

Diagnosing the pattern effect in the atmosphere-ocean coupled system through linear response theory

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: The energy surplus resulting from radiative forcing causes warming of the Earth system. This initial warming drives a myriad of changes in, for example, ice coverage, cloud properties and sea surface temperatures (SSTs), leading to different radiative feedbacks. Understanding the relationship between the radiative feedbacks and the pattern of SST changes is often referred to as the “pattern effect”. The main current approach to study the pattern effect relies on Green’s function experiments, diagnosing the response of atmosphere-only models to perturbations in the SST boundary condition. Here, we argue that the fluctuation-dissipation relation (FDR) and response theory, together with careful considerations about coarse-graining procedures, can be a computationally cheap and theoretically grounded alternative to model experiments. We introduce a protocol based on FDR to study the pattern effect directly from data and present its application in a state-of-the-art coupled climate model. By focusing on the coupled dynamics, rather than atmosphere-only models as in previous studies, we unveil the role of the slow ocean component in setting the pattern effect. We present a new “sensitivity map”, representing a first, qualitative prediction of the response of the global mean top-of-the-atmosphere (TOA) radiative flux to local perturbations in the SST field in fully coupled climate models. We find negative sensitivity throughout the tropics, in contrast to the current understanding where the eastern and western tropical Pacific respectively show positive and negative sensitivity. On the other hand, we show that if only the fastest time scales are considered, then the system’s response is dominated by the atmospheric variability and we recover results in qualitative agreement with the literature. Therefore, the difference between the sensitivity map proposed in this and previous studies, largely comes from the inclusion of the atmosphere-ocean coupling rather than methodological details. The framework offers a conceptually novel perspective on the pattern effect: feedbacks in the coupled climate system, are encoded in a temporally and spatially dependent response operator, rather than time-independent maps as for studies involving atmosphere-only models.

1. Introduction

The response of Earth’s surface temperature to a change in atmospheric CO₂ 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 CO₂, i.e. “equilibrium climate sensitivity”, ranges between 1.8 – 5.6°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 [$W/(m^2K)$] 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), largely 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 CO₂ forcing (e.g., Murphy (1995); Senior and Mitchell (2000); Williams et al. (2008); Andrews et al. (2015)). 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 a very different radiative feedback. 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 quantifying 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 surface temperature field (Zhou et al. 2017; Zhang et al. 2023; Dong et al. 2020, 2019; Alessi and Rugenstein 2023). Despite

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

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) utilizing relatively large perturbations (e.g., $1 - 4K$) in the SST field when constructing the Green’s functions. The relatively large perturbations are needed to detect a clear response with a shorter integration time, but may lead to additional issues as large perturbations can result in nonlinear responses (Williams et al. 2023), invalidating assumption (i) stated above. As we will argue, the protocol proposed in this paper, provides a new perspective on the pattern effect by relaxing assumptions (ii) and (iii) outlined above. In addition to the Green’s function approach, there have also been efforts to find a response operator statistically from existing simulations, 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. For example, Bloch-Johnson et al. (2020) used multiple linear regression to generate an operator from the natural variability of a pre-industrial control run to explain the forced response in coupled climate models. Regression approaches successfully eliminate the computational expenses of previous Green’s function methods, but are not optimally designed to infer spatiotemporal-dependent causal links among climate fields.

In this work, we present a method for diagnosing the pattern effect based on the Fluctuation-Dissipation Relation (FDR), also often referred to as Fluctuation-Dissipation Theorem (FDT) (see, e.g., Majda et al. 2005), and response theory (Marconi et al. 2008). Here, we are going to refer to FDR or FDT interchangeably. Roughly speaking, 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 (Marconi et al. 2008; Majda et al. 2005). The Earth’s climate is a multi-scale, complex dynamical system for which the application of the FDR 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. 2022). The protocol proposed here, building on the framework recently presented in Falasca et al. (2024), leverages the FDR formalism and dimensionality reduction techniques to infer a response operator in a coarse-grained representation of the climate system. The success of linear response theory in climate science 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, (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 FDR 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 balances the strengths of the Green’s function protocol (Bloch-Johnson et al. 2024) and previous statistical methods Zhou et al. (2017); Bloch-Johnson et al. (2020). In particular, similar to the Green’s function approaches derived from atmospheric models (Zhang et al. 2023), the framework infers the causal linkages among climate fields in the paradigm of responses to perturbations (Baldovin et al. 2020), and it can be applied directly from data, as in previous statistical approaches. The proposed protocol contributes to the pattern effect literature in three main ways:

- The method gives qualitative predictions of responses to small perturbations in the coupled climate system. In the coupled dynamics, both the atmosphere and ocean would respond to perturbations in the SST field across a wide range of spatiotemporal scales. By accounting for such a complex array of spatial and temporal interactions, we provide a new estimate of the response (sensitivity) of the TOA fluxes to *external* SST perturbations.
- We eliminate the need of prescribing large perturbations to the system to construct the Green’s function operator, therefore ensuring the assumption of linearity when studying the response of TOA fluxes to SST perturbations. The FDT response operator is derived in the limit of infinitesimally small perturbations, for which the linearity assumption holds (Majda et al. 2005; Marconi et al. 2008). Evaluation tests in this paper will show that in the specific context of the pattern effect, and given the coarse-graining procedure proposed in this paper, the FDT formalism can also give good predictions (in terms of global mean) of responses to finite amplitude perturbations.
- The proposed protocol is computationally cheap, allowing for *qualitative* predictions of how the system would respond to perturbations solely as a function of a long control run of a coupled climate model. Most modeling centers have already generated a control run and therefore the proposed method can be applied across every existing models without the need to initiate new model runs.

We note that inferring the response operator in the limit of infinitesimal perturbations is, in general, a potential limitation for predictability studies as climate change often requires studying the response to finite perturbations (Boffetta et al. 2003). However, if the task is to infer the causal relationships among different fields, then the choice of small perturbations is desirable as it allows us to study the dynamics of the system on the attractor itself (Aurell and Del Ferraro 2016; Baldovin et al. 2020)

We leverage the proposed framework in the context of the GFDL-CM4 model (Held et al. 2019). First, we infer the response operator in a long control run. We then evaluate the methodology against two 150 years-long forced runs, respectively, with (i) an incremental 1 percent per year and (ii) an abrupt 4 times increase of CO₂ concentration. Specifically, we show that it is possible to reconstruct the global mean changes in the net TOA radiative flux, solely as a function of the response operator and the forced SST field. We then present a prediction of the response of the TOA radiative flux to perturbations in the SST field. We compare the predicted sensitivity map with the current literature (Zhang et al. 2023) and highlight that the main difference is the strong negative sensitivity of TOA fluxes to SST perturbations all throughout the tropics. We argue that, regardless of the methodology, the discrepancies with the existing literature mainly come from focusing on the atmosphere-ocean coupled system rather than atmosphere-only models. To demonstrate our point, we explore the dependency of our results on the range of time scales considered. We show that we can recover previous results if only the shortest time scales are considered, for which the slow ocean response is largely absent.

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. In Section 5, we present results from the application of the method to the GFDL-CM4 model. In Section 6, we compare the proposed protocol presented here with the Green’s function protocol. The key contribution of this work to diagnose the pattern effect can at first be understood by Sections 2a and 3, respectively focusing on linear response theory and data preprocessing, and the results section. More in-depth information about the method, protocol, and main assumptions are supplied in all other sections. In particular, considerations for the correct application of FDT to realistic climate data, namely coarse-graining, are key to the success of the proposed framework; these are outlined in detail in Appendix Ab and are especially relevant for future applications of FDT.

2. Methods

In what follows, we present the method utilized in this work, which focuses on computing a response operator from the unforced fluctuations of the climate system. The operator, together with suitable convolution formulas, allows us to find the linear response of the system to external perturbations (see Section 2a). In practice, to meet the necessary assumptions in linear response theory, we compute the response operator in a coarse-grained (low-dimensional) representation of the relevant state variables, by focusing on large scale averages and on a limited range of time scales. We detail two dimensionality reduction techniques in Section 2b and later the data preprocessing step in Section 3.

a. Linear response theory and the Fluctuation-Dissipation Relation

Consider a dynamical system written as:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}), \quad (1)$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_N(t)]$ is the state vector. N is the dimension of the system, and is theoretically infinite in the case of spatiotemporal systems. In the case of a General Circulation Model (GCM), N will be of the order of millions, accounting for all variables at all grid points. $\mathbf{F}(\mathbf{x})$ represents the complex dynamical processes advancing the state vector in time.

We now add a perturbation $\delta\mathbf{f}(t)$ on the right side of Eq. (1) as:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) + \delta\mathbf{f}(t), \quad (2)$$

where $\mathbf{f}(t)$ is a generic spatiotemporal perturbation and δ is a small parameter controlling the strength of the perturbation. Given the new system in Eq. (2), the goal is to estimate the average change $\delta\langle x_k(t) \rangle$ of each $x_k(t)$ with respect to the unperturbed (stationary and unforced) system in Eq. (1). Here, the brackets $\langle x_k(t) \rangle$ stand for the ensemble average of the observable $x_k(t)$ ¹. If $\delta\mathbf{f}(t)$ is very small, the leading order response in terms of ensemble average is linear and equal to the following convolution:

$$\delta\langle \mathbf{x}(t) \rangle = \int_0^t \mathbf{R}(\tau) \delta\mathbf{f}(t-\tau) d\tau, \quad (3)$$

with $\mathbf{R}(t) \in \mathbb{R}^{N,N}$ and $R_{k,j}(t)$ representing the response, in terms of ensemble average, of each $x_k(t)$ at time t , given

¹In this paper, we are going to focus on the identity case where the observables of interest are the state variables of our system, $O(x_k(t)) = x_k(t)$. However, we note that the formalism is general and it allows us to study changes in the ensemble average of different functionals, for example the second moment $O(x_k(t)) = x_k^2(t)$. Examples can be found in (Majda et al. 2005; Gritsun et al. 2008)

a small, impulse perturbation of $x_j(0)$. The response at time t is therefore determined by the cumulative effect of the perturbations at all earlier times. In case of a constant, (i.e. step function) external perturbation, expression (3) simplifies even further as a dot product:

$$\delta\langle\mathbf{x}\rangle = \left(\int_0^\infty d\tau \mathbf{R}(\tau) \right) \delta\mathbf{f}. \quad (4)$$

Both Eq. (3) and Eq. (4) will be considered in this paper.

The Fluctuation-Response formalism can be leveraged to establish a rigorous link between the variability of the *unperturbed*, stationary system and its response to external perturbations (Marconi et al. 2008; Hairer and Majda 2010). Given the stationary probability distribution $\rho(\mathbf{x})$ of the dynamical system described in Eq. (1), $\rho(\mathbf{x})$ being sufficiently smooth and non-vanishing, the following result holds:

$$R_{k,j}(t) = \lim_{\delta x_j(0) \rightarrow 0} \frac{\delta\langle x_k(t) \rangle}{\delta x_j(0)} = - \left\langle x_k(t) \frac{\partial \ln \rho(\mathbf{x})}{\partial x_j} \Big|_{\mathbf{x}(0)} \right\rangle. \quad (5)$$

This is the most general form of the Fluctuation-Dissipation theorem (FDT) and is valid for both deterministic and stochastic systems. Eq. (5) allows us to diagnose the response of any dynamical system to infinitesimally small *external* perturbations solely from the unperturbed dynamics of the system.

Importantly, recently Aurell and Del Ferraro (2016) 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, given a dynamical system $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_N(t)]$, cause-effect relations can be inferred by probing the system and examining its response, as done in physical experiments. Specifically, the link $x_j(t) \rightarrow x_k(t + \tau)$ is inferred by studying how an *external* perturbation at variable $x_j(t)$ propagates along the system, inducing *on average* a change in variable $x_k(t + \tau)$ (Baldovin et al. 2020). In the case of small perturbations, the Fluctuation-Dissipation relation shown in Eq. (5) allows us to do so in a straightforward way, by inferring what the response would have been if we had perturbed the system². We refer to the papers of Aurell and Del Ferraro (2016) and Baldovin et al. (2020) for

²We point out the important, but often ignored, difference between methodologies for “causal discovery”, aiming at reconstructing a causal graph from time series, and the more general task of causal inference. Specifically, given three variables $\{x, y, z\}$ and, for example, a causal graph such as $x \rightarrow y \rightarrow z$, the causal discovery method’s goal 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. With this in mind, the variability of both variables x and y will cause the variability of z . We refer to Ay and Polani (2008) and Runge et al. (2015) for details.

further details and comparisons with other methods for causality, such as Granger Causality and Transfer Entropy (Granger 1969; Schreiber 2000). The link between causality and linear response theory is of course not limited to the Fluctuation-Dissipation formalism. We mention the contributions by Lucarini and Colangeli (2012); Lucarini (2018); Tomasini and Lucarini (2021); Lembo et al. (2020) on linear response and causality in the general framework proposed by Ruelle (Ruelle 1998, 2009). Such approaches are very relevant also for future studies on the pattern effect and a brief summary of such tools are detailed in Appendix A. Additionally, we refer to Lucarini and Chekroun (2023) for an excellent review/perspective on Ruelle response theory as a general causal tool to understand the climate response to forcing.

