On the theory and optimal design of emulators for climate impact assessment

by

Christopher B. Womack

B.S. Aerospace Engineering, Massachusetts Institute of Technology (2021)
S.M. Aerospace Engineering, Massachusetts Institute of Technology (2024)
S.M. Technology and Policy, Massachusetts Institute of Technology (2024)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN
AEROSPACE COMPUTATIONAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2026

Authored by: Christopher B. Womack
Department of Aeronautics and Astronautics
May 15, 2026

Certified by: Noelle E. Selin
Director of the Center for Sustainability Science and Strategy,
Professor of Atmospheric Chemistry, Thesis Supervisor

Certified by: Glenn R. Flierl
Professor of Physical Oceanography, Thesis Supervisor

Accepted by: Jonathan P. How
Ford Professor of Engineering
Chair, Graduate Program Committee

THESIS COMMITTEE

Thesis Supervisors

Noelle E. Selin
Director of the Center for Sustainability Science and Strategy
Professor of Atmospheric Chemistry
Massachusetts Institute of Technology

Glenn R. Flierl
Professor of Physical Oceanography
Massachusetts Institute of Technology

Committee Chair

David L. Darmofal
Vice Chancellor for Graduate and Undergraduate Education
Jerome B. Wiesner Professor of Aeronautics and Astronautics
Massachusetts Institute of Technology

Committee Members

Sebastian D. Eastham
Associate Professor in Sustainable Aviation
Imperial College London

Claudia Tebaldi
Earth Scientist
Joint Global Change Research Institute,
Pacific Northwest National Laboratory and University of Maryland

Thesis Readers

Raffaele Ferrari
Cecil and Ida Green Professor of Oceanography
Massachusetts Institute of Technology

William E. Chapman
Assistant Professor of Atmospheric and Oceanic Sciences
University of Colorado Boulder

On the theory and optimal design of emulators for climate impact assessment

by

Christopher B. Womack

Submitted to the Department of Aeronautics and Astronautics
on May 15, 2026
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN
AEROSPACE COMPUTATIONAL ENGINEERING

ABSTRACT

Earth System Models (ESMs) are our most comprehensive tools for projecting future climate impacts across the land, ocean, and atmosphere, yet their extreme computational costs limit their ability to survey the vast space of potential emissions trajectories. Climate emulators—reduced-order models that reproduce the statistics of these full-scale models in a fraction of the time—are poised to fill this scenario-assessment gap. Despite the rapid uptake of emulators in many domains, motivated in part by the broader machine learning (ML) revolution, many questions around their theoretical underpinnings, physical consistency, and ultimate utility for areas like impact assessment remain.

In this thesis, we first address the lack of a comprehensive theoretical basis for emulators by developing a framework that enables fundamental methodological comparisons. This framework connects disparate emulation techniques via ideas from statistical mechanics and stochastic calculus, and we apply it to understand potential sources of emulator error, focusing on memory effects, hidden variables, system noise, and nonlinearities. We discuss optimal use cases for a number of emulation techniques in light of these potential sources of error, along with implications for ESMs based on our pedagogical model results. Based on these findings, we then address emulator physical consistency and extrapolative skill. While efforts to improve emulator generalizability typically focus on the design of more complex ML architectures, we show that the training data itself is a major bottleneck for predictive skill. We introduce a method to generate optimal training data by iteratively updating an initial emissions trajectory to maximize emulator skill, showcasing applications to simple and intermediate-complexity climate models. An emulator trained on just one or two of these optimized scenarios outperforms one trained on six standard ScenarioMIP pathways. We achieve higher predictive skill despite training on a smaller dataset, and find that our emulators successfully isolate the distinct physical behaviors of different climate forcing agents (e.g., greenhouse gases vs. aerosols) without training on single-forcing runs. To support these theoretical and methodological improvements, we conclude by applying a novel, generative AI climate emulator to capture compound climate hazards like wet-bulb temperature. By coupling the MIT Emissions Predictions and Policy Analysis model to this emulator, we rapidly generate realizations of spatially- and cross-correlated climate fields. We utilize this framework to assess local sensitivities to various emissions scenarios, including an early assessment of the projected ScenarioMIP-CMIP7 protocol. Furthermore, we demonstrate that temperature overshoot pathways result in substantially higher cumulative heat stress risks compared to stabilization pathways with similar end-of-century outcomes. The improvements presented herein democratize access to computational science and detailed climate projections, enabling the probabilistic assessment of compound climate hazards essential for robust adaptation planning.

Thesis Supervisor: Noelle E. Selin
Title: Director of the Center for Sustainability Science and Strategy,
Professor of Atmospheric Chemistry

Thesis Supervisor: Glenn R. Flierl
Title: Professor of Physical Oceanography

Acknowledgements

We are what we repeatedly do. Excellence, then, is not an act, but a habit.

— Will Durant

I have been *incredibly* fortunate throughout my time in grad school[1] to have been surrounded by some truly amazing communities and loved ones. It takes a village to complete a PhD[2], and it has been a humbling experience to have a village made up of such thoughtful, compassionate, and silly people. This section is dedicated to the many, many people who helped me along the way.

It feels appropriate to start my acknowledgements with some unusual revelations I had at a Friendsgiving last year[3]. To my phone: thank you for allowing me to stay connected with my family and friends, despite being spread across the world. To airplanes: thank you for enabling me to wake up in Cambridge and be back in Redlands before the sun goes down[4]. To my plants: thank you for adding some color to my life, my home, and my office, and for reminding me to spend more time outside. Grad school has taught me to appreciate the little things a whole lot more; these daily affirmations, along with everyone that I plan to write about in this section, have made the grind of grad school, thesis writing, and defense prep much more manageable.

Despite being a primarily technical document, I found a (small) bit of wiggle-room for creativity in my thesis through the quotes at the start of each chapter. Though they all have some personal value, the quote at the start of this section is by far the most meaningful to me[5]; it is the closest thing I have to a mantra, and is something I have come back to frequently throughout grad school. To Mom and Dad: thank you for everything. Your love, support, and willingness to go the extra mile for me have never gone unnoticed. Being able to attend grad school at all is a very privileged position to be in, and I hope I can take the lessons you have taught me and put some good out into the world. I cherish my daily phone calls home[6], and I hope you know how much I love and appreciate everything you have done for me. To Kevin, Catherine, Charlotte, and Evelyn; to Garrett, Morgan, Maddie, Kate, and Connor; to Artie, Bandit, and Zoomer[7]; thank you for being there for me through the many ups and downs of life, and especially for sharing the good times with me[8]. To the Calbreaths; to the (other) Womacks; to the Daniels; to the Güttler clan[9]; and to every other family member: thank you. Words cannot capture how much you all mean to me.

I started grad school in a strange time; the world was still reeling from the pandemic, and I spent my first several months wearing a mask in a mostly empty lab[10]. Despite the strange atmosphere, I found my first community thanks to AeroAstro, the ACDL[11], and the Darmofal Research Group. To Dave: thank you for supporting me, teaching me, and encouraging me to pursue a PhD, even though it meant moving on from your group. I cannot thank you enough for keeping my excitement for research high through the slow churn of grad school. To Marshall and Steve: your technical expertise was invaluable to my development as a researcher,

1: Reading these acknowledgements back, I've been incredibly fortunate for my entire life!

2: Or something like that?

3: Thanks Mike and Camilla for giving me a forum to appreciate some smaller things in life

4: Only in the summer though, as it still is a long travel day

5: Fun fact: I thought this quote belonged to Aristotle, but I found out while writing my thesis that this is a misattribution! It is actually a summary of his philosophy written by author Will Durant in his 1926 book *The Story of Philosophy*.

6: Though sometimes it's more like every other day

7: And, as Darya likes to remind me, the million other miscellaneous animals we had growing up

8: Such as being an uncle ×5!!!

9: My newest family members!

10: Shout out to the 37-312 cluster and to Loek for being the only other person in that empty room

11: Now ACSEL, but forever ACDL!

for sure,[12] but it was your openness and humor that helped me feel at ease in spite of an uncertain world[13]. To Cory and Pam: thank you for your mentorship and guidance both in and out of the lab. To Josh, Mike, Sarah[14], and Saba: thank you for the amazing, incredibly silly times. I miss (and will miss) you all dearly[15]. To Emily, Loek, Danny, Joanna, Kelvin, and the other members of the ACDL: thank you for Halloween parties, karaoke nights, and so much more. To Adriana, Carter, Spencer, and the many other AeroAsteroids and members of GA3: thank you for making my time in AeroAstro incredible.

To Langdon Winner[16]: thank you for your seminal work[1], which encouraged me to broaden my horizons from the purely technical and apply to MIT's Technology and Policy Program (TPP). Not only did TPP provide me perspective on where my research sat in the broad, ever-shifting landscape of policy, it also connected me with some of the most influential people I have ever had the pleasure of meeting. To Seb: thank you for connecting me with an entirely new world; I owe my entire PhD transition to you and your willingness to take a chance on me[17]. To Noelle: thank you for keeping my research grounded and always reminding me to ask, "okay, so what?"[18] As someone who tends to lose the forest for the trees, your constant big-picture point of view has helped me take my research far beyond what I could have achieved from a purely technical perspective. To Glenn[19]: thank you for the *numerous* conversations about any topic I could come up with. Having someone to bounce ideas off of any time even the slightest bit of inspiration[20] struck has been incredible. Thank you for your support, guidance, and cat stories. To Mike, Nirmal, Helena, Graham, Maya, Hannah, and all the other wonderful folks I met through TPP: thank you for the insightful[21] conversations. To Shahine, Paolo, Andre, Anthony, Kevin, Björn, and the rest of the BC3 team: thank you for supercharging my PhD and for some *very* memorable AGU experiences. To the Selin Group[22]: thank you for being so welcoming; it meant the world to me to find a home in the Green Building. To my lovely office mates, Jessie and Dianna: thank you for the chit chat, for being sounding boards for my life, and for being such genuine, amazing people. I will miss our daily conversations greatly. To Jen, Adam, Andrei, and the massive ensemble of CS3 affiliates: thank you for showing me how to use research to inspire real change[23]. To Schmidt Sciences, LLC and the MIT Climate Grand Challenges: thank you for funding this work and for making my PhD experience what it is today.

After officially entering the PhD portion of grad school, I had to assemble my committee. Though many of the aforementioned incredible people were clear choices, I still had a few spots to fill. To Claudia: thank you for your advice, both in research and in life. Especially in this last year of my PhD, the future has been plagued with uncertainty. Your levelheadedness and calm demeanor[24] always put me at ease, and it never ceases to amaze me how you found time and patience for all our discussions, no matter how busy things got[25]. To Raf and Will: thank you for your actionable feedback during the stressful crunch to finish my thesis. Your willingness to read this behemoth document for someone outside of your direct group speaks volumes to your character, and I cannot thank you enough for your support during this final push.

Grad school is tough. It requires long nights, dealing with the frustration of experiments not going as planned[26], and mountains of reading. I'm very fortunate, then, to be in a place like MIT, where I can be a part of such wonderful communities outside of the office that bring so much joy. To our 2cans: thank you for the smiles, the laughs, the many dinners, Spooktober,

12: Still not sure I understand lifting operators, though...

13: I'll be implementing the "Friday slide" wherever I go next!

14: Gone from MIT, but not forgotten

15: Let's just say... I value your friendship more than evil Mike hates when someone spills his drink

16: Whom I have unfortunately never met personally

[1] Winner, 1980

17: I'll never forget the way you went from chill to hyper-attentive in our very first meeting when I mentioned I worked on adaptive FEM

18: I think you may have phrased it more as "why should we care about these findings?"

19: Speaking of technical expertise!

20: Or confusion!

21: And sometimes questionable

22: And a special thanks to Eric for always letting me come upstairs to bother him!

23: Some might even say *global* change!

24: Even in the face of some insurmountable odds, I might add

25: And also how you know *everyone* in the CMIP community

26: Or of installing SANS

Thanklsgiving, and so much more. We hope you know how much you mean to us, and how difficult it is going to be to move on from this community[27]. Watching you grow into the amazing people you are today is an experience we will treasure forever. To the BC GRAs and House Team: thank you for always providing support, even in the toughest situations. Your kindness is appreciated immensely, and I cannot thank you enough for the amazing experience you all provided throughout grad school. To my Phi Delts brethren: despite being separated by a distance, thank you for always showing up for the big moments. To my wonderful dance partner, Maggie: thank you for your patience, kindness, and willingness to adapt our practice plan on any given day[28]. To Peter, Francesca, Hamid, Cindy, Raluca, Allen, Kat, Mike, Anne, Tanya, and the rest of MIBDT: thank you for making up such a vibrant community and for helping me take my mind off work for a few hours each day. To Charlotte, Armin, Mark, Didi, Jan, Ben, Fil, and Esther: thank you for pushing me to be the best I can be and for changing my outlook on the world. Thanks to you, I will always invite new experiences and keep my head up.

Finally, to Darya[29]: thank you for being you. You are my best friend, my confidant, my shoulder to cry on, my couch co-op buddy, my everything. I'm grateful every day that you came into my life, and even more grateful that I get to spend the rest of my life with you; I could complete one hundred PhDs[30] and being with you would still be the greatest thing I do in life[31]. Thank you for your love and support, even in the depths of coming back from the office at 3am. Thank you for bringing me dinner when I'm knee-deep in paper revisions. Thank you for still finding ways to be silly through all of it. I'm the luckiest person in the world and I'm so excited for this next chapter of our lives together.

27: After we're gone, please make sure you're getting enough sleep, drinking enough water, and eating enough fiber!

28: Maybe you'll be able to get more dancing done without me yapping all the time

29: Dizzler Much; Bingus, my Belovéd; beb; sweetness; and so much more

30: Or one thousand, or one million...

31: Though I suppose I'll one-up myself by marrying you!

Contents

List of Figures

List of Tables

Introduction

It is unequivocal that human influence has warmed the atmosphere, ocean and land. Widespread and rapid changes in the atmosphere, ocean, cryosphere and biosphere have occurred... The scale of recent changes across the climate system as a whole—and the present state of many aspects of the climate system—are unprecedented over many centuries to many thousands of years...

— Intergovernmental Panel on Climate Change

We live in a moment of unique opportunity. As the climate crisis continues to worsen[3], it serves as an unprecedented catalyst for innovation. It demands we redesign the human experience across all sectors, changing everything from energy grids and urban spaces to agricultural systems and water management[4–14], all in the pursuit of a more sustainable future. Though we can be hopeful in the face of such a challenge, we must not be blind to its effects. Disease, famine, mass migration, and loss in biodiversity[15–22]—these are all real damages we see today that are only projected to increase as we continue emitting. Though our global atmospheric commons are threatened by human activities, the impacts of climate change are anything but equitable. The Global South—responsible for only a small fraction of historical emissions—is projected to bear the brunt of these impacts[19,23–26], while the mid- and high-latitude nations largely responsible for climate change may actually benefit[27]. This further complicates discussions of climate justice, with debates surrounding both the role of polluters in supporting adaptation and mitigation efforts and how that role may shift over time as benefits are redistributed in a warming world[28]. Despite the scientific consensus that climate change *will* impact humans negatively[3], uncertainty remains around key details, such as *where*, *when*, and *to what extent*.

[3] Intergovernmental Panel on Climate Change (IPCC), 2023

[4] Howden et al., 2007; [5] Crawley, 2008; [6] Clastres, 2011; [7] Dulal, Brodnig, and Onoriose, 2011; [8] Anwar et al., 2013; [9] Schlosser et al., 2014; [10] Döll et al., 2015; [11] Jiang et al., 2017; [12] Perera et al., 2020; [13] Yalew et al., 2020; [14] Hultgren et al., 2025

[15] Piguet, Pécoud, and Guchteneire, 2011; [16] Bellard et al., 2012; [17] Doney et al., 2012; [18] D'Amato et al., 2014; [19] Hasegawa et al., 2016; [20] Kaczan and Orgill-Meyer, 2020; [21] Habibullah et al., 2022; [22] Semenza, Rocklöv, and Ebi, 2022

[19] Hasegawa et al., 2016; [23] Patz et al., 2005; [24] Samson et al., 2011; [25] Tol, 2018; [26] Dellink, Lanzi, and Chateau, 2019

[27] O'Brien and Leichenko, 2003

[28] Mintz-Woo and Leroux, 2021

Resolving these spatial and temporal uncertainties is a prerequisite for equitable adaptation, but doing so requires an understanding of the complex physics of our planet. While many of the natural laws that govern the Earth system are known and can be predicted with high certainty, the system as a whole is chaotic; infinitesimal perturbations can lead to vastly different outcomes over time[29]. Projecting even the average behavior of such a system requires many perturbed realizations to isolate the forced response from internal chaotic noise[30–32]. Unfortunately, we live in and experience only a single realization of the Earth system, upon which we cannot perform controlled experiments. We are instead restricted to computational methods to develop our understanding of the climate. Earth System Models (ESMs) are our most comprehensive tools to that end.

[29] Lorenz, 1972

[30] Gregory et al., 2004; [31] Maher et al., 2019; [32] Lembo, Lucarini, and Ragone, 2020

These simulacra of our natural world are massive systems of partial differential equations, coupling the atmosphere, land, and ocean to simulate how anthropogenic emissions of greenhouse gases and aerosols impact the evolution of climate fields such as temperature, precipitation, relative humidity, and wind speeds over time[33–35]. The complexity and stochasticity of the coupled climate system lead to three main sources of

[33] Flato, 2011; [34] Flato et al., 2014; [35] Jeevanjee et al., 2017

uncertainty in climate projections[36]:

1. **Internal variability:** Uncertainty arising from the inherent chaos of the climate system, which dominates on annual timescales.
2. **Model structural uncertainty:** Uncertainty in the representation of the physics and dynamics within a model, which dominates on annual-to-decadal timescales.
3. **Scenario uncertainty:** Uncertainty due to human factors, such as population growth and energy demand, which dominates on decadal-to-centennial timescales.

[36] Hawkins and Sutton, 2009

Efforts such as the Coupled Model Intercomparison Project (CMIP) have made great strides towards quantifying model structural uncertainty under a common set of anthropogenic forcings[37,38], while ensemble modeling strategies enable us to assess both present and future ranges of internal variability[31,39,40]. However, ESMs are computationally expensive. For example, simulations with high-resolution ESMs may achieve only 10-20 simulated years per day on several thousand CPU cores, implying that a 250-year simulation with five ensemble members can require on the order of weeks of wall-clock time even on large supercomputing systems[41]. When accounting for the time required for model setup, output processing, and the general delays and restarts that often occur when running models of this complexity, this estimate quickly balloons to several months or more[42,43]. This severely limits the utility of ESMs in the context of climate impact assessment, where self-consistent projections of economic and climate outcomes are used to quantify future risks under large scenario uncertainty across sectors such as energy, water resources, and human health[9,24,25,44–48]. To alleviate the computational bottleneck imposed by the use of ESMs, impact assessment typically relies on reduced-order models.

[37] Taylor, Stouffer, and Meehl, 2012; [38] Eyring et al., 2016

[31] Maher et al., 2019; [39] Shiogama et al., 2023; [40] King et al., 2024

[41] Müller et al., 2018

[42] Balaji et al., 2017; [43] Balaji et al., 2022

[9] Schlosser et al., 2014; [24] Samson et al., 2011; [25] Tol, 2018; [44] Edmonds, Wise, and MacCracken, 1994; [45] Riahi, Grübler, and Nakicenovic, 2007; [46] Ciscar et al., 2011; [47] Calvin et al., 2019; [48] Schlosser et al., 2023

Simple Climate Models (SCMs) are one popular choice for rapid assessment of globally averaged scenario uncertainty. These computationally efficient models represent only a subset of climate processes, generally through idealized assumptions (e.g., only considering ice-albedo feedback and diffusive heat transport)[49–54]. While these idealized models may provide useful scientific insights[55,56], their relevance for impact assessment is limited because their aggregation of climate processes does not allow for easily interpretable regional outcomes. Earth system Models of Intermediate Complexity (EMICs) attempt to bridge the gap between computational efficiency and complexity by providing a more complete representation of climate processes than an SCM, but at a lower computational cost than an ESM—a compromise generally at the expense of spatial resolution[57–61]. Resolving spatially explicit impacts from an EMIC then requires an additional downscaling step to map from zonal to regional outputs[62,63].

[49] North, 1975; [50] North, 1990; [51] Meinshausen, Raper, and Wigley, 2011; [52] Millar et al., 2017; [53] Smith et al., 2018; [54] Leach et al., 2021

[55] Armour, Bitz, and Roe, 2013; [56] Giani et al., 2025

[57] Claussen et al., 2002; [58] Weber, 2010; [59] Holden et al., 2016; [60] Platov et al., 2017; [61] Ruggieri et al., 2024

[62] Monier et al., 2013; [63] Eby et al., 2013

> *It may very well be that entirely different methods for reducing the computational burden will have to be found in which the full physical complexity can be retained as needed.*
>
> — Schneider and Dickinson, 1974

[64] Tebaldi et al., 2025

Seemingly a direct response to this call, climate emulators—data-driven models that aim to reproduce the statistics of climate fields from full-scale ESMs—have surged in popularity[64]. Their usage in impact assessment is rapidly expanding[65–68], alongside applications such as attribution and net-zero pathway comparisons[69–73].

[65] Shiogama, Takakura, and Takahashi, 2022; [66] Munday et al., 2025; [67] Polonik, Burney, and Ricke, 2025; [68] Varney et al., 2026

[69] Beusch et al., 2022; [70] Kitsios, O'Kane, and Newth, 2023; [71] Schwaab et al., 2024; [72] Schöngart et al., 2025; [73] Quilcaille et al., 2025

Climate emulators first emerged in the early 1990s with the advent of pattern scaling, a simple linear regression of local climate variables against global mean temperature[74]. Researchers have consistently improved the original technique over time, introducing variations that represent different mixes of greenhouse gases and spatially heterogeneous forcings (e.g., aerosols)[75], capture seasonal anomalies[76], include a land-sea contrast term[77], and incorporate zonal temperatures in the regression[78]. Despite its simplicity, pattern scaling has stood the test of time[79,80], constituting the backbone of major present-day emulation efforts[81,82]. This approach produces accurate projections assuming exponential and fixed-pattern forcing, along with linear, time-independent dynamics; these criteria are roughly satisfied in a number of CMIP experiments[56]. Temperature overshoot scenarios violate these assumptions, causing pattern scaling to break down in many decision-relevant storylines[2].

[74] Santer et al., 1990

[75] Schlesinger et al., 2000

[76] Mitchell, 2003

[77] Herger, Sanderson, and Knutti, 2015

[78] Gao, Sokolov, and Schlosser, 2023

[79] Tebaldi and Arblaster, 2014; [80] Wells et al., 2023

[81] Beusch et al., 2022; [82] Mathison et al., 2025

[56] Giani et al., 2025

[2] Womack et al., 2026

Impulse response functions, commonly referred to as either response or Green's functions, fill this gap by incorporating forcing history into the emulator, rather than relying only on instantaneous forcing. They can accurately represent a number of climate processes[83–85] and were used in the earliest efforts to incorporate climate outcomes into an Integrated Assessment Model (IAM) to assess the efficacy of short- and long-term climate policies[86–88]. More recently, response functions have been utilized to develop better estimates of effective radiative forcing in climate models[89,90], as well as to drive explicit climate emulation efforts that prioritize overshoot performance[91,92]. Beyond overshoot scenarios, response functions offer a methodological formalism derived from statistical mechanics and linear response theory, enabling the analysis of climate behaviors that pattern scaling cannot replicate[32,93,94]. Despite their utility, neither pattern scaling nor response functions can represent the stochasticity of the climate system or generate spatially correlated climate fields; these capabilities are required for understanding compound climate risks due to co-occurring events[95,96].

[83] Joos and Bruno, 1996; [84] Orbe et al., 2018; [85] Cimoli et al., 2023

[86] Hasselmann et al., 1997; [87] Hasselmann, 2001; [88] Hasselmann et al., 2003

[89] Fredriksen, Rugenstein, and Graversen, 2021; [90] Fredriksen et al., 2023

[91] Womack et al., 2025; [92] Sandstad et al., 2025

[32] Lembo, Lucarini, and Ragone, 2020; [93] Lucarini, Ragone, and Lunkeit, 2017; [94] Lucarini and Chekroun, 2024

[95] Zscheischler et al., 2020; [96] Mathison et al., 2023

Machine Learning (ML) and statistical emulation techniques offer a promising pathway to overcome these hurdles, particularly due to their efficiency and ability to capture nonlinear effects. Relatively simple statistical models demonstrated skill in predicting the forced and unforced components of climate variability[70,97]. However, questions regarding their physical significance and extrapolative ability limited their widespread adoption. As the field shifted toward more complex architectures like convolutional neural networks and diffusion models (see Bracco et al. (2024)[98] for an overview), new challenges emerged. Specifically, concerns regarding overfitting on internal variability[99], a lack of comprehensive benchmarking datasets[100,101], and broader issues of interpretability have hindered their practical adoption[102], though this is an active area of research in climate[103,104]. Despite this, recent advances in generative modeling are shifting the paradigm. By demonstrating high skill in emulating climate fields with accurate spatial and cross-variable correlations[105,106], these generative approaches effectively bridge the gap between computational efficiency and cross-variable coherence, making them well-suited for impact assessment. Although other autoregressive ML techniques exist to emulate the entire atmospheric or oceanic states[107–112], many of these techniques are unstable on climatic timescales and remain unproven for impact assessment[113,114]. Consequently, they fall outside the scope of this thesis.

