Author response to RC1 for “A framework for assessing and understanding sources of error in Earth System Model emulation”
By Christopher B. Womack et al.

We first thank the two anonymous reviewers for their constructive comments and careful engagement with our substantial manuscript. Their effort has helped us improve the framing, quality, and clarity of our contribution.

We now respond to RC1, which is reproduced in black text below. Our responses follow immediately in red text, and any additions to the manuscript are included in italic red text.

Anonymous Referee #1

The authors present a framework for comparing emulation techniques. They do so by showing the theoretical connections between several existing emulation methods and relating them to two types of linear operators. These operators are shown to explain the same information about the system, demonstrating a link among all methods considered. The authors then test these methods’ abilities to predict four forcing response scenarios in four simplified toy models of either the climate system or the Lorenz convection approximation. Response function methods outperform both pattern scaling and attempts to directly estimate the linear operator in these example tests. The discussion around modeled results in the various tests is thorough and the connections to a common set of linear operators will likely be useful when considering how different emulators might perform. I have experience with pattern scaling, FDT, and ridge regression (which is how the deconvolution method has been practically implemented), though less so with much of the emulator-specific background cited here. As such, I will limit my comments to how this work fits with understanding ESMs more broadly.

Specific comments:

My main comment covers the goal and applicability of this work. I understand that the intent of the paper is to establish a “framework”, by which the authors mean the ability to frame each of these emulators as a variation or simplification on the paired linear response operators Fokker-Planck/Koopman. What is less clear to me is how directly the link can be made to “sources of error in Earth System Model emulation”. Generally, I understand if this paper is laying the groundwork for ESM testing, but in that case I felt that the writing did not make that intention clear. As presented, it reads as offering a tool that is directly applicable to evaluating emulators with respect to ESMs. The tests get at particular challenges in ESMs: memory effects, hidden variables, noise, and

nonlinearities. However, the reader does not see the actual interaction between these methods and errors in ESM emulation.

Based on this and comments from Reviewer #2, we agree the manuscript has a framing issue relative to its treatment (or lack thereof) of ESMs. The value in the theoretical framing is that we can use it to assess and improve emulators by analyzing where errors arise. We can evaluate assumptions present in many common emulation techniques, along with what types of error those assumptions invite and how this is problematic for ESMs. To help clarify these points, we will restructure the manuscript slightly, separating the theory (now Sect. 2) from our simplified experiments (now Sect. 3 + results in Sect. 4). We will also make the following changes to our abstract and introduction to clarify our experimental setup, along with other minor changes throughout to ensure continuity with these structural changes.

Addition to abstract: To support our theoretical contributions, we provide practical implementation details for each technique, along with discussion on the relative utility of these emulation methods. We evaluate emulator performance using simplified climate models, including box models and a modified version of the Lorenz 63 model, across a series of experiments designed to highlight different potential sources of error.

Changes to introduction (final paragraph): Section 2 first presents our theoretical framework, highlighting that the goal of many emulation techniques is to simplify complex climate dynamics into a linear set of modes associated with the Fokker-Planck and Koopman operators. We then apply this framework to identify potential sources of error within six emulation techniques, analyzing them from both a theoretical and practical perspective (Sect. 2.3). In Sect. 3, we introduce a series of experiments using simplified climate models and forcing scenarios designed to stress test and evaluate each emulator; these experiments include box models and a modified version of the Lorenz 63 system. Section 4 contains experimental results, showing that response functions consistently outperform other emulators across potential high-error scenarios. We conclude by discussing optimal use cases for each emulator, along with implications for ESMs based on our pedagogical model results (Sect. 5).

The reviewer is correct in that we do not emulate ESMs in this work. We will add an “Implications for ESMs” subsection in the discussion to explicitly address the utility of our framework in that context.