The main practical issue with the formulation in Eq. (5) comes from the fact that the functional form of $\rho(\mathbf{x})$ is not known a priori and it can be highly nontrivial in high-dimensional systems. We note that the machine learning literature has shown promising results in the estimation of the gradients $\nabla_{\mathbf{x}} \ln \rho(\mathbf{x})$ directly from data via generative models (Giorgini et al. 2024), and this can be considered in future applications. However, commonly, the strategy has been to approximate at first order $\rho(\mathbf{x})$ as a Gaussian distribution (Leith 1975). In the case of Gaussian distributions, Eq. (5) reduces to:

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{C}(0)^{-1}, \quad (6)$$

where 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. (6) is valid for linear systems and has been referred to as the “quasi-Gaussian approximation” (Leith 1975; Majda et al. 2010, 2005). The form of FDT shown in Eq. (6) allows us for a first-order estimation of responses in linear systems and motivated many studies in climate applications such as Gritsun and Branstator (2007); Ring and Plumb (2008); Majda et al. (2010); Hassanzadeh and Kuang (2016a,b); Christensen and Berner (2019) mainly focusing on the implications or limitations of this formalism. Furthermore, it has been shown that the quasi-Gaussian approximation in Eq. 6 has high skills for predicting the response in ensemble mean in non-Gaussian regimes (i.e., nonlinear systems) (Gritsun and Branstator 2007; Gershgorin and Majda 2010; Baldovin et al. 2020) and it is highly relevant in climate studies after after spatial and temporal coarse-graining of the system (Sardeshmukh and Sura 2009).

The theoretical tools presented above require computing correlations through ensemble averages, which is impossible in climate applications. The common way to overcome this is through the assumption of ergodicity so that $\overline{O(\mathbf{x})} = \langle O(\mathbf{x}) \rangle$ in the limit $T \rightarrow \infty$; where $\overline{O(\mathbf{x})}$ indicates the time average of a generic physical observable $O(\mathbf{x})$

(Castiglione et al. 2008). Covariance matrices are then computed using temporal averages $C_{i,j}(t) = \overline{x_i(\tau+t)x_j(\tau)}$. In doing so, we always expect 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 independence of any pair of variables x_k and x_j , with $k \neq j$, it is possible to analytically derive a null Gaussian distribution of the response operator $\mathbf{R}(t) = \mathbf{C}(t)\mathbf{C}(0)^{-1}$ as a function of the autocorrelation ϕ_i , the standard deviation σ_i and the length T of each time series $x_i(t)$. The null distribution proposed in Eq. 8 in Falasca et al. (2024) is the analytic form of a red noise null test, commonly used in climate analysis of univariate time series (Dijkstra 2013), in the specific case of (i) multivariate time series and (ii) the Fluctuation-Dissipation theorem in Eq. 6; we refer to Falasca et al. (2024) for details and its derivation. In this paper, confidence bounds of the response operator, are always defined at the $\pm 3\sigma$ confidence level, allowing for uncertainty estimation as a function of both sample size and autocorrelation of the data. The confidence bounds for the response operator allow us to overcome important problems in applications of FDT by extending the computation of the integral such as in Eq. (3) to very large lags. In particular, previous studies such as Majda et al. (2010); Hassanzadeh and Kuang (2016b) evaluated the integral such as Eq. (3) while focusing on upper bounds as short as 30 days (Majda et al. 2010) or by tuning it to have the best performance of FDT (Hassanzadeh and Kuang 2016b). Another source of error comes from spurious results at shorter time scales, which, even if small, can still contribute to biases and accumulation of errors in convolution integrals. The confidence bounds adopted here are leveraged to neglect spurious terms and ensure robust estimations of linear responses.

b. Dimensionality reduction through community detection and Empirical Orthogonal Functions

The formulas proposed in the previous section cannot be applied to the high-dimensional, original system. The dimensionality reduction step is important for a few reasons. First, in the case $N > T$, i.e. the number of grid points is larger than the length of the time series, the covariance matrix $\mathbf{C}(0)$ is not full rank and therefore not invertible; more generally, even if $N < T$, neighboring time series (e.g. $x_i(t)$ and $x_{i+1}(t)$) will be very highly correlated leading to ill-conditioned matrices $\mathbf{C}(0)$ and large errors in its inverse (Gritsun and Branstator 2007; Hassanzadeh and Kuang 2016b). Second, focusing on large scale averages is an important step of the coarse graining procedure, which has been shown to be necessary for FDT to work in practice (e.g., Colangeli et al. 2012;

Sardeshmukh and Sura 2009). Finally, as the climate lives on a low-dimensional attractor (Ghil and Lucarini 2020), we aim (at least in theory) to study the climate in its *effective* dimension (Cvitanović et al. 2016). This allows us to move away from the “gridded-representation” of the system and consider resolution independent modes/patterns as the main blocks of the framework (Cvitanović et al. 2016; Dubrulle et al. 2022).

We are going to consider two methodologies for dimensionality reduction: the scheme recently proposed in Falasca et al. (2024) and the more traditional Empirical Orthogonal Functions (EOF or PCA) (Hotelling 1933). As further explained in Sections 4 and Appendix A, the community detection/clustering approach will be preferred here and the EOF method will be considered as a useful baseline. We assume the reader to be familiar with EOF analysis and refer to Section IIA of Bueso et al. (2020) for details. Here, we very briefly outline the scheme proposed in Falasca et al. (2024) but we refer to Appendix B and to the original paper for more 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) allows us to partition this N dimensional field in terms of n , non-overlapping patterns $c_1, c_2, c_3, \dots, c_n$, with $n \ll N$. The methodology utilizes a community detection algorithm, Infomap (Rosvall et al. 2009) and therefore we are going to refer to the low dimensional variables as patterns, regions or communities interchangeably. Each c_j represent a two-dimensional region defined as a *spatially contiguous* set of time series with large average pairwise correlation. 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) given its previous use in similar studies on the pattern effect (e.g., Zhang et al. 2023) and an available long control run, used for trustworthy computations of the response operator in Eq. 6. 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 also a land component (LM4) and a sea-ice component

(SIS2). Here, we focus on just the GFDL-CM4 model for a proof of concept of the method, but the full protocol can be applied to any model and it will be detailed in Section 4.

We consider three simulations of CM4: a pre-industrial control (piControl), and two idealized scenarios of CO₂ increase, namely 1pctCO₂ and 4×CO₂. The piControl run is a 650 years long, stationary run of CM4 with constant CO₂ forcing at the preindustrial level. The 1pctCO₂ and 4×CO₂ are idealized experiments simulating the climate system under a 1% CO₂ increase per year and an abrupt increase of 4 times CO₂ concentration respectively. Both the 1pctCO₂ and 4×CO₂ experiments start from the preindustrial CO₂ concentration and are run for 150 years. The linear response operator $\mathbf{R}(t)$, shown in Eq. (6), will be 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 (SST) and net radiative flux at the top of the atmosphere (TOA). The TOA flux, hereafter referred to simply as “TOA”, is computed as the (incident shortwave) - (reflected shortwave) - (upward radiative longwave) fluxes³, with all fluxes computed at the top of the atmosphere. All fields are remapped to 2.5° by 2° resolution and only grid cells in the latitudinal range $[-60^\circ, 60^\circ]$ are considered to avoid sea-ice covered areas. The original temporal resolution is 1 day, but our analysis focuses on monthly averages, therefore excluding fast variability at the daily temporal scale and focusing on the slow dynamics (see also Appendix A). Importantly, our analysis focuses only on ocean-covered regions for both the SST (by definition) and TOA variables, as we initially wanted to explore differences and similarities with other approaches, see e.g. Zhang et al. (2023). Utilizing regions above land for the TOA variable is conceptually a trivial extension of the framework, and future studies will be focusing on the temperature-at-the-surface variable rather than SST. We note that our reported global mean results of TOA change, for example, throughout the rest of this paper, are, in fact, for the global mean excluding the poles, which is approximately but not exactly equal to the full global mean.

a. Data preprocessing for control run

Consider a field saved as a data matrix $\mathbf{y} \in \mathbb{R}^{N,T}$ from the stationary, 650-year piControl run; the preprocessing steps outlined below are independent on whether $\mathbf{y} = \text{SST}$ or $\mathbf{y} = \text{TOA}$. N is the number of grid points, T is the length of each time series at monthly frequency, and we refer to $y_i(t)$ as the time series at grid point i . We now 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 $y_i(t)$, calculated as the temporal mean $\mu_i = \overline{y_i(t)}$.
- (iii) 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).
- (iv) High-pass filter the data with a cut-off frequency of 10^{-1} years to remove slow (e.g., multidecadal) oscillations. We stress that the highpass filtering step is performed only in the control run.

After this preprocessing, the resultant time series 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 E). In subsequent sections, control run data will be denoted as $\mathbf{y} \in \mathbb{R}^{N,T}$ and is assumed to have been preprocessed as above.

b. Data preprocessing for forced experiments

We now consider the 150-years long 1pctCO₂ and 4×CO₂ runs. As in the section above, we describe the preprocessing steps for a generic field $\mathbf{y}^f \in \mathbb{R}^{N,T}$, where here f stands for “forced”. We preprocess the forced data in the following way:

- (i) Compute and store the climatology of each time series $y_i^f(t)$, calculated as the temporal mean $\mu_i^f = \overline{y_i^f(t)}$.
- (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).
- (iii) Add the difference in means between the forced and control runs to retain the mean state difference despite removing seasonality; i.e., update $y_i^f(t) \leftarrow y_i^f(t) + \mu_i^f - \mu_i$.

Finally, in the forced experiments, we remove the contribution to the radiative forcing coming from an increase in CO₂ concentration alone, allowing us to isolate and study the TOA radiative feedback to changes in SSTs. The change in radiative forcing driven by changes in CO₂ alone scales 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×CO₂ 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 1pctCO₂ run, where C_0 and C_t are the concentration of

³Each 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”.

CO₂ in the control run and forced experiment, respectively. We fit the α parameter by the change of $\sim 8Wm^{-2}$ in the $4\times CO_2$ simulation. Thus, we have an additional step:

- (iv) Remove the contribution to radiative forcing coming from the CO₂ concentration alone.

Data from forced experiments in subsequent sections should be assumed to have been preprocessed following steps (i)-(iv) above.

4. Proposed data-driven protocol to study the pattern effect

The main steps of the framework 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 two fields: the SST field \mathbf{y}^{SST} and the net radiative flux at the TOA field \mathbf{y}^{TOA} . The linear response operator $\mathbf{R}(t)$ is inferred in the piControl run. Each field is saved as a data matrix $\mathbf{y}^{\text{SST}} \in \mathbb{R}^{N,T}$ and $\mathbf{y}^{\text{TOA}} \in \mathbb{R}^{N,T}$, with N being the number of grid points and T the length of the simulation. Given the spatial and temporal resolution considered this accounts for $N = 5932$ (points on top of the ocean) and $T = 7200$ months (600 years at monthly resolution).

a. Main assumption

The theory in Section 2 assumes that the state vector \mathbf{x} is the whole climate system. This is practically impossible because of the many variables and spatiotemporal scales involved. In line with the ideas first proposed in Hasselmann (1976); Frankignoul and Hasselmann (1977), we are going to exploit the role of time scales separation and switch our discussion from a deterministic to a stochastic description (Penland 1989; Lucarini and Chekroun 2023). The focus is going to be on the SST and net TOA flux variables coarse-grained in both space and time. Therefore, the underlying assumption will be that these variables, averaged over large regions and saved as monthly averages, represent the slow dynamics of the system. The integrated effect of (linear or nonlinear) processes active at (i) small spatial scales and (ii) fast time scales is then considered as noise (Penland 1996; Majda et al. 2005; Penland 2019; Lucarini and Chekroun 2023). Such choices should be kept in mind especially in the causal attribution step (Baldovin et al. 2020), as subsurface ocean variability can also play a role in the pattern effect by shaping the evolution of the SST field. We refer the reader to Appendix A for further discussions on these points.

b. Protocol through community detection

We consider the SST and TOA fields, $\mathbf{y}^{\text{SST}} \in \mathbb{R}^{N,T}$ and $\mathbf{y}^{\text{TOA}} \in \mathbb{R}^{N,T}$ in the stationary piControl run. In order to compute response formulas, we proceed as follows:

- We run the community detection presented in Section 2b for the \mathbf{y}^{SST} field only. We choose this strategy as the simplest option⁴. This step reduces the N dimensional field \mathbf{y}^{SST} into n , *regionally constrained* patterns c_j .
- Given each pattern c_j , we compute its SST time series, $X^{\text{SST}}(c_j, t)$, by computing the integrated SST anomaly inside it as:

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

where θ_i is the latitude at grid point i . The same is done for the TOA field. This step defines two new fields $\mathbf{X}^{\text{SST}} \in \mathbb{R}^{n,T}$ and $\mathbf{X}^{\text{TOA}} \in \mathbb{R}^{n,T}$ where $n \ll N$. In Appendix E, we show that the inferred time series $X^{\text{SST}}(c_j, t)$ and $X^{\text{TOA}}(c_j, t)$ are well approximated by Gaussians, justifying the use of the approximation presented in Eq. 6.