[70] Kitsios, O'Kane, and Newth, 2023; [97] Castruccio et al., 2014

[98] Bracco et al., *Machine Learning for the Physics of Climate*, 2024

[99] Lütjens et al., 2025

[100] Watson-Parris et al., 2022; [101] Christensen et al., 2024

[102] Rudin, 2019

[103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [104] Winkler and Sierra, 2025

[105] Bassetti et al., 2024; [106] Bouabid, Souza, and Ferrari, 2026

[107] Watt-Meyer et al., 2023; [108] Duncan et al., 2024; [109] Kochkov et al., 2024; [110] Chapman et al., 2025; [111] Cresswell-Clay et al., 2025; [112] Duncan et al., 2025

[113] Clark et al., 2025; [114] Rucker et al., 2025

Applying ML techniques to physical systems is challenging, particularly regarding generalizability to unseen scenarios and adherence to physical

laws. Efforts to create physically consistent ML models have largely focused on architectural improvements, such as embedding governing equations into loss functions or enforcing hard physical constraints[115–121]. Alongside these pure ML advancements, hybrid approaches such as NeuralGCM demonstrate that coupling ML directly with physical solvers can yield computational savings without sacrificing predictive skill[109]. However, while these architectural improvements are critical, data design plays an equally important role in determining whether ML models capture underlying physics rather than interpolating between observed states. These techniques include physics-informed feature engineering (e.g., using nondimensional quantities such as the Reynolds number instead of raw velocity fields)[122], physics-guided data augmentation that exploits known invariances or linearity properties[123], and synthetic data generation via active learning to place new samples in regions of large physical error or high model uncertainty[124,125].

[115] Greydanus, Dzamba, and Yosinski, 2019; [116] Raissi, Perdikaris, and Karniadakis, 2019; [117] Mohan et al., 2020; [118] Cai et al., 2021; [119] Karniadakis et al., 2021; [120] Satorras, Hoogeboom, and Welling, 2021; [121] Cuomo et al., 2022

[109] Kochkov et al., 2024

[122] Fazliani, Frangella, and Udell, 2025

[123] Li, Pang, and Shan, 2022

[124] Shields et al., 2023; [125] Guo et al., 2024

Assessing whether climate emulators respect physical constraints remains challenging, as demonstrating physical consistency requires extrapolating to forcing trajectories distinct from those seen in training. In practice, however, most studies emphasize in-sample and within-range performance—where Global Mean Surface Temperature (GMST) or emissions trajectories lie within the training range—with limited emphasis on structurally out-of-distribution tests[64,99,100]. This gap stems from the high temporal and computational costs of running ESMs, limiting emulator developers to the data available via CMIP for training and evaluation. Concurrently, ESM scenario design for future CMIP efforts is moving towards emissions-driven experiments and a broader set of forcing combinations[126], underscoring the need for emulators responsive to individual forcing agents rather than solely to aggregate emission pathways. Previous work suggests that ScenarioMIP-like experiments, although standard for emulator training[82,103,127–129], are not necessarily optimal[2]. While running ESMs solely to generate training data offers a promising alternative[64,130], high simulation costs impede both the exploration and adoption of this approach.

[64] Tebaldi et al., 2025; [99] Lütjens et al., 2025; [100] Watson-Parris et al., 2022

[126] Van Vuuren et al., 2026

[82] Mathison et al., 2025; [103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [127] Beusch, Gudmundsson, and Seneviratne, 2020; [128] Tebaldi, Snyder, and Dorheim, 2022; [129] Geogdzhayev et al., 2026

[2] Womack et al., 2026

Reliable, physically consistent climate projections are crucial for agriculture[4,8,14], the built environment[5,7], energy systems[6,12,13], and the finance and insurance sectors[131,132], all of which face substantial physical and transition risks from climate change. In this context, emulators have demonstrated remarkable skill in reproducing impact-relevant variables such as near-surface air temperature, precipitation, relative humidity, and wind speed across annual, monthly, and daily timescales[51,64,91,97,103,105,106,127,133]. Crucially, impact assessment requires both accurately representing local climate processes and quantifying their uncertainty. Capturing compound climate hazards—such as heat stress (measured by wet-bulb temperature) and fire risk (driven by Vapor Pressure Deficit (VPD))—which directly impact metrics like labor productivity[134,135], requires the sampling of individual realizations with accurate spatial correlations between variables. Traditional frameworks, such as the MIT Integrated Global Systems Model (IGSM)[62,78], utilize an ensemble pattern scaling approach to capture model structural uncertainty. While this framework links human activity and physical climate response in a self-consistent manner, it does not allow for the sampling of individual realizations. Furthermore, this methodology requires the ability to run an EMIC prior to downscaling, limiting its accessibility.

[64] Tebaldi et al., 2025; [130] Van Katwyk et al., 2026

[4] Howden et al., 2007; [8] Anwar et al., 2013; [14] Hultgren et al., 2025

[5] Crawley, 2008; [7] Dulal, Brodnig, and Onoriose, 2011

[6] Clastres, 2011; [12] Perera et al., 2020; [13] Yalew et al., 2020

[131] Collier, Elliott, and Lehtonen, 2021; [132] Zhou, Endendijk, and Botzen, 2023

[51] Meinshausen, Raper, and Wigley, 2011; [64] Tebaldi et al., 2025; [91] Womack et al., 2025; [97] Castruccio et al., 2014; [103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [105] Bassetti et al., 2024; [106] Bouabid, Souza, and Ferrari, 2026; [127] Beusch, Gudmundsson, and Seneviratne, 2020; [133] Sudakow, Pokojovy, and Lyakhov, 2022

[134] Stull, 2011; [135] Williams et al., 2019

[62] Monier et al., 2013; [78] Gao, Sokolov, and Schlosser, 2023

Motivated by the need for accessible, physically consistent, and spatially resolved climate data for impact assessment, this thesis explores the potential of climate emulators to bridge this gap, addressing three main questions:

1. Under what conditions do structural assumptions cause emulators to fail, and what trade-offs emerge across different emulation techniques?

Chapter 1 introduces the first concerted effort to develop a theoretical framework to understand the relative strengths and weaknesses of various climate emulation techniques. Drawing from concepts in statistical mechanics and stochastic calculus, we perform a fundamental methodological comparison of emulators, applying our framework to understand potential sources of emulator error: memory effects, hidden variables, system noise, and nonlinearities. We consider popular emulation techniques such as pattern scaling and response functions, relating them to less commonly used methods, such as Dynamic Mode Decomposition (DMD) and the Fluctuation Dissipation Theorem (FDT). To support our theoretical contributions, we provide practical implementation guidance for each technique. Using pedagogical examples including idealized box models and a modified Lorenz 63 model, we illustrate the expected errors from each emulation technique considered. We find that response function-based emulators outperform other techniques, particularly pattern scaling, across all scenarios tested. Potential benefits and trade-offs from incorporating statistical mechanics in climate emulation through the use of the FDT are discussed, along with the importance of designing future scenarios for ESMs with emulation in mind. We argue that large-ensemble experiments utilizing the FDT could benefit climate modeling and impacts communities. We conclude by discussing optimal use cases for each emulator, along with implications for ESMs based on our pedagogical model results.

2. To what extent do varying scenario structures constrain or enhance an emulator's predictive skill?

Chapter 2 builds on a key finding of Chapter 1: the popular Shared Socioeconomic Pathways (SSPs) are not necessarily the best scenario choice to train an emulator on. Here, we demonstrate that the information content of the training data itself is a bottleneck for generalization, introducing a methodology to generate optimal training data. The approach leverages a differentiable simple climate model to calculate the sensitivity of emulator loss to input emissions trajectories. We apply this framework to both simple and intermediate complexity climate models, and compare an emulator trained using a dataset optimized for emulator training against a baseline emulator trained using the proposed ScenarioMIP protocol for the 7th phase of the Coupled Model Intercomparison Project. Results show that training on just one or two optimal scenarios outperforms an emulator trained on the ScenarioMIP Priority 1 protocol, achieving higher predictive skill for a lower computational cost. Combining scenarios from distinct optimization initializations additionally yields super-linear improvements in skill. Crucially, emulators trained on these optimized trajectories successfully learn the distinct physical behaviors of individual forcing agents (e.g., greenhouse gases vs. aerosols) despite never observing them in isolation. This suggests that stress-testing climate models with structurally diverse, high-frequency forcings enables emulators to learn robust physical patterns that standard scenarios mask. We discuss extensions of this technique to other domains and suggest that climate modeling centers consider dedicating resources to scenarios that are well-suited for emulator development.

3. How can operationalizing emulators enhance the assessment of impact-relevant metrics?

Finally, Chapter 3 explores the utility of emulators in bridging the gap between the socio-economic detail of integrated assessment models and the physical fidelity of ESMs. We present an impact assessment framework coupling the MIT Emissions Prediction and Policy Analysis (EPPA) model with a generative diffusion-based climate emulator. This approach rapidly generates realizations of spatially correlated climate fields at a fraction of the computational cost of an ESM. Benchmarking against a pattern scaling technique utilized by the MIT IGSM confirms the generative emulator's ability to reproduce several impact-relevant climate variables. We utilize this framework to assess regional sensitivities to various socio-economic pathways, including an early assessment of the proposed ScenarioMIP-CMIP7 protocol. Results show that internal variability masks spatially explicit differences between globally distinct scenarios (e.g., 2°C vs. 1.5°C). Furthermore, we demonstrate that temperature overshoot pathways result in substantially higher cumulative heat stress risks compared to stabilization pathways with similar end-of-century outcomes. This framework democratizes access to high-resolution climate data, enabling the rapid, probabilistic assessment of compound climate hazards essential for robust adaptation planning.

A theoretical framework to understand sources of error in Earth System Model emulation[2]

1

If the flap of a butterfly's wings can be instrumental in generating a tornado, it can equally well be instrumental in preventing a tornado.

— Edward N. Lorenz

Earth System Models (ESMs) are our most comprehensive tools to simulate the climate system, yet their high computational cost limits the range and number of scenarios that can be investigated[33,41]. Growing demand for high-quality climate projections which differ from the scenarios considered within the Coupled Model Intercomparison Project (CMIP) drives a need for computationally efficient alternatives[38]. Climate emulators—reduced-order models that reproduce the outputs of full-scale climate models—have seen a surge in popularity as they can be many orders of magnitude faster than the parent models[133]. Their low computational costs also make them an appealing tool to disseminate climate information to audiences beyond the climate science community.

[33] Flato, 2011; [41] Müller et al., 2018

[38] Eyring et al., 2016

[133] Sudakow, Pokojovy, and Lyakhov, 2022

Because of sensitivity to initial conditions, predicting the instantaneous state of the atmosphere (i.e., the weather) is infeasible beyond short time horizons[29]. Climate emulators must therefore target the statistics of climate variables, such as means, variances, or higher moments, rather than simulating chaotic dynamics when approximating the state of the atmosphere on longer timescales[127,129,136]. Many emulation techniques exist to estimate the mean state and/or probability distribution of climate variables[51,54,77,97,100,103,105,137,138], and in this work we explore methods that emulate the mean state of the system. In a recent review, Tebaldi et al. (2025)[64] distinguished between five main categories of climate emulators, including linear pattern scaling, statistical approaches, and machine learning algorithms. Following their categorization, we focus on linear pattern scaling and its immediate extensions along with dynamical system/impulse response theory emulators.

[29] Lorenz, 1972

[127] Beusch, Gudmundsson, and Seneviratne, 2020; [129] Geogdzhayev et al., 2026; [136] Wang et al., 2025

[51] Meinshausen, Raper, and Wigley, 2011; [54] Leach et al., 2021; [77] Herger, Sanderson, and Knutti, 2015; [97] Castruccio et al., 2014; [100] Watson-Parris et al., 2022; [103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [105] Bassetti et al., 2024; [137] Tebaldi and Knutti, 2018; [138] Addison et al., 2024

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

In the climate context, the most commonly used emulation technique is pattern scaling[74]. This approach relies on a simple linear regression of local climate variables (e.g., temperature or precipitation anomaly) onto the global mean temperature anomaly, where the local variables are typically averaged over an ensemble of many realizations to capture the mean state of the predicted climate attractor. Pattern scaling has been used and studied extensively since its development[56,76,79,80], with variations that capture seasonal anomalies, different mixes of greenhouse gases, and spatially heterogeneous forcings such as aerosols[75,77,82]. This approach produces accurate projections assuming exponential and fixed-pattern forcing, linear feedbacks, and linear and time-independent dynamics, criteria that are roughly satisfied in a number of CMIP experiments[56]. Memory effects in overshoot scenarios (forcing history, rather than only instantaneous forcing, affecting a future state) represent one way these assumptions can be violated, causing this approach to break down for many decision-relevant scenarios.

[74] Santer et al., 1990

[56] Giani et al., 2025; [76] Mitchell, 2003; [79] Tebaldi and Arblaster, 2014; [80] Wells et al., 2023

[75] Schlesinger et al., 2000; [77] Herger, Sanderson, and Knutti, 2015; [82] Mathison et al., 2025

[56] Giani et al., 2025

Impulse response methods, commonly referred to as either response or Green's functions, address this limitation by encoding forcing history into the emulator, rather than relying only on the instantaneous forcing. These techniques have been studied thoroughly in the contexts of dynamical systems and climate science[83,84,86,93,139,140], and are an active area of research[104]. Response functions are popular due to their ease of interpretability and improvement in skill over pattern scaling in capturing realistic dynamics[91]. Pure linear response functions cannot account for nonlinear effects, which are distinct from memory effects, though hybrid schemes that incorporate machine learning (ML) may help resolve this issue[104].

[83] Joos and Bruno, 1996; [84] Orbe et al., 2018; [86] Hasselmann et al., 1997; [93] Lucarini, Ragone, and Lunkeit, 2017; [139] Freese et al., 2024; [140] Giorgini et al., 2024

[104] Winkler and Sierra, 2025

[91] Womack et al., 2025

[104] Winkler and Sierra, 2025

Pattern scaling and linear response functions are prevalent in climate emulation literature, yet these approaches are only two methods among a broad spectrum of emulators, with each technique offering trade-offs in terms of complexity, data requirements, and interpretability. For example, quasi-equilibrium emulation is closely related to pattern scaling, though only a handful of studies explore the utility of this principle beyond the traditional choice of global mean temperature as emulator input[141,142]. Other techniques, such as Dynamic Mode Decomposition (DMD) and its variants, are generally not classified as emulators despite their potential to identify and predict modes of variability in the climate system[143–146].

[141] Huntingford and Cox, 2000; [142] Cao et al., 2015

[143] Kutz, Fu, and Brunton, 2016; [144] Gottwald and Gugole, 2020; [145] Navarra, Tribbia, and Klus, 2021; [146] Mankovich et al., 2025

We consider climate emulators as defined in Tebaldi et al. (2025)[64], excluding Simple Climate Models (SCMs) and Earth system Models of Intermediate Complexity (EMICs), though they share similarities with emulators. We also do not examine ML emulators such as FourCastNet and NeuralGCM – while these techniques are promising for weather prediction, they currently lack the stability required for reliable climate prediction[109,147]. Several studies have employed ML techniques to instead target the statistics of the climate, rather than weather[105,106,136,148], but these works focus on emulator implementation rather than theoretical analysis.

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

[109] Kochkov et al., 2024; [147] Pathak et al., 2022

[105] Bassetti et al., 2024; [106] Bouabid, Souza, and Ferrari, 2026; [136] Wang et al., 2025; [148] Lewis et al., 2017

In this chapter, we develop a framework connecting a spectrum of emulators through the Koopman and Fokker-Planck operators, which govern the evolution of stochastic processes. In doing so, we identify a gap in the[64] emulator typology: operator-based emulators, an area largely unexplored in existing climate emulator literature. While previous work has connected operator frameworks with the Fluctuation Dissipation Theorem and thus, linear response theory[32,93,149–151], our contribution explicitly demonstrates its utility in the context of climate emulation. Section 1.1 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. 1.1.3). In Sect. 1.2, 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 1.3 contains the results of these simplified climate model experiments, 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. 1.4).

[64] Tebaldi et al., 2025

[32] Lembo, Lucarini, and Ragone, 2020; [93] Lucarini, Ragone, and Lunkeit, 2017; [149] Cooper and Haynes, 2011; [150] Zagli et al., 2024; [151] Giorgini, Falasca, and Souza, 2025

1.1 Theoretical framework for climate emulation

In this section, we outline a theoretical framework for climate emulation based on the Koopman and Fokker-Planck operators. Section 1.1.1 introduces our emulation target, a general, stochastic system, outlining potential sources of error when emulating this system. Section 1.1.2 then formalizes two complementary emulation strategies: emulating the full probability distribution, or emulating a collection of statistical moments (e.g., mean, variance). We conclude this section by connecting theoretical and practical (i.e., implementation) details for the six emulators of interest (Sect. 1.1.3). See Fig. 1.2 for a conceptual roadmap of emulator theory and Table 1.1 for an overview of selected methods.

Throughout this section, we denote scalars with lowercase characters, vectors with lowercase, boldface italic characters, matrices with uppercase, boldface characters, and operators with script characters (e.g., $\mathcal{N}$ or $\mathcal{L}$). We use $\boldsymbol{x}$ and n_x to denote the spatial coordinate and its dimensionality, along with t and n_t to denote the temporal coordinate and its dimensionality. Our examples focus on climate anomalies relative to a background state, though these techniques are applicable to general chaotic dynamical systems.

1.1.1 Problem setup

A full-scale climate model is a deterministic, albeit chaotic, system. This chaos results in extreme sensitivity to initial conditions, requiring emulation of the system's statistics, rather than its dynamics[29]; here we assume the climate system has a chaotic attractor that is predictable. To understand the statistics of the system and how they may change over time, we follow Hasselmann (1976)[152] in modeling the evolution of a single climate variable using a stochastic differential equation (SDE) (Fig. 1.2, box 1). We assume time-scale separation between slow climate processes (e.g., ocean, cryosphere, land vegetation) and other, faster sources of variability.

[29] Lorenz, 1972

[152] Hasselmann, 'Stochastic climate models Part I. Theory', *Tellus*, 1976

In this framework, the climate is regarded as the statistical mean of a process that appears stochastic in individual realizations. We treat variations occurring either on timescales shorter than climate change (such as short-term weather fluctuations and interannual variability) or in different realizations as stationary, stochastic noise. This allows us to parameterize their influence on the statistics of the chaotic system:

$$\frac{\partial w}{\partial t} = \mathcal{N}(w) + F(t) + \varepsilon\xi(t), \tag{1.1}$$

where w is the climate variable (or set of variables) of interest (e.g., temperature), F is an external forcing (e.g., CO_2), $\mathcal{N}$ is the operator governing the evolution of that variable (under slow climate processes), ξ is a white noise term (aggregated fast effects, including weather and interannual variability), and ε is the noise standard deviation. The treatment of the aggregated fast variables as white noise relies on timescale separation between weather and climate. Under this assumption, the low-frequency red-noise spectra characteristic of climate variables naturally emerge from the integration of this white-noise forcing by the slow system dynamics[152,153].

[152] Hasselmann, 1976; [153] Frankignoul and Hasselmann, 1977

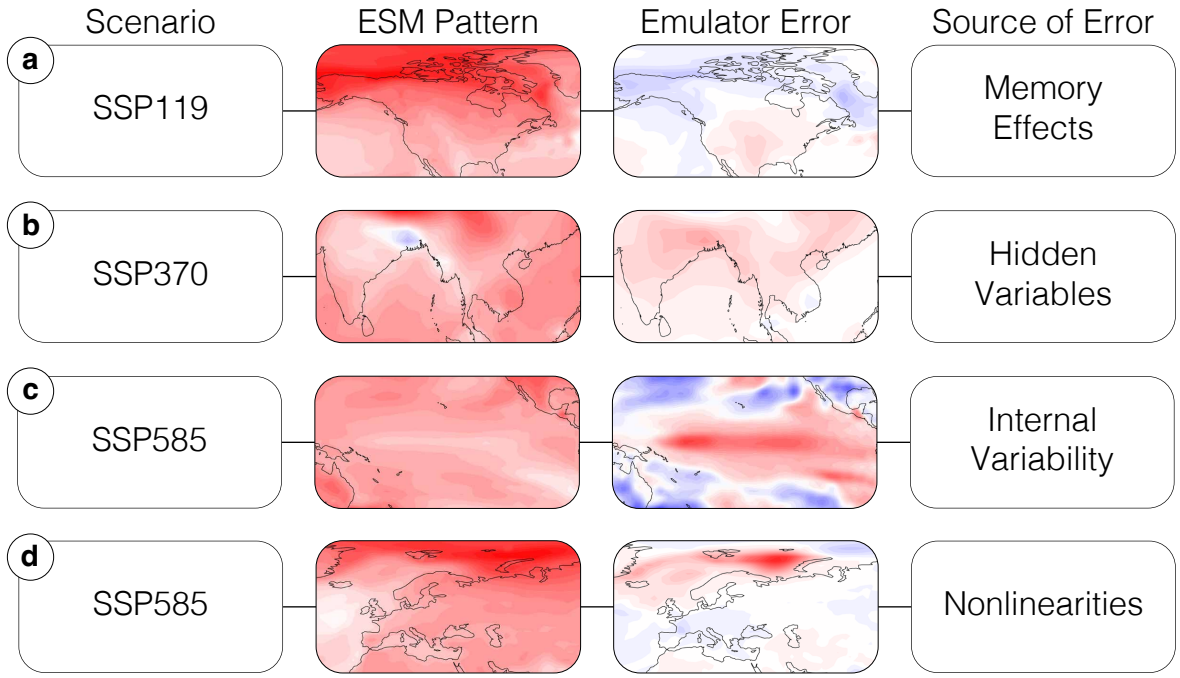

Figure 1.1: Potential sources of emulator error by scenario. Emulator errors shown here are meant for illustrative purposes only; we introduce experiments which reproduce these errors in simplified climate models (e.g., box models) in Sect. 1.2. (a) Pattern scaling emulator trained on *historical* and *SSP585*, tested against *SSP119* in 2100; error over northern North America results from memory effects. (b) Pattern scaling emulator trained on *historical*, tested against *SSP370* in 2050; error over India and SE Asia results from hidden variables (aerosols not contained in training data). (c) High-order polynomial pattern scaling emulator trained on *historical*, tested against *SSP585* in 2020; error results from overfitting on internal variability. (d) Pattern scaling emulator trained on *historical*, tested against *SSP585* in 2100; error results from nonlinear feedbacks in the Arctic. All ScenarioMIP data shown are taken from the MPI Grand Ensemble[31,154].

We consider variables of interest to be anomalies relative to some base state (e.g., temperature anomaly with respect to preindustrial conditions). $\mathcal{N}$ may involve both linear and nonlinear terms in one or several fields, and we cannot directly represent this operator; this parameterization aggregates the effects of processes such as heat and momentum transfers. The operator may also be influenced by variables we observe as well as unobserved hidden variables (e.g., aerosol forcing in a pattern scaling emulator with only global mean temperature as an input). The noise standard deviation can also be state dependent, though we treat it as independent for this exploration.

Climate emulators approximate Equation 1.1, either implicitly (pattern scaling) or explicitly (Dynamic Mode Decomposition), rendering them vulnerable to several potential sources of error. Figure 1.1 provides an overview of the sources of error we consider across a range of scenarios: Errors can enter from the forcing if an emulator assumes only the instantaneous forcing is significant and not the forcing history (Fig. 1.1 (a) - memory effects in an overshoot scenario). The presence of hidden variables can lead to errors in some techniques (Fig. 1.1 (b) - localized aerosol effects when assuming well-mixed forcings), while other techniques are sensitive to noise (Fig. 1.1 (c) - overfitting on internal variability). Finally, any linear emulation technique will break down in the presence of nonlinearities (Fig. 1.1 (d) - ice-albedo feedbacks).

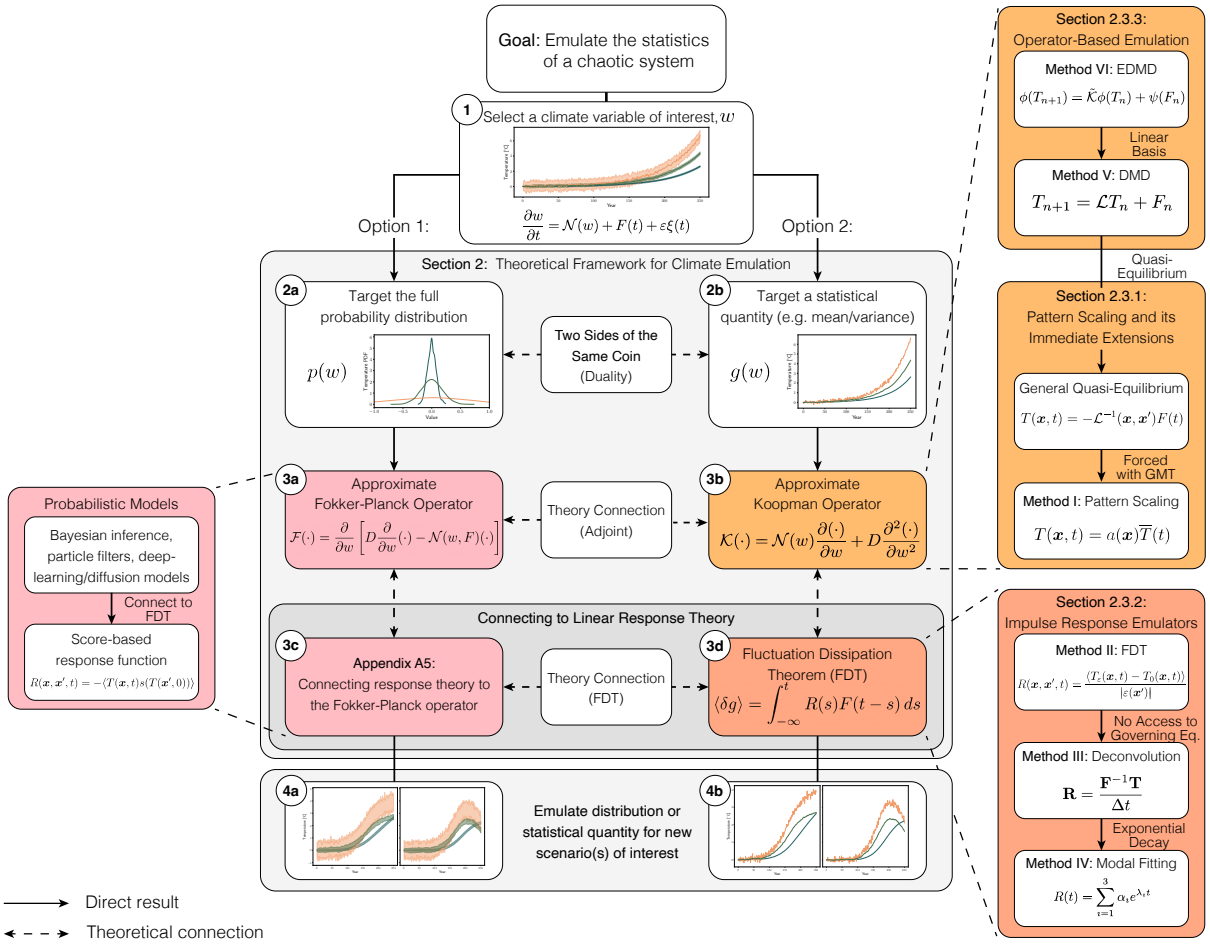

Figure 1.2: Conceptual flowchart for building an emulator through the joint Fokker-Planck/Koopman operator framework. Pop-outs show specific emulation techniques, while the shaded color indicates which concept a class of emulators relates to. Dashed arrows indicate conceptual/theoretical connections and solid arrows indicate a direct pathway. The overall process is as follows: (1) Select a climate variable of interest, w, such as temperature, here parameterized as the output of a stochastic differential equation. (2) Choose an emulation target, either the full probability distribution (option 1; 2a, 3a, 3c, 4a) or a statistical quantity such as the mean or variance (option 2; 2b, 3b, 3d, 4b). (3) Construct an emulator by selecting an approximation for either the Fokker-Planck or Koopman operator, including their response function representations; these options are connected through duality and are directly linked to linear response theory. (4) Given a new scenario of interest, emulate either the probability distribution or statistical quantity. A summary of emulation techniques explored in this work (right side of this figure) can be found in Table 1.1.