Implications for ESMs: While the lack of a common conceptual baseline has historically hindered comparisons between emulator classes, our framework takes an important step towards resolving this. Efforts such as ClimateBench, which provide a common

training and evaluation benchmark, have been useful to that end (Watson-Parris et al., 2022), but emulator structural differences prevent this framework from being applied to all existing emulation techniques. Additionally, the high computational burden of running scenarios beyond those in the CMIP archive (for training or evaluation), prevents rigorous assessment of emulator capability (e.g., emulating the impact of individual forcings) and generalizability (accuracy beyond ScenarioMIP). Results from experiments such as the Detection and Attribution MIP (DAMIP) and Regional Aerosol MIP (RAMIP) can help fill these gaps (Gillett et al., 2016; Wilcox et al., 2023), but the field of ESM emulation is currently data-constrained. Our theoretical framework provides value in this data-limited setting, as it allows us to evaluate the assumptions present in many common emulators. Our results illustrate the potential sources of error different emulator structural assumptions invite, giving us tools to assess and improve emulation techniques independently of ESM results. As ESM outputs improve with CMIP7 and beyond, this framework can help ensure emulators are prepared to train on those new results.

Our pedagogical experiments provide a useful tool to isolate and examine individual sources of error when emulating ESMs (Fig. 1). Though our simplified models are limited in that they lack much of the complexity of full-scale ESMs, our experiments highlight that emulator errors can be proactively resolved through structural changes, regardless of the parent model. For example, our results further support the growing body of literature on the utility of response functions (Freese et al., 2024; Womack et al., 2025; Winkler and Sierra, 2025). Response functions offer improvements over traditional pattern scaling, particularly when considering memory effects in decision-relevant scenarios. They may also emulate longer (post-2100) scenarios by accounting for regional pattern shifts, though longer ESM runs, such as the extensions proposed in ScenarioMIP for CMIP7, are required to test this (Van Vuuren et al., 2025). Existing emulators of ESMs may also benefit from incorporating response functions, c.f., recent work into hybrid emulation using generative machine learning methods in addition to pattern scaling (Bouabid et al., 2025).

Several promising emulation techniques explored here, including the Fluctuation Dissipation Theorem (FDT), Dynamic Mode Decomposition (DMD), and Extended DMD (EDMD), have seen uses in climate science but have yet to be applied directly as emulators of ESM outputs as defined by Tebaldi et al. (2025). An intermediate step for either the FDT or EDMD and DMD may be to first emulate an EMIC, helping determine useful training scenarios without the cost of a full ESM. Our results suggest further research into these techniques is warranted, as they may represent more complex dynamics than other methods. In this context, the FDT stands apart as the most promising technique for emulating general dynamical systems, as evidenced by its skill

in this and other recent work (Giorgini et al., 2025b). However, using the FDT to derive response functions through perturbations requires a full initial condition ensemble for every perturbed grid cell/region (Lucarini et al., 2017; Lembo et al., 2020), similar to the Green's Function MIP (Bloch-Johnson et al., 2024), and is likely prohibitively expensive for full ESMs. The score-based FDT (Sect. 2.3) provides a remedy, using statistical learning methods to learn the score function and thus the system response (Giorgini et al., 2025b). Regardless of the derivation method, our results suggest response functions are the dominant emulation technique both in terms of accuracy and interpretability.

Most work studying climate emulation focuses on developing and implementing new approaches in an application-specific manner. Our results show the utility of an operator-based framework for systematic analysis and comparison of climate emulation techniques. The main benefit of this framework is providing a toolkit for understanding trade-offs between emulator complexity and performance while connecting emulation techniques to fundamental principles of statistical mechanics and stochastic systems. We find that memory effects, internal variability, hidden variables, and nonlinearities are potential error sources, and that response function-based emulators consistently outperform other methods, such as pattern scaling and DMD, across all experiments. Emulator performance varies by experimental setup, particularly through the choice of training data, and further work is required to fully characterize these effects. This framework currently relies on simple experiments, and further work is needed to determine if operator-based methods like EDMD can be practically realized to emulate nonlinear processes in full-scale climate models. Our analysis also highlights the FDT's potential for deriving robust, physically-interpretable response functions, though its computational cost is a potential barrier. As interpretability is an ongoing discussion in the emulator community, investing resources in physically-grounded methods like the FDT may go a long way towards increasing the utility of emulators not just for emulation, but for linear system analysis.

