrom, 2009: The map equation. *Eur. Phys. J. Spec. Top.*, **178**, 13–23.

Rosvall, M., and C. Bergstrom, 2007: An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. USA*, **104**, 7327–7331.

Rosvall, M., and C. Bergstrom, 2008: Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, **105**, 1118 –1123.

Ruelle, D., 1998: General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium. *Physics Letters A*, **245 (3)**, 220–224, https://doi.org/https://doi.org/10.1016/S0375-9601(98)00419-8.

Ruelle, D., 2009: A review of linear response theory for general differentiable dynamical systems. *Nonlinearity*, **22**, 855, https://doi.org/DOI10.1088/0951-7715/22/4/009.

Runge, J., V. Petoukhov, J. Donges, and Coauthors, 2015: Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat Commun*, **6**, 8502, https://doi.org/https://doi.org/10.1038/ncomms9502.

Sardeshmukh, P., and P. Sura, 2009: Reconciling Non-Gaussian Climate Statistics with Linear Dynamics. *Journal of Climate*, **22**, 1193–1207.

Sarracino, A., and A. Vulpiani, 2019: On the fluctuation-dissipation relation in non-equilibrium and non-Hamiltonian systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **29 (8)**, 083 132, https://doi.org/10.1063/1.5110262, https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/1.5110262/14623408/083132_1_online.pdf.

Senior, C. A., and J. F. Mitchell, 2000: The time-dependence of climate sensitivity. *Geophysical Research Letters*, **27 (17)**, 2685–2688.

Smiljanić, J., C. Blöcker, A. Holmgren, D. Edler, M. Neuman, and M. Rosvall, 2023: Community Detection with the Map Equation and Infomap: Theory and Applications. *arXiv:2311.04036*, https://doi.org/https://doi.org/10.48550/arXiv.2311.04036.

Smith, L., 2002: What might we learn from climate forecasts? *Proc. Natl Acad. Sci.*, **99**, 2487–2492.

Stevens, B., S. C. Sherwood, S. Bony, and M. J. Webb, 2016: Prospects for narrowing bounds on earth's equilibrium climate sensitivity. *Earth's Future*, **4 (11)**, 512–522.

Takens, F., 1981: *Detecting strange attractors in turbulence, in Dynamical Systems and Turbulence*, Vol. 898, 21–48. Springer, Berlin,Heidelberg.

Taylor, K., M. Crucifix, P. Braconnot, C. Hewitt, C. Doutriax, A. Broccoli, J. Mitchell, and M. Webb, 2006: Estimating Shortwave Radiative Forcing and Response in Climate Models. *Journal of Climate*, **20**, 2530–2543.

Thompson, E., and L. Smith, 2019: Escape from model-land. *Economics: The Open-Access, Open-Assessment E-Journal*, **13**, 1–15.

Tomasini, U., and V. Lucarini, 2021: Predictors and predictands of linear response in spatially extended systems. *Eur. Phys. J. Spec. Top.*, **230**, 2813–2832.

Wang, C., 2019: Three-ocean interactions and climate variability: a review and perspective. *Clim Dyn*, **53**, 5119–5136, https://doi.org/https://doi.org/10.1007/s00382-019-04930-x.

Wielicki, B., B. Barkstrom, E. Harrison, R. Lee III, G. smith, and J. Cooper, 1996: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment. *Bulletin of the American Meteorological Society*, 853–868, https://doi.org/https://doi.org/10.1175/1520-0477(1996)077⟨0853:CATERE⟩2.0.CO;2.

Williams, A. I., N. Jeevanjee, and J. Bloch-Johnson, 2023: Circus tents, convective thresholds, and the non-linear climate response to tropical ssts. *Geophysical Research Letters*, **50 (6)**, e2022GL101 499.

Williams, K., W. Ingram, and J. Gregory, 2008: Time variation of effective climate sensitivity in gcms. *Journal of Climate*, **21 (19)**, 5076–5090.

Zelinka, M. D., T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor, 2020: Causes of higher climate sensitivity in cmip6 models. *Geophysical Research Letters*, **47 (1)**, e2019GL085 782.

Zhang, B., M. Zhao, and Z. Tan, 2023: Using a Green's Function Approach to Diagnose the Pattern Effect in GFDL AM4 and CM4. *Journal of Climate*, **36 (4)**, 1105–1124, https://doi.org/10.1175/JCLI-D-22-0024.1.

Zhao, M., 2022: An investigation of the effective climate sensitivity in GFDL's new climate models CM4.0 and SPEAR. *Journal of Climate*, **35**, 5637 5660, https://doi.org/DOI:10.1175/JCLI-D-21-0327.

Zhao, M., and Coauthors, 2018a: The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, **10**, 691–734, https://doi.org/https://doi.org/10.1002/2017MS001208.

Zhao, M., and Coauthors, 2018b: The GFDL global atmosphere and land model am4.0/LM4.0: 2. Model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, **10**, 735–769, https://doi.org/https://doi.org/10.1002/2017MS001208.

Zhao, Y., and A. Capotondi, 2024: The role of the tropical Atlantic in tropical Pacific climate variability. *npj Clim Atmos Sci*, **7 (140)**.

Zhou, C., M. D. Zelinka, and S. A. Klein, 2017: Analyzing the dependence of global cloud feedback on the spatial pattern of sea surface temperature change with a green's function approach. *Journal of Advances in Modeling Earth Systems*, **9 (5)**, 2174–2189, https://doi.org/https://doi.org/10.1002/2017MS001096, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017MS001096.