1.1.2 Operator framework for emulators

Our operator framework simplifies complex, possibly nonlinear climate dynamics into a linear set of modes with associated decay rates. We use the term operator to refer to an update rule that advances the system one timestep for a quantity of interest. An emulator attempts to approximate these modes, which are physically interpretable; for temperature, the decay rates correspond to heat-uptake timescales.

Table 1.1 summarizes emulation techniques discussed in this section, providing a short conceptual description of each method along with their key assumptions. We focus on linear emulation techniques that target the mean state of a climate variable when averaged over many realizations: pattern scaling, the Fluctuation Dissipation Theorem (FDT), deconvolution, modal fitting, Dynamic Mode Decomposition (DMD), and Extended DMD (EDMD). The FDT, deconvolution, and modal fitting emulators are all response function-based emulators, while EDMD and DMD are operator-based emulators.

Table 1.1: Summary of emulation techniques discussed in this work including a short description and their key assumptions; a conceptual overview of these methods can be found in Fig. 1.2. Fluctuation Dissipation Theorem assumptions are shared with deconvolution and modal fitting emulation techniques. All techniques except the Fluctuation Dissipation Theorem additionally assume no hidden variables.

Technique	Short Description	Key Assumptions	Pros	Cons
Method I: Pattern Scaling (Pattern Scaling and its Immediate Extensions)	Time-invariant pattern based on global mean temperature	Climate is always near equilibrium; response is instantaneous; fixed spatial pattern	Computationally efficient	Structurally biased with irreducible errors
Method II: Fluctuation Dissipation Theorem (Dynamical System/Impulse Response Theory)	Response functions derived through perturbation ensemble experiments	Perturbations are small; data come from linear response regime	Gives interpretable physical response	Requires nonstandard, computationally expensive scenarios
Method III: Deconvolution (Dynamical System/Impulse Response Theory)	Response functions solved for from any general experiment	Quasi-equilibrium initial condition; influence of noise is small	Applicable to any scenario	Sensitive to noise, can give non-physical responses
Method IV: Modal Fitting (Dynamical System/Impulse Response Theory)	Response functions fit from any general experiment	Response is a decaying exponential; few significant modes	Applicable to any scenario	Requires initial guess, can give non-physical responses
Method V: Dynamic Mode Decomposition (DMD) (Operator-based Emulation)	Approximating system dynamics with a linear operator	Dynamics are approx. linear; training data capture relevant dynamics	Gives interpretable spatiotemporal information	Strong assumption of linearity
Method VI: Extended DMD (Operator-based Emulation)	Approximating system dynamics with nonlinear basis functions	Basis functions span Koopman operator; dynamics are approx. linear in new basis	Can theoretically reproduce any system behavior	Requires selection of basis functions

Emulating a probability distribution. Our governing system, Equation 1.1, simulates a variable of interest, w, forward in time under a stochastic forcing. The trajectory of the time evolution of w is characterized by the probability distribution, $p(w,t)$. We therefore focus our efforts on emulating $p(w,t)$ via the Fokker-Planck operator. This is a mathematical tool to evolve the probability distribution of a stochastic system forward in time. As this operator is linear, emulating it is equivalent to approximating a series of eigenvalues and eigenfunctions.

As described by Hasselmann (1976)[152], the time evolution of $p(w,t)$ is given by the Fokker-Planck equation corresponding to the governing SDE

[152] Hasselmann, 'Stochastic climate models Part I. Theory', *Tellus*, 1976

$$\frac{\partial}{\partial t}p(w,t) = -\frac{\partial}{\partial w}\left[p(w,t)\left(\mathcal{N}(w)+F(t)\right)\right] + D\frac{\partial^2}{\partial w^2}p(w,t), \qquad (1.2)$$

where D is a diffusion coefficient set by the noise term, $D = \varepsilon^2/2$. The Fokker-Planck equation describes how the probability density evolves in time and can be viewed as an advection-diffusion process.

Advection, which shifts the mean of $p(w,t)$, occurs due to the deterministic action of the governing operator and the external forcing. Because the advective term acts on the flux, it both shifts the mean and reshapes the density. Diffusion, which increases the variance in $p(w,t)$, is driven by system noise. Integrating Equation 1.2 forward diffuses the probability distribution, initially increasing the variance of w until balanced by the mean-reverting drift $(\mathcal{N}(w) + F(t))$. It is common practice to write a Fokker-Planck equation directly from an SDE, as there exists a general relationship between any SDE and its corresponding Fokker-Planck equa-

tion; the full general derivation can be found in Denisov, Horsthemke, and Hänggi (2009)[155].

[155] Denisov, Horsthemke, and Hänggi, 'Generalized Fokker-Planck equation: Derivation and exact solutions', *The European Physical Journal B*, 2009

Importantly, the right hand side of Equation 1.2 is linear in the derivatives of w, allowing us to rewrite it in terms of the linear Fokker-Planck operator, $\mathcal{F}$,

$$\mathcal{F}(\cdot) = \frac{\partial}{\partial w}\left[D\frac{\partial}{\partial w}(\cdot) - (\cdot)\left(\mathcal{N}(w) + F(t)\right)\right], \tag{1.3}$$

where the notation $\mathcal{F}(\cdot)$ means the Fokker-Planck operator is acting on some arbitrary variable (in our case, $p(w,t)$ in Equation 1.2). The Fokker-Planck operator (Fig. 1.2, box 3a) gives us a linear method to represent the time evolution of the probability distribution. Linearity additionally allows us to decompose $\mathcal{F}$ into eigenvalues and eigenfunctions (continuous eigenvectors). These are the target of our emulator, and our emulator skill is directly proportional to how well it can approximate those eigenvalues and eigenfunctions, along with our estimate of $p(w,0)$. This eigendecomposition is given by

$$\mathcal{F} f_{\mathcal{F}} = \lambda_{\mathcal{F}} f_{\mathcal{F}}, \tag{1.4}$$

where $\lambda_{\mathcal{F}}$ denotes an eigenvalue and $f_{\mathcal{F}}$ denotes an eigenfunction of the Fokker-Planck operator. Assuming the operator and boundary conditions permit a complete discrete spectrum, the collection of $\lambda_{\mathcal{F}}$ and $f_{\mathcal{F}}$ fully characterizes the system's behavior. In such cases, our stochastic system evolves as a linear combination of probability distributions, $f_{\mathcal{F}}$, each decaying at rate $\lambda_{\mathcal{F}}$; the real part of the eigenvalues controls the decay rate, while any imaginary components result in oscillations over time. In the advection-diffusion analogy, each eigenfunction is a probability parcel that is carried and spread by the flow. The imaginary parts of the eigenvalues transport this parcel (shifting the mean) while the real parts act like an effective diffusivity (increasing the variance). This tells us which physical behaviors dominate and on what timescales they matter for climate prediction.

Unfortunately, in most cases we cannot obtain an explicit representation of the Fokker-Planck operator due to $\mathcal{N}$ being nonlinear; see Appendix A.3 for an analytic example of when this is possible. Because it acts on functions, the operator is infinite dimensional with infinitely many eigenpairs. This poses an immediate issue since computers have a finite amount of memory. Finite dimensional matrix approximations of the Fokker-Planck operator have been studied (often framed through the more general Perron-Frobenius operator)[156–161], but require a large amount of data to reliably estimate the operator. For climate emulation this poses an additional issue, as generating large enough ensembles to resolve $p(w,t)$ is prohibitively expensive. Because of these difficulties, little work exists studying the Fokker-Planck/Perron-Frobenius operator in the climate context[145], though methods that reconstruct the full probability distribution of a climate variable using statistical methods (e.g., diffusion models and Gaussian processes) implicitly represent it[103,105,136].

[156] Klus, Koltai, and Schütte, 2016; [157] Klus et al., 2018; [158] Kaiser, Kutz, and Brunton, 2019; [159] Souza, 2024; [160] Souza, 2024; [161] Souza and Silvestri, 2024

[145] Navarra, Tribbia, and Klus, 2021

[103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [105] Bassetti et al., 2024; [136] Wang et al., 2025

Emulating a statistical quantity. In practice, it is often easier to emulate statistical quantities, such as the mean or variance of a climate variable. Many common emulation techniques (e.g., pattern scaling and response functions) target only the mean of a single variable[77,80,139], though other work extends this to approximate second-order moments[127,136]. Relating these techniques requires the use of Koopman operator theory (Fig. 1.2,

[77] Herger, Sanderson, and Knutti, 2015; [80] Wells et al., 2023; [139] Freese et al., 2024

[127] Beusch, Gudmundsson, and Seneviratne, 2020; [136] Wang et al., 2025

box 3b), a linear framework for propagating statistical quantities (usually referred to in the Koopman literature as statistical observables) forward in time[162,163]. Emulator studies rarely link their methods to Koopman theory, while literature that explicitly connects to the theory does not use the same emulator terminology[145,164], though they accomplish similar prediction tasks. The Koopman operator allows for an exact representation of nonlinear dynamics using a linear operator, making it appealing when studying complex systems. We show how it can be used to emulate climate variables, simplifying nonlinear processes to the linear problem of emulating physically interpretable eigenvalues and eigenfunctions.

[162] Mezić, 2013; [163] Otto and Rowley, 2021

[145] Navarra, Tribbia, and Klus, 2021; [164] Slawinska, Szekely, and Giannakis, 2017

To derive the Koopman operator, we first define a general statistical quantity, $g(w)$, whose expectation, $\langle\cdot\rangle$, is given by

$$\langle g(w)\rangle = \int g(w)p(w,t)\,dw, \tag{1.5}$$

We then take the time derivative of this expression, moving the partial derivative inside the integral to act only on p since $g(w)$ is independent of time. This allows us to substitute the resulting expression into the right hand side of Equation 1.2. Integrating this by parts twice gives

$$\frac{\partial}{\partial t}\langle g(w)\rangle = \left\langle \left[\mathcal{N}(w) + F(t)\right]\frac{\partial}{\partial w}g(w)\right\rangle + D\left\langle \frac{\partial^2}{\partial w^2}g(w)\right\rangle, \tag{1.6}$$

where the diffusivity, $D = \varepsilon^2/2$, is identical to the Fokker-Planck case. This form allows us to define the Koopman operator, $\mathcal{K}$. It is linear in its derivatives of w, and we rewrite it as

$$\mathcal{K}(\cdot) = \mathcal{N}(w)\frac{\partial(\cdot)}{\partial w} + D\frac{\partial^2(\cdot)}{\partial w^2}, \tag{1.7}$$

where the notation $\mathcal{K}(\cdot)$ means the Koopman operator is acting on some arbitrary variable ($g(w)$ in Equation 1.7). Substituting this into Equation 1.6 gives

$$\frac{\partial}{\partial t}\langle g(w)\rangle = \langle \mathcal{K}g(w)\rangle + F(t)\left\langle \frac{\partial}{\partial w}g(w)\right\rangle, \tag{1.8}$$

This expression applies to any arbitrary statistical quantity (of which there are infinitely many), thus it can be used to integrate every statistical quantity forward in time; it is an alternate way to represent the complete probability distribution by representing each individual statistic. A useful choice is to select $g(w) = w$, giving

$$\frac{\partial}{\partial t}\langle w\rangle = \langle \mathcal{K}w\rangle + F(t), \tag{1.9}$$

which we will refer back to later.

Analogously to the Fokker-Planck operator, the Koopman operator provides a linear method to represent the time evolution of our entire collection of statistical quantities. As before, we can perform an eigendecomposition on the Koopman operator

$$\mathcal{K}f_{\mathcal{K}} = \lambda_{\mathcal{K}}f_{\mathcal{K}}, \tag{1.10}$$

where $\lambda_{\mathcal{K}}$ denotes an eigenvalue and $f_{\mathcal{K}}$ denotes an eigenfunction. Assuming the operator admits a complete discrete spectrum, the time evolution of our statistical quantity of interest is a linear combination of these eigenpairs. These can be used to identify dominant system dynamics

and on what timescales they emerge. Training an emulator is equivalent to approximating eigenpairs; reproducing these pairs accurately emulates the behavior of the system.

However, approximations of the Koopman operator are limited by the same finite memory constraint as the Fokker-Planck case and deriving analytic solutions is dependent on the exact form of $\mathcal{N}$; see Appendix A.3 for an example of when analytic approximations are possible. Matrix approximations of the Koopman operator are nevertheless more prevalent than their Fokker-Planck counterparts[162,163,165,166]. Variants of these methods have recently been implemented in the climate context to identify dominant modes of variability in the system (e.g., El Niño-Southern Oscillation or Pacific decadal oscillation)[145,146,167], but have not been applied for the purpose of climate emulation. We outline two of these methods explicitly in Sect. 1.1.3.

[162] Mezić, 2013; [163] Otto and Rowley, 2021; [165] Schmid, 2010; [166] Williams, Kevrekidis, and Rowley, 2015

[145] Navarra, Tribbia, and Klus, 2021; [146] Mankovich et al., 2025; [167] Navarra et al., 2024

Two sides of the same coin. The Koopman operator advances all statistical quantities of interest, and provides an alternative to the Fokker-Planck description of a distribution's time evolution. For distributions that are uniquely determined by their moments, knowing every statistic is equivalent to knowing the full distribution. Access to either operator fully characterizes our system, allowing us to emulate it. Mathematically, these operators are dual (adjoint) under the appropriate choice of functional spaces, where duality refers to two mathematical objects that contain alternate descriptions of the same information; this property is how we derived the Koopman operator in the previous section. This is analogous to, but physically and mathematically distinct from adjoint methods in climate modeling. There, adjoints to dynamics (rather than statistics as is the case for the Koopman/Fokker-Planck approach) are exploited to calculate gradients with respect to input parameters more efficiently, which can be used to tune parameters and compute output sensitivities[168–170].

[168] Thuburn, 2005; [169] Henze, Hakami, and Seinfeld, 2007; [170] Lyu et al., 2018

Estimating the full probability distribution of a variable requires large initial condition ensembles, incurring significant computational cost. This is necessary to fully sample the climate system's internal variability, which comprises unforced fluctuations that arise from interactions between the coupled components of the Earth system. This computational expense is exacerbated for variables such as precipitation, where internal variability masks the forced response to a greater degree[171]. Reliably estimating the full distribution at each timestep to approximate the Fokker-Planck operator from relatively coarse data is impractical. However, under additional assumptions of quasi-ergodicity, we bolster our sampling power by assuming that the statistics do not change sufficiently quickly over a given time period. We thus focus on emulating lower-order statistical quantities, presenting those techniques in Sect. 1.1.3.

[171] Blanusa, López-Zurita, and Rasp, 2023

Connecting to linear response theory. Linear response theory states that the climate system's forced response (assuming perturbations are small) is encoded by a response function, $R(t)$. The response function is generated by the Koopman operator, $\mathcal{K}$, where each eigenpair of the operator determines the characteristic timescales of the system. Considering temperature anomaly as an example variable, fast modes map to land and shallow ocean heat uptake, while slow modes capture deep ocean heat uptake[172]. Response functions have been applied to a variety of climate problems[83–85,88,173], including climate emulation[91,92,139], though often without addressing the formal response theory underlying these techniques. As was the case with the Koopman operator, more formal applications of response theory to climate science often do not

[172] Caldeira and Myhrvold, 2013

[83] Joos and Bruno, 1996; [84] Orbe et al., 2018; [85] Cimoli et al., 2023; [88] Hasselmann et al., 2003; [173] Joos et al., 2013

[91] Womack et al., 2025; [92] Sandstad et al., 2025; [139] Freese et al., 2024

share the same language as climate emulators despite the shared goal of predicting the climate's forced response[32,93,150].

[32] Lembo, Lucarini, and Ragone, 2020; [93] Lucarini, Ragone, and Lunkeit, 2017; [150] Zagli et al., 2024

To make the relationship between response theory and the Koopman operator explicit in the context of emulation, we first consider the continuous-time dynamics of the system. When the system is subjected to a small external perturbation, the continuous-time generator of the Koopman operator, $\mathcal{K}$, can be split into an unperturbed, time-independent component, $\mathcal{K}_0$, and a perturbation induced by the forcing, $\delta\mathcal{K}(t)$. As a result, the time evolution of the expectation value of a statistical quantity g can be written as an unperturbed baseline plus a linear correction, governed by the differential equation:

$$\frac{\partial}{\partial t}\delta\langle g\rangle = \delta\langle \mathcal{K}_0 g\rangle + \langle \delta\mathcal{K}(t) g\rangle_0, \tag{1.11}$$

where $\langle\cdot\rangle_0$ denotes the expected value under the unperturbed state.

A general solution for this linear correction is provided by Ruelle's response theory. By treating the perturbation $\delta\mathcal{K}(t)$ as a forcing term, we can integrate the differential equation above. This yields a convolution where the response function, $R(t)$, defines how the baseline system (described by $\mathcal{K}_0$) propagates the effects of the perturbation. For systems in a statistical steady state (i.e., at equilibrium), this framework simplifies to the Fluctuation Dissipation Theorem (FDT)[174]. The FDT describes how a system (e.g., the Earth system) responds to perturbations (anthropogenic CO_2 emissions) relative to some baseline state (preindustrial conditions). The change in the ensemble average field, $\delta\langle g\rangle$, is obtained by convolving a forcing, $F(t)$, with the system's response function, $R(t)$

[174] Lucarini et al., 2026

$$\delta\langle g\rangle = \int_{-\infty}^{t} R(t')F(t-t')\,dt'. \tag{1.12}$$

Formally, the response function is calculated by computing the temporal autocorrelation between the statistical quantity g and the system's score function, s,

$$R(t) = \langle g(t'=t)s(t'=0)\rangle = \langle (\mathcal{K}^t g)s\rangle, \tag{1.13}$$

where $\mathcal{K}^t$ is the Koopman semigroup advancing the statistical quantity g in time. $s = \nabla \ln p_0$ is the general form of the score function of the steady-state distribution which encodes how a small perturbation alters the system's statistics; see Giorgini et al., Giorgini, Falasca, and Souza (2024, 2025)[140,151] for more details. This expression directly connects the system's response to the Koopman operator framework[150].

[140] Giorgini et al., *Response Theory via Generative Score Modeling*, 2024; Giorgini, Falasca, and Souza, 'Predicting forced responses of probability distributions via the fluctuation–dissipation theorem and generative modeling', *Proceedings of the National Academy of Sciences*, 2025

[150] Zagli et al., 2024

Equation 1.12 is one way to state the Fluctuation Dissipation Theorem (FDT, Fig. 1.2, box 3d), a tool widely used in statistical mechanics and one of the main features of linear response theory[32,93]. The FDT predicts the first-order response of a statistical quantity due to external perturbations and is defined in terms of an ensemble average over a quantity of interest. As written, this form does not account for state- or time-dependent effects (i.e., one could consider the alternate formulation: $R = R(w, t, t')$), though extensions to capture these effects and higher-order statistical moments have been proposed[104,151,175,176].

[32] Lembo, Lucarini, and Ragone, 2020; [93] Lucarini, Ragone, and Lunkeit, 2017

[104] Winkler and Sierra, 2025; [151] Giorgini, Falasca, and Souza, 2025; [175] Metzler, Müller, and Sierra, 2018; [176] Giorgini, Bischoff, and Souza, 2025

Response function emulators approximate the left hand side of Equation 1.13 using a variety of techniques, which we outline in more detail in Sect. 1.1.3. Their emulation goal is typically either to fit the eigenpairs which make up $\mathcal{K}$ explicitly[92], or to find a direct representation of $R(t)$ (i.e.,

[92] Sandstad et al., 2025

an implicit representation of $\mathcal{K}$)[32,91,139]. The former may be more easily interpretable through analyzing the explicit eigenpairs, while the latter offers flexibility in allowing for parametric forms other than a decaying exponential.

[32] Lembo, Lucarini, and Ragone, 2020; [91] Womack et al., 2025; [139] Freese et al., 2024

Response theory builds upon the operator frameworks presented in the previous sections by providing a method to illustrate how a given quantity responds to small changes in forcing. While the Fokker-Planck and Koopman perspectives offer complete characterizations of the statistics of the system over time, response theory offers a practical approach to use this information to predict how a quantity shifts under perturbations, described by the FDT.

1.1.3 Connecting emulators to theory

Following the framework from the previous section, we introduce several emulation techniques targeting the mean of a climate variable (Fig. 1.2, pop-outs on right hand side). We use the example of estimating the expected (or annual-average) temperature anomaly, $T(\boldsymbol{x}, t)$, given an external forcing, $F(t)$ (e.g., CO_2 or other GHG emissions), though these techniques can be applied to any climate field. Each technique relates explicitly to the Fokker-Planck or Koopman operator and/or the Fluctuation Dissipation Theorem (FDT). We begin with methods that impose strong assumptions on the underlying data and progressively lift those assumptions until we are left with the most general emulation techniques; headings follow the taxonomy of Tebaldi et al. (2025)[64] when possible.

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

Pattern scaling and its immediate extensions

Method I: Pattern Scaling. Pattern scaling is arguably the most well-known climate emulation technique[56,74,76,79,80,137,177]; it is formally derived via the Koopman operator, and is a specific case of a more general quasi-equilibrium emulation framework. It assumes that, at any given moment, the climate is in a quasi-equilibrium, rather than a transient, state and that changes in the forcing are small enough and/or the response of the system is fast enough to neglect system memory. Pattern scaling also assumes that the response does not depend on the background climate state, only the instantaneous forcing. Despite work showing that there are measurable differences between transient and quasi-equilibrium climate responses depending on the transient warming rate[178], the success of pattern scaling has led to its continued use.

[56] Giani et al., 2025; [74] Santer et al., 1990; [76] Mitchell, 2003; [79] Tebaldi and Arblaster, 2014; [80] Wells et al., 2023; [137] Tebaldi and Knutti, 2018; [177] Kravitz et al., 2017

[178] King et al., 2021

We first restate Equation 1.9 in terms of the quasi-equilibrium assumption and our climate variable of interest as

$$\frac{\partial}{\partial t} T(\boldsymbol{x}, t) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{x}') T(\boldsymbol{x}', t) + F(t) \approx 0, \tag{1.14}$$

where $\mathcal{L}$ indicates that this is no longer the true Koopman operator and $\boldsymbol{x}$ and $\boldsymbol{x}'$ indicate summation over spatial interactions, i.e., how one location, $\boldsymbol{x}$, is influenced by all other locations (including itself), $\boldsymbol{x}'$; a more detailed description of the transition from Equation 1.9 to 1.14 can be found in Appendix A.1.4. We additionally assume $T(\boldsymbol{x}, t)$ here refers to the ensemble mean temperature, which has the practical advantage of reducing the impact of internal variability on our emulator. Inverting

this equation gives

$$T(\boldsymbol{x}, t) = -\mathscr{L}^{-1}(\boldsymbol{x}, \boldsymbol{x}')F(t), \tag{1.15}$$

which is a more general formulation of pattern scaling based on a generic forcing, $F(t)$. Alternate definitions of pattern scaling have been explored previously, with a handful of studies developing extensions based on alternatives to global mean temperature such as radiative forcing or a combination of factors[141,142]. A traditional pattern scaling formulation makes the further assumption that the forcing is proportional to the global mean temperature anomaly (acting as a proxy for the integrated history of emissions and subsequent radiative forcing), $F(t) \propto \overline{T}(t)$, and replaces $\mathscr{L}^{-1}$ with a low-order polynomial, leading to

[141] Huntingford and Cox, 2000; [142] Cao et al., 2015

$$T(\boldsymbol{x}, t) = a_0(\boldsymbol{x}) + a_1(\boldsymbol{x})\overline{T}(t) + \frac{1}{2}a_2(\boldsymbol{x})\overline{T}^2(t) + \dots, \tag{1.16}$$

where $a_i(\boldsymbol{x})$ indicates the spatially varying pattern, and we typically keep only the first-order ($a_1(\boldsymbol{x})$) term. While some work has incorporated higher-order terms, such models are limited in their general extrapolative ability[77]. Other studies indicate that quadratic terms may be necessary to capture end-of-century warming behavior[129].

[77] Herger, Sanderson, and Knutti, 2015

[129] Geogdzhayev et al., 2026

Although pattern scaling implicitly attempts to approximate the Koopman operator - the perfect linear representation of the system - it is limited by its assumption of time-invariant, quasi-equilibrium dynamics. Truncating the operator with a finite dimensional approximation and using only a single predictive field (here, annual-mean temperature) further reduces its skill. Pattern scaling's inability to reproduce the pattern effect—where changes in the spatial distribution of surface warming over time dynamically alter global climate feedbacks—and other nonlinear/state-dependent feedbacks illustrates these limitations[56,179]. In Sect. 1.1.3, we explore alternative low-order approximations of the Koopman operator to resolve these issues.