530: While the 2- and 3-box models are frequent approximations to the climate system, they lack many of the physical mechanisms that make the climate system difficult to model. The parameters in these models are fit to ESMs, so are themselves simplified estimates of the actual behavior. I felt that the link between ability to emulate these examples and the ability to emulate ESMs deserved more discussion. I would have found this conceptually more useful than the level of technical detail included for the linear operators and each emulation model in the main text.

On the limitations of box models, we agree that these lack many of the physical mechanisms present in full-scale climate models and that is a limitation of this work. We will add discussion around this point to the Implications for ESMs section.

From paragraph two of Implications for ESMs: Our pedagogical experiments provide a useful tool to isolate and examine individual sources of error when emulating ESMs (Fig. 1). Though our simplified models are limited in that they lack much of the complexity of full-scale ESMs, our experiments highlight that emulator errors can be proactively resolved through structural changes, regardless of the parent model.

On the applicability of this methodology to ESMs, we agree that a more explicit discussion around this topic is necessary for this manuscript. The additional Implications for ESMs subsection covers these points more explicitly.

846: “This framework currently relies on simple experiments, and further work is needed to determine if operator-based methods like EDMD can be practically realized to emulate nonlinear processes in full-scale climate models.”: this sentence to me suggests that the step of showing that this framework is useful for ESMs is left to future work. I can see that there is some value in being able to connect the different models through a common framework in the way the authors use it to diagnose differences in the toy model. This may be more in line with a proof of concept for the framework rather than demonstrating how the framework applies to ESMs. However, if the goal is for this framework to be used by others and applied to ESMs, this seems like an important step to include. This may just be a framing issue.

We agree that this first draft suffers from a framing issue. While we do not apply our framework directly to ESMs in this manuscript, formalizing these ideas through our idealized experiments constitutes necessary foundational work towards that goal. The previous structural changes to the manuscript will help highlight the utility of our contribution.

From paragraph one of Implications for ESMs: Our theoretical framework provides value in this data-limited setting, as it allows us to evaluate the assumptions present in many common emulators. Our results illustrate the potential sources of error different emulator structural assumptions invite, giving us tools to assess and improve emulation techniques independently of ESM results. As ESM outputs improve with CMIP7 and beyond, this framework can help ensure emulators are prepared to train on those new results.

Figure 4: If the results suggest that directly estimating response operators is the most prone to error, does this challenge the response operator framework as the most useful common link for the different emulation methods? This seems to suggest the Koopman operator is not the most useful simplification of the climate system.

This is a great observation, as directly estimating these operators is a nuanced, data-intensive task. While EDMD and DMD attempt to learn the Koopman operator, they are extremely simplified representations and in most cases do not closely approximate the true operator. Despite this, the Koopman and Fokker-Planck operators provide the most useful theoretical basis as they offer a way to directly link vastly different forms of emulators. We will add text to clarify the differences between the theoretical and data-derived Koopman operators in the discussion.

Addition to discussion: While EDMD and DMD attempt to approximate the Koopman operator, they are simplified representations and in many cases do not closely approximate the true operator. Despite this, the Koopman and Fokker-Planck operators provide the most useful theoretical basis as they offer a way to directly link disparate forms of emulators.

Minor technical:

42: “Impulse response (response/Green’s function) methods” this wording is confusing, how is “response” an example of “impulse response”

We agree, the original wording here was unclear. We will clarify the intended meaning of this phrase.

Change to introduction: Impulse response methods, commonly referred to as either response or Green’s functions,…