- The new fields \mathbf{X}^{SST} and \mathbf{X}^{TOA} are then standardized by their standard deviation, respectively $\sigma_{\mathbf{X}^{\text{SST}}}$ and $\sigma_{\mathbf{X}^{\text{TOA}}}$. Importantly, the computation of standard deviations of the two fields incorporate the latitudinal weighting as in Eq. (7). The low-dimensional fields are then standardized as $\mathbf{X}^{\text{SST}}/\sigma_{\mathbf{X}^{\text{SST}}}$ and $\mathbf{X}^{\text{TOA}}/\sigma_{\mathbf{X}^{\text{TOA}}}$. To simplify the notation, we are going to keep referring to the new standardized fields as \mathbf{X}^{SST} and \mathbf{X}^{TOA} .
- The new standardized fields \mathbf{X}^{SST} and \mathbf{X}^{TOA} are then joined together to form a single state vector of dimensionality $2n$ as $\mathbf{x}(t) = [\mathbf{X}^{\text{SST}}, \mathbf{X}^{\text{TOA}}](t)$. At each instant in time t , the state of the system is then encoded in the state vector $\mathbf{x}(t)$. The system is $\mathbf{x} \in \mathbb{R}^{2n,T}$.
- The response operator $\mathbf{R}(t)$ is computed in the reduced space using Eq. (6). We leverage the statistical test in shown in Eq. 8 in Falasca et al. (2024) so to neglect later on spurious responses in the computation of convolutions such as Eq. (3). The confidence level considered here is $\pm 3\sigma$.
- To compute integrals such as Eq. (3), we also need a low-dimensional representation of the perturbation pattern $\delta \mathbf{f}(t)$. In our tests, perturbations will

⁴Note that there are potentially other options such as (i) reducing the dimensionality of \mathbf{y}^{SST} and \mathbf{y}^{TOA} separately or (ii) performing the dimensionality reduction after embedding the two fields as $[y_i^{\text{SST}}, y_i^{\text{TOA}}](t)$, where i is a grid point. Here, we choose to reduce the dimensionality of \mathbf{y}^{SST} and \mathbf{y}^{TOA} in the simplest and most interpretable way.

always be applied to the SST field, referred to as $\delta \mathbf{f}^{\text{SST}} \in \mathbb{R}^{N,T}$. For a given perturbation, we compute its low-dimensional representation as done in the case of the state vector $\mathbf{x} \in \mathbb{R}^{2n,T}$, see above. We are going to refer to this perturbation as $\delta \mathbf{f}^{r,\text{SST}} \in \mathbb{R}^{n,T}$, where r stands for reduced. The SST perturbation is concatenated with a zero perturbation field, representing the perturbation in the TOA field. The total perturbation then is made by embedding $\delta \mathbf{f}^{r,\text{SST}}$ and $\delta \mathbf{f}^{r,\text{TOA}}$. The total perturbation field as $\delta \mathbf{f}^r \in \mathbb{R}^{2n,T}$ ⁵.

- Response formulas (Section 2) are computed for the low-dimensional system $\mathbf{x} \in \mathbb{R}^{2n,T}$ using the convolution in Eq. (3). The resulting response is referred to as $\delta \langle \mathbf{x}(t) \rangle \in \mathbb{R}^{2n}$.
- We multiply the first n responses by the standard deviation of the original field $\sigma_{\mathbf{x}^{\text{SST}}}$ to obtain the response $\delta \langle \mathbf{x}^{\text{SST}}(t) \rangle \in \mathbb{R}^n$ of the SST field. The second n responses are scaled by $\sigma_{\mathbf{x}^{\text{TOA}}}$ to obtain the response $\delta \langle \mathbf{x}^{\text{TOA}}(t) \rangle \in \mathbb{R}^n$ of the TOA field. The global mean change in TOA at time t is computed as $\frac{\sum_i^n \delta \langle x_i^{\text{TOA}}(t) \rangle}{A}$, with $A = \sum_i^n \cos(\theta_i)$.

c. Protocol through Empirical Orthogonal Functions

A similar protocol can be proposed by using Empirical Orthogonal Functions (EOFs) see for example Majda et al. (2010); Gritsun and Branstator (2007); Hassanzadeh and Kuang (2016b) among many others. In Appendix C, we present the main steps for studying the pattern effect problem via EOFs.

In this study, we consider the results from community detection as more trustworthy. The preference for community detection will be supported by a few tests, presented in Sections 5 and 6, showing good agreement of our predictions with known responses simulated by the climate model. Hassanzadeh and Kuang (2016b) provides a detailed analysis on how reducing the dimensionality through EOFs can alone lead to major biases in linear response theory studies. Additionally, we note that the choice of how many modes to retain in the EOF step is not obvious: the more modes are retained, the more variance is solved. This could lead to wrong results in FDT applications, where coarse-graining plays an important role (Sardeshmukh and Sura 2009; Colangeli et al. 2012; Penland 2019). Retaining a large number of EOF modes would allow us to resolve finer scales, therefore

accounting for less coarse-graining⁶. Physically, retaining more and more EOF modes, would allow us to resolve finer spatial scales and project any external perturbation down to the resolution of the dataset. However, in doing so, (i) the temporal resolution of the data would also have to adapt by solving faster processes and (ii) variables previously parametrized as noise should now also be included in the state vector. In this study, we choose 50 EOFs respectively for the SST and TOA field, solving respectively 65% and 38% of the variance per each field.

Despite the limitations of EOFs, we include this analysis in our work as: (i) EOFs allow us to visualize responses down to the grid scale, making it easier to qualitatively compare our results to previous ones, such as (Zhou et al. 2017; Bloch-Johnson et al. 2020; Kang et al. 2023; Bloch-Johnson et al. 2024); (ii) by including two dimensionality reduction schemes, we can explore the robustness of our own results across different methodologies.

d. Metrics

The analysis in Section 5 will explore the relation between patterns of SST warming and changes in TOA fluxes through two main metrics: (i) ‘‘sensitivity maps’’ and (ii) cumulative responses. In this section, we describe the computation of the metrics in the context of community detection (see Appendix C for the EOFs case).

(i) *Sensitivity maps.* 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 usually 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 . Such maps show values in units of $[W/(m^2K)]$: positive sensitivity will correspond to positive global radiative feedbacks to local SST forcings. Such positive feedbacks will increase the downward global radiative forcing, amplifying the initial temperature changes. The opposite scenario will be true for negative sensitivity. In the case of community detection where our variables are not at the grid scale, we define an equivalent metric as follows: given a pattern (i.e. community) c_j , with $j = 1, \dots, n$, 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:

$$\Delta T_j = \sum_{i \in c_j} (1K) \cos(\theta_i); \text{ for } t > 0. \quad (8)$$

⁵Note: we refer to the perturbation as $\delta \mathbf{f}$, with the notation implying an infinitesimal perturbation (i.e. $\delta \ll 1$), for consistency with Section 2; however, in practice, we are going to also test the framework outside its range of validity and consider finite perturbation.

⁶In the extreme case of all EOF modes being retained, the EOF step will not account for any coarse-graining but simply for a change of basis.

Where θ_i is the latitude at grid point i . The perturbation field is defined as $\delta \mathbf{f}^r \in \mathbb{R}^{2n, T}$, where the j -th entry of $\delta f_j^r = \Delta T_j$, and all other entries equal to zero (as shown in Section 4b). We then compute the equilibrated linear response of the system to this constant external perturbation as shown in Eq. (4):

$$\delta \langle \mathbf{x} \rangle = \mathbf{\Theta} \delta \mathbf{f}^r; \text{ with } \mathbf{\Theta} = \int_0^{\tau_\infty} d\tau \mathbf{R}(\tau). \quad (9)$$

The upper bound of the integral in Eq. (9), τ_∞ , should be as large as possible, and the theory dictates that $\tau_\infty = \infty$, see Eq. (4). In practice, the implementation should use a time scale τ_∞ that is much larger than the characteristic time of the response. As anticipated in Section 2, one of the main issues in past applications of FDT has been to limit the analysis to very short time scales (e.g., $\tau_\infty \approx 1$ month) in order to avoid spurious results. Here we leverage the confidence bounds shown in Eq. 8 in Falasca et al. (2024) and study the responses at time scales as long as $\tau_\infty = 10$ years. The 10 year time scale allows us to capture the equilibrated response of the system as we will show in Section 5.

Here, $\delta \langle \mathbf{x} \rangle$ gives the response of the whole state vector \mathbf{x} given the perturbation $\delta \mathbf{f}^r$. We then extract the response in the TOA flux $\delta \langle \mathbf{x}^{\text{TOA}} \rangle$ by considering the last n entries of $\delta \langle \mathbf{x} \rangle$, see Section 4b. The sensitivity map $\mathbf{S} \in \mathbb{R}^N$ is a gridded map of the same dimensionality N of 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 :

$$S_i = \frac{\langle \delta \langle \mathbf{x}^{\text{TOA}} \rangle \rangle_G}{\Delta T_j}; \forall i \in c_j, \quad (10)$$

where the brackets $\langle \rangle_G$ refer to the global average. The units of the sensitivity map are $[W/(m^2K)]$.

We anticipate here an important difference with the pattern effect literature. The sensitivity map proposed here integrates all time scales of the response. In the pattern effect literature, similar maps are used to define the radiative feedback as $\partial \overline{\text{TOA}} / \partial \text{SST}_i$, where $\overline{\text{TOA}}$ represent the the global average net flux at the TOA and SST_i is the SST at grid point i . The dot product between such maps and the SST field at time t is then computed as a tool to define the TOA at the same time t . This can be a reasonable step in the case of atmosphere-only model where SST is a boundary condition. Differently, in this study the ‘‘feedback’’ among all variables in the fully coupled system, is considered to be time- and spatially-dependent and encoded in the response operator. The sensitivity map has to be understood as a cumulative feedback/response. This point will be further discussed in

Section 6.

Finally, we remind the reader that the operator $\mathbf{R}(t)$ is a causal estimator (Baldovin et al. 2020). Therefore, even if referring to the maps in Eq. 10 as ‘‘sensitivity maps’’, such results should be understood as a cumulative (over time) *causal* relation of the mean TOA radiative fluxes to perturbations in SST.

(ii) *Cumulative responses.* As shown in Baldovin et al. (2020), the cumulative response shown in Eq. (4) allows us to define a cumulative degree of causation between variables of the state vector $\mathbf{x}(t)$. Theoretically, this metric is preferred as it allows us to analyze the main object at play in the Fluctuation-Dissipation formalism: the response operator $\mathbf{R}(t)$, and it meets to theoretical assumptions by not dealing with finite perturbations. We are going to analyze the cumulative responses of a few patterns j , with every other pattern k , i.e. $\mathcal{D}_{j \rightarrow k} = \int_0^{\tau_\infty} d\tau R_{k,j}(\tau)$ with the goal of showcasing the potential of the framework to diagnose the causal links between local patterns of SST and local changes in TOA. As for the case of sensitivity maps, we are going to consider the $\tau_\infty = 10$ years as large enough to capture the true dynamical response of the system.

5. Results

a. Global responses: testing the framework

We test the proposed framework on the 1pctCO₂ and 4xCO₂ forced experiments presented in Section 3. For both forced experiments, we consider the sea surface temperature (SST) monthly field $\delta \mathbf{f}^{\text{SST}} \in \mathbb{R}^{N, T}$ as a perturbation pattern and aim to predict the monthly change in the global mean top-of-the-atmosphere (TOA) flux as outlined in Section 4. In Figure 1(a), we show the regionally constrained patterns identified by the community detection method (see Section 2). Figure 1(b,c) shows the predicted global mean TOA flux response given the SST perturbation pattern $\delta \mathbf{f}^{\text{SST}}$ in the 1pctCO₂ and 4xCO₂ respectively. Specifically, we show the ensemble mean response of the system $\delta \langle \mathbf{x}^{\text{TOA}}(t) \rangle$ predicted by the convolution of the response operator $\mathbf{R}(t)$ with the SST field (Eq. (3)). The blue curve shows the true TOA values from the coupled climate model. The black and orange lines represent the reconstruction through the two protocols based on community detection patterns and EOFs, respectively (see Section 2). For the 1pctCO₂ experiment, the predicted change in the TOA flux agrees well with the model truth, when using the community detection dimensionality reduction (black line in Fig. 1(b)), the results are degraded when using EOFs. Quantitatively, the trend in the global mean TOA net flux in the model has been found to be equal to $-0.044 \text{ W m}^{-2} \text{ yr}^{-1}$. Trends predicted by the linear response theory are $-0.047 \text{ W m}^{-2} \text{ yr}^{-1}$, $-0.09 \text{ W m}^{-2} \text{ yr}^{-1}$

for the reconstruction using the community detection and EOF method, respectively. For the $4\times\text{CO}_2$, the limits of the framework are more apparent, and the reconstruction fails in both the community detection and EOF cases while still capturing the qualitative tendency of TOA changes. The failure in the case of $4\times\text{CO}_2$ is expected as such an abrupt large change may cause strongly nonlinear changes in the system. On the other hand, more realistic incremental changes, as in the case of the 1pct CO_2 experiment, are predicted well even when reaching a concentration of $4\times\text{CO}_2$ at year 140.

The testing procedure proposed here for assessing the skill of the operator $\mathbf{R}(t)$ corresponds to the one used to test the Green's functions built from atmospheric models (e.g., Zhang et al. 2023). Therefore, we consider Figure 1 as evidence of the validity of the linear response theory framework for pattern effect studies. Furthermore, this also shows that (i) the variables chosen, (ii) the patterns, and (iii) the range of time scales considered are qualitatively reasonable choices to approximate the system as Markovian and meet the assumptions of FDT (see discussion in Appendix A). Another test for the methodology is to compute the correlation between the changes in global mean TOA net flux as predicted by the linear response and simulated by the climate model. To do so, we consider the 1pct CO_2 run for which the community detection protocol gives good predictions. We remove a linear trend and compute the correlation from blue and black curves in Figure 1(b); we obtain a correlation coefficient of $r = 0.36$ and show the detrended time series in Figure 2(a). We remind the reader, that linear response allows us to predict changes in TOA in the ensemble average sense rather than single realizations. A more robust comparison is then obtained by removing noisy month-to-month variability and calculating the correlation after computing a 1-year running mean of the simulated and predicted TOA changes. We then find a much higher correlation of $r = 0.66$ and show the two time series in Figure 2(b). The framework is therefore suitable to study the variability of the system, in terms of ensemble mean, at least in the context of pattern effect studies.