[56] Giani et al., 2025; [179] Stevens et al., 2016

Pattern scaling could be extended to the Fokker-Planck operator by shifting and rescaling the full probability distribution based on global mean temperature, but this faces several limitations. Reliably estimating probability distributions requires large ensembles, which are computationally expensive. An alternate approach is to use long preindustrial control runs to generate the initial probability distribution and attempt to learn the linear scaling factor through the shorter SSP experiments. However, a simple linear shift may not capture scenario-dependent changes in the shape of the distribution; recent emulation work with Gaussian process regression suggests these distributional shifts may be complex[136]. When applying pattern scaling to the Fokker-Planck operator, we must also ensure the process does not violate the normalization of the distribution (i.e., the area under the curve must equal one).

[136] Wang et al., 2025

We implement pattern scaling by calculating the global mean temperature anomaly and solving

$$\min_{a(\boldsymbol{x})} \|T(\boldsymbol{x}, t) - a(\boldsymbol{x})\overline{T}(t)\|^2. \tag{1.17}$$

In Appendix A.1.1 we show that pattern scaling has two irreducible sources of error when trained on a ScenarioMIP-like forcing: (1) an equilibrium term, where pattern scaling converges to the wrong steady-

state value when forcing plateaus and (2) a memory term, where pattern scaling breaks down when the system responds slowly compared to changes in the forcing. The former stems from the mismatch between training pattern scaling in a transient regime and attempting to use it to project an equilibrium condition. The latter cannot be accounted for within the pattern scaling framework, motivating the need for methods that explicitly capture memory.

Dynamical system/impulse response theory

Emulators that represent the climate system through response functions connect to fundamental principles of statistical mechanics and the Koopman/Fokker-Planck framework[32,83–87,89–93,139,180]. Response function emulators relax the quasi-equilibrium assumption, assuming instead that the current transient climate state is close to some baseline climate state that is in statistical equilibrium (generally preindustrial conditions). Perturbations to a field of interest are assumed to be small relative to magnitude of that field. These methods enable us to capture memory effects by integrating the entire forcing time history rather than only using the instantaneous forcing. One major benefit of this is that we can use them to represent regional shifts in surface warming patterns over time (the pattern effect)[181].

[32] Lembo, Lucarini, and Ragone, 2020; [83] Joos and Bruno, 1996; [84] Orbe et al., 2018; [85] Cimoli et al., 2023; [86] Hasselmann et al., 1997; [87] Hasselmann, 2001; [89] Fredriksen, Rugenstein, and Graversen, 2021; [90] Fredriksen et al., 2023; [91] Womack et al., 2025; [92] Sandstad et al., 2025; [93] Lucarini, Ragone, and Lunkeit, 2017; [139] Freese et al., 2024; [180] Farley et al., 2026

[181] Bloch-Johnson et al., 2024

The use of different methods to derive response functions affects their utility as an emulator. A key assumption behind the Fluctuation Dissipation Theorem, for example, is that we have access to the governing equation, i.e., we are free to run large ensembles as needed. We begin this section assuming this is true, and relax this assumption later.

Method II: The Fluctuation Dissipation Theorem. In the case of a fully deterministic system with a zero initial condition, simply forcing our system with a spatially explicit unit impulse ($F(\boldsymbol{x}, t) = \delta(\boldsymbol{x}, t)$) is used to find the system's response function

$$T(\boldsymbol{x}, \boldsymbol{x}', t)|_{F(\boldsymbol{x}',t)=\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t)} = R(\boldsymbol{x}, \boldsymbol{x}', t), \tag{1.18}$$

where perturbations are applied at each spatial location, $\boldsymbol{x}'$, to determine their influence on a location of interest, $\boldsymbol{x}$; pulses can also be applied at alternate times, t', to determine how different time lags impact the response (e.g., seasonality), but we neglect these effects to simplify our analysis.

In this case, we can derive our response function directly without the need for an ensemble of simulations, but real systems are not this simple. Utilizing an impulse forcing naively in a chaotic system may lead to a single realization with behavior far from the expected forced response. For our nonlinear SDE, we use the Fluctuation Dissipation Theorem (FDT), to calculate a response function from an ensemble. Our system's response to a perturbation of magnitude ε is given by

$$R(\boldsymbol{x}, \boldsymbol{x}', t) = \frac{\langle T_\varepsilon(\boldsymbol{x}, t) - T_0(\boldsymbol{x}, t)\rangle}{|\varepsilon(\boldsymbol{x}')|}, \tag{1.19}$$

where $T_0(\boldsymbol{x}, t)$ and $T_\varepsilon(\boldsymbol{x}, t)$ correspond to unperturbed and perturbed initial condition ensembles, respectively. More detail on this expression can be found in Marconi et al. (2008)[182].

[182] Marconi et al., 'Fluctuation–dissipation: Response theory in statistical physics', *Physics Reports*, 2008

With this definition, we estimate the response function through the Fluctuation Dissipation Theorem by first spinning up a simulation to get a steady state distribution from which we draw an ensemble of initial conditions, $T_0(\boldsymbol{x}, t)$. We then create a copy of the initial condition ensemble with an additional small perturbation, ε, applied to each member, $T_\varepsilon(t)$, and simulate every member from both ensembles for a scenario of interest. Applying Equation 1.19 then gives us the response function, which we can use to emulate a variable of interest by convolving it with a forcing from a new scenario (Equation 1.12).

Both the stochastic and deterministic approaches only yield an accurate estimate of the true response function when the system is perturbed from a quasi-equilibrium rather than a transient state. For climate models, this is typically done with step change CO_2 experiments after a spin-up period. This method is common in the literature around climate response functions and linear response theory[32,93,139], though methods from the former two citations have not been applied to climate emulation and the latter does not reference formal response theory. Repeating this perturbation exercise at multiple background climate states can produce state-dependent response functions, but it is prohibitively expensive in practice.

[32] Lembo, Lucarini, and Ragone, 2020; [93] Lucarini, Ragone, and Lunkeit, 2017; [139] Freese et al., 2024

Analogously to our discussion of using the Koopman vs. Fokker-Planck operator, there also exists an extension of the FDT to probability distributions. This relationship is given by

$$R(\boldsymbol{x}, \boldsymbol{x}', t) = -\langle T(\boldsymbol{x}, t) s(T(\boldsymbol{x}', 0)) \rangle, \tag{1.20}$$

where $s(w) = \frac{\partial}{\partial w} \ln p(w)$ is the score function of the steady-state distribution and encodes how a small perturbation alters the system's dynamics; more details can be found in[140].

[140] Giorgini et al., 2024

The score function captures the direction a distribution shifts in response to a perturbation, and correlating it with a climate variable explains how the expectation of that variable shifts. Appendix A.1.5 outlines the link between this approach and the Fokker-Planck operator. Analytical expressions for the score function are unavailable for most systems, necessitating machine learning techniques to learn it. This approach has achieved high skill in representing the response function for several systems[140], though it has not yet been applied to the full climate system. We do not explore it further in this work because of the machine learning infrastructure required to implement it.

[140] Giorgini et al., 2024

The FDT faces accessibility issues in practice. First, there are high costs associated with this technique: a large ensemble of ESM runs is often prohibitively expensive. Second, there are also some configurations we simply cannot access: formal response theory assumes perturbations can be applied in a straightforward manner, which is not always the case. Because response functions are defined as a mapping from some perturbed input variable (e.g., CO_2 or radiative forcing) to an output variable of interest (e.g., temperature or precipitation), applying the FDT requires the ability to manually perturb a variable. Climate models may not be configured to accommodate e.g., radiative forcing as an input. The FDT therefore cannot be applied to derive radiative forcing response functions, though this is possible through other methods[91].

[91] Womack et al., 2025

Method III: Deconvolution. Without access to the true system to run specific perturbation experiments to find $R(\boldsymbol{x}, \boldsymbol{x}', t)$, data-driven approaches can estimate it. Deconvolution has been used to calculate response functions in the climate emulation context to derive spatially explicit response functions mapping effective radiative forcing to temperature[91]. It implicitly approximates the Koopman operator by deriving response functions that nominally correspond to Equation 1.12. To derive the deconvolution algorithm, we assume the data we have (e.g., annual temperature anomaly) are taken from an ensemble average of a general scenario. We begin from the FDT (Equation 1.12), assuming that our experiment begins from a quasi-equilibrium initial condition

[91] Womack et al., 2025

$$T(\boldsymbol{x}, t) = \int_0^t R(\boldsymbol{x}, t')F(t - t')\, dt'. \tag{1.21}$$

Treating this expression discretely, we rewrite it as a matrix expression and invert to solve for $R(\boldsymbol{x}, t)$ from any general scenario

$$\mathbf{R} = \frac{\mathbf{F}^{-1}\mathbf{T}}{\Delta t}, \tag{1.22}$$

where $\mathbf{F}$ is a lower-triangular matrix with $F_{t=0}$ along the diagonal, $F_{t=1}$ on the first off-diagonal, and so on (a Toeplitz matrix), and $\mathbf{T}$ is a matrix of temperature values with rows corresponding to the time dimension and columns corresponding to the spatial dimension. A more in-depth exploration of this process can be found in Womack et al. (2025)[91]. As written here, deconvolution aggregates spatial interactions (i.e., does not include an $\boldsymbol{x}'$ term), cutting down on data requirements. Extensions of this procedure can account for spatial interactions, though they require additional experiments with varying spatial forcings.

[91] Womack et al., 'Rapid Emulation of Spatially Resolved Temperature Response to Effective Radiative Forcing', *Journal of Advances in Modeling Earth Systems*, 2025

In practice, noisy data require us to apply regularization to Equation 1.22 to ensure matrix stability. We instead solve

$$\min_{\mathbf{R}} \|\mathbf{RF} - \mathbf{T}\|^2 + \alpha\|\mathbf{R}\|^2, \tag{1.23}$$

where α is the hyperparameter denoting the strength of our ridge regression. This simple ridge regression is equivalent to placing a Gaussian prior on the response function and assuming that the simulated temperature data we collect are corrupted by Gaussian noise. We discuss the rationale of Gaussian noise further in Appendix A.2 and outline our approach to tune the hyperparameter α through *maximum a posteriori* optimization.

Deconvolution can be applied to any general scenario that begins from a quasi-equilibrium initial condition. However, since we require an explicit matrix inverse to perform deconvolution, it is sensitive to the frequency spectrum of the forcing data. If the eigenvalues of the matrix $\mathbf{F}$ are very small (corresponding to near-zero frequencies) or the system is very noisy (corresponding to large differences in magnitudes between frequencies), the matrix becomes ill-conditioned, leading to an unstable response function. To illustrate these challenges, an explicit frequency-based derivation is included in Appendix A.1.2. In practice, we regularize the system to avoid these issues (see Appendix A.2 for details).

Method IV: Modal Fitting. Modal fitting is another data-driven technique to calculate response functions that retains some physical interpretability by explicitly representing the climate's response to a forcing as a series of decaying exponentials. The decay rates then represent the various timescales of the climate system (e.g., shallow vs. deep ocean heat uptake) and the modes represent how those timescales interact spatially. It has been used for tasks such as estimating effective radiative forcing and recently for climate emulation[89,90,92].

[89] Fredriksen, Rugenstein, and Graversen, 2021; [90] Fredriksen et al., 2023; [92] Sandstad et al., 2025

To connect this approach to our framework, we begin from the same set of assumptions as deconvolution, but make the additional assumption that our response function is exactly a decaying exponential; in this case, our response function is exactly a Green's function as described in Appendix A.1.3. We start from a restatement of Koopman response function definition (Equation 1.12)

$$G(\boldsymbol{x}, \boldsymbol{x}', t) = e^{\mathscr{L}(\boldsymbol{x},\boldsymbol{x}')t}, \tag{1.24}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ track spatial interactions as before. We assume we can represent the Koopman operator with a finite, linear operator, $\mathscr{L}$ (Appendix A.1.4).

We then diagonalize the matrix $\mathscr{L}$ though an eigenvalue decomposition, giving

$$G(\boldsymbol{x}, \boldsymbol{x}', t) = e^{V(\boldsymbol{x},n)\Lambda(n,n)V^{-1}(n,\boldsymbol{x}')t}, \tag{1.25}$$

where $\Lambda(n, n)$ and $V(\boldsymbol{x}, n)$ are matrices containing the system's eigenvalues and eigenvectors, respectively, and n is the mode number. Since the matrix exponential respects similarity transformations, we rewrite this exactly as the summation

$$G(\boldsymbol{x}, \boldsymbol{x}', t) = \sum_{i=1}^{k} v(\boldsymbol{x}, n_i) e^{\lambda_i t} v^{-1}(\boldsymbol{x}', n_i), \tag{1.26}$$

where k is equal to the total number of eigenvalues in the system. In the case of a climate model, the dimension of k is equivalent to the number of spatial dimensions. This may be much higher than the true number of modes that are significant in determining, e.g., the temperature response of the system. Instead of the explicit form above, we typically see an alternate implementation, such as that in Fredriksen, Rugenstein, and Graversen (2021)[89], Fredriksen et al. (2023)[90] and Sandstad et al. (2025)[92]. These show that one can fit an alternate form given simply by

$$G(t) \approx R(t) = \sum_{i=1}^{3} \alpha_i e^{\lambda_i t}, \tag{1.27}$$

where using just three timescales (inter-annual, inter-decadal, and inter-centennial) is sufficient to represent the global mean behavior of the climate system; these methods specify a range/initial guess of timescales to initialize the optimization routine. Here, the scalar coefficients α_i are derived by integrating the spatial interaction of the i-th mode, defined as $\alpha_i(\boldsymbol{x}, \boldsymbol{x}') = v(\boldsymbol{x}, n_i)v^{-1}(\boldsymbol{x}', n_i)$, over the relevant spatial domain. As we are implementing this at a grid cell level, we opt for a hybrid approach, given by

$$R_i(t) = \sum_{j=1}^{3} \alpha_{i,j} e^{\lambda_i t}, \tag{1.28}$$

where i indicates the grid cell/region of interest, and j denotes the

[89] Fredriksen, Rugenstein, and Graversen, 'Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response', *Journal of Geophysical Research: Atmospheres*, 2021

[90] Fredriksen et al., '21st Century Scenario Forcing Increases More for CMIP6 Than CMIP5 Models', *Geophysical Research Letters*, 2023

[92] Sandstad et al., 'METEORv1.0.1: a novel framework for emulating multi-timescale regional climate responses', *Geoscientific Model Development*, 2025

contribution from each timescale in a given region. We use the three timescales given above as the initial guess for each lambda, along with an initial guess for $\alpha_{i,j}\ \forall\ i = j$, assuming that one mode is dominant for each box.

We thus need to solve

$$\tilde{T}(\alpha_{i,j}, \lambda_i) = \int_{-\infty}^{t} R_i(s) F(t-s)\, ds, \tag{1.29}$$

$$\min_{\alpha_{i,j}, \lambda_i} \|T - \tilde{T}(\alpha_{i,j}, \lambda_i)\|^2. \tag{1.30}$$

For climate applications, the decay rates (λ_i) can span several orders of magnitude, which are difficult for the optimizer to identify, even with normalization. This is exacerbated by the need to solve for the eigenvectors simultaneously, which are also likely to have values that span several orders of magnitude; using more sophisticated optimization techniques than we apply in our test case could potentially resolve this issue. When implementing this algorithm, we follow Fredriksen, Rugenstein, and Graversen (2021)[89], providing an initial guess of the correct order of magnitude to our optimizer.

[89] Fredriksen, Rugenstein, and Graversen, 'Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response', *Journal of Geophysical Research: Atmospheres*, 2021

Modal fitting has two major benefits. First, by truncating the leading modes, we reduce the dimensionality of the problem without the need for e.g., Empirical Orthogonal Functions (EOFs) or a Singular Value Decomposition (SVD). Second, we require all $\Re(\lambda_i) < 0$ (the real component of λ_i) to ensure response functions to decay to zero as $t \to \infty$, a requirement not imposed on e.g., deconvolution and DMD. Because it is a best-fit problem, it naturally damps noise, making it well suited to systems with strong internal variability. However, this method can also be sensitive to local minima, requiring multiple iterations or a stochastic fitting procedure to alleviate this issue. Fitting may also be expensive on fine grids, since the number of eigenpairs scales with grid size, though we may not require all eigenpairs to accurately emulate the system.

Operator-based emulation

The most general class of emulators are those that aim to directly approximate the Koopman operator. Every previous emulator can be thought of as a specific case of this general operator framework. Tebaldi et al. (2025)[64] do not include operator-based emulators in their classification, as they are not typically referred to explicitly as emulators. However, we classify them as such to facilitate communication across disciplines with similar prediction goals.

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

The most common data-driven approximations of the Koopman operator are Dynamic Mode Decomposition (DMD) and Extended DMD (EDMD)[165,166]. Schmid (2010)[165] developed DMD to extract dynamic information from fluid flows, and it has since been used to identify dominant modes of variability within the climate system, including El Niño–Southern Oscillation, North Atlantic Oscillation, and Pacific Decadal Oscillation[143–146,167,183]. Under specific conditions, DMD provides a finite-dimensional approximation of the Koopman operator[184]. EDMD expands this idea to approximate Koopman eigenvalues and eigenfunctions directly[166]. The bulk of the work surrounding EDMD is theoretical[185,186], as in practice it has several limitations that we outline later in this section.

[165] Schmid, 2010; [166] Williams, Kevrekidis, and Rowley, 2015

[165] Schmid, 'Dynamic mode decomposition of numerical and experimental data', *Journal of Fluid Mechanics*, 2010

[143] Kutz, Fu, and Brunton, 2016; [144] Gottwald and Gugole, 2020; [145] Navarra, Tribbia, and Klus, 2021; [146] Mankovich et al., 2025; [167] Navarra et al., 2024; [183] Franzke, Gugole, and Juricke, 2022

[184] Schmid, 2021

[166] Williams, Kevrekidis, and Rowley, 2015

[185] Haseli and Cortés, 2019; [186] Netto et al., 2021

Method V: Dynamic Mode Decomposition (DMD). DMD assumes that the climate response is linear in w with respect to an operator. If this is the true Koopman operator, this assumption holds by definition, provided it acts on the entire infinite space of statistical climate fields, $g(w)$. In practice, this leads to limitations based on how accurate the assumption of linearity is, which depends on the choice of variables; this approximation may hold better for a variable such as temperature, rather than precipitation. To derive DMD, we begin from Equation 1.9 applied to our variable of interest

$$\frac{\partial}{\partial t}T(\boldsymbol{x},t) = \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')T(\boldsymbol{x}',t) + F(\boldsymbol{x},t). \tag{1.31}$$

DMD assumes that we separate our data in discrete snapshots, $T(\boldsymbol{x},t_0)$, $T(\boldsymbol{x},t_1),\ldots,T(\boldsymbol{x},t_n)$, which we assume are linearly related

$$T_{n+1} = \mathscr{L}T_n + F_n, \tag{1.32}$$

where we have used the subscript n as shorthand for t_n and omitted the spatial dimension for conciseness. By discretizing, we are no longer solving for the exact Koopman operator (as in the previous case), which we now denote $\mathscr{L}$. This notation is standard in DMD literature. The traditional DMD algorithm assumes autonomous dynamics, omitting the forcing term. Equation 1.32 is referred to as DMD with control (DMDc)[187], and has only recently been studied in the climate context[146].

[187] Proctor, Brunton, and Kutz, 2016

[146] Mankovich et al., 2025

To implement DMD, we collect our snapshots into matrices and invert this system, solving for $\mathscr{L}$

$$\mathscr{L} = [\mathbf{T}_{n+1} - \mathbf{F}_n]\,\mathbf{T}_n^{+}, \tag{1.33}$$

where the superscript + denotes the Moore-Penrose pseudo-inverse of a matrix (required as it is unlikely **T** will be a square matrix) and **F** denotes a forcing matrix with the same dimension as our data; assuming well-mixed forcing means each row is identical in the forcing matrix. This is the simplest form of DMD, though in practice the Singular Value Decomposition (SVD) is often used to further reduce the dimensionality of the problem. This also increases the algorithm's robustness relative to real-world systems that are subject to noise[165].

[165] Schmid, 2010

This approach suffers mainly from its strong assumption of linear dynamics, which can break down for complex systems. Its success in identifying the dominant modes of variability in the climate suggests it may have utility as an explicit emulation technique[143,144,183]; future work will apply DMD to a full scale climate model to test this hypothesis. Unfortunately, DMD only provides a reliable estimate for the Koopman operator if it acts on a large set of statistical fields (more than simply the temperature anomaly when considering the full climate system). This is because the Koopman operator relies on an infinite-dimensional space of statistical fields to fully linearize nonlinear dynamics; a small, finite set of physical variables rarely forms a Koopman-invariant subspace[188]. Furthermore, DMD only functions properly if the dynamics governing the evolution of that quantity (or quantities) is linear, which is not the case in general. While the dynamics producing the base climate state are nonlinear, the success of methods such as pattern scaling suggest the dynamics of anomalies may be close to linear. DMD assumes all hidden variables are accounted for and the observed quantities fully describe the (linear) dynamics of our anomaly of interest. For example, the atmospheric temperature may be significantly influenced by heat uptake in the deep

[143] Kutz, Fu, and Brunton, 2016; [144] Gottwald and Gugole, 2020; [183] Franzke, Gugole, and Juricke, 2022

[188] Brunton et al., 2016

ocean, which, if it is not explicitly accounted for, will lead to errors when applying DMD. This motivates the need for a better algorithm for approximating the Koopman operator.

Method VI: Extended DMD (EDMD). As the baseline DMD algorithm is only able to approximate the Koopman operator in specific contexts, EDMD instead frames the problem such that we are deliberately trying to approximate the eigenvalues and eigenfunctions of the Koopman operator. This, ideally, leads to more reliable approximation than DMD and thus, a better emulator.

EDMD was introduced by Williams, Kevrekidis, and Rowley (2015)[166] as an explicit attempt to approximate the Koopman operator. The EDMD procedure involves projecting variables of interest into a higher dimensional space that has a richer representation of the system dynamics. As an example, we consider the problem of emulating precipitation anomaly using global mean temperature anomaly as the forcing. Precipitation may depend on the global mean temperature, $\overline{T}(t)$, but it also may depend on higher-order or nonlinear terms, such as $(\overline{T}(t))^2$, $\cos(\overline{T}(t))$, $\tanh(\overline{T}(t))$, etc. To implement EDMD, the user must select a set of basis functions, $\phi(\cdot)$, such as these, that provide a better representation of the system dynamics than in the purely linear DMD case. Typical choices of basis functions as described by the original EDMD manuscript are Hermite polynomials, radial basis functions, and discontinuous spectral elements[166].

[166] Williams, Kevrekidis, and Rowley, 'A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition', *Journal of Nonlinear Science*, 2015

[166] Williams, Kevrekidis, and Rowley, 2015

After choosing a set of basis functions, the EDMD problem statement is exactly the same as the original DMD algorithm. Solve for $\tilde{\mathcal{K}}$ from

$$\phi(T_{n+1}) = \tilde{\mathcal{K}}\phi(T_n) + \psi(F_n), \tag{1.34}$$

where $\psi(\cdot)$ is the basis chosen for the forcing, and can be the same or different than the forcing for the quantity of interest. We use $\tilde{\mathcal{K}}$ here as we are explicitly trying to approximate the Koopman operator. We ensure the basis includes the physical field of interest, e.g., $\phi(\mathbf{T}) = \left[\mathbf{T},\ \mathbf{T}^2,\ \mathbf{T}^3,\ \ldots\right]$, where the first entry is the physical field. As in the case with DMD, we solve this as

$$\tilde{\mathcal{K}} = \left[\phi(\mathbf{T}_{n+1}) - \psi(\mathbf{F}_n)\right]\phi^{+}(\mathbf{T}_n), \tag{1.35}$$

which we can use an SVD to solve more efficiently and reduce the influence of noise on the system. When applying this method, we first use Equation 1.34 with an appropriate initial condition to emulate the solution in our high-order basis. We then must project our solution back into physical space. Since we chose our basis to include the original physical coordinate, this is done by truncating the emulator output and keeping only the entries corresponding to $\mathbf{T}$.

This method has seldom been applied to climate problems[167], likely due to the limitations acknowledged in Navarra, Tribbia, and Klus (2021)[145], particularly the dimensionality of the problem. For a full climate model, DMD requires a matrix solve of dimension $(N_{\text{lat}} \times N_{\text{lon}})^2$ for a single variable, which is extremely costly. In the case of EDMD, this dimension grows with every basis function used. To accurately represent the Koopman operator for the climate system, we potentially require many more variables and many basis functions, causing the problem to rapidly increase in complexity, though this may be alleviated by emulating EOFs rather than gridded data. As with DMD, EDMD implicitly assumes no hidden variables, though the choice of basis function can help alleviate this issue; e.g., if the hidden variables are higher-order terms, EDMD

[167] Navarra et al., 2024

[145] Navarra, Tribbia, and Klus, 'Estimation of Koopman Transfer Operators for the Equatorial Pacific SST', *Journal of the Atmospheric Sciences*, 2021

may be able to represent them accurately. The selection of basis functions typically requires some experimentation though, as it can be difficult to predict which set of functions will be best suited for a given application; exploiting physical relationships such as the logarithmic relationship between CO_2 concentration and temperature may help alleviate this issue, however. More work is required to fully characterize the utility of EDMD for the climate system.

1.2 Experimental overview

Here we outline a set of experiments which reproduce the sources of error seen in Fig. 1.1, using them to evaluate the emulation techniques introduced in Sect. 1.1.3. We outline a climate box model with a simple local energy balance ODE in Sect. 1.2.1 and Sect. 1.2.2, followed by a nonlinear, cubic Lorenz system in Sect. 1.2.3. Experiments using these two simple models highlight the following potential sources of error: (1) memory effects, Fig. 1.1 (a); (2) hidden variables, Fig. 1.1 (b); (3) noise, Fig. 1.1 (c); (4) weak nonlinearities, Fig. 1.1 (d). We then describe forcing scenarios applied to each system in Sect. 1.2.4.

1.2.1 Experiments 1 and 2: Climate Box Model

A classical box model is a standard, easily interpretable model for temperature evolution. We use this idealized box model as it is the simplest system that includes the pattern effect and it is not necessarily meant to replicate CMIP experiments. We assume the form of this model is given by a simple local energy balance

$$C(\boldsymbol{x})\frac{\partial T(\boldsymbol{x},t)}{\partial t} = \lambda(\boldsymbol{x})T(\boldsymbol{x},t) + R(\boldsymbol{x},t) + \nabla\cdot\mathbf{F}(\boldsymbol{x},t), \tag{1.36}$$

similar to Armour, Bitz, and Roe (2013)[55] and Giani et al. (2025)[56]. $C(\boldsymbol{x})$ is the local effective heat capacity, $T(\boldsymbol{x},t)$ is the local temperature anomaly, $\lambda(\boldsymbol{x})$ is the local feedback parameter, $R(\boldsymbol{x},t)$ is the forcing function, and $\nabla\cdot\mathbf{F}(\boldsymbol{x},t)$ is the anomaly in heat flux divergence; parameters for this model are listed in Table 1.2. Furthermore, we assume that the forcing function can be linearly decomposed as a constant-amplitude spatial pattern and a variable time series: $R(\boldsymbol{x},t) = r(\boldsymbol{x})R(t)$.

[55] Armour, Bitz, and Roe, 'Time-Varying Climate Sensitivity from Regional Feedbacks', 2013

[56] Giani et al., 'Origin and Limits of Invariant Warming Patterns in Climate Models', *Journal of Climate*, 2025

We consider two configurations for our box model. The first corresponds to a horizontally coupled three box system representing atmospheric boxes over land, low-latitude ocean, and high-latitude ocean. In the continuous system, the heat flux vector is parameterized by $\mathbf{F}(\boldsymbol{x},t) = -k(\boldsymbol{x})\nabla T(\boldsymbol{x},t)$. To transition to our discrete box model, we replace the continuous divergence term with the net heat exchange between adjacent boxes. Applying a 1D forward difference spatial discretization and absorbing the spatial step into our constant, the net flux term for box i is modeled as $-k\left(T_{i+1}(t) - T_i(t)\right)$, where k represents an effective, constant inter-box heat transfer coefficient. We assume uniform forcing into each box, and use this configuration for experiments one and three (memory effects and noise; noise details can be found in Sect. 1.2.2). The second configuration corresponds to a vertically coupled two box system representing the atmosphere and the ocean; this has the same form as the previous case, with the caveat that there is no forcing applied into the oceanic box. We use this configuration for experiment two (hidden

Table 1.2: Parameters for the three box model, adapted from Giani et al. (2025)[56]. The heat capacity of each box is given in terms of the effective water depth, $h(\boldsymbol{x})$: $C(\boldsymbol{x}) = \rho_w c_w h(\boldsymbol{x})$, where ρ_w and c_w are the density and specific heat capacity of water, respectively. *Land*, *Low*, and *High* refer to atmospheric boxes over land, low-latitude ocean, and high-latitude ocean, respectively.

Parameter	Symbol	*Land*	*Low*	*High*
Effective Water Depth (m)	$h(\boldsymbol{x})$	5	150	1500
Local Feedback ($Wm^{-2}K^{-1}$)	$\lambda(\boldsymbol{x})$	-0.86	-2.0	-0.67

variables). We begin this system from a zero initial condition, aiming to simulate the temperature anomaly, rather than the absolute temperature.

1.2.2 Experiment 3: Noisy Box Model

As the default configuration for our box model is purely deterministic, we add a stochastic noise term to the forcing to replicate the impact of inter-annual variability on the real climate system. To ensure the impact of this variability is similar to that of the true system, we use CMIP6 *piControl* experiments to estimate the magnitude of the variability. Namely, we compute the standard deviations of piControl runs for three climate models (ACCESS-ESM1-5, MIROC6, MPI-ESM2-LR) and set the magnitude of the variability as the multi-model average $\sigma = 0.117K$[189–191].

[189] Tatebe and Watanabe, 2018; [190] Dix et al., 2019; [191] Wieners et al., 2019

1.2.3 Experiment 4: Cubic Lorenz System

As the previous experiments are all defined by an operator which is linear in the quantity of interest, we additionally implement a weakly nonlinear, cubic Lorenz system. This provides a representation of the atmosphere that includes chaos, allowing us to test the limits of these emulation techniques. In the standard Lorenz equations that represent a simplified model of atmospheric convection[192], the steady state is a linear function of ρ, and the mean heat flux ($\langle XY \rangle = \langle Z \rangle$) is very nearly linear[193]. We modify the system to the cubic form shown below to illustrate another failure mode of simple pattern scaling: the quasi-equilibrium value may not be a linear function of the forcing.

[192] Lorenz, 1963

[193] Souza and Doering, 2015

The cubic Lorenz equations are defined by the system

$$\frac{\partial}{\partial t} X = \sigma(Y - X), \tag{1.37}$$

$$\frac{\partial}{\partial t} Y = -(Z + \alpha Z^3)X + \rho(t)X - Y, \tag{1.38}$$

$$\frac{\partial}{\partial t} Z = XY - \beta Z, \tag{1.39}$$

with $\alpha = 1/1000$. Due to the system's invariance under the spatial transformation $(X, Y, Z) \to (-X, -Y, Z)$, the steady-state mean of both X and Y are zero, while the steady-state behavior of $\langle Z \rangle$ is determined by $\rho(t)$. Values for $\rho(t)$ are chosen such that nonlinearities are weak, as all linear methods are expected to break down in the presence of strong nonlinearities. These vary between experiments and are outlined in Table 1.4. We initialize this system through an initial condition ensemble starting from $\rho(t) = 28$ with white noise applied to perturb the starting positions of each ensemble member.

1.2.4 Scenarios

We consider four scenarios of interest for both the box model and cubic Lorenz system, focusing on scenarios which have CMIP analogues: (1) *Abrupt*, an abrupt increase in forcing, (2) *High Emissions*, an exponential increase in forcing, (3) *Plateau*, an exponential increase in forcing that levels off, and (4) *Overshoot*, a forcing that sharply increases and decreases. Descriptions of each scenario are given in Table 1.3. Figure 1.3 shows ODE-integrated solutions for each scenario in each experiment, and descriptions of experimental parameters can be found in Tables 1.4 and 1.5.

Table 1.3: Conceptual overview of forcing scenarios considered in this work. These scenarios are used in all experiments outlined in Sect. 1.2, and lists of experiment-specific parameters for each scenario can be found in Tables 1.4 and 1.5.

Scenario	Short Description
Abrupt	An abrupt doubling of CO_2 concentration; corresponds roughly to the *Abrupt2xCO2* CMIP experiment.
High Emissions	An exponential increase of CO_2 concentration in time; corresponds roughly to *SSP585*.
Plateau	An increase in CO_2 concentration in time that follows a hyperbolic tangent, increasing exponentially and then tapering off; corresponds roughly to *SSP245*.
Overshoot	An increase in CO_2 concentration in time that follows a Gaussian profile, increasing and decreasingly rapidly; inspired by *SSP119*, but decreases more quickly.

1.2.5 Evaluation

To evaluate each emulation technique, we utilize Normalized Root Mean Square Error (NRMSE, Equation 1.40) given as a percentage, as our primary evaluation metric:

$$\text{NRMSE} = \frac{100}{\overline{g}(w_k)} \sqrt{\frac{\sum_{k=1}^{N_{\text{years}}} \left(g(w_k) - \hat{g}(w_k)\right)^2}{N_{\text{years}}}}. \tag{1.40}$$

$\overline{g}(w_k)$ indicates the mean of our quantity of interest over the period error is calculated over. We calculate NRMSE with respect to the entire time series. To compare performance across training datasets, we train each emulator on one scenario at a time, testing against the others which are held out from the training (e.g., train on *Abrupt* and test on *High Emissions*).

We implement an alternate protocol for the cubic Lorenz system as there is no ground-truth to compare with due to chaos. Instead, we compare the skill of each emulator when training on only a subset of the ensemble members for that experiment. For example, given n_{ensemble} ensemble members for a given experiment, we construct a subset of n ensemble members without replacement, where $n = 1 : n_{\text{ensemble}} - 1$, and train our emulator from that subset. We then test the emulator's skill in emulating the mean response given the ensemble average forcing. We repeat this subsampling exercise 10 times, recording the average performance over those trials. For the noisy three box model, we use the same protocol, additionally presenting the ground truth of emulating the noiseless three box model.

Scenario	Functional Form
Abrupt	$F(t) = F_{abr} H(t)$
	$\rho(t) = \rho_{0,abr} + \rho_{1,abr} \tanh(t - \eta_{abr})$
High Emissions	$F(t) = F_{high} \exp(t/\tau_{high})$
	$\rho(t) = \rho_{0,high} + \rho_{1,high} \exp(t/\eta_{high})$
Plateau	$F(t) = F_{plat} + F_{plat} \tanh(\omega_{plat}(t - \tau_{plat}))$
	$\rho(t) = \rho_{0,plat} + \rho_{1,plat} \tanh(\omega_{plat}(t - \tau_{plat}))$
Overshoot	$F(t) = F_{over} \exp(-(t - \tau_{over})^2/(2\sigma^2))$
	$\rho(t) = \rho_{0,over} + \rho_{1,over} \exp(-(t - \eta_{over})^2/(2\sigma^2))$

Table 1.4: Forcing scenarios for each experiment, with the upper half of each row corresponding the box model and the lower half of each row corresponding to the cubic Lorenz system. Parameters for the box model experiments are based on Giani et al. (2025)[56] and Armour et al. (2013)[55] and parameters for the cubic Lorenz system are chosen such that the system exhibits weakly nonlinear behavior. $H(t)$ is the Heaviside step function, and parameters for these scenarios are listed in Table 1.5.

Box Model		Cubic Lorenz System	
Parameter	Value	Parameter	Value
-	-	ρ_0	[45, 28, 40, 28]
F_{abr}	$3.7\ \mathrm{W\,m^{-2}}$	η_{abr}	10
		$\rho_{1,abr}$	17
F_{high}	$\frac{8.5\ \mathrm{W\,m^{-2}}}{\exp(\tau_f/\tau_{high})}$	$\rho_{1,high}$	$\frac{30}{\exp(\eta_f/\eta_{high})}$
τ_f	250 yr	η_f	250
τ_{high}	50 yr	η_{high}	50
$F_{0,plat}$	$2.25\ \mathrm{W\,m^{-2}}$	$\rho_{1,plat}$	$\frac{12}{\tanh(5)}$
$F_{1,plat}$	$\frac{2.25\ \mathrm{W\,m^{-2}}}{\tanh(\omega_{plat}\tau_{plat})}$	τ_{plat}	150
ω_{plat}	$1/50\ \mathrm{yr^{-1}}$	ω_{plat}	1/50
F_{over}	$4\ \mathrm{W\,m^{-2}}$	$\rho_{1,over}$	30
τ_{over}	200 yr	η_{over}	200
σ_{over}	42.47	σ	50

Table 1.5: Scenario parameters used for the experiments in this study. Values for ρ_0 are listed in the order *Abrupt*, *High Emissions*, *Plateau*, and *Overshoot*. Box-model parameters have physical units to output temperature; the cubic-Lorenz parameters are dimensionless.

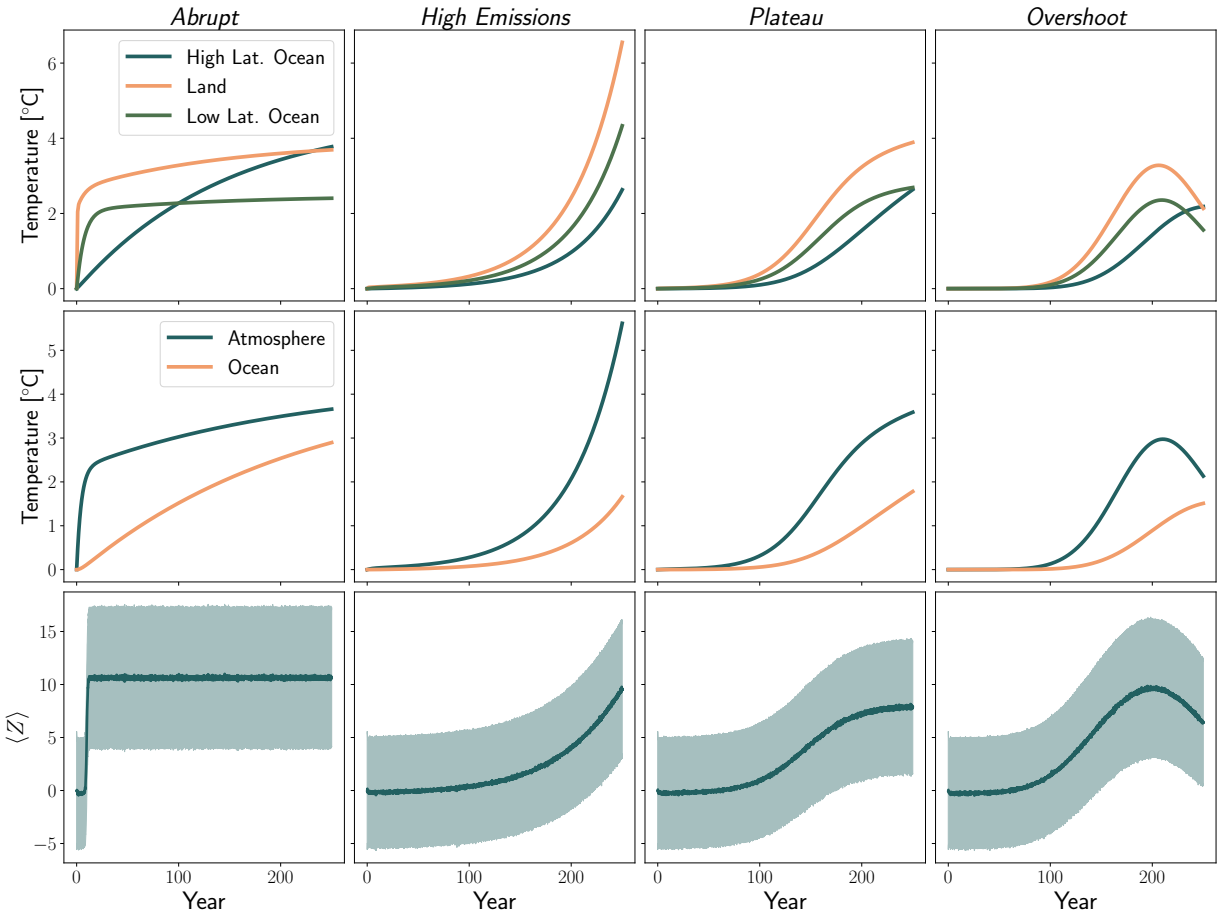

Figure 1.3: ODE-integrated solutions for the three box model (top), two box model (middle), and cubic Lorenz system (bottom) for the (from left to right) *Abrupt*, *High Emissions*, *Plateau*, and *Overshoot* scenarios. $D = 0.55$ [$\mathrm{Wm^{-2}K^{-1}}$] for the three box experiment and $D = 0.7$ [$\mathrm{Wm^{-2}K^{-1}}$] for the two box experiment. For the cubic Lorenz problem we show the mean value of Z over 5,000 ensemble members as a line, and the shaded region indicates its standard deviation. Values shown are anomalies relative to a baseline of $T = 0$ (experiments one through three) or $\rho = 28$ (experiment four).

1.3 Results

Section 1.3.1 presents a summary of results across each of the emulation techniques outlined in Sect. 1.1.3 when emulating the simplified climate systems presented in Sect. 1.2, with subsequent sections highlighting key results from individual experiments. Section 1.3.2 contains the results for the three box model with significant memory effects (Fig. 1.1 (a)); the three boxes represent atmospheric boxes over the land, low-latitude ocean and high-latitude ocean. We then report emulator performance on the restricted two box model in Sect. 1.3.3. In this case we highlight the issue of hidden variables (Fig. 1.1 (b)) by only giving the emulators access to the temperature anomaly in only one of the two boxes during training; the two boxes represent an atmospheric and oceanic box (forcing only into the atmosphere). This is followed by a version of the three box model with a stochastic forcing to test the robustness of each method to noise (Fig. 1.1 (c)). Finally, we showcase results for the nonlinear, cubic Lorenz system in Sect. 1.3.5 (Fig. 1.1 (d)), which tests emulator performance in the presence of chaos and weak nonlinearities. In the case of models with multiple regions (boxes), we present only a single evaluation score, as relative performance across boxes was consistent for all cases analyzed.

1.3.1 Overall emulator performance

Figure 1.4 summarizes emulator performance in terms of Normalized Root Mean Square Error (NRMSE) across all four experiments. For each experiment, there are four possible train/test scenarios (*Abrupt*, *High Emissions*, *Plateau*, and *Overshoot)*. We test on one scenario and train against the remaining three, showing median NRMSE over all train/test combinations. For experiments two and four, the pattern scaling emulator is trained to map forcing to quantity of interest, as these experiments do not have a global mean temperature equivalent. Results for deconvolution are shown using the regularization presented in Appendix A.2. Error values are calculated with a constant 40 ensemble members for experiment three and 4,000 ensemble members for experiment four.

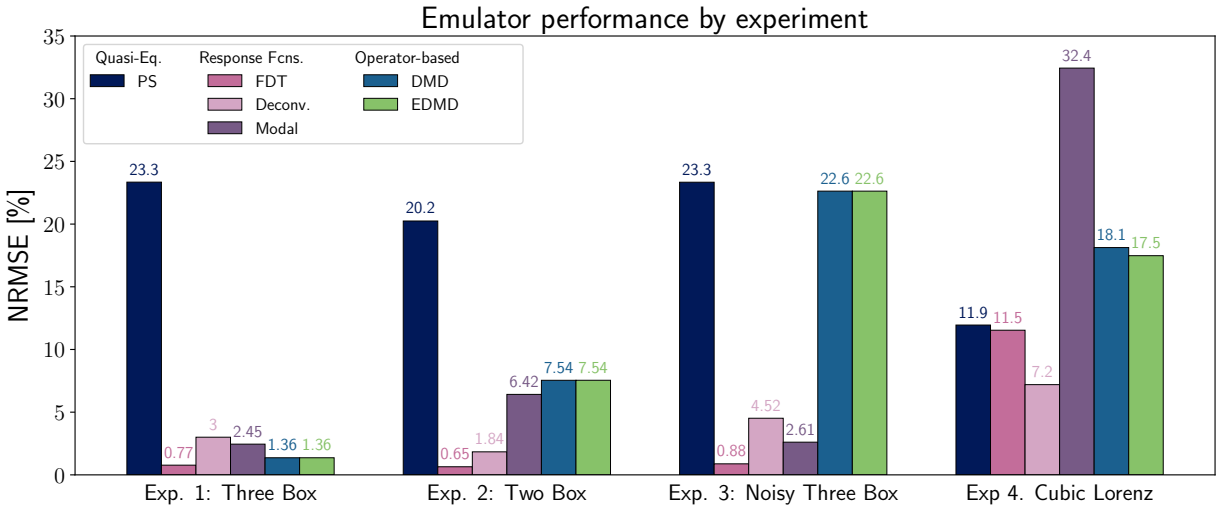

Figure 1.4: Summary of emulator performance over all experiments considered in this work. For each experiment, there are four scenarios. We show the median NRMSE value across all scenario train and test combinations, excluding the trivial case of training and testing on the same dataset. Error values are calculated with 40 ensemble members for experiment three and 4,000 ensemble members for experiment four. Emulator abbreviations are as follows: PS - Pattern Scaling, FDT - Fluctuation Dissipation Theorem, Deconv. - Deconvolution, Modal - Modal Fitting, DMD - Dynamic Mode Decomposition, EDMD - Extended DMD.

Response function based emulators (the FDT, deconvolution, and modal fitting methods) generally outperform other approaches, demonstrating consistently lower NRMSE across most experiments. The FDT is particularly reliable relative to all other methods, yielding consistently low errors across all four test cases, indicating its robustness regardless of scenario; while it has higher error in the cubic Lorenz case, this is primarily a function of ensemble size (see Sect. 1.3.5). As FDT response functions are, in principle, equation-driven rather than data-driven, they provide the perfect solution given a linear system (experiments one through three) or enough realizations (experiment four). Deconvolution similarly performs well across all experiments, while modal fitting has high performance in experiments one, two, and three; both of these methods exhibit higher errors in experiment four. For deconvolution, this is due to its sensitivity to noise as discussed in Sect. 1.1.3, while modal fitting suffers because of an inability to reliably separate timescales and the need for an accurate initialization for its unknown parameters, which we discuss in Sect. 1.3.4.

In contrast, pattern scaling consistently underperforms, exhibiting the highest error in all experiments except for the cubic Lorenz case. This is most likely due to the presence of strong memory effects in the box models, which pattern scaling cannot capture by definition. DMD and EDMD outperform pattern scaling in experiments one and two, but exhibit much more variable performance in experiments three and four. For the first three experiments, DMD and EDMD produce identical results. This is because the models in these experiments are purely linear, and the use of any higher-order basis for EDMD leads to a drop in skill. These methods struggle with the noisy three box model, and more in-depth results can be found in Sect. 1.3.4. While theory suggests DMD/EDMD would not be well-suited for the restricted two box problem due to the presence of hidden variables, they outperform pattern scaling in practice. This is likely due to the simplicity of the problem, and more complex dependencies on hidden variables would likely lead to further decreases in skill. The main advantage of EDMD over DMD begins to become apparent in the cubic Lorenz experiment, where moving to a third-order Hermite polynomial basis allows it to slightly outperform its linear counterpart, though the variability in the system (Fig. 1.3) is a greater magnitude than this improvement in skill.

1.3.2 Experiment 1: Three Box Model

The three box model experiment is meant to benchmark the baseline performance of each technique in the presence of strong memory effects (Fig. 1.1 (a)). Figure 1.5 summarizes the results of four emulation techniques (pattern scaling, deconvolution, modal fitting, and DMD) when trained and tested on different scenario combinations, while Fig. 1.6 compares the true (ODE-integrated) solution to that obtained using the Fluctuation Dissipation Theorem.

Pattern scaling (Method I) consistently underperforms relative to the other techniques presented in this section, exhibiting the highest NRMSE values for all train/test combinations. It fails across almost every scenario due to the influence of long timescales on the global mean temperature (strong memory effects). This experiment highlights pattern scaling's brittleness when key assumptions, such as exponential forcing[56], are violated. These assumptions are consistent in most ScenarioMIP experiments however, leading to higher performance in practice relative to this simple example[80].

[56] Giani et al., 2025

[80] Wells et al., 2023

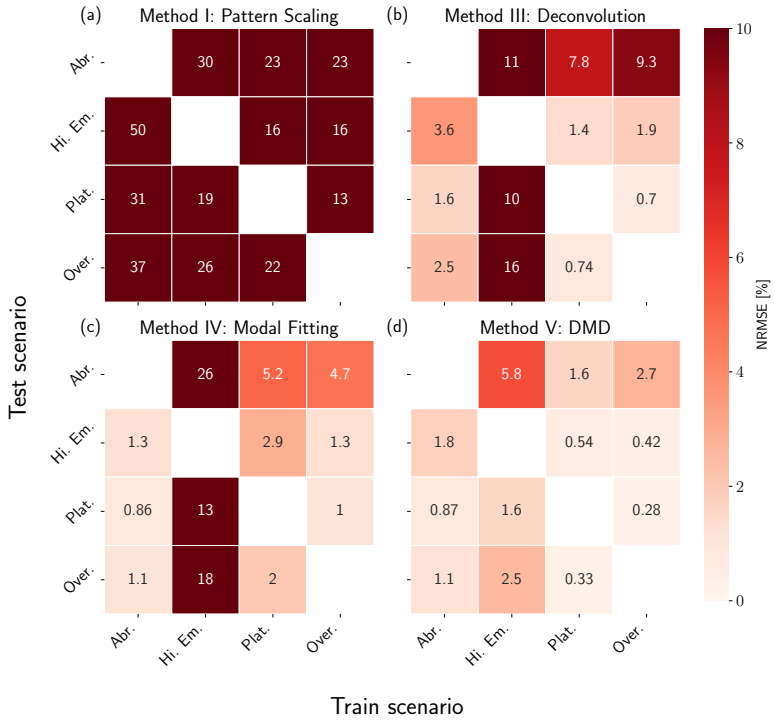

Figure 1.5: NRMSE heatmaps for pattern scaling (a), deconvolution (b), modal fitting (c), and DMD (d) emulators trained and tested against the three box model. Results are shown in percentages, where lighter values correspond to lower error (higher performance) and darker values correspond to higher error (lower performance). Scenarios used for training are shown on the x-axis, while scenarios used for testing are shown on the y-axis. We do not include results for training and testing on the same dataset.

Applying deconvolution (Method III) leads to much higher performance than pattern scaling when trained on either *Abrupt*, *Plateau*, or *Overshoot*, but sees a drop in performance when trained on *High Emissions*. This is because the true solution is an eigenfunction of the forcing (i.e., both the temperature response and forcing are exponentials), so the system is effectively characterized by a single timescale, that of the forcing. Deconvolution loses skill due to difficulties identifying all the timescales in the system, leading to extrapolation errors when training on this scenario. When trained on either *Plateau* or *Overshoot*, we see errors in emulating *Abrupt*, meaning that the emulator has not learned the true system response despite relatively high performance in emulating the other scenarios. This is due to ill-conditioning of the $\mathbf{F}$ matrix in these scenarios, leading to a response function that overfits these data; we discuss the limitations of training deconvolution with these scenarios further in Sect. 1.4.