Results in Figure 1 have been computed using the convolution in Eq. (3). Theoretically, the response operator $\mathbf{R}(t)$ will include all time scales, in this case, responses at each month from year 0 to year 150. However, the response operator itself is defined as the response to a small impulse perturbation. Therefore, at long time scales ($t \rightarrow \infty$), 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 and in Figure 3 we explore its value by recomputing the results shown in Figure 1 while setting to zero every $R_{k,j}(t)$ for $t > \tau_\infty$, with $\tau_\infty = 1, 12, 60, 120$ months. We do not utilize $\tau_\infty > 10$ years as the data in the piControl run have been high-pass filtered in the preprocessing step (see Section 3) with a 10^{-1} years cut-off frequency. Specifically, Figures 3(a,b), shows the results obtained through the community detection. The same conclusions have been found in the EOF analysis. Independently of the methodology (EOF analysis is not shown here) and of the CO_2 forcing experiment, a larger τ_∞ always corresponds to better results as expected. $\tau_\infty = 1$ month shows a very poor reconstruction in global mean TOA changes. Increasing τ_∞ up to 1 year already results in much better performance, up to ~ 70 years. Results show convergence for $\tau_\infty = 5$ years. This analysis shows that $\tau_\infty \gtrsim 5$ years is long enough to capture the characteristic time scales of the system's response. In the analysis that will follow, $\tau_\infty = 10$ years will be used to compute the upper bound of integrals such as in Eq. (4).

b. Sensitivity of global mean TOA fluxes to local SST perturbation patterns

Here, we focus on the main question: What is the relationship between patterns of SST and changes in the global average TOA radiative fluxes? A common way to investigate this relationship is through the sensitivity maps proposed in Section 4d (see for example, Zhang et al. 2023; Bloch-Johnson et al. 2024). The key idea is to plot the cumulative response of the global mean TOA fluxes induced by a constant, step function perturbation $\delta\mathbf{f}$ of 1 Kelvin at each grid point i at the sea surface. Positive sensitivity values at grid point i will imply positive radiative feedbacks, amplifying the original temperature change; the opposite is true for negative sensitivities. As shown in Section 4d (and in Appendix C in the case of EOFs), the analysis requires us to compute integrals of the form $\left(\int_0^{\tau_\infty} d\tau \mathbf{R}(\tau)\right)\delta\mathbf{f}$ and we use $\tau_\infty = 10$ years, as discussed in the previous section.

The sensitivity map obtained in this work is shown in Figure 4. The two-dimensionality reduction techniques agree on a large negative sensitivity in the tropics, especially across the whole tropical Pacific. This means that for the same global mean warming, if tropical basins warm up more than the higher latitudes, then the system will counteract the warming more by radiative cooling in the GFDL-CM4 model. As noted in Section 4c, we assign more weight to results coming from the community detection as it showed the best skill in reconstructing the global mean TOA response in the forced experiments. Additionally, in contrast to EOFs, community detection allows us to perfectly prescribe the location of the

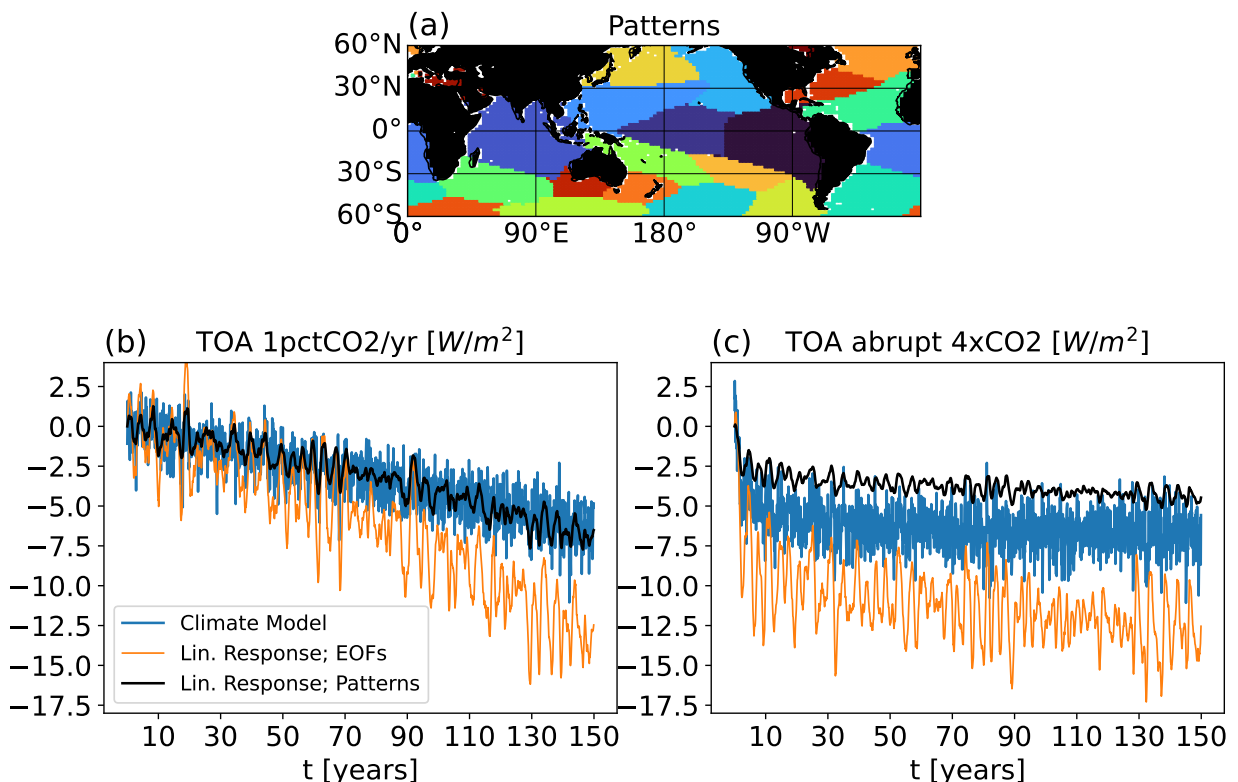


FIG. 1. Panel (a): Low-dimensional representation of the system via community detection as “patterns”. Our analysis focuses on monthly climate variability integrated in each individual region shown here, see Sections 2 and 4. Different colors are used to distinguish amongst different patterns. Panels (b-c): changes in global mean radiative flux at the top of the atmosphere (TOA) in the 1pctCO2 (b) and 4xCO2 (c) experiments (see Section 3). TOA as simulated by the fully coupled GFDL-CM4 model in the two idealized climate change experiments is in blue. A prediction of TOA by linear response theory solely as a function of the changes in the sea surface temperature spatiotemporal field is in black. TOA using EOF analysis, which replaces the community detection as the dimensionality reduction step is in orange. Linear responses have been computed using the convolution in Eq. (3). In panel (b), the trend in the global mean TOA net flux is equal to: $-0.044 \text{ Wm}^{-2}\text{yr}^{-1}$ for the model, $-0.047 \text{ Wm}^{-2}\text{yr}^{-1}$ and $-0.09 \text{ Wm}^{-2}\text{yr}^{-1}$ respectively for the reconstruction from the community detection and EOF method. For the analysis of the month-to-month correlation, see Appendix E.

perturbation patterns up to the resolution of the patterns themselves, whereas EOFs can result in a non-local projection of the perturbation. As opposed to previous studies (e.g., Zhang et al. 2023), the higher latitudes also appear to play an important role in the radiative feedback and generally display positive feedback, with small disagreements depending on the dimensionality reduction method adopted. Such contribution should be explored more in future studies; here, instead, we mainly focus on the robust and large negative sensitivity in the tropics.

The sensitivity maps shown in Figure 4 differ substantially from what has been found in Green’s model experiments (e.g., Bloch-Johnson et al. 2024), and in statistical studies, for example, as in (Kang et al. 2023). An important difference comes from the marked negative sensitivity in the equatorial central to eastern Pacific,

which is absent in previous studies. Apart from the details of the methodology, we hypothesize that the main qualitative difference comes from the absence of atmosphere-ocean coupling in previous Green’s function experiments, where the TOA radiative flux response is computed after iteratively perturbing the SST boundary conditions (Dong et al. 2019). As we demonstrated in Figure 3, the response operator $\mathbf{R}(t)$ can be considered as equal to zero after temporal scales of $\tau_\infty = 5$ years, and results in Figure 4 have been computed with $\tau_\infty = 10$ years. This means that $\tau_\infty = 5$ to 10 years is long enough for the atmosphere-ocean coupled system to equilibrate after a step function perturbation imposed on the SST field, at least given the precesses and time scales considered here. Under the assumption of a fast response of the atmosphere compared to the ocean, we attempt to study sensitivity maps in the absence of the atmosphere-ocean coupling by assuming $R_{k,j}(t) = 0$ for $t > \tau_\infty$, with $\tau_\infty = 1$ month (i.e.,

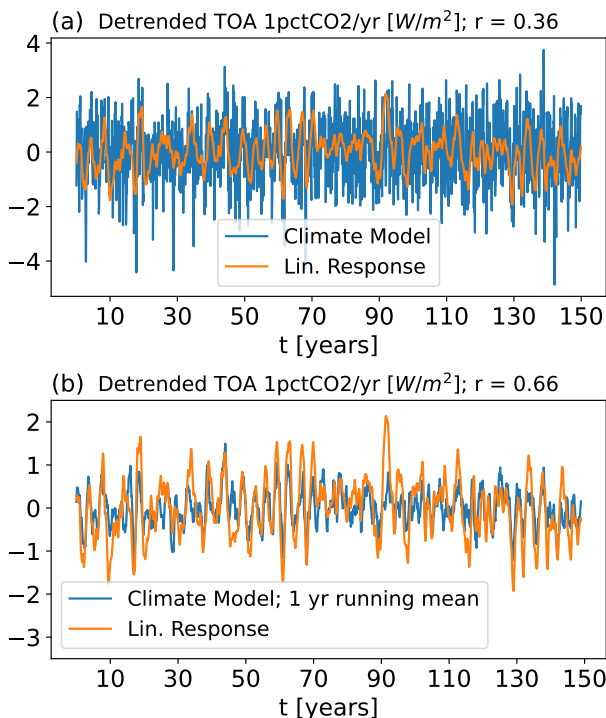


FIG. 2. Panel (a): Detrended global mean changes in net radiative flux at the TOA as reconstructed by linear response theory and simulated by the climate model in the 1pctCO₂ run. Panel (b): same as panel (a) but after smoothing the timeseries from the climate model by computing a 1-year running mean. For the linear response theory reconstruction, we focus on the results from using the protocol with community detection as the dimensionality reduction step (i.e., the black curve in Figure 1(b).)

the shortest time scale considered in this work). We then explore how the integration of different processes at time scales $\tau_\infty = 1$ and $\tau_\infty = 5$ years contribute to the sensitivity map in Figure 4.

We start by discussing the shortest possible time scale of $\tau_\infty = 1$ month as shown in Figure 5(a,b). This time scale is too short to observe a significant response of the global ocean to *local* SST perturbations and can be seen as qualitatively similar to the system explored in previous Green’s function approaches involving atmosphere-only models (Bloch-Johnson et al. 2024; Zhang et al. 2023). Specifically, we note a close resemblance between our results and Kang et al. (2023) who used Ridge Regression (compare our Figure 5(b) to Extended Data Fig. 7(b) in Kang et al. (2023)). 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 Bloch-Johnson et al. (2020); Zhang et al. (2023); Bloch-Johnson et al. (2024). We find that results are

qualitatively independent of the dimensionality reduction.

The marked dipole in sensitivities found in the tropical Pacific at shorter time scales is no longer present as we integrate over longer time scales. Figure 5(c,d) depicts the sensitivity map computed with $\tau_\infty = 1$ year. Independent of the dimensionality reduction technique, the tropical Pacific, with the epicenter in the central Pacific, emerges as the largest contributor to changes in the global TOA flux and has negative sensitivity. Finally, the integration of responses up to $\tau_\infty = 5$ years (Figure 5(e,f)) qualitatively agree in terms of sensitivity patterns with the one computed with $\tau_\infty = 10$ years (Figure 4), further confirming that $\tau_\infty = 5$ years is long enough for equilibration of the system to a SST perturbation, as shown for the global mean in Figure 3.

The marked difference between sensitivity maps computed by integrating over only the short (i.e., $\tau_\infty \sim 1$ month) or overall (i.e., $\tau_\infty \geq 5$ years) time scales, as discussed in this Section, has a physical interpretation directly related to coupled climate dynamics. A temperature perturbation pattern prescribed at the ocean surface will impact the radiative balance at both the TOA and surface at very short time scales. However, at longer time scales, the initial SST perturbation will also impact the ocean itself, interacting with the ocean dynamics and subsequently resulting in non-local coupled processes such as teleconnection patterns. Such processes result in a cascade of additional surface perturbation patterns at later times and in basins far away from the originally perturbed location, acting effectively as new SST perturbations driven by the dynamics of the system itself. Regardless of the location where the SST perturbation is prescribed, introducing an *external* perturbation in the coupled climate system can produce nontrivial feedbacks among modes of variability across a vast range of spatial and temporal scales (Ghil and Lucarini 2020). Such a cascade of changes is not independent of the initial perturbation pattern: it is *caused* by it. The linear response theory formalism adopted here allows us to integrate these effects across time and spatial scales when diagnosing radiative feedbacks.

c. Exploring a few local-to-local cumulative responses

We end the results section by presenting a brief qualitative analysis of the response of the whole system (SST and TOA fluxes) to perturbations in SST in specific regions: the eastern, central and western tropical Pacific and a region in the Southern Ocean. This analysis is performed to showcase a useful way to use the proposed framework to study causal linkages among climate variables, following Baldovin et al. (2020) and Falasca et al. (2024). The

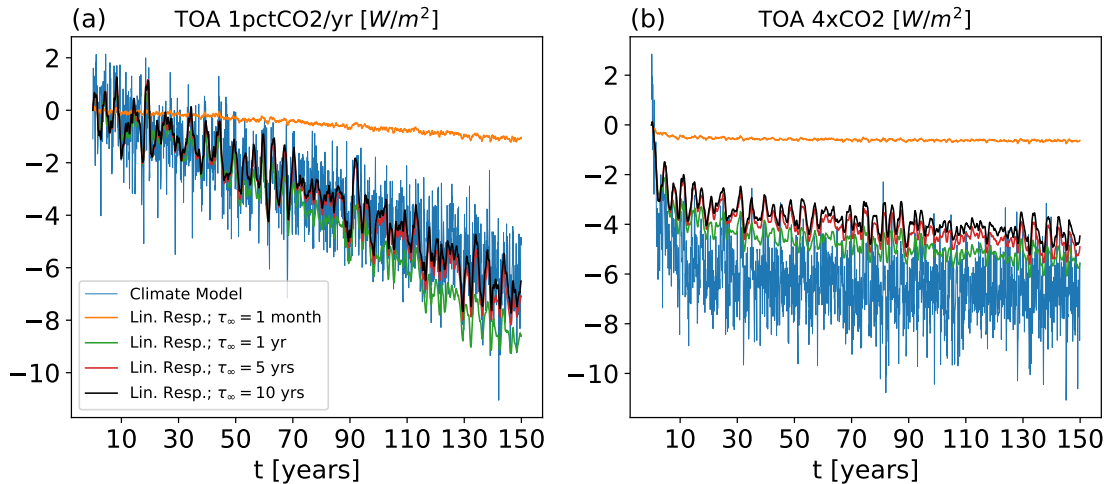


FIG. 3. Exploration of the characteristic time scales of the TOA response by predicting the change in the net radiative flux at the top of the atmosphere (TOA) as a function of the parameter τ_{∞} . We compute the convolution in Eq. (3) after setting $R_{k,j}(t) = 0$ if $t > \tau_{\infty}$, with $\tau_{\infty} = 1, 12, 60, 120$ months. The TOA flux, as simulated by the fully coupled GFDL-CM4 model, is in blue. The other curves show a prediction of the TOA flux solely as a function of the changes in the sea surface temperature field using the convolution in Eq. (3). In panels (a) and (b), we show the analysis via the community detection protocol for the 1pctCO2 and 4xCO2 experiments. The analysis shows that $\tau_{\infty} = 5$ or 10 years is long enough to capture the characteristic time scales of the system’s response, and it can be, therefore, used as the upper bound $t \rightarrow \infty$ in integrals such as in Eq. (4). A similar result of panels (a) and (b) holds for when using EOFs rather than community detection for dimensionality reduction.

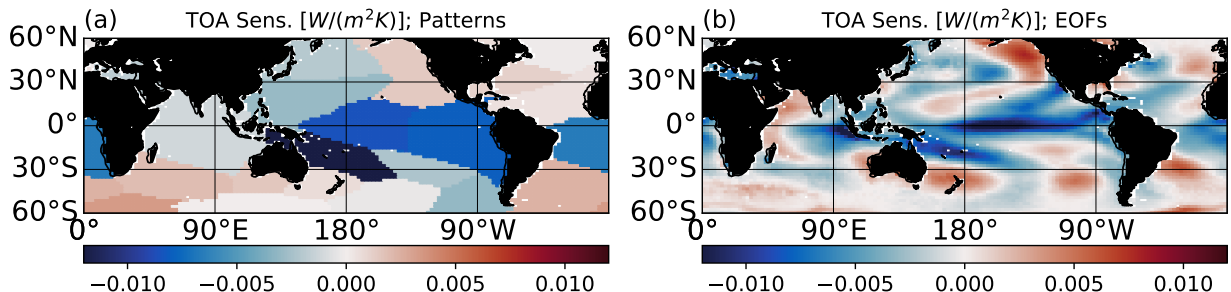


FIG. 4. Sensitivity maps \mathbf{S} . Panel (a): map obtained via the community detection protocol. Panel (b): map obtained via the EOF-based protocol. At each grid point i , we plot the equilibrated, global average response of the net radiative flux at the TOA given a constant perturbation of 1 Kelvin in SST imposed at that point. Positive sensitivity corresponds to a positive radiative feedback, therefore amplifying the initial temperature changes; the opposite is true for negative sensitivity. Results have been computed as shown in Section 4d. The two methods show qualitative agreement. The parameter τ_{∞} in Eq. 9 has been set to $\tau_{\infty} = 10$ years. In Figure 5, we show that 10 years is a long enough time to capture the characteristic time scale of the system’s response.

analysis shown here, does not claim or attempt to propose a mechanistic view of each linkage as this is not a focus of this paper. However, as argued first in Baldovin et al. (2020), in this section, we stress that the response operator $\mathbf{R}(t)$ alone, is a useful starting point to diagnose the dynamics behind climate interactions which could be considered mechanistically in future studies. The analysis shown here focuses on cumulative causal linkages, computed by integrating the $\mathbf{R}(t)$ in time (see Section 4d). The analysis reinforces the considerations on time scales shown in the Section above, while also showing the response of both fields to specific regional perturbations in SST. The

results, shown in Figures 6, 7 and in Appendix D, Figures D1, D2, outline the following qualitative picture:

- If only fast time scales (i.e., $\tau_{\infty} = 1$ month) are considered, a local perturbation in the SST field will propagate non-locally in the TOA field, but remain largely local in the SST field. This is clear in Figure 6. For example, the SST response to a perturbation applied to the central Pacific (Figure 6(c,d)) is largely constrained to the location of the perturbation; however, due to fast atmospheric dynamics, there are some non-local responses in the TOA flux field. For this time scale, we obtain the worst reconstruction in terms of

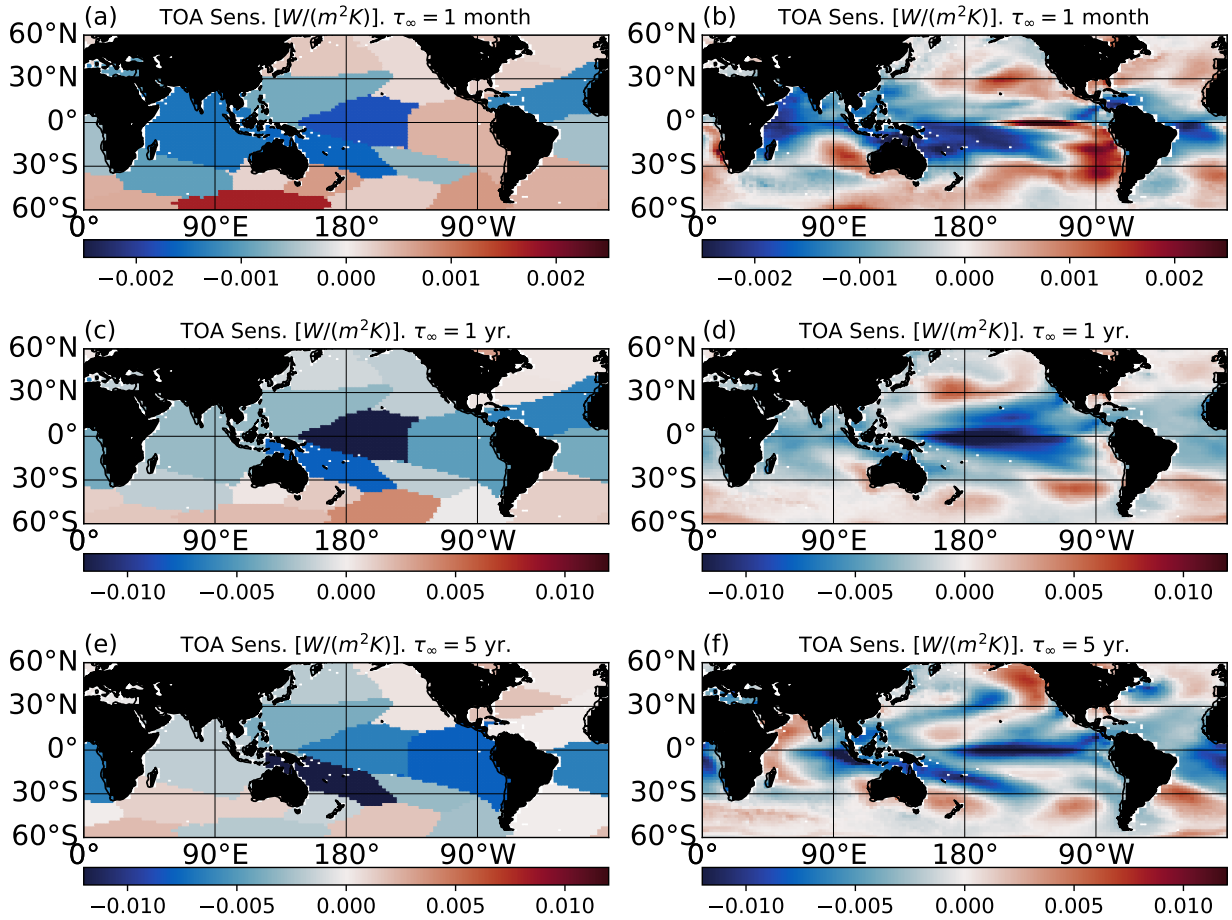


FIG. 5. Exploration of the cumulative contribution to the sensitivity maps in Figure 4 as coming from different time scales. Sensitivity maps are computed as shown in 4d but by setting the response operator $R_{k,j}(t) = 0$ to zero if $t > \tau_\infty$, with $\tau_\infty = 1, 12, 60$ months. Results shown for $\tau_\infty = 5$ years, i.e. Panels (e,f), are very similar to the ones shown in Figure 4 with $\tau_\infty = 10$ years. Therefore $\tau_\infty = 5$ years is long enough to capture the characteristic time scales of the system’s response, and it can be therefore used as the upper bound $t \rightarrow \infty$ in integrals such as in Eq. (4). The analysis confirms the findings shown in Figure 3. Physically, focusing on the shortest time scales, as for $\tau_\infty = 1$ month, allows us to approximately explore what the sensitivity map would have looked in the absence of atmosphere-ocean coupling. In panels (a,c,e) and (b,d,f), we show the results obtained via community detection or EOF-based protocols. The two methods show qualitative agreement.

global mean TOA changes in the 1pctCO₂ and 4xCO₂ forced experiments, see Figure 3, as this is far from the theory which necessitates $\tau_\infty \rightarrow \infty$. However, to first order, this is the picture that could arise in atmospheric models only, where the SST is a boundary condition rather than a fully interacting component of the system.

- If responses are integrated up to $\tau_\infty = 1$ year, teleconnection patterns, mostly driven by tropical wave dynamics, cause local perturbations to spread over the whole SST field, further impacting the TOA fluxes by effectively perturbing remote SST regions at later times. This is shown in Figure 7. Cumulative responses in the Southern Ocean, see Figure 6(h), are larger than in $\tau_\infty = 1$ month case, but not as large as

what we observe for tropical Pacific domains. We hypothesize that this may a general case for high-latitudes variability. For this case, we observe a large improvement in the skill of our reconstruction of global mean TOA changes in the 1pctCO₂ and 4xCO₂ forced experiments, see Figure 3. The reconstruction is good initially but starts deteriorating after ~ 70 years as $\tau_\infty = 1$ year may be too short for processes active at the high latitudes to significantly impact TOA fluxes.

- If responses are integrated up to $\tau_\infty = 5$ years, we observe a larger cumulative response in the TOA flux driven by perturbations applied to the Southern Ocean. See Figures D1 and D2 in Appendix D. We hypothesize that such considerations may apply

generally to perturbations in higher latitudes, but this requires future work focused solely on teleconnection patterns instead of the general qualitative analysis shown here. Such processes, active at longer time scales, correct the global mean TOA reconstruction in the 1pctCO₂ and 4xCO₂ forced experiments as shown in Figure 3.

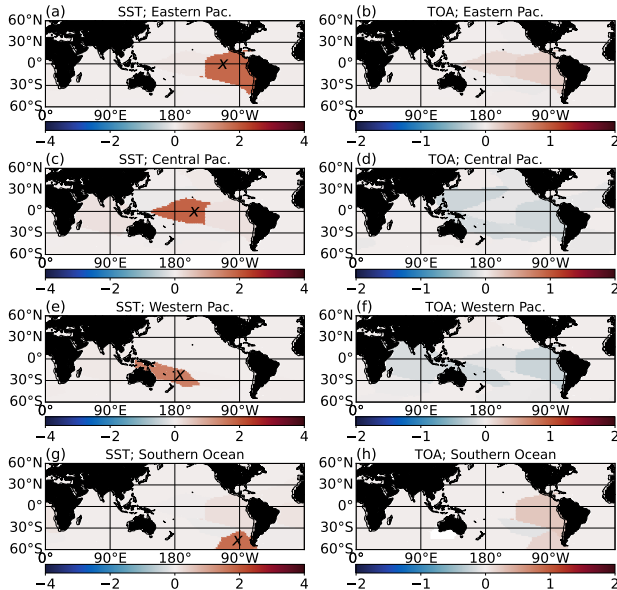


FIG. 6. Cumulative responses $\mathcal{D}_{j \rightarrow k} = \int_0^{\tau_{\infty}} d\tau R_{k,j}(\tau)$ (see Section 4d) for the shortest time scale, $\tau_{\infty} = 1$ month. Physically, focusing on this shortest time scale allows us to approximately explore responses in the absence of atmosphere-ocean coupling, at least at first order. Panels (a,b): cumulative response of the sea surface temperature (SST) and radiative flux (TOA) flux fields to an infinitesimally small impulse (rather than a constant) perturbation prescribed on pattern in the eastern Pacific region, denoted with an “x”. Panels (c,d), (e,f), (g,h): same as panel (a,b) but when the perturbation is applied to the central and western Pacific and a region in the Southern Ocean, respectively. Values are unitless: positive values represent positive responses to perturbations in the location marked by the “x”.