Modal fitting (Method IV) exhibits two interesting properties: (1) training on *High Emissions* leads to poor extrapolative capability and (2) training on *Abrupt* leads to the highest performance overall. The first is also caused by the solution being an eigenfunction of the forcing. It is difficult for the optimization routine to determine the correct timescales, even when initialized near the true values. This is true to a lesser degree in *Plateau* and *Overshoot*, which also do not display clean separation of time scales like *Abrupt*.

DMD (Method V) is able to capture all relevant timescales and interactions regardless of the scenario, with a maximum of 5.8% NRMSE across all train/test combinations; this level of error results from training on *High Emissions* and testing on *Abrupt*, as was the case with the modal fitting emulator. The method's high skill here is due to the governing dynamics being purely linear and there being no hidden variables, meaning all assumptions for applying DMD are accurate. Results for EDMD (Method VI) are omitted from this section as they are identical to DMD.

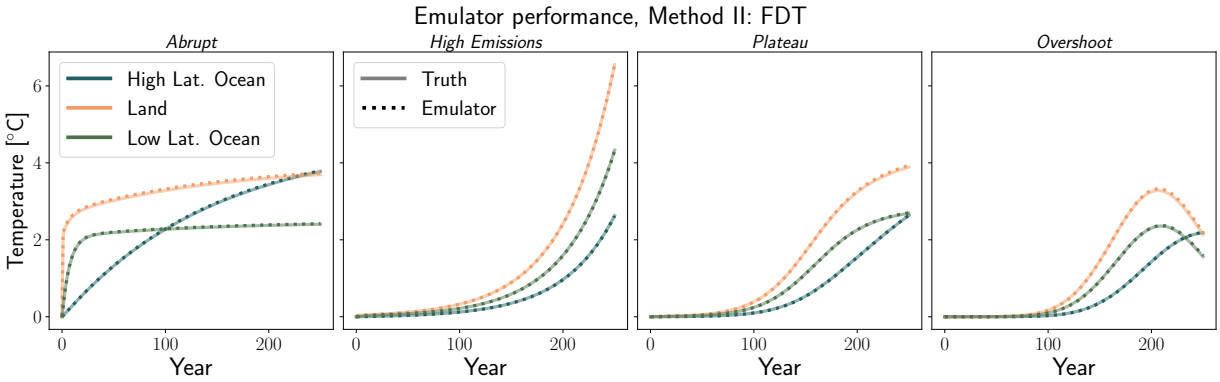

Figure 1.6: Fluctuation Dissipation Theorem emulator performance for three box model scenarios. The solid, lighter line shows ground truth (ODE-integrated) solution, while the dotted, darker line shows emulated solution. The high performance of the FDT results in the emulated and ground-truth curves overlapping closely.

The Fluctuation Dissipation Theorem (Method II) has consistently high performance across all scenarios considered, with NRMSE values of 0.80%, 0.50%, 0.75%, and 1.29% for the four scenarios shown in Fig. 1.6 (NRMSE values given by scenario from left to right). These values are lower than any other technique on average. These errors are due to the integration scheme with which we derive the FDT response function, as we only use a first-order integrator. Since it requires us to simulate two scenarios (one perturbed and one unperturbed), error can accumulate between these simulations; decreasing the integrator time step or using a higher-order integrator (not shown) increases accuracy for this method. Despite this, the FDT gives us, up to the precision of our integrator, the system's true response function, which is a major advantage compared to the other techniques which may or may not provide a physically interpretable solution. The full implementation of the FDT requires a spatially explicit response matrix with multiple perturbation runs, but for a more even comparison to the other techniques, we only consider the well-mixed case here.

1.3.3 Experiment 2: Restricted Two Box Model

The restricted two box model investigates the impact of hidden variables (Fig. 1.1 (b)). This experiment is meant to test if an emulator can learn the true system response if not all information is included in the training data. Figure 1.7 summarizes the results of four emulation techniques (pattern scaling, deconvolution, modal fitting, and DMD) when trained and tested on different scenario combinations. Restricting the data means there is only one temperature series, rather than the three in the previous case. We therefore cannot calculate a global mean, and use a modified definition of pattern scaling in this section, mapping from forcing to temperature anomaly. As the FDT (Method II) has roughly equivalent performance to the previous section and is not impacted by the introduction of hidden variables, we omit it from this section.

For all methods except deconvolution (Method III), we see a sharp drop in performance when introducing a hidden variable into the system. Deconvolution exhibits the same failure mode when training on *High Emissions* as before but to a greater degree, along with the ill-conditioning failure mode when training on *Plateau* and *Overshoot*. Because this method treats each region as independent, it is more robust to the addition of hidden variables. It is able to capture the aggregate response of the

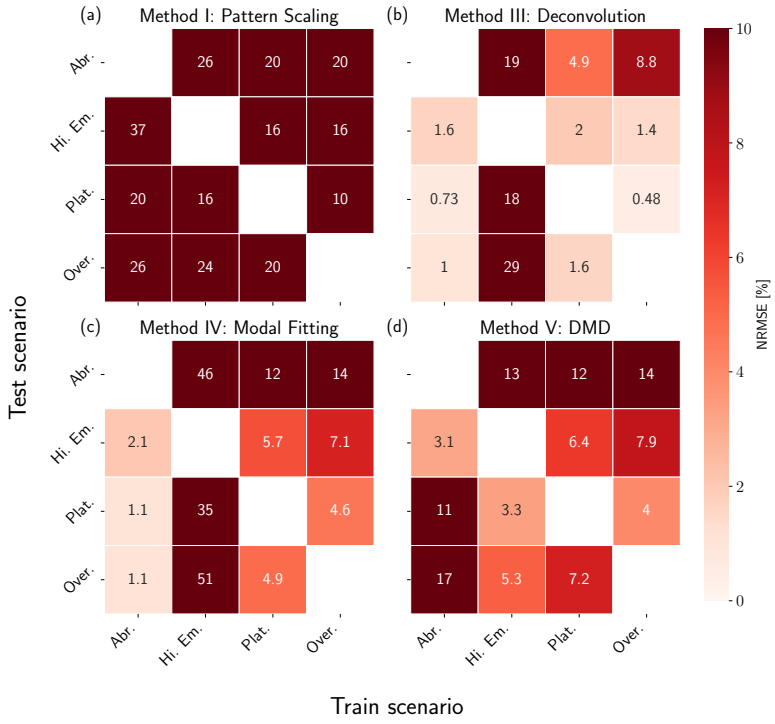

Figure 1.7: NRMSE heatmaps for pattern scaling (a), deconvolution (b), modal fitting (c), and DMD (d) emulators trained and tested against the restricted two box model. Results are shown in percentages, where lighter values correspond to lower error (higher performance) and darker values correspond to higher error (lower performance). Scenarios used for training are shown on the x-axis, while scenarios used for testing are shown on the y-axis. We do not include results for training and testing on the same dataset.

atmospheric box that includes the influence of the ocean, but would not be able to separate those effects; i.e., the response function we derive is somewhat non-physical, though it can emulate the system effectively.

For the modal fitting emulator (Method IV), we initialize the optimization routine with guesses for both dominant modes (the fast atmospheric response and slower oceanic response). It is largely unsuccessful in identifying these modes, except in the case of training with *Abrupt*. This scenario is unique in that both modes are visible in the atmospheric box alone (see the leftmost plot in the middle row of Fig. 1.3). Training on either *High Emissions* or *Overshoot* appears promising at first, but neither can extrapolate to *Abrupt*, meaning it effectively overfits on these scenarios and loses extrapolative capabilities. As before, we see that training on *High Emissions* leads to the worst performance overall, as this scenario is characterized by only one effective timescale.

DMD (Method V) and by extension, EDMD (Method VI), experiences the sharpest decline in performance, with errors increasing by several orders of magnitude in some cases. Both methods see lower error in emulating scenarios similar to the training data (e.g., *High Emissions* vs. *Plateau*), but rapidly increasing error outside that regime. In addition to learning timescales like the previous two methods, DMD and EDMD are attempting to learn spatial interactions as well, meaning they are disproportionately affected by the hidden variable. We can also frame this issue theoretically by stating that hidden variables violate one of the fundamental assumptions of EDMD and DMD: the quantities we emulate are representative of all relevant system dynamics. By hiding the oceanic box, neither algorithm can learn the true physical behavior of the system. With EDMD, increases in polynomial order lead to further decreases in performance (not shown).

1.3.4 Experiment 3: Noisy Three Box Model

Results of the noisy three box model show how noise affects each emulator (Fig. 1.1 (c)). Figure 1.8 summarizes the results of four emulation techniques (pattern scaling, deconvolution, modal fitting, and DMD) when trained only on *Abrupt* and tested against the other three scenarios; we choose to train only on *Abrupt* as it yielded high performance across all methods (except pattern scaling), and we want to isolate the impact of noise. See Fig. 1.4 in Sect. 1.3.1 for performance metrics across all train/test combinations with a constant ensemble size. Since the noise is added linearly, taking the difference between the perturbed and unperturbed ensembles effectively removes the noise when using the FDT (Method II). This leads to constant performance regardless of ensemble size, which is shown in Fig. 1.4. We additionally omit EDMD (Method VI) as it gives no improvements over DMD (Method V) in this linear case.

For these results, we evaluate performance relative to their noiseless baseline, rather than the absolute value of NRMSE; although *Abrupt* led to high performance for most methods, each method has a different baseline and some methods (e.g., pattern scaling) performed poorly when trained on this scenario. All methods exhibit decreased performance in the noisy case relative to the noiseless baseline.

Pattern scaling (Method I) experiences no change in performance as the number of ensemble members is increased, as the linear regression smooths the data, reducing the impact of noise regardless of the ensemble size. With both deconvolution (Method III) and modal fitting (Method IV), there is an almost random change in performance depending on the number of ensemble members. This is because both methods regularize the data. Deconvolution requires extra regularization when the system is noisy, or else the algorithm overfits on the noise, leading to extremely

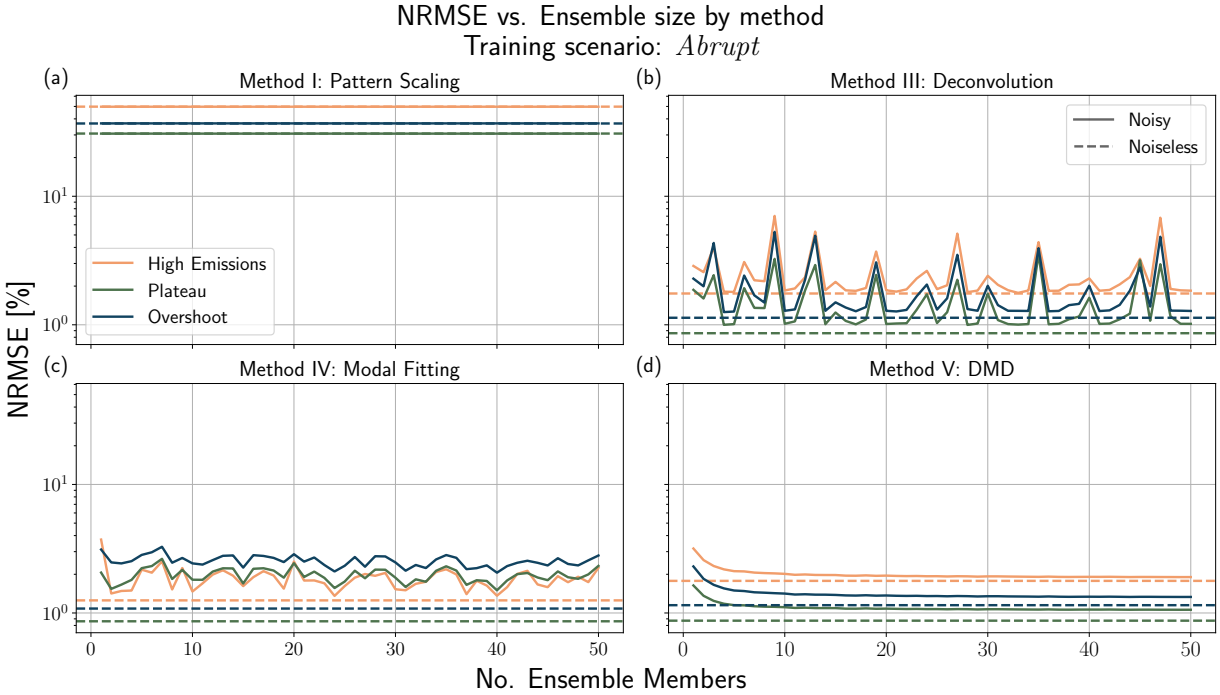

Figure 1.8: NRMSE vs. number of ensemble members for pattern scaling (a), deconvolution (b), modal fitting (c), and DMD (d) emulators trained on *Abrupt* and tested against the three remaining scenarios. Solid lines indicate the error in training/testing with noisy data, while the dashed lines indicate error in training/testing with noiseless data.

high error ($> \mathcal{O}(10^{10})$). The regularization has a similar effect to pattern scaling in making the expected performance of these algorithms more robust to noise. The variation in performance is due to the random sampling of ensemble members, with combinations that exhibit high error skewing the overall results. The error in DMD (Method V) is monotonically decreasing with ensemble size, though the presence of noise leads to a drop in performance relative to the noiseless baseline.

1.3.5 Experiment 4: Cubic Lorenz System

While the underlying cubic Lorenz system is strongly nonlinear and chaotic, the response of the ensemble mean to our applied forcing is only weakly nonlinear. This allows us to jointly investigate the impact of underlying chaos and weak nonlinearities on our emulators (Fig. 1.1 (c) and (d)). We run a 5,000 member ensemble as the variation in this experiment is much higher than the previous noisy case. As in experiment two, we use a slightly modified definition of pattern scaling, mapping from forcing to quantity of interest (the ensemble mean of Z). Figure 1.9 summarizes emulator performance against the number of ensemble members, while Fig. 1.10 shows the response function derived using the FDT.

Similar to the previous noisy experiment (Sect. 1.3.4), pattern scaling (Method I) exhibits a constant level of performance independent of the number of ensemble members. The linear fitting process creates a strong artificial smoothing effect on the data, diminishing the potential impact of noise. This is also the case with both deconvolution (Method III) and the modal fitting (Method IV) approach, both of which have little variability based on the number of ensemble members. The modal fitting approach

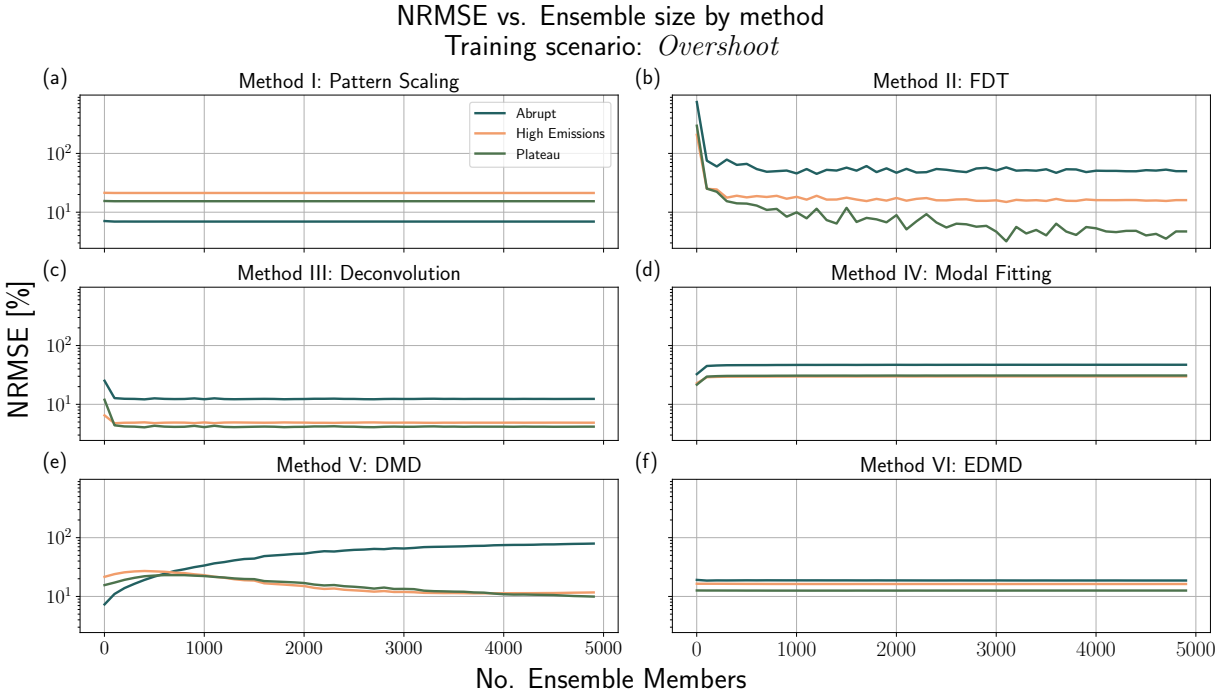

Figure 1.9: NRMSE vs. number of ensemble members for all emulators trained on *Overshoot* and tested against the three remaining scenarios. Emulators are shown as pattern scaling (a), FDT (b), deconvolution (c), modal fitting (d), and DMD (e), and EDMD (f). The FDT is trained on separate perturbation scenarios, and is therefore tested against all four scenarios. Unlike experiment three, there is no baseline/noiseless skill to compare against.

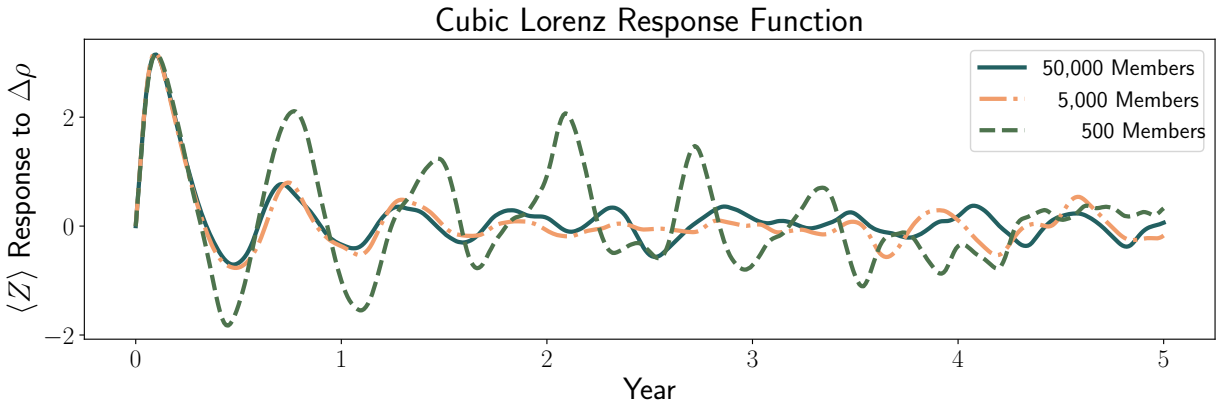

Figure 1.10: Response function for the cubic Lorenz system derived using the Fluctuation Dissipation Theorem with three sets of ensemble members: 500, 5,000, and 50,000. We use $\Delta t = 0.01$ and $\delta = 50\,\Delta t$ applied to the Y component of the system.

additionally requires an imaginary component to enforce oscillations in the response function similar to those in the FDT result (Fig. 1.10). All approaches except DMD additionally show increased skill for smaller perturbations, i.e., higher skill in predicting *Plateau* than *Abrupt*. This is likely because smaller forcings lead to smaller deviations from the theoretical limit of response theory, which assumes small perturbations from the background state.

The performance of the FDT (Method II) is strongly dependent on the number of ensemble members. Figure 1.10 illustrates this point by showing how the response function derived using the FDT changes based on ensemble size. We treat the 50,000 member ensemble as our point of comparison, as further increases in ensemble size did not result in notable performance improvements. Key features, such as the initial magnitude of the response along with the time to reach that magnitude are consistent across all ensemble sizes, but the three cases deviate after this initial peak. All three cases exhibit a similar frequency of oscillation over the time period tested, with noise in the 500 member ensemble influencing the longer-term behavior of that response (between years 3-5). There are deviations from the 50,000 member response in the 5,000 member case as well, though it is generally more in-phase than the 500 member ensemble. The NRMSE between the 50,000 and 5,000 member ensembles is 166.22%, while the NRMSE between the 50,000 and 500 member ensembles is 546.06%. Both responses are far from the ground truth, but the 5,000 member ensemble is much closer than the 500 member ensemble. Because the 5,000 member ensemble has such high error relative to the 50,000 member ensemble, the predictive skill shown in Fig. 1.4 and Fig. 1.9 does not tell the full story. By further increasing ensemble size, we expect to see commensurate increases in accuracy when emulating this system with the FDT.

Despite the fact that this experiment violates the linearity assumption of DMD (Method V), it has relatively stable performance of a similar order to the other methods tested. Predictive skill on *High Emissions* and *Plateau* increases with the number of ensemble members, as one would expect as noise is averaged out, but skill on *Abrupt* decreases, which seems to be counterintuitive. In this case, we may not be introducing any further information about the coherent, underlying dynamics, which is supported by other methods showing consistent performance in these regimes. Increasing the ensemble size is leading to further refinement of the emulator's parameters for *Overshoot* and its more closely related

scenarios (*High Emissions* and *Plateau)*. A deeper investigation is required to assess DMD's suitability for Lorenz-like systems. EDMD (Method VI) does not exhibit this behavior, instead performing with consistent skill across all combinations. This is likely because the third-order Hermite polynomial used as the basis is well-suited to train on this scenario, illustrating the need for careful selection of basis functions.

1.4 Discussion and conclusions

While emulators of Earth System Models (ESMs) have recently surged in popularity, uncertainty regarding their performance under a variety of scenarios and the lack of a comprehensive theoretical framework for analysis have posed problems for efforts at fundamental methodological comparisons. Our framework for emulator design and analysis builds on ideas from statistical mechanics and stochastic calculus, facilitating analysis of several emulation techniques from a theoretical and practical perspective. Our experiments based on simplified representations of the climate stress test a suite of emulators, including pattern scaling, response functions, and operator-based emulators, in the presence of memory effects, hidden variables, noise, and nonlinearities. Response function emulators consistently outperform other techniques, and the Fluctuation Dissipation Theorem (FDT) provides a robust method to derive them, though it also requires its own experimental ensemble. Section 1.4.1 describes emulator performance and key findings from our pedagogical examples, while Sect. 1.4.2 discusses the implications of our findings for ESMs. Table 1.6 additionally summarizes our experimental findings, focusing on the robustness of different emulators to different sources of error.

Table 1.6: Summary of emulator capability by technique based on the results from Sect. 1.3. An 'X' indicates a technique possess the listed capability, while a '∼' indicates may meet this requirement if other conditions are met; we discuss these capabilities explicitly in Sect. 1.4. *Memory* refers to an emulator's ability to capture memory effects (Fig. 1.1 (a), experiment one), *Hidden* refers to an emulator's skill in the presence of hidden variables (Fig. 1.1 (b), experiment two), *Noise* refers to an emulator's robustness to simulation noise (Fig. 1.1 (c), experiment three), and *Nonlin.* refers to an emulator's ability to capture weak nonlinear effects (Fig. 1.1 (d), experiment four).

Technique	*Memory*	*Hidden*	*Noise*	*Nonlin.*
Method I: Pattern Scaling			X	
Method II: Fluctuation Dissipation Theorem	X	X	∼	∼
Method III: Deconvolution	X	X	∼	∼
Method IV: Modal Fitting	X	∼	X	∼
Method V: Dynamic Mode Decomposition (DMD)	X		∼	
Method VI: Extended DMD	X		∼	∼

1.4.1 Emulator performance and trade-offs

Each emulation technique considered in this work belongs to a spectrum of methods as defined by the joint Fokker-Planck/Koopman operator framework. Some emulators on this spectrum demand strict assumptions (quasi-equilibrium/pattern scaling), while others are much more general (EDMD). There is a trade-off between the strictness of assumptions and emulator complexity, and relaxing these assumptions can shift the emulator's optimal use case. More general techniques may require specifically designed experiments, and decreasing structural emulator error may

come at the price of increased computational costs (e.g., the Fluctuation Dissipation Theorem). Using this framework additionally identifies a gap in the current emulator typology as defined by Tebaldi et al. (2025)[64], as we need to consider the potential role operator-based emulators can play in this ecosystem; e.g., characterizing physical behavior in the system in addition to emulating it, as in Navarra et al. (2024)[167].

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

[167] Navarra et al., 'Variability of SST through Koopman Modes', *Journal of Climate*, 2024

Pattern scaling is a popular emulation technique because it is easy to implement, fast to apply, and its limits are well understood empirically[76,79,80]. Its efficiency makes it the method of choice particularly for assessments of mean annual temperature in monotonic forcing scenarios (e.g., SSP5-8.5, 3-7.0, or 2-4.5) and for understanding first-order trends of climate signals, even in the presence of internal variability. Previous work has shown this approach is valid only when the forcing is exponential and has a fixed spatial pattern, along with linear dynamics and feedbacks[56]. Our results additionally show that pattern scaling exhibits two sources of irreducible error: a mismatch between the true and predicted patterns at equilibrium and the assumption that the climate must respond instantaneously to external forcings. If forcing history is important, such as in centennial-scale or strong overshoot experiments, the single-pattern approximation breaks down, misrepresenting shifts in regional warming over time. This is also the case with highly variable fields such as precipitation, where the first-order approximation may not capture significant trends. More general quasi-equilibrium approaches show promise (e.g., mapping from forcing to temperature in experiments two and four), but have yet to be widely explored in the context of full-scale ESMs. Pattern scaling's limitations push us towards emulation techniques that can capture more complex dynamics.

[76] Mitchell, 2003; [79] Tebaldi and Arblaster, 2014; [80] Wells et al., 2023

[56] Giani et al., 2025

Response functions are increasing in popularity as they can capture many processes of interest that are missed by pattern scaling, such as the pattern and memory effects[91,92,104,139]. This makes them ideal for representing decision-relevant, non-monotonic forcing scenarios, such as temperature overshoots. Response function approaches assume a linear relationship between the input forcing and output variable interest and that perturbations to the system are small[93]. As a result, they are able to capture weakly nonlinear effects, so long as perturbations remain within the linear response regime. They must be used with caution when nonlinear effects are dominant or (depending on the technique) when internal variability is significant.