6. Comparison with Green’s function model experiments

In this paper, we have introduced a comprehensive protocol to investigate the pattern effect. Specifically, given the operator $\mathbf{R}(t)$, estimated as in Eq. (6) only from a piControl, stationary climate model run, the theory allows us to compute the climate response to external perturbations by evaluating convolution integrals, see Eq. (3). In this section, we detail a few key differences between our approach and the standard Green’s function protocol as in (Dong et al. 2019; Bloch-Johnson et al. 2024).

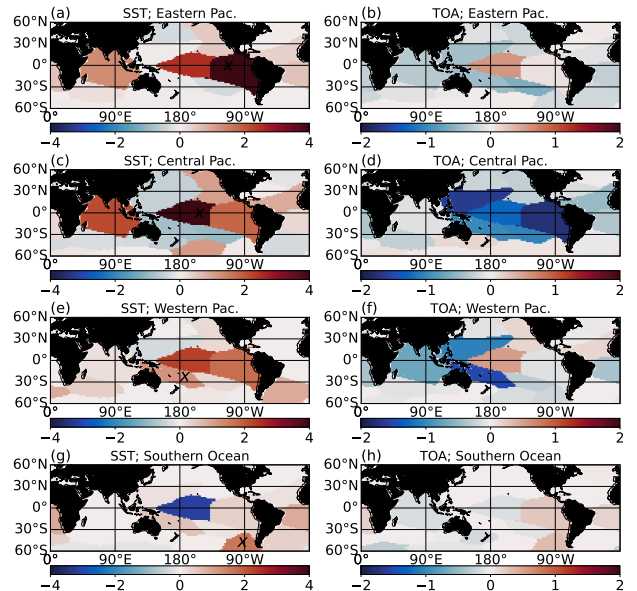


FIG. 7. As in Figure 6, but with $\tau_{\infty} = 1$ year.

a. Conceptual difference

The framework proposed in this paper is conceptually different from other approaches for the pattern effect proposed in the literature. One of the main differences relates to the sensitivity maps presented in Section 5. In the “Green’s function protocol” (Bloch-Johnson et al. 2024; Zhang et al. 2023; Dong et al. 2019), and in its data-driven versions (Bloch-Johnson et al. 2020; Kang et al. 2023), such maps are what is commonly referred to as the radiative feedback and defined at each grid point i by $\frac{\partial \overline{\text{TOA}}}{\partial \text{SST}_i}$, where $\overline{\text{TOA}}$ is the globally averaged net radiative flux at the TOA and SST_i the SST at point i . This map can be a useful tool as its dot product with an imposed SST pattern at time t will return back an estimate of the globally averaged change in net flux $\overline{\text{TOA}}$ at the same time t . This also means that an imposed change in SST patterns at time t cannot drive changes in average TOA fluxes at later times $t + \tau$.

The main object of our proposed protocol is instead 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 $\delta x_j(t) \rightarrow 0$. The first clear difference is given by the presence of the brackets $\langle \rangle$ stating that predictions of response theory are done in terms of ensemble averages rather than individual realizations. Second, the 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. Third, and more important, the operator $R_{k,j}(\tau)$ 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 $\delta\mathbf{x}(t)$ at time t is then computed by the integrated effect of the perturbation patterns across all previous time scales $t - \tau$ (Christensen and Berner 2019). The sensitivity map found in this study, defined in Section 4d and shown in Figure 4, represents the “cumulative response/feedback” after reaching equilibration, and should not be considered as a map to be taken the dot product with the SST field at time t .

In what follows, we then show that the same ideas of the “Green’s function protocol” can be reformulated using the sensitivity maps proposed in Figure 5(a,b), therefore providing a bridge between the two frameworks.

b. Relating the two approaches

The “Green’s function protocol” presented in Bloch-Johnson et al. (2024) relates the SST change $\Delta T_i(t)$ at grid point i and time t , with the global mean change in TOA flux $\Delta\overline{\text{TOA}}(t)$ at the same time t (see Eq. 5 in Bloch-Johnson et al. (2024)). Importantly, the SST is imposed as a boundary condition for the atmosphere, and it cannot respond to perturbations coming from either atmospheric or oceanic fields (as it would in a true coupled system). In such a setup, the atmospheric model will equilibrate to a step function SST perturbation pattern very quickly, and the τ_∞ in Eq. (9) in our paper can be approximated with very small τ_∞ . Thus, it is reasonable that linear regression approaches have shown similar skills for the same task; see, for example, Bloch-Johnson et al. (2020); Kang et al. (2023). Here, we show that the same approach can be formulated with the results obtained in this paper and without the need for expensive model runs, therefore providing a bridge across methods. Specifically, we consider the sensitivity map \mathbf{S} at the shortest time scale, $\tau_\infty = 1$ month, as shown in Figures 5(a,b) respectively for the community detection and EOF dimensionality reduction approaches. Under the assumption of a fast response of the atmosphere compared to the ocean, the shortest time scale of $\tau_\infty = 1$ month will correspond, at very first order, to a system without the atmosphere-ocean coupling, similar to the protocol in Bloch-Johnson et al. (2024). We then write the changes in global mean TOA fluxes at time t ($\Delta\overline{\text{TOA}}(t)$) simply as a dot product of the sensitivity map \mathbf{S} with changes in the temperature pattern. Formally: $\Delta\overline{\text{TOA}}(t) = \sum_i S_i \Delta T_i(t)$, which exactly mirrors Eq. 5 in Bloch-Johnson et al. (2024). Results are shown in Figures 8(a,b) for the 1pctCO2 and 4xCO2 forced experiments. We stress that the theoretically and physically sound approach using linear response theory is to perform convolution integrals of the response operator as in Eq. (3); the test proposed here is only to explore similarities and provide a bridge across different perspectives on the pattern effect. The regression approach provides an instantaneous mapping in between the SST field and global

mean TOA fluxes. Such mapping is useful in the context of atmosphere-only models where SST is a boundary condition, and, as shown below, it can be a very powerful tool for statistical reconstruction studies. If we limit the methodology to a regression approach, we then find an excellent reconstruction of the global mean TOA flux, outperforming existing reconstructions, see for example (Zhang et al. 2023). Surprisingly, the best reconstruction comes from using the sensitivity map defined in Figure 5(a) where the regression coefficient S_i is the same in large ocean areas rather than the more detailed analysis at the grid level as in Figure 5(b). We give the following qualitative explanation: 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(a) allows us to effectively linearize the mapping between temperatures changes $\Delta T_i(t)$ and global mean TOA fluxes $\Delta\overline{\text{TOA}}(t)$. In other words, focusing on the right observables, defined here by integrating changes in temperature over large spatial patterns, allows us to (i) average out small and nonlinear contributions at the grid scale and (ii) focus on the right spatial scales given the temporal resolution of our data.

7. Conclusion and discussion

The dependence of the TOA radiative fluxes on the patterns of SST warming, known as the pattern effect, determines the temporal evolution of the climate feedback parameter and, thus, future climate sensitivity (Armour et al. 2013; Zhao 2022). In this work, we developed a protocol based on coarse-graining and the Fluctuation-Response formalism to diagnose and understand the pattern effect from data in fully-coupled climate models. Two main classes of methods have been utilized previously to investigate the pattern effect: (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). 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 studies the response of the atmosphere to SST perturbations, and it can be applied using only a model’s control run, as in previous statistical approaches. Linear response theory allows us to infer what the response of a dynamical system to small perturbations would have been without actually perturbing the system. 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). An important difference with previous studies is of conceptual nature:

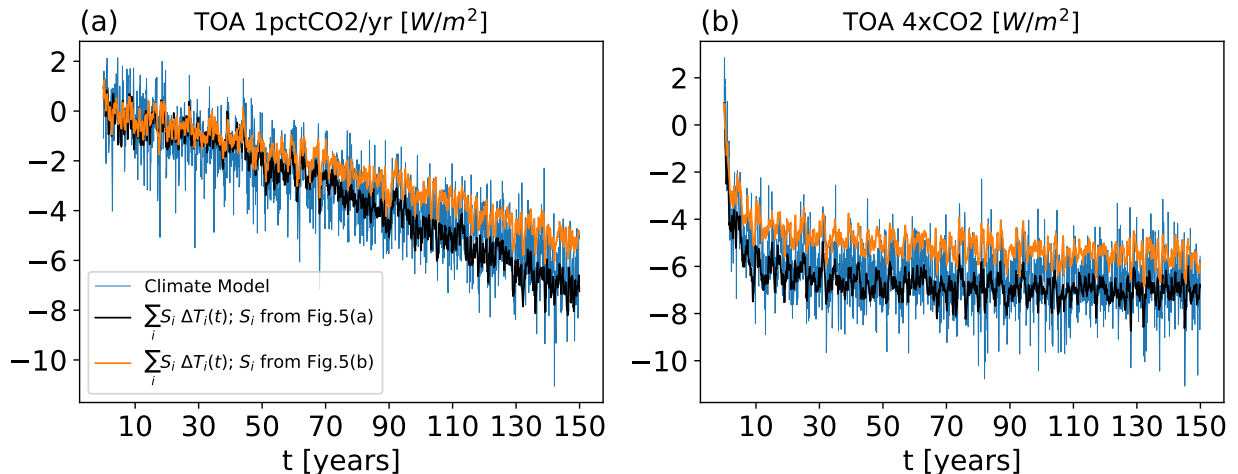


Fig. 8. Analysis showing that the approach proposed in the “Green’s function protocol”, as in Zhang et al. (2023); Bloch-Johnson et al. (2024), can be qualitatively reproduced using the sensitivity maps for the shortest time scale $\tau_\infty = 1$ month, shown in Figure 5(a,b). Given the sensitivity maps \mathbf{S} shown in Figures 5(a,b) and the spatial sea surface temperature field $\Delta\mathbf{T}(t)$ at time t , for the 1pctCO2 and 4xCO2 forced experiments, we reconstruct the global mean change in TOA flux at time t by a regression approach $\Delta\text{TOA}(t) = \sum_i S_i \Delta T_i(t)$ as done in Bloch-Johnson et al. (2024).

what is referred to as “feedback” here is encoded in a time- and spatially-dependent response operator $\mathbf{R}(t)$ derived from the coupled climate system. Sensitivity maps, are then computed from the operator as cumulative (in time) feedbacks.

The proposed approach is general and simple to apply in practice to infer the response of climate fields to small *external* perturbations from data. The method requires two main ingredients: (i) coarse-graining the system (spatially and temporally) in terms of physically relevant, projected dynamics; (ii) computing a response operator utilizing the variability of a long and stationary control simulation only. In our application, we utilize two forms of dimensionality reduction: community detection (Falasca et al. 2024) and EOFs. Previous studies using the Fluctuation-Response relation have typically used EOFs (e.g., Majda et al. 2010; Gritsun and Branstator 2007). However, it has been shown how EOF dimensionality reduction alone can lead to large errors in FDT applications (e.g., Hassanzadeh and Kuang 2016b). We argue that the dimensionality reduction technique using community detection can overcome some of the errors driven entirely by the utilization of EOFs.

The main outcome of the framework is a first qualitative prediction of how the TOA radiative fluxes would respond to SST changes in perturbation experiments performed in fully coupled climate models. Such cumulative responses, summarized in the sensitivity map in Figure 4, shows a negative sensitivity of the global mean TOA radiation fluxes to SST warming all throughout the tropics. Physically, this means that for a given global mean surface warming, a constant ocean warming in the tropics rather

than in the higher latitudes will lead to negative radiative feedbacks, therefore partially counteracting the initial warming through radiative cooling. On average, the opposite (with smaller absolute magnitude) result holds for warming in higher latitudes, with details dependent on the dimensionality reduction method adopted. This differs from previous results, where an important feature of the sensitivity TOA fluxes to SST changes is a dipole in the tropical Pacific, with positive and negative sensitivity respectively on the east and west side of the basin. However, in the limit of considering only the shortest time scales, here $\tau_\infty = 1$ month, we recover similar regional sensitivities as in the existing literature. We give the following explanation: if only the shortest time scales are considered, then the system’s response is dominated by fast time scale processes in the atmosphere, similar to the atmosphere-only model with the ocean acting only as a boundary condition. Therefore, we conclude that the difference between the sensitivity map proposed in this study and presented in the literature, come largely from the inclusion of the atmosphere-ocean coupling rather than methodological differences.

The results in this work also 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. In fact, we highlight in Appendix A that the validity and performance of linear response theory through FDT, is largely dependent on the spatial and temporal scales and the fields considered. Given a careful consideration of the coarse-graining steps, the FDT can be utilized throughout climate science beyond what is exam-

ined 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, including outlining limitations of the fluctuation-dissipation theorem (or relation, i.e., FDT or FDR). Then we point out a few considerations that allowed us to successfully apply the fluctuation-response formalism in the climate context and in the specific case of the pattern effect. The considerations that follow are not only technical details, but important steps for correct applications of the FDT formalism for 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 (Leith 1975; Gritsun and Branstator 2007; Majda et al. 2005; Majda and Qi 2018; Falasca et al. 2024) and (ii) its response to external forcings (Lucarini and Chekroun 2023; Lucarini et al. 2017; Lucarini and Colangeli 2012; Lembo et al. 2020; Gritsun and Lucarini 2017; Basinski-Ferris and Zanna 2023). 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 lots of 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 (Sarracino and Vulpiani 2019; Ghil and Lucarini 2020; Christensen and Berner 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 CO₂ 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), therefore allowing us for causal attribution of climate change signals (Lucarini and Chekroun 2023).

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. A few notes on the application of FDT in climate science

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) 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 in fact 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.

In what follows, we argue that a few of the drawbacks of the FDT, in its quasi-Gaussian implementation (see Section 2), can be avoided if focusing on the proper variables and a limited range of spatiotemporal scales through coarse-graining. The observations that follow should be carefully considered (and stated) for a correct application of the framework and theory presented in this work.

Choosing the proper variables. It is well known that the performance of FDT depends strongly on the observables of interest (Lucarini and Colangeli 2012; Gritsun and Lucarini 2017). Theoretical studies of FDT often focus on the perfect scenario where we have access to the full state vector. This is far from true in real-world cases where we can only access a few variables (Baldovin et al. 2020; Falasca et al. 2024). A solution in climate studies is offered by exploiting the role of time scale separation as first explored by Hasselmann (1976); Frankignoul and Hasselmann (1977) and studying the system as a stochastic

dynamical system, where the noise plays the role of unresolved physical degrees of freedom (Penland 1996; Majda et al. 2005). As noted in Penland (2019), the processes active at (i) small spatial scales and (ii) fast time scales can still be very nonlinear, however their integrated effect over the coarse-grained variables can be sometimes well-approximated by white noise⁷. Our application considers the coarse-grained SST and radiative fluxes at the TOA as the important, slow deterministic components to understand the pattern effect. Focusing on the relationship between these two variables to diagnose the pattern effect is justified by previous studies (e.g., Dong et al. 2019; Bloch-Johnson et al. 2020; Zhang et al. 2023). However, this is a simplification of the original problem (involving the whole system) and it should be kept in mind especially in the causal attribution step (Baldovin et al. 2020). The spatial and temporal (discussed in the next paragraph) coarse-graining is an essential step in choosing the proper variables. Climate studies using FDT overwhelmingly focused on EOFs as dimensionality reduction. However, Hassanzadeh and Kuang (2016b) noted that the EOF step alone can be a major source of errors in FDT applications, and as noted in the main text, many contributions focused on retaining a large number of EOF modes therefore undermining the coarse-graining procedure. In Figure 1 we showed a much better skill of the response operator inferred when considering the dimensionality reduction method proposed in Falasca et al. (2024). Qualitatively, given the large size of the patterns in Figure 1(a), and the temporal resolution of 1 month, the system can be considered at first order as Markovian (Baldovin et al. 2020) and the methodology in Section 2 is then relevant. Finally, we briefly note that the Takens embedding theorem (Takens 1981), commonly used for reconstructing the attractor of a (deterministic) dynamical system from partial observations, cannot be a valuable option for high-dimensional system and refer to Baldovin et al. (2018, 2020); Lucente et al. (2022) for further details.

Assumption of linearity. The proposed framework utilizes a particularly simple form of FDT presented in Eq. (6), 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 clearly nonlinear, and it is therefore not obvious why Eq. (6) should work. First, we note that the Gaussian assumption gives qualitative good results even for nonlinear systems (Gritsun and Branstator 2007; Gershgorin and Majda 2010; Baldovin et al. 2020).

⁷Note that the focus on white noise here is a simplification. Different contributions have focused on more complicated proposals, such as correlated additive–multiplicative noise (e.g., Sardeshmukh and Sura 2009; Martinez-Villalobos et al. 2018) or temporally correlated and spatially coherent noise through multilevel regression modeling (e.g., Kravtsov et al. 2005; Chekroun et al. 2011), among others.

Second, and most important, it has been shown that linear frameworks such as linear inverse models can be competitive with complex nonlinear models when working with anomalies (Penland 1989; Penland and Sardeshmukh 1995; Sardeshmukh and Sura 2009). The main reasons for this are coarse-graining procedures as discussed above, mainly: (i) dimensionality reduction, (ii) the choice of specific fields and (iii) focusing on a restricted range of time scales. Regarding points (i) and (ii): integrating climate anomalies over large regions (or projecting over a few modes) often results in quasi-Gaussian distributions; additionally, the choice of fields is clearly important and, for example, considering precipitation fields would invalidate the Gaussian assumption. Regarding point (iii): the range of time scale to focus on has involved a two-step preprocessing, see Section 3. First, we focused on monthly averages, rather than the original data with temporal resolution of 1 day, as Penland and Sardeshmukh (1995) (among many others) showed quasi-Gaussian probabilities of coarse-grained SST monthly anomalies. Second, we high-pass filtered the data with a cut-off frequency of $f = 1/(10 \text{ years})$ to remove the presence of multidecadal oscillations, which (i) cause departures from Gaussianity at higher latitudes and (ii) are very low-frequency events sampled a few times even in a 600 years long run. This last step is performed only in the control run. The range of time scales in the control run (see Section 3) is then between 1 month and 10 years. Given this preprocessing and the dimensionality reduction step, the probability distributions of our data have a strong Gaussian component, see Appendix E.

Eliminating spurious results. The confidence bounds for FDT presented in Eq. 8 in Falasca et al. (2024) are leveraged to remove spurious terms, present at both long and short time scales, in the response operator allowing for more trustworthy computation of responses. We refer to Section 2 and to Falasca et al. (2024) for more details on the adopted statistical test.

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 covariance 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 covariance 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 $q_\eta = 0.1$. 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 $j = 1, \dots, n$. 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. We stress that no tuning has been performed for these parameters when evaluating the protocol. In fact, the main point here is that differences in the number and size of identified patterns (in a reasonable range) could result in small quantitative differences, but the qualitative picture will remain equivalent.
- 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 $n \ll N$. As in Falasca et al. (2024), we consider correlation, rather than covariance matrices, in the dimensionality reduction step, with no qualitative differences. We note that this choice of simple and robust dimensionality reduction methods is further motivated in Appendix A of Falasca et al. (2024).

APPENDIX C

Details on protocol through Empirical Orthogonal Functions

a. Computations of linear response in EOF space

We consider the SST and net flux at the TOA (hereafter refer to as TOA only) fields, $\mathbf{y}^{\text{SST}} \in \mathbb{R}^{N,T}$ and $\mathbf{y}^{\text{TOA}} \in \mathbb{R}^{N,T}$ in the 600 years long, stationary piControl run (see Section 3). In order to compute response formulas, we proceed as follows:

- Each time series $y_i(t)$, whether in \mathbf{y}^{SST} or \mathbf{y}^{TOA} , is weighted by its latitudinal weight as $y_i(t) \cos(\theta_i)$; θ_i being the latitude at grid point i . To ease the formalism, in what follows we are going to keep referring to time series $y_i(t)$, but the reader should remember that each $y_i(t)$ is weighted by its latitudinal weight.
- Fields \mathbf{y}^{SST} and \mathbf{y}^{TOA} are then standardized by their standard deviation. We remind the reader that both fields are anomalies, already defined to be at zero mean. Therefore, $\mathbf{y}^{\text{SST}}/\sigma_{\mathbf{y}^{\text{SST}}}$ and $\mathbf{y}^{\text{TOA}}/\sigma_{\mathbf{y}^{\text{TOA}}}$. Here, for each field $\mathbf{y} \in \mathbb{R}^{N,T}$ (whether SST or TOA) its standard deviation is defined as $\sigma_{\mathbf{y}} = \left(\frac{\sum_i^N \sum_t^T y_i(t)^2}{T * A} \right)^{1/2}$ where the global area is defined by $A = \sum_i^N \cos(\theta_i)$.
- We then reduce the dimensionality of both fields separately through Empirical Orthogonal Functions (EOFs; i.e., singular value decomposition of the covariance matrix, (Hotelling 1933)). We retain the first m EOFs (singular vectors) for both fields. We refer to the EOFs of the temperature and TOA fields respectively as $\mathbf{U}_{\text{SST}} \in \mathbb{R}^{m,N}$ and $\mathbf{U}_{\text{TOA}} \in \mathbb{R}^{m,N}$. The new set of time series are the principal components for the SST and TOA field, i.e. $\mathbf{x}^{\text{SST},r} \in \mathbb{R}^{m,T}$ and $\mathbf{x}^{\text{TOA},r} \in \mathbb{R}^{m,T}$, where the supercript r stands for “reduced”. Note that another way to proceed would have been to reduce the dimensionality after embedding the two fields together in the same state vector, i.e. $[\mathbf{y}^{\text{SST}}, \mathbf{y}^{\text{TOA}}](t)$; one problem we found in that case is that it is difficult to project a perturbation/forcing in SST into the low-dimensional space without impacting the TOA field too.
- The principal components $\mathbf{x}^{\text{SST},r} \in \mathbb{R}^{m,T}$ and $\mathbf{x}^{\text{TOA},r} \in \mathbb{R}^{m,T}$ are then joined together to form a single state vector of dimensionality $2m$, $\mathbf{x}^r(t) = [\mathbf{x}^{\text{SST},r}, \mathbf{x}^{\text{TOA},r}](t)$. The system is $\mathbf{x}^r \in \mathbb{R}^{2m,T}$.
- The response operator $\mathbf{R}(t)$ is computed in the reduced space using Eq. (6). This means that covariance functions are computed using the principal components x^r : $C_{i,j}(t) = \overline{x_i^r(\tau+t)x_j^r(\tau)}$, where the overline stands for temporal average.

- We then leverage the statistical test in Eq. 8 in Falasca et al. (2024) and set to zero all responses that are not statistical significant. Confidence bounds are here considered as $\pm 3\sigma$.
- Importantly, to compute integrals such as Eq. (3) we need to project also the forcing/perturbation into the low-dimensional spaces. In our experiments, the time dependent perturbation field is going to be the SST field in the 1pctCO2 and 4xCO2 runs. For a given SST, time dependent perturbations, $\delta \mathbf{f}^{\text{SST}} \in \mathbb{R}^{N,T}$, we first weight each δf_i^{SST} by its latitudinal weight $\cos(\theta_i)$. We then standardize the perturbation field $\delta \mathbf{f}^{\text{SST}}$ by its own standard deviation $\sigma_{\delta \mathbf{f}^{\text{SST}}}$ as shown before. We then define a fictitious perturbation field in the TOA field $\delta \mathbf{f}^{\text{TOA}} \in \mathbb{R}^{N,T}$. The perturbation field $\delta \mathbf{f}^{\text{TOA}}$ is defined by zeros at all times, i.e. no perturbation. In the case of step function perturbations, we are going to focus on constant perturbations in the SST of 1 Kelvin at a single grid point. The steps are the same as the one proposed above, apart from the standardization step for which we simply use the standard deviation $\sigma_{\mathbf{y}^{\text{SST}}}$ of the original SST field. In what follows, we are going to focus on the time-dependent perturbation.
- The (weighted and scaled) forcing fields $\delta \mathbf{f}^{\text{SST}}$ and $\delta \mathbf{f}^{\text{TOA}}$ are then projected onto their low-dimensional spaces, as $\delta \mathbf{f}^{\text{SST},r} = \delta \mathbf{f} \mathbf{U}_{\text{SST}}^T$ and $\delta \mathbf{f}^{\text{TOA},r} = \delta \mathbf{f} \mathbf{U}_{\text{TOA}}^T$. The projected perturbations are then embedded in the same vector as $\delta \mathbf{f}^r(t) = [\delta \mathbf{f}^{\text{SST},r}, \delta \mathbf{f}^{\text{TOA},r}](t)$. Therefore $\delta \mathbf{f}^r \in \mathbb{R}^{2m,T}$.
- The response $\delta \langle \mathbf{x}^r(t) \rangle \in \mathbb{R}^{2m}$ is computed in the low-dimensional space by the convolution integral defined in Eq. (3). We simply approximate these integrals as Riemann sums.
- The response $\delta \langle \mathbf{x}^r(t) \rangle \in \mathbb{R}^{2m}$ is then projected back into the high-dimensional, original space. To do so, we have to distinguish between responses in the SST or TOA fields. The first m entries of vector $\delta \langle \mathbf{x}^r(t) \rangle$ correspond to the SST field response, $\delta \langle \mathbf{x}^{\text{SST},r}(t) \rangle$, the second m to the TOA field $\delta \langle \mathbf{x}^{\text{TOA},r}(t) \rangle$. We then project both these response in the high-dimensional field by $\delta \langle \mathbf{y}^{\text{SST}}(t) \rangle = \delta \langle \mathbf{x}^{\text{SST},r}(t) \rangle \mathbf{U}_{\text{SST}}$. The same is done for the TOA response.
- The two responses need to be scaled back. We do so by multiplying $\delta \langle \mathbf{y}^{\text{SST}}(t) \rangle$ by the standard deviation of the original field $\sigma_{\mathbf{y}^{\text{SST}}}$ as defined in the piControl run. The same is done for the responses in the TOA field.
- The global mean change in TOA at time t can be then computed as $\frac{\sum_i^N \delta \langle y_i^{\text{TOA}}(t) \rangle}{A}$, with $A = \sum_i^N \cos(\theta_i)$; same for the SST field.

- We finally note that in case of plotting the response $\delta \langle \mathbf{y}(t) \rangle$ of field \mathbf{y} (whether SST or TOA) at a given time t , it is important to rescale every value $y_i(t)$ by its own latitudinal weight as $\frac{y_i(t)}{\cos \theta_i}$.

b. Computations of the sensitivity map metric

We briefly present the computation of sensitivity maps as in Section 4d but in the case of the EOF dimensionality reduction. We recommend reading Section 4d first.

The main idea is to compute the equilibrated response of the net flux at the TOA given a constant, step function perturbation of 1 Kelvin in the SST field. Differently from the community detection method, EOFs allows us to do this by perturbing each grid point. A perturbation of 1 Kelvin is then iteratively prescribed at each grid point i as $\Delta T_i = (1K) \cos(\theta_i)$; for $t > 0$. Where $\cos(\theta_i)$ is the latitudinal weighting of grid point i . The response to a step function perturbation is then computed using the formula in Eq. (4). The upper bound of the integral in Eq. (4) is considered to be $\tau_\infty = 10$ years. The sensitivity map $\mathbf{S} \in \mathbb{R}^N$ is a gridded map of the same dimensionality N of the original space. The map is defined by plotting at each grid point i the global mean TOA response caused by the

perturbation ΔT_i , as: $S_i = \frac{\langle \delta \langle \mathbf{x}^{\text{TOA}} \rangle \rangle_G}{\Delta T_i}$, where the brackets $\langle \rangle_G$ refer to the global average. The units of the sensitivity map are $[W/(m^2K)]$.