[91] Womack et al., 2025; [92] Sandstad et al., 2025; [104] Winkler and Sierra, 2025; [139] Freese et al., 2024

[93] Lucarini, Ragone, and Lunkeit, 2017

Despite its computational costs, deriving response functions with the Fluctuation Dissipation Theorem (FDT) offers a benefit over other response function techniques: it generates the system's exact linear response. Deconvolution and modal-fitting, by contrast, can produce non-physical output. As the FDT states, the response to small perturbations can be captured by $R(t)$ if the system statistics are approximately stationary and the dynamics drive the weakly perturbed system back to the unperturbed state. The concept of climate is predicated on assuming the latter is true, further cementing the FDT's utility in this context. Because FDT-based response functions are physically interpretable, they support linear analyses of Earth system processes and serve as a reliable foundation for climate emulators[94].

[94] Lucarini and Chekroun, 2024

Emulators that seek an explicit representation of the Koopman operator are potentially powerful tools as they are founded on rigorous theory and are interpretable[166,184,194]. They can, in principle, reproduce any behavior the climate system might exhibit. In practice, however, their utility is

[166] Williams, Kevrekidis, and Rowley, 2015; [184] Schmid, 2021; [194] Tu et al., 2014

constrained by several factors. Both Dynamic Mode Decomposition (DMD) and Extended DMD (EDMD) require the input and output variables of interest (e.g., radiative forcing and temperature) to completely characterize the dynamics of the system, rendering them sensitive to hidden variables. DMD additionally requires linearity between inputs and outputs, which is often violated in practice[165]. EDMD relaxes this assumption by using a higher-dimensional space at the cost of selecting an appropriate (and often problem-specific) set of basis functions[166]. The choice of basis functions is a major consideration with this method, and we may have been able to improve our implementation of EDMD further with a different choice. Solving the resulting large eigenvalue problems with either algorithm can be computationally demanding, and EDMD and DMD can be sensitive to noise, potentially overfitting to data. Despite these challenges, operator methods allow us to identify dominant modes of variability in the climate system. They can also, in theory, be used to capture state-dependent and non-stationary processes, though this again requires a careful selection of basis functions and a large amount of training data. 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. These techniques have the potential to be highly generalizable to scenarios beyond the training data as they can reproduce the system's true dynamics, but further research is required to determine the potential of using operator-based methods directly for climate emulation.

[165] Schmid, 2010

[166] Williams, Kevrekidis, and Rowley, 2015

Emulator performance varies depending on the experimental setup, highlighting that emulators are often designed to be application specific and not completely general. Figure 1.4 provides an overview of these results, but each emulator had the potential for high performance depending on the application. For example, pattern scaling performs poorly on all experiments, but shows high skill regardless of the experiment when trained and tested against *High Emissions*; this is not shown, as the case where the training and testing datasets are the same is trivial (near zero error) for all emulation techniques. However, this illustrates that pattern scaling has utility if used on scenarios with exponential forcing, more akin to ScenarioMIP[154], or scenarios with linearly increasing forcing (after a short initial transient period); see Giani et al. (2025)[56] for further discussion. Future work will further examine the role training data plays in emulator development.

[154] O'Neill et al., 2016

[56] Giani et al., 'Origin and Limits of Invariant Warming Patterns in Climate Models', *Journal of Climate*, 2025

Whether emulators learn physically interpretable representations of the system they are emulating remains an open question, though our process of testing an emulator's extrapolative capability suggests that some techniques do learn the system's true behavior. The clearest example of this is the FDT, which performed consistently well across all scenarios. This is to be expected as the theory behind the FDT shows that it calculates the physical impulse response of the system[93,140]. Pattern scaling on the other hand, by definition, does not learn realistic behavior unless the system is fully determined by the pattern scaling coefficients. For other techniques, the results are less clear. For example, the modal fitting approach is able to extrapolate successfully in any of the first three experiments when trained on *Abrupt*, but not when trained on *High Emissions*, further supporting the need for an effort focused on quantifying the impact of training data on climate emulators. Deconvolution and DMD also exhibit mixed levels of extrapolative skill, leading to difficulties

[93] Lucarini, Ragone, and Lunkeit, 2017; [140] Giorgini et al., 2024

in making a consistent argument about interpretability from our results. This is especially the case for DMD, as the $\mathcal{L}$ matrix we derive is not easily mappable to the true underlying parameters of e.g., the coupled three box model, as this problem is effectively underdetermined; we are solving for twelve DMD parameters, whereas the full system is determined by three heat capacities, three feedback parameters, and one diffusion coefficient. Future work will investigate the possibility of learning true system parameters from these emulated representations.

1.4.2 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[100], but emulator structural differences prevent it 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[195,196], but the field of ESM emulation is currently data-constrained. Our theoretical framework and pedagogical experiments provide value in this data-limited setting, as they allow 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 ESMs improve, this framework can help ensure emulators are prepared to train on those new results.

[100] Watson-Parris et al., 2022

[195] Gillett et al., 2016; [196] Wilcox et al., 2023

Our pedagogical experiments provide a useful tool to isolate and examine individual sources of error relevant to emulating ESMs (Fig. 1.1). 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 in emulation, regardless of the parent model. For example, our results further support the growing body of literature on the utility of response functions[91,104,139]. Response functions offer improvements over pattern scaling, particularly when considering memory effects in decision-relevant scenarios. They may also better 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[126]. Existing emulators of ESMs may also benefit from incorporating response functions. For example, recent work into hybrid emulation using a generative model conditioned on pattern scaling could be extended by conditioning on response functions instead[106].

[91] Womack et al., 2025; [104] Winkler and Sierra, 2025; [139] Freese et al., 2024

[126] Van Vuuren et al., 2026

[106] Bouabid, Souza, and Ferrari, 2026

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)[64]. An intermediate step for either the FDT or EDMD may be to first emulate an EMIC, helping determine useful training scenarios without the cost of a full ESM. Though EMICs are much less computationally expensive than ESMs and therefore are

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

not as beneficial to emulate, they potentially offer a higher-dimensional, more rigorous way than an SCM to evaluate the emulation techniques discussed here. 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[151]. However, using the FDT to derive response functions through perturbations requires a full initial condition ensemble for every perturbed grid cell/region[32,93], similar to the Green's Function MIP[181], and is likely prohibitively expensive for full ESMs. The score-based FDT (Sect. 1.1.3) provides a remedy, using statistical learning methods to learn the score function and thus the system response[151]. Regardless of the derivation method, our results suggest response functions are a highly effective emulation technique both in terms of accuracy and interpretability.

[151] Giorgini, Falasca, and Souza, 2025

[32] Lembo, Lucarini, and Ragone, 2020;
[93] Lucarini, Ragone, and Lunkeit, 2017

[181] Bloch-Johnson et al., 2024

[151] Giorgini, Falasca, and Souza, 2025

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.

Optimal scenario design for climate emulation: How to train your emulator

2

The means of obtaining as much variety as possible, but with the greatest possible order... is the means of obtaining as much perfection as possible.

— Gottfried Wilhelm Leibniz

While machine learning (ML) models exhibit immense utility in interpolating complex physical systems, their ability to generalize to unseen, out-of-distribution scenarios while adhering to physical laws remains a fundamental challenge. Efforts to enforce physical consistency typically focus on model architecture and include Physics-Informed Neural Networks (PINNs) that embed governing equations into the loss function[116,118,119,121], operator learning approaches that map directly between function spaces[197,198], and hard constraints (e.g., enforcing conservation or symmetries)[115,117,120]. Hybrid techniques such as NeuralGCM further demonstrate that combining physical and statistical components can achieve significant computational savings without sacrificing predictive skill[98,109].

[116] Raissi, Perdikaris, and Karniadakis, 2019; [118] Cai et al., 2021; [119] Karniadakis et al., 2021; [121] Cuomo et al., 2022

[197] Li et al., 2021; [198] Lu et al., 2021

[115] Greydanus, Dzamba, and Yosinski, 2019; [117] Mohan et al., 2020; [120] Satorras, Hoogeboom, and Welling, 2021

[98] Bracco et al., 2024; [109] Kochkov et al., 2024

Beyond architectural constraints, the design of the training data itself dictates whether an ML model learns the underlying physics or interpolates between observed states. Data design methods include physics-informed feature engineering (e.g., using nondimensional quantities such as the Reynolds number instead of raw velocity fields)[122], physics-guided data augmentation that exploits known invariances or linearity properties[123], and synthetic data generation via active learning to place new samples in regions of large physical error or high model uncertainty[124,125]. Such methods may be particularly impactful in climate science, as the high computational cost of large-scale simulations restricts the availability of training data[42,199].

[122] Fazliani, Frangella, and Udell, 2025

[123] Li, Pang, and Shan, 2022

[124] Shields et al., 2023; [125] Guo et al., 2024

[42] Balaji et al., 2017; [199] Keller, Alerany Solé, and Acosta, 2025

In climate science, ML emulators address the demand for spatially explicit projections beyond the standard suite of realistic emissions scenarios simulated as part of the the Coupled Model Intercomparison Project (CMIP)[38,126]. Following Tebaldi et al. (2025)[64], we define emulators as statistical surrogates for physical models, distinct from process-based Simple Climate Models (SCMs) and Earth system Models of Intermediate Complexity (EMICs). Reliable climate projections are crucial for areas such as agriculture[14], the built environment[5], energy systems[13], and the finance and insurance sectors[131,132], all of which face substantial physical and transition risks from climate change. Emulators have demonstrated skill in reproducing variables such as near-surface air temperature, precipitation, relative humidity, and wind speed across annual, monthly, and daily timescales[51,64,91,97,103,105,106,127,133].

[38] Eyring et al., 2016; [126] Van Vuuren et al., 2026

[64] Tebaldi et al., 'Emulators of Climate Model Output', *Annual Review of Environment and Resources*, 2025

[14] Hultgren et al., 2025

[5] Crawley, 2008

[13] Yalew et al., 2020

[131] Collier, Elliott, and Lehtonen, 2021; [132] Zhou, Endendijk, and Botzen, 2023

[51] Meinshausen, Raper, and Wigley, 2011; [64] Tebaldi et al., 2025; [91] Womack et al., 2025; [97] Castruccio et al., 2014; [103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [105] Bassetti et al., 2024; [106] Bouabid, Souza, and Ferrari, 2026; [127] Beusch, Gudmundsson, and Seneviratne, 2020; [133] Sudakow, Pokojovy, and Lyakhov, 2022

Assessing whether emulators respect physical constraints remains challenging, as demonstrating physical consistency requires extrapolating to emissions trajectories distinct from those seen in training. In practice, however, most studies emphasize in-sample and within-range performance—where Global Mean Surface Temperature (GMST) or emissions trajectories lie within the training range—with limited emphasis on structurally

out-of-distribution tests[64,99,100]. This gap stems from the high temporal and computational costs of running full-scale Earth System Models (ESMs), typically limiting emulator developers to the data made available via CMIP for training and evaluation. As a result, emulators are largely trained on aggregate emission pathways[82,103,127–129]. Previous work demonstrates that training on ScenarioMIP-like pathways is not necessarily optimal[2], as it restricts our ability to test and train emulators that accurately respond to emissions of individual forcing agents (e.g., anthropogenic greenhouse gases and aerosols). This shortcoming is particularly pressing given that ESM scenario design for future CMIP efforts is moving towards a broader set of forcing combinations[126]. One solution is to run the ESM for each individual forcing to generate a broader set of training data[64,130], but high simulation costs and the potential for nonlinear interactions when combining forcings currently impede both the exploration and adoption of this approach. Consequently, there is a need for an approach that yields highly informative training data at a low computational cost.

[64] Tebaldi et al., 2025; [99] Lütjens et al., 2025; [100] Watson-Parris et al., 2022

[82] Mathison et al., 2025; [103] Bouabid, Sejdinovic, and Watson-Parris, 2024; [127] Beusch, Gudmundsson, and Seneviratne, 2020; [128] Tebaldi, Snyder, and Dorheim, 2022; [129] Geogdzhayev et al., 2026

[2] Womack et al., 2026

[126] Van Vuuren et al., 2026

[64] Tebaldi et al., 2025; [130] Van Katwyk et al., 2026

Here, we introduce a method to generate optimal emissions scenarios that improve both overall emulator performance and the ability to emulate the climate response to individual forcing agents. By framing training data generation as a problem of optimal experimental design[200], we directly optimize the emissions scenarios themselves to maximize emulator predictive skill; high-level and detailed descriptions of this procedure are given in Section 2.1 and Appendix B.1, respectively. Leveraging a differentiable model based on the Finite Amplitude Impulse Response (FaIR) SCM[54], our approach calculates the sensitivity of emulator predictive skill with respect to the training data, enabling iterative updates of the training data to minimize a user-defined skill metric (Fig. 2.1). Using simple and intermediate complexity climate models as proxies for ESM-simulated data, we demonstrate that training on a single optimized scenario outperforms a baseline emulator trained on a suite of six standard socio-economic scenarios (ScenarioMIP-CMIP7). Furthermore, the optimized training data yields increases in emulator skill when extrapolating to structurally out-of-distribution scenarios, indicating a more robust statistical mapping from emissions to temperature. We validate the scalability of our approach by running optimized scenarios generated by an SCM with the MIT Earth System Model (MESM), a zonally resolved EMIC, demonstrating that this method is transferrable to models of higher complexity. Finally, we discuss implications for designing ESM scenarios specifically for emulator training, along with the potential for extending our approach to other data-constrained domains of machine learning for physical systems.

[200] Fedorov, 2010

[54] Leach et al., 2021

Significance: Machine learning climate emulators promise to revolutionize impact assessments by rapidly projecting future warming, yet they often fail to generalize to types of scenarios not seen during training. While current research emphasizes improving emulator architectures, we identify the training data itself as a potential bottleneck. We introduce a method to directly optimize training datasets, enabling emulators to learn more robust physical dynamics. Training scenarios developed on computationally inexpensive models successfully transfer to train skillful emulators of higher-complexity models. This cross-model transferability circumvents the prohibitive cost of discovering optimal scenarios directly on full-scale Earth System Models, suggesting climate modeling centers could simulate scenarios tailored for emulator training over standard projections.

2.1 Results

We compare the performance of two emulator configurations: a baseline emulator trained on ScenarioMIP Priority 1, and an emulator trained on optimized data (hereafter referred to as the optimized emulator). As this work focuses on the impact of training data on predictive skill rather than emulator architecture (i.e., emulator structure and feature design), both configurations use the simplest possible neural network emulator: a multi-layer perceptron. The emulator predicts temperature time series resulting from input emissions trajectories. We generate optimized training data through a four-part iterative procedure (Fig. 2.1) that treats the training trajectory as a set of tunable parameters (see Appendix B.1 for

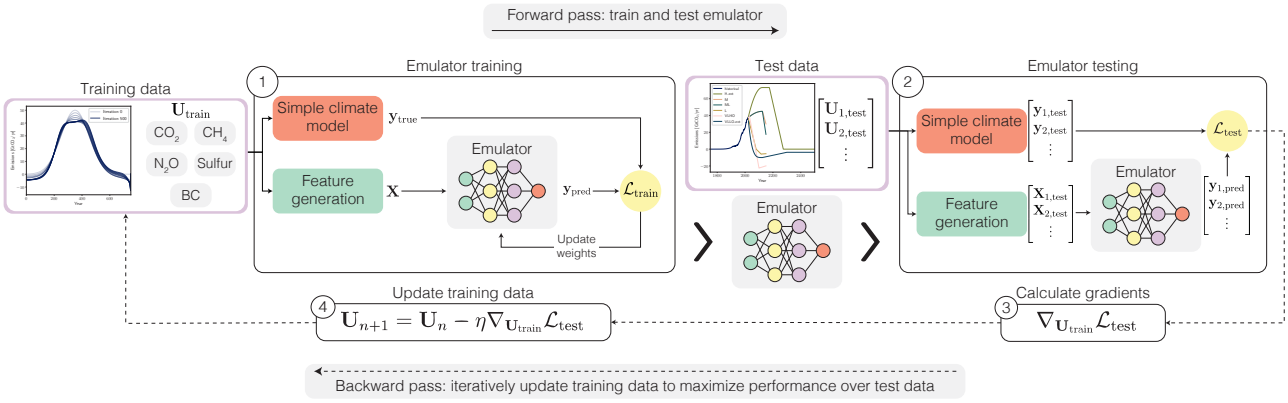

Figure 2.1: Overview of the training data optimization process for an emulator that maps from emissions (input) to global mean surface temperature (output). We iteratively update training emissions pathways through four steps: (1) train a base emulator on an initial emissions trajectory; (2) test predictive skill on target scenarios; (3) compute the sensitivity of the test loss to the training data via automatic differentiation; and (4) update the training data via stochastic gradient descent. We repeat this until convergence, performing a final independent evaluation on held-out datasets. For more details on this procedure and emulator architecture, see Appendices B.1 and B.3, respectively.

a technical description). First, we simulate the temperature response to an initial emissions time series to train a base version of our emulator (Fig. 2.11). Second, we test the emulator's performance by measuring its predictive skill in terms of Normalized Root Mean Square Error (NRMSE) over a fixed test dataset (e.g., ScenarioMIP-CMIP7 Priority 1, Fig. 2.12); skill is normalized by maximum scenario GMST to avoid overemphasizing performance on high-warming scenarios (see Equation B.7 in Appendix B.1.2). Third, we backpropagate through the testing, training, and data generation processes to calculate the sensitivity of the test loss to perturbations in the training data (Fig. 2.13). Finally, we use stochastic gradient descent to iteratively update the training emissions trajectory to maximize performance (Fig. 2.1.4). In this context, 'optimizing for a scenario' strictly means iteratively updating a training emissions trajectory to maximize the resulting emulator's ability to accurately reproduce the temperature response of that specific target scenario (or set of scenarios). To calculate sensitivities, we implement a differentiable SCM (Appendix B.2) based on the FaIR SCM[54], which includes a subset of anthropogenic forcing agents (CO_2, CH_4, N_2O, sulfur and black carbon); this limited set allows us to focus on the dominant drivers of future warming while retaining a tractable parameter space.

[54] Leach et al., 2021

We evaluate the emulators' ability to reproduce temperature anomalies predicted by an SCM and an EMIC under individual forcing (e.g., CO_2-only) and combined forcings. To test the emulators' performance across different dynamical regimes, we evaluate them against several sets of emissions scenarios. These include realistic future socio-economic policy projections (the proposed ScenarioMIP-CMIP7* Priority 1 and 2 protocol[126], and the 2025 MIT Global Change Outlook[201]), along with idealized experiments from the CMIP DECK designed to display model feedback response characteristics[38]. When training the emulators to reproduce the effect of multiple active forcing agents, we additionally evaluate the emulators' skill in reproducing the effects of isolated historical forcings (Detection and Attribution MIP (DAMIP)[195,202]) and a

[126] Van Vuuren et al., 2026

[201] Paltsev et al., 2025

[38] Eyring et al., 2016

[195] Gillett et al., 2016; [202] Gillett et al., 2025

* At the time of performing this investigation and writing this manuscript, the final version of ScenarioMIP-CMIP7 was not yet published. As a result, we use the scenarios outlined in the preprint manuscript, not including the *High-to-Low* scenario added in the final version.

climate intervention pathway implementing sulfur injection to cool the climate (Geoengineering MIP (GeoMIP)[203,204]). Because our SCM calculates sulfur's radiative forcing contribution as parameterized aerosol-cloud interactions, the sulfur emissions of the GeoMIP analogue are much larger than the true GeoMIP protocol (i.e., unrealistic) and instead serve as a strongly out-of-distribution test for the emulator.

[203] Kravitz et al., 2015; [204] Visioni et al., 2026

We generate optimized training emissions trajectories to maximize predictive skill on each set of scenarios individually, along with a configuration optimized over all scenarios simultaneously (Table 2.1). Furthermore, because optimizing over all scenario sets at once inherently introduces information leakage (i.e., evaluation data influences training), we perform an additional, independent evaluation. We feed the optimized scenarios generated by our differentiable SCM as input to the EMIC, training an emulator to reproduce the EMIC's zonal temperature response under an identical evaluation protocol.

We first present the results of emulating GMST from the SCM, followed by the results of emulating zonal temperatures from the EMIC. Complete descriptions of emulator architecture, emissions scenarios, and evaluation protocol can be found in Appendices B.2 - B.5.

2.1.1 SCM results: individual forcing agents

We first focus on CO_2-only experiments, as the optimization results are qualitatively consistent across most agents (Appendix B.7). Fig. 2.2 provides an illustrative example of the optimization process when maximizing predictive skill for a high-warming emissions scenario (ScenarioMIP-CMIP7 Priority 1 *H-ext*). While training an emulator on a naive, constant emissions time series (50 $GtCO_2/yr$) yields poor initial predictions (green dot-dash line, Fig. 2.2c), iterative updates to the training data drive the emulator's temperature predictions to near-perfect agreement with the SCM-projected targets. This convergence is robust across forcing agents, albeit at varying rates (Fig. 2.3). The optimized emissions trajectory differs structurally from the ground-truth emissions trajectory (compare Fig. 2.2a and b). While the optimized input shares some features with the ground truth, such as sign changes in the slope and concavity, it does not simply reconstruct it. This distinction suggests the optimization process (Fig. 2.1) successfully isolates the physically salient features required for emulation, rather than memorizing a specific trajectory.

Iterative optimization of training data yields higher predictive skill for individual forcing agents compared to the baseline emulator trained

Table 2.1: Summary of the experimental protocol utilized in this work. For each climate model, we train a baseline emulator, along with multiple optimized emulator configurations as described in the optimization column.

Climate model	Baseline scenarios	Optimization (training data generation)	Evaluation scenarios	Emulator targets
Differentiable SCM	ScenarioMIP-CMIP7 Priority 1	Iteratively updated to maximize predictive skill when tested on: 1. Individual sets (ScenarioMIP-CMIP7, DECK, CS3, DAMIP, and GeoMIP) 2. All scenario sets simultaneously	Evaluated against all individual scenario sets	Global Mean Surface Temperature (GMST)
EMIC (MESM)	ScenarioMIP-CMIP7 Priority 1	*No direct optimization.* Uses the optimal emissions trajectories generated by the SCM optimized for performance over all scenarios	Independent evaluation across all single-forcing scenario sets (ScenarioMIP-CMIP7, DECK, CS3)	Zonal Temperatures

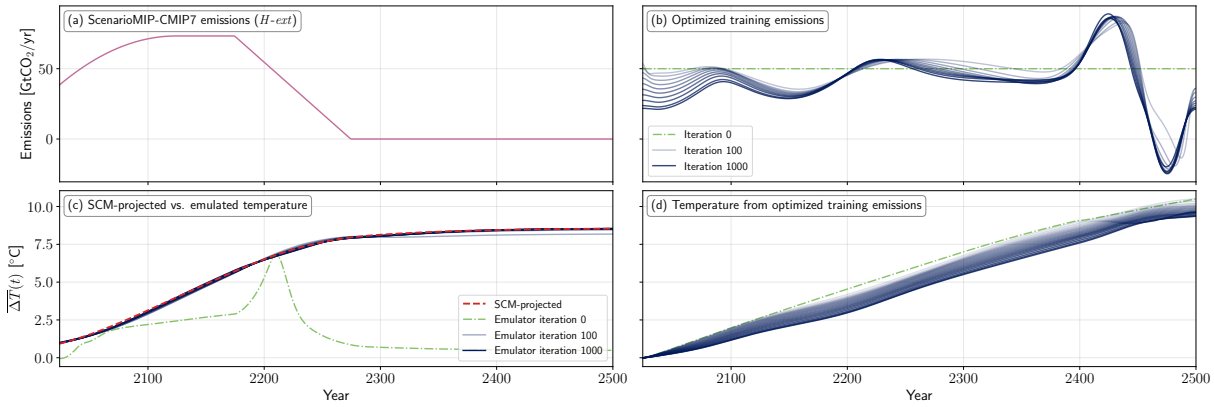

Figure 2.2: Optimization results for a single CO_2-only high-warming scenario (ScenarioMIP-CMIP7: *H-ext*). (a) Ground-truth emissions trajectory. (b) Evolution of optimized training emissions over 1000 iterations, beginning from a constant initial condition (green dot-dash line). (c) Comparison of SCM-projected (dashed red line) vs. emulated GMST predictions (green dot-dash and solid blue lines). (d) Temperature trajectories corresponding to the emissions in (b). Faint lines trace intermediate states every 100 iterations in (b)-(d). The emulator is trained on the synthetic input-output pair (b, d) and tested by predicting the response to ground-truth input (a), as shown in (c).

on standard socio-economic scenarios (Fig. 2.3). Because the baseline emulator is evaluated against its own training set (ScenarioMIP-CMIP7 Priority 1), this evaluation represents its theoretical error lower bound. Despite this, our optimized emulator achieves lower normalized error (NRMSE) for all agents except sulfur, crossing below the baseline emulator's error threshold (dark blue vs. dashed orange lines, Fig. 2.3). This performance gap demonstrates that standard baseline scenarios are sub-optimal for training, lacking the feature diversity necessary to capture all potential system behaviors. For sulfur, where baseline emulator error is already minimal ($\mathcal{O}(10^{-2})$ vs. $\mathcal{O}(10^{-1})$ for other agents), the error in the optimized emulator decreases monotonically, suggesting eventual convergence. Transient spikes observed in the error trajectory (e.g., CH_4-only experiment) reflect the inherent trade-offs in multi-objective optimization, where aggregate skill gains across the full dataset may temporarily degrade performance on individual scenarios.

Fig. 2.4a summarizes the change in performance between the baseline and optimized emulator configurations for CO_2-only experiments, where positive values indicate improvement. Overall, optimizing for any of the realistic socio-economic pathways (Opt. Priority 1, Priority 2, or CS3), or for the combined dataset (Opt. All) consistently increases average emulator skill. Optimizing for the baseline Priority 1 scenarios yields the largest mean improvement (44.3%). Notably, simultaneous optimization over all datasets yields broad performance gains across all evaluation datasets without overfitting to any specific scenario. While specialized optimization targets achieve the highest skill on their respective evaluation sets (e.g., optimizing for Priority 1 yields a 47.2% increase when predicting Priority 1, compared to 34.4% for the combined dataset), combined optimization ensures the emulator can generalize across scenario structures.

A clear trade-off emerges, however, regarding the idealized forcing scenarios (DECK). Optimizing for slowly varying socio-economic pathways (Priority 1, Priority 2, or CS3) yields little to no improvement, or even degrades performance, on the idealized scenarios. Conversely, optimizing for the DECK reduces skill across all other datasets. This bifurcation stems from the idealized scenarios' unique forcing structure, specifi-

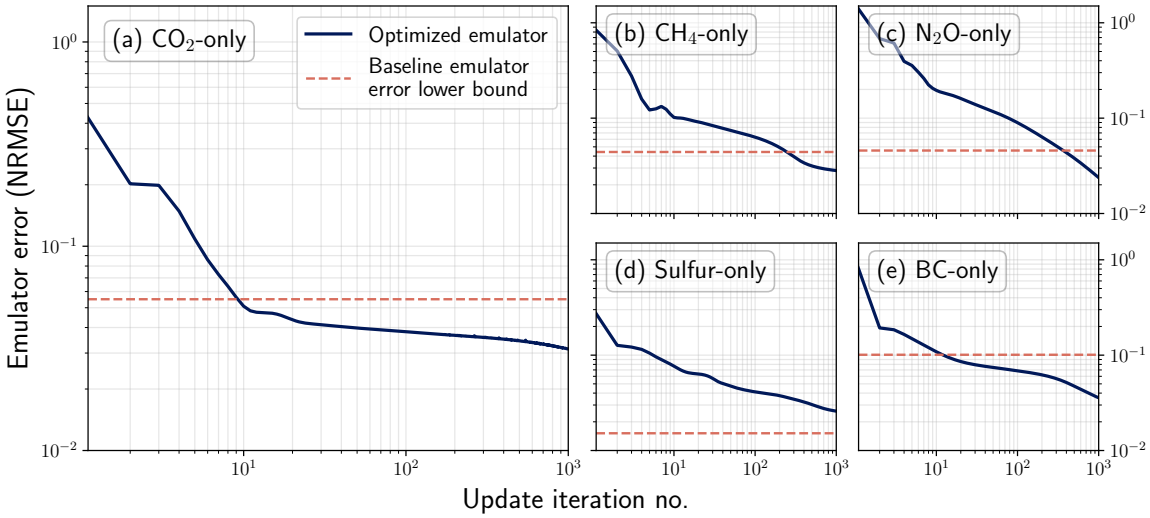

Figure 2.3: Error in emulating single-forcing experiments. Evolution of evaluation loss (NRMSE) when reproducing SCM-projected GMST anomalies for ScenarioMIP-CMIP7 Priority 1 single-forcing scenarios (e.g., (a) CO_2-only, (b) CH_4-only). The solid dark blue line tracks the optimized emulator's performance, while the dashed orange line indicates the baseline emulator's error lower bound (evaluated on its own training data).

cally the abrupt quadrupling of CO_2 (*abrupt-4xCO2*), which features a pulse-and-decline emissions trajectory driving rapid warming ($\mathcal{O}$(50 years) to reach 4°C compared to $\mathcal{O}$(200 years) in other high-warming scenarios). While an idealized step-forcing yields a skillful emulator for many data-driven approaches[2], it acts as a statistical outlier during optimization. Minimizing emulator error on this shock without including the gentle gradients that characterize realistic socio-economic emissions pathways, coupled with the idealized dataset's small sample size (two scenarios), promotes overfitting. In contrast, the comparably small CS3 dataset shares structural similarities with Priority 2, allowing for successful extrapolation. Because the physical features required to emulate an emissions pulse conflict with those needed for more realistic emissions pathways, including the DECK in the combined optimization creates competing objective functions. Consequently, the average improvement in predictive skill across all scenarios is slightly lower when optimizing over all datasets (41.0%) compared to optimizing solely for the Priority 1 baseline (44.3%). However, it is necessary to include the idealized scenarios for generalization, as optimizing over all datasets successfully increases predictive skill on the abrupt scenario, whereas optimizing only for realistic pathways yields no such improvement.

[2] Womack et al., 2026

2.1.2 SCM results: multiple forcing agents

Consistent with the single-agent results, the optimized emulator with all forcing agents active outperforms the baseline emulator when tested against standard socio-economic projections (Priority 1), which represents the baseline emulator's theoretical error lower bound (Fig. 2.5). The optimization process begins with a performance plateau attributable to small gradient magnitudes from the constant initialization (Appendix B.6). It then enters a phase of monotonic error reduction, where fluctuations in error convergence reflect sensitivity to the fixed learning rate; future stability improvements may be achieved through learning rate scheduling[205]. Panels (b) and (c) of Fig. 2.5 display the optimized time series for well-mixed and aerosol forcing agents. In the ground-truth realistic emissions pathways (Priority 1), all forcing agents follow highly

[205] Li and Arora, 2019

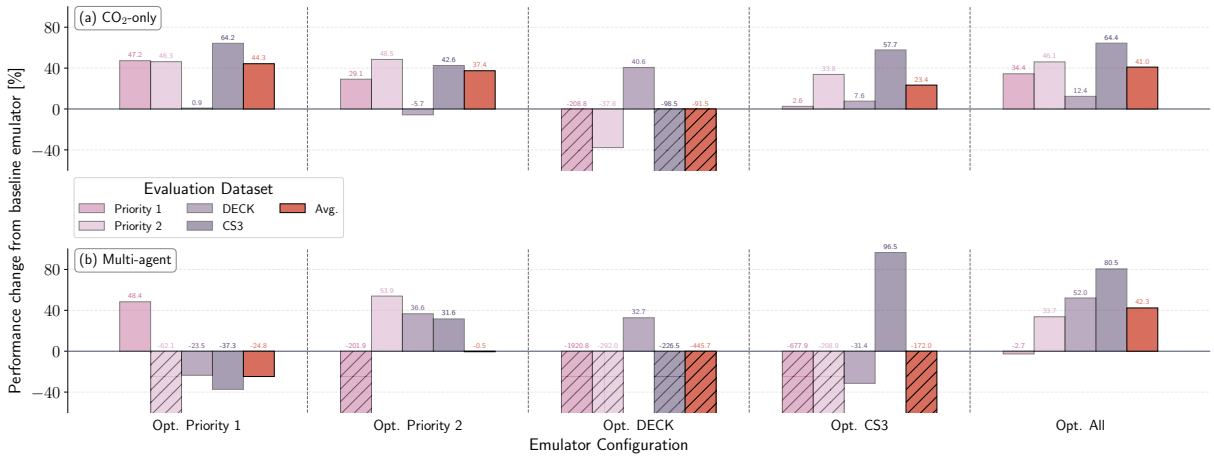

Figure 2.4: Performance of optimized emulators relative to baseline configuration across several evaluation datasets. Change in predictive skill (NRMSE) from the baseline emulator for (a) CO_2-only and (b) multi-agent forcing experiments. Positive values indicate improved accuracy (reduced error). Bars represent mean performance across all scenarios in the specified evaluation dataset. Optimization targets include realistic policy pathways (ScenarioMIP-CMIP7 Priority 1 and 2, CS3), idealized forcing scenarios (DECK) and the full combined dataset (All). Hatched bars indicate a performance decrease exceeding y-axis limits.

correlated trajectories (e.g., CO_2 and CH_4 follow the same pattern of increase and decrease over time). This allows the baseline emulator to achieve high in-sample skill by learning aggregate forcing behavior rather than individual agent dynamics. While the optimized pathways we generate are structurally distinct from standard scenarios, they exhibit consistent low-frequency features across all agents that are overlaid with high-frequency variations.

While optimizing for performance over individual datasets may lead to trade-offs in extrapolative skill (Fig. 2.4b), simultaneous optimization over the full scenario set yields performance gains across every evaluation dataset. This result suggests that optimization isolates fundamental physical features independent of specific scenario structures; Appendix B.6 demonstrates that potentially infinite valid features exist, depending on the optimizer's initialization. Incorporating a diverse set of scenarios during optimization can yield higher predictive skill on a specific target than optimizing exclusively for that target. For example, when evaluated on the idealized DECK scenarios, the emulator optimized over the combined dataset (Opt. All) outperforms the baseline emulator by 52.0%, providing an additional 15.4% improvement over the emulator optimized solely for the DECK (which achieves only a 36.6% increase). This effect is also present, though less pronounced, when optimizing for the longer, more structurally diverse Priority 2 scenarios, further supporting the need for diverse optimization targets. Conversely, restricting the number of optimization targets degrades extrapolative performance relative to the single-agent case. This is likely due to the increased complexity of emulating multiple agents and disaggregating their responses. Overfitting is most prevalent for the idealized DECK and realistic CS3 datasets, where small sample sizes (two scenarios each) and limited agent diversity (the DECK scenarios are CO_2-only) fail to adequately constrain the parameter space. Similarly, optimizing exclusively for standard, aggregate emissions pathways reduces extrapolative skill by roughly 25%, highlighting the limitations of scenarios dominated by aggregate forcing pathways.

Training an emulator with a scenario optimized for performance over all scenario types simultaneously (the realistic policies, idealized forcings, isolated historical forcings, and climate interventions described

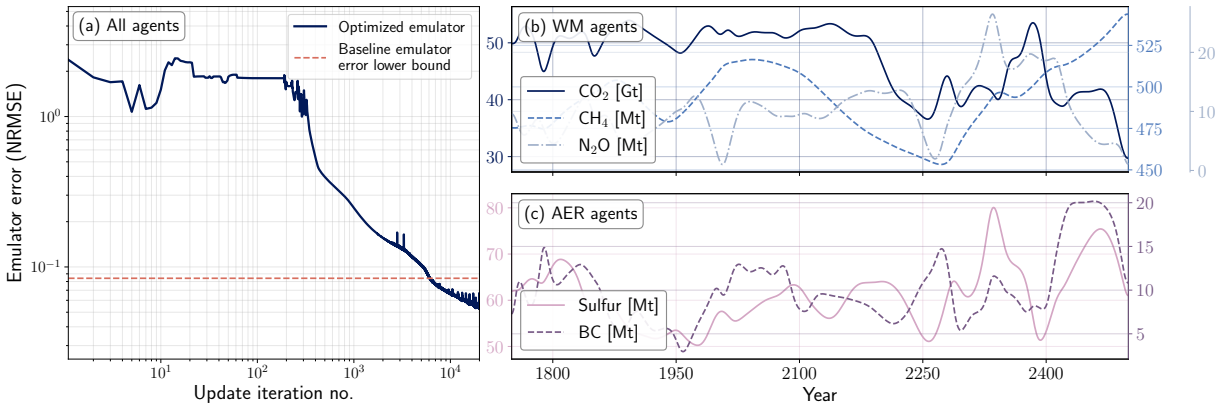

Figure 2.5: Emulator error and forcing trajectories for multi-forcing experiments (a) Evolution of evaluation loss (NRMSE) when reproducing SCM-projected GMST anomalies for realistic emissions pathways (ScenarioMIP Priority 1 with all forcing agents active). The solid dark blue line tracks the optimized emulator; the dashed orange line indicates the baseline emulator's error lower bound (evaluated on its own training data). (b) Optimized emissions time series for well-mixed forcing agents (CO_2, CH_4, and N_2O). (c) Same as (b), but for aerosol agents (sulfur and black carbon).

in Appendix B.5) enables us to accurately reproduce both individual and aggregate forcing agent dynamics, correcting biases present in the baseline emulator (Fig. 2.6). The emulator optimized over the combined dataset achieves high accuracy when evaluated on the out-of-distribution isolated forcing and climate intervention subsets; emulating DAMIP and GeoMIP yields $R^2 = 0.97$ (Fig. 2.6d). In contrast, emulators optimized for or trained on highly correlated aggregate emissions pathways (e.g., the realistic Priority 1 scenarios) fail to generalize to these unseen datasets, exhibiting systematic errors. Neither the baseline emulator nor the Priority 1-optimized emulator accurately captures the distribution of warming and cooling effects between individual agents. For example, the baseline emulator systematically overestimates the cooling effect of sulfur. This failure is most evident when emulating *G6sulfur*, a high-emissions climate intervention scenario that utilizes sulfur injection to limit warming. These emulators capture the aggregate trends prior to the geoengineering intervention but underestimate subsequent warming once sulfur injection begins. Only the emulator optimized on the full, diverse scenario set eliminates this bias, accurately predicting temperature anomalies across the full range of individual and aggregate effects.

2.1.3 Intermediate complexity model (MESM) results

To validate our approach and demonstrate its scalability, we perform an independent evaluation using an intermediate complexity climate model that outputs zonal temperatures (MESM). By utilizing the SCM from the previous section to generate optimized training scenarios that we then simulate with the intermediate complexity model, we both verify our optimized scenarios are useful for the more complex task of emulating zonal temperatures and prevent any information leakage during training. Due to operational constraints associated with running MESM in emissions-driven mode, we limit our evaluation to CO_2-only scenarios. As before, we compare a baseline emulator trained on six realistic emissions scenarios (ScenarioMIP-CMIP7 Priority 1) against emulators trained on optimized scenarios, now using either one or two scenarios for training (derived from constant and sinusoidal initializations, Fig. 2.7a and b). Our results demonstrate that training on these optimized scenarios yields performance that matches or exceeds the six-scenario baseline emulator

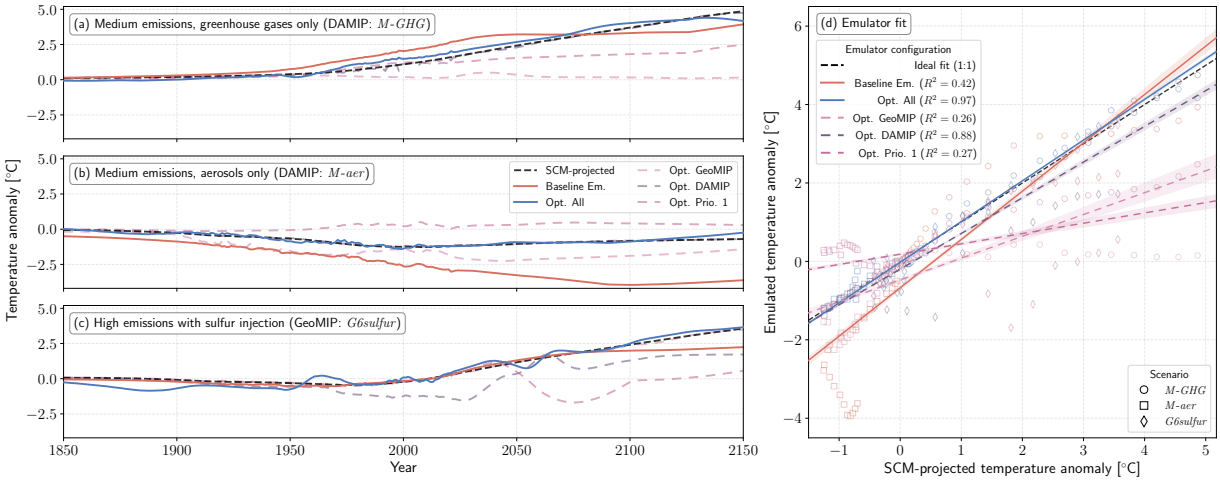

Figure 2.6: Emulator extrapolative performance on structurally distinct forcing scenarios (isolated historical forcings – DAMIP; climate interventions – GeoMIP). (a)-(c) GMST anomaly trajectories relative to 1750 for (a, b) isolated historical forcings and (c) climate intervention scenarios. Lines compare SCM-projected (dashed black), baseline emulator predictions (solid orange), and emulator optimized over all scenarios (solid blue) against emulators optimized for specific training subsets (dashed colored lines). (d) Linear fit of emulated vs. SCM-projected anomalies for the scenarios in (a)-(c). The black dashed line marks the ideal 1:1 relationship. Colors denote training configuration; scatter markers denote scenario (sampled every 15 years). Shaded regions indicate 95% confidence interval of the linear fit.

across both alternate policy projections (Priority 2) and idealized forcing scenarios (DECK). While the baseline emulator inherently retains the highest skill on its own training data (Priority 1), our optimized emulators demonstrate broad extrapolative improvements.

When emulating the intermediate complexity model, optimizing from a sinusoidal initial emissions trajectory generally yields higher predictive skill compared to a constant initialization, capturing a wider array of long-term physical dynamics; the choice of initialization dictates which physical features the optimizer can isolate. Because the sinusoidal initial condition produces a trajectory with extended periods of decreasing and net-negative carbon emissions (Fig. 2.7b), it provides more informative features for extrapolating to new scenarios that exhibit these behaviors. Specifically, the centennial-scale oscillations present in the sinusoidal trajectory likely enable the optimization process to better constrain the characteristic timescales of the climate system; these temporal modes are required to accurately emulate delayed warming or cooling associated with physical processes like deep ocean heat uptake. This translates to increased skill, leading to a 12.5% average improvement over the baseline emulator on Priority 2 scenarios and a 15.6% improvement on the idealized DECK. In contrast, optimizing from a constant initial condition produces a high-emissions trajectory (Fig. 2.7a) that lacks a substantial period of net-negative emissions. As a result, though the constant-initialized emulator marginally outperforms the sinusoidal model on shorter, positive-emissions pathways (e.g., *M*, *ML*, *L*, *M-ext*, and *L-ext*), it struggles to capture overshoot pathways like *VLLO-ext* and *H-ext-OS*, and suffers a 28.9% decrease in skill on the idealized DECK. Whereas several initial conditions yield similar performance improvements over the baseline emulator in the SCM case (Appendix B.6), the initialization of an optimized scenario plays a major role in the case of emulating the intermediate complexity model.

Including optimized scenarios from both initial conditions in the emulator training dataset (i.e., concatenating the outputs of two separate optimization runs into a single expanded training dataset) yields, in sev-

eral cases, an improvement in skill that surpasses the performance of the individual configurations combined (e.g., *H-ext*, *1pctCO2*, and *VLHO-ext*). Training on these two complementary scenarios drives a 37.9% average improvement in extrapolative skill on Priority 2 scenarios, indicating the combined dataset captures a broader spectrum of physical responses than either the six-scenario baseline or the individual optimized scenarios. However, combining multiple trajectories can occasionally lead to destructive interference. For example, performance on the current trends policy scenario (CS3 *CT*) degrades relative to the individual optimized configurations. This likely occurs because the emulator attempts to average the distinct physical feedbacks triggered by the high-warming constant initialization and the more moderate sinusoidal initialization, effectively interpolating to a non-physical intermediate state. While this failure mode is restricted to this pathway and could potentially be resolved through scenario reweighting during optimization, our results indicate more broadly that optimizing across multiple initial conditions provides a robust pathway for training emulators that generalize across a wide range of future climates.

2.2 Discussion

Generating maximally informative training data offers major utility for ML models of physical systems, particularly where data generation is computationally expensive. Our approach optimizes the training data for a climate emulator directly using a low-cost surrogate simple climate model, decoupling the computational cost of the optimization process from the run-time of full-scale Earth System Models (ESMs). While training data for ML surrogate models are typically generated by costly numerical simulations[206,207], our method produces optimal trajectories efficiently. This approach shares similarities with dataset distillation[208–210], but differs fundamentally as we aim to generate a maximally informative dataset based on a specific target rather than identify salient features from an existing dataset. Our application to an intermediate complexity climate model (MESM) validates the scalability of this approach, showing increased predictive skill across structurally dissimilar scenarios.

[206] Ganti and Khare, 2020; [207] Zhang and Zhao, 2021

[208] Wang et al., 2020; [209] Nguyen et al., 2021; [210] Cazenavette et al., 2022

As this study utilizes a simple multi-layer perceptron, our results provide a conservative estimate of the method's potential. While employing more complex sequence-based or attention-driven architectures would likely yield higher absolute predictive skill by more effectively capturing temporal dynamics, the simple neural network architecture highlights the relative benefit of the optimized training data itself. As demonstrated by the sensitivity analysis to changes in the neural network architecture (Appendix B.6, sinusoidal initial condition), the optimized time series generated by our approach are largely consistent across architectures. This suggests that the features extracted by our method are physically salient, rather than artifacts associated with a specific architecture. Further work across alternate ML architectures, domains, and systems with stronger nonlinearities is required to fully characterize our method's performance, but it could apply more generally to any ML approach integrated with a differentiable synthetic data generation pipeline, highlighting its potential for the design of parsimonious training data.

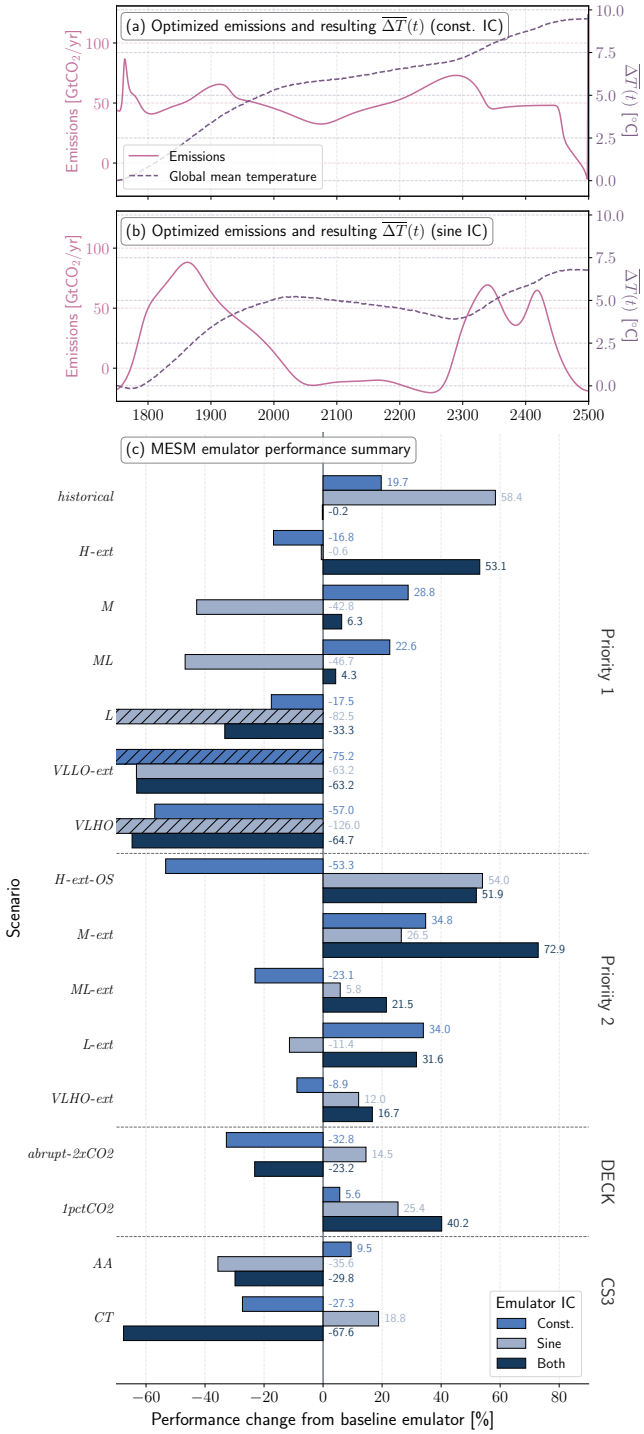

Figure 2.7: Training data and performance of optimized emulators relative to baseline configuration across several evaluation datasets when emulating an intermediate complexity climate model (MESM). (a) Emissions and GMST trajectories resulting from optimizing for predictive skill over all scenarios when initialized from a constant emissions trajectory. (b) Same as (a), but initialized from a sinusoidal emissions trajectory. While GMST is shown here for illustrative purposes, the baseline and optimized emulator configurations are trained to reproduce zonal temperature anomalies. (c) Change in predictive skill (NRMSE) from the baseline emulator for CO_2-only scenarios across realistic emissions pathways (ScenarioMIP Priority 1 and 2, CS3) and idealized scenarios (DECK). Positive values indicate improved accuracy (reduced error). Bars represent global average (latitude-weighted) performance. Const. and sine refer to optimized scenarios initialized from constant and sinusoidal trajectory, respectively. Both indicates the emulator was trained on both sets of initial conditions.

Scenarios characterized by high structural diversity are better suited for both emulator training and understanding system behavior than baseline scenarios. The unconventional, rapidly varying emissions trajectories generated by our optimization process (Figs. 2.2, 2.5, and 2.7) are highly informative for identifying a system's response, as suggested by system identification techniques[211]. Our results show that potentially many alternate choices for climate scenarios are more informative than the current choice of standard policy projections (ScenarioMIP), leading to higher predictive skill despite training on a smaller dataset. This is supported by Giani et al. (2025)[56] and Womack et al. (2026)[2], which show that traditional high-warming emissions scenarios used for emulator training (e.g., *SSP585*) can cause the temperature response to reduce to a single timescale, rendering the emulator unable to learn the full system dynamics. Paired with the sensitivity analysis in Appendix B.6, the generated trajectories illustrate that there is no single optimal scenario for training, but rather a family of optimal scenarios for a given application. For example, training on multiple scenarios generated from distinct initial conditions (constant and sinusoidal) can yield an improvement in extrapolative skill that surpasses the performance of the individual configurations combined (Fig. 2.7). Optimizing over all scenarios additionally increases average emulator performance regardless of the number of forcing agents present (Fig. 2.4). This includes learning both individual and aggregate forcing behavior from a single scenario (Fig. 2.6), with emulator performance validated on out-of-distribution scenarios. This skill has not been explicitly demonstrated by other emulation techniques, illustrating the potential utility of our method.

[211] Kravitz et al., 2017

[56] Giani et al., 'Origin and Limits of Invariant Warming Patterns in Climate Models', *Journal of Climate*, 2025

[2] Womack et al., 'A theoretical framework to understand sources of error in Earth System Model emulation', *Earth System Dynamics*, 2026

We demonstrate the generalizability of our approach through the direct transfer of optimized scenarios between structurally distinct climate models (