The procedure to compute responses to step function perturbations requires to first specify perturbation patterns in the high-dimensional, original field and then project fields and perturbations in a low-dimensional field spanned by a few EOFs. Response formulas are computed in the low-dimensional field and results are then projected back into the high-dimensional original field. All these steps are defined in the Section above.

APPENDIX D

Cumulative response. Case of $\tau_\infty = 5, 10$ years.

Here, we show the response of the whole globe to perturbations in four given regions with varying τ_∞ . Thus, Figures D1 and D2 are analogous to Figures 6 and 7 but τ_∞ is set to 5 and 10 years respectively.

APPENDIX E

Probability distributions

In Figures E1 and E2, we show the histogram of each signal $X^{\text{SST}}(c_j, t)$ and $X^{\text{TOA}}(c_j, t)$ of the SST and TOA variables integrated over a pattern c_j (see Eq. 7). The

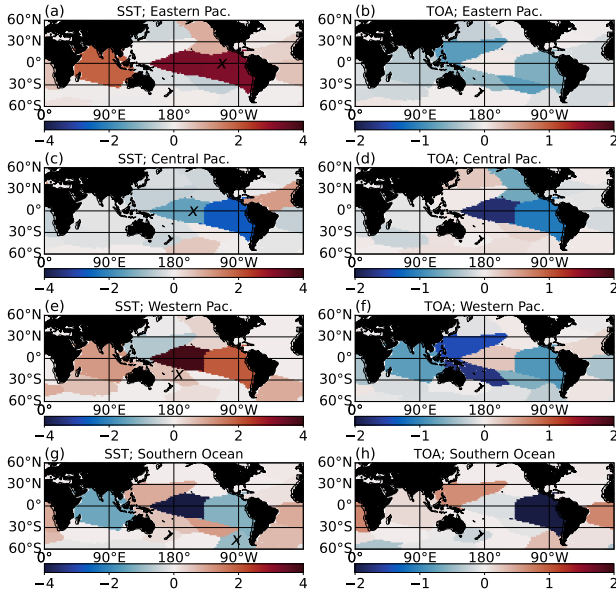


FIG. D1. As in Figure 6, but with $\tau_\infty = 5$ years.

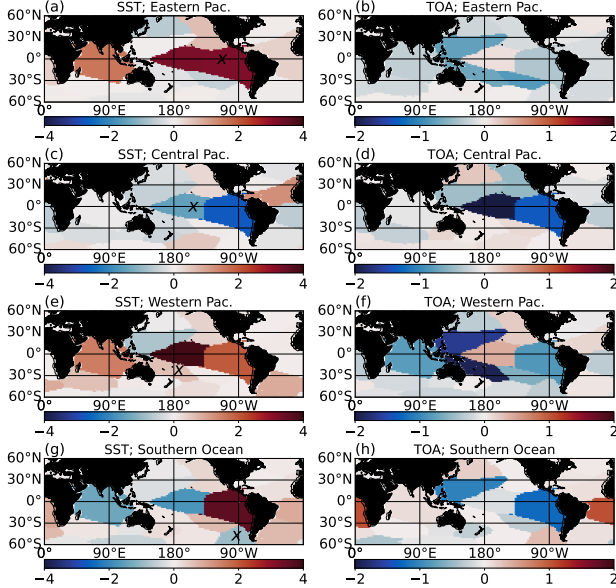


FIG. D2. As in Figure 6, but with $\tau_\infty = 10$ years.

signals are considered in the stationary, piControl run (see Section 3). Each time series have been scaled to zero mean and unit variance. A standard normal distribution is also shown in red for comparison. Figures E1 and E2 demonstrate that the quasi-Gaussian approximation shown in Eq. (6) 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 find-

ings of previous papers such as Sardeshmukh and Sura (2009), regarding the linearity argument, and Colangeli et al. (2012), regarding the regularity of distributions in the projected dynamics. Specifically: (i) we are considering only a range of time scales by focusing on monthly averages and removing variability longer than 10 years (i.e. multidecadal time scales) and (ii) we are integrating over large spatial regions through the dimensionality reduction processes.

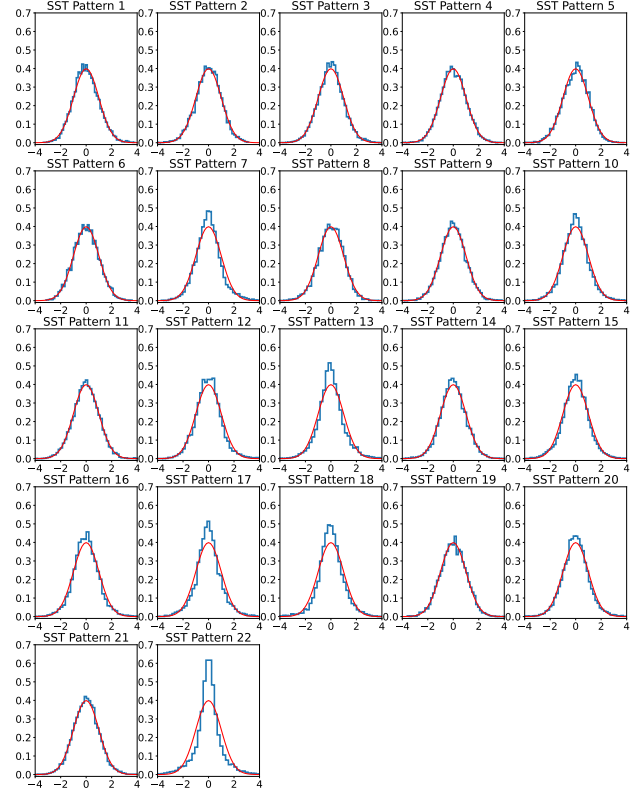


FIG. E1. Probability distributions of the cumulative time series of *sea surface temperature* in each pattern in Figure 1. Each signal had the mean removed and has been standardized to unit variance. A Gaussian fit with zero mean and unit variance is shown in red on top of each histogram.

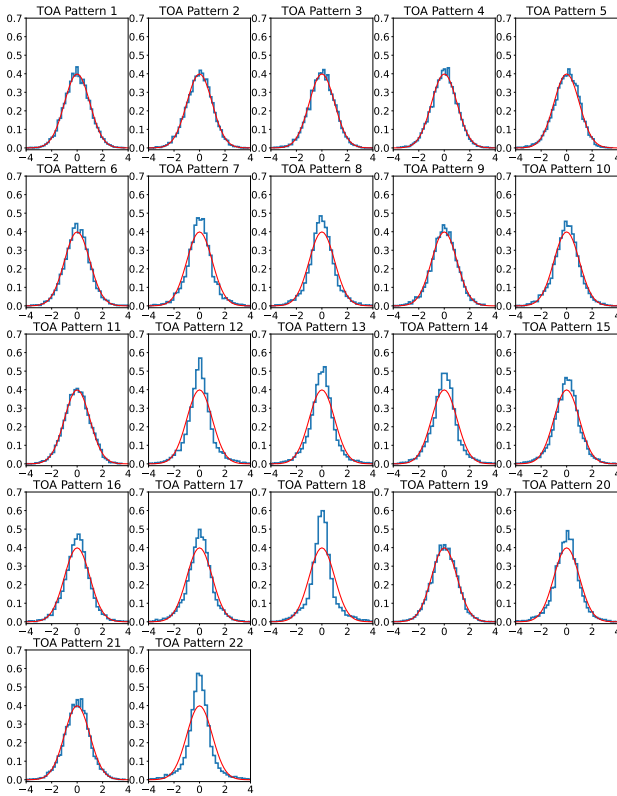


FIG. E2. Probability distributions of the cumulative time series of the net radiative flux at the TOA in each pattern in Figure 1. Each signal had the mean removed and has been standardized to unit variance. A Gaussian fit with zero mean and unit variance is shown in red on top of each histogram.

Acknowledgments. This work was supported by NOAA grant NOAA-OAR-CPO-2019-2005530, by the KITP Program “Machine Learning and the Physics of Climate” supported by the National Science Foundation under Grant No. NSF PHY-1748958. This research received support through Schmidt Sciences, LLC.

Data availability statement. Codes and materials will be soon available at <https://github.com/FabriFalasca/FDT-Pattern-Effect>. Codes for the community detection and Fluctuation-Dissipation Theorem only, 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), e2023GL105795.
- 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**, 012002, <https://doi.org/doi.org/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., F. Cecconi, M. Cencini, A. Puglisi, and A. Vulpiani, 2018: The role of data in model building and prediction: A survey through examples. *Phys. Rev. Res.*, **20**, <https://doi.org/doi.org/10.3390/e20100807>.
- Baldovin, M., F. Cecconi, A. Provenzale, and A. Vulpiani, 2022: Extracting causation from millennial-scale climate fluctuations in the last 800 kyr. *Scientific Reports*, **12**, 15320.
- Baldovin, M., F. Cecconi, and A. Vulpiani, 2020: Understanding causation via correlations and linear response theory. *Physical Review Research*, **2**, 043436.
- 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, 2023: Estimating freshwater flux amplification with ocean tracers via linear response theory. *Earth System Dynamics preprint*, <https://doi.org/https://doi.org/10.5194/esd-2023-14>.
- 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 (gf mip) protocol. *Journal of Advances in Modeling Earth Systems*, **16** (2), e2023MS003700, <https://doi.org/https://doi.org/10.1029/2023MS003700>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023MS003700>.
- Boffetta, G., G. Lacorata, S. Musacchio, and A. Vulpiani, 2003: Relaxation of finite perturbations: Beyond the fluctuation-response relation. *CHAOS*, **13** (3), 806–811.
- Bueso, D., M. Piles, and G. Camps-Valls, 2020: Nonlinear pca for spatio-temporal analysis of earth observation data. *IEEE Transactions on Geoscience and Remote Sensing*, **58** (8), 5752–5763.
- Castiglione, P., M. Falcioni, A. Lesne, and A. Vulpiani, 2008: *Chaos and coarse-graining in statistical mechanics*. Cambridge University Press.
- Chekroun, M., D. Kondrashov, and M. Ghil, 2011: Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *Proceedings of the National Academy of Sciences*, **108** (29), 11766–11771.
- Chenal, J., B. Meyssignac, A. Ribes, and R. Guillaume-Castel, 2022: Observational constraint on the climate sensitivity to atmospheric co₂ concentrations changes derived from the 1971–2017 global energy budget. *Journal of Climate*, **35** (14), 4469–4483.
- 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), L04002, <https://doi.org/doi.org/10.1088/1742-5468/2012/04/L04002>.
- 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/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/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**, 044202, <https://doi.org/10.1103/PhysRevE.109.044202>.
- Frankignoul, C., and K. Hasselmann, 1977: Stochastic climate models, Part II Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, **29** (4), 289–305, <https://doi.org/doi:10.3402/tellusa.v29i4.11362>.
- 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**, 035002, <https://doi.org/10.1103/RevModPhys.92.035002>.
- Giorgini, L., K. Deck, T. Bischoff, and A. Souza, 2024: Response Theory via Generative Score Modeling. *arxiv*, <https://doi.org/https://doi.org/10.48550/arXiv.2402.01029>.
- Granger, C., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424.
- 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/10.1175/JAS3943.1](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.
- Gutiérrez, M. S., and V. Lucarini, 2022: On some aspects of the response to stochastic and deterministic forcings. **55** (42), 425002, <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>.
- 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.org/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. 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>.
- Hottelling, H., 1933: Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6), 417.
- Ismael, J., 2023: Reflections on the asymmetry of causation. *Interface Focus*, **12**, 20220081.
- 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.
- Kravtsov, S., D. Kondrashov, and M. Ghil, 2005: Multilevel Regression Modeling of Nonlinear Processes: Derivation and Applications to Climatic Variability. *Journal of Climate*, **18**, 4404–4424.
- 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. Colangeli, 2012: Beyond the linear fluctuation-dissipation theorem: the role of causality. *J. Stat. Mech.*, **05**, P05013.
- 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**, 043103, <https://doi.org/10.1103/PhysRevResearch.4.043103>.
- Majda, A., B. Gershgorin, and Y. Yuan, 2010: 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., and D. Qi, 2018: Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Review*, **60** (3), 491–549, <https://doi.org/10.1137/16M1104664>.
- 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>.
- Martinez-Villalobos, C., D. Vimont, C. Penland, M. Newman, and J. Neelin, 2018: Calculating State-Dependent Noise in a Linear Inverse Model Framework. *Journal of the Atmospheric Sciences*, **75**, 479–496.
- 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>.
- 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](https://doi.org/https://doi.org/10.1016/0167-2789(96)00124-8).
- Penland, C., 2019: The Nyquist Issue in Linear Inverse Modeling. *Monthly Weather Review*, **147**, 1341–1349, <https://doi.org/https://doi.org/10.1175/MWR-D-18-0104.1>.
- 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>.
- 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](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), 083132, <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.
- Schreiber, T., 2000: Measuring information transfer. *Phys. Rev. Lett.*, **85**, 461.
- 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>.
- 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.
- Tomasini, U., and V. Lucarini, 2021: Predictors and predictands of linear response in spatially extended systems. *Eur. Phys. J. Spec. Top.*, **230**, 2813–2832.
- Williams, A. I., N. Jeevanjee, and J. Bloch-Johnson, 2023: Circus tents, convective thresholds, and the non-linear climate response to tropical sts. *Geophysical Research Letters*, **50** (6), e2022GL101499.

- 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), e2019GL085782.
- 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. **35**, 5637–5660, 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>.
- 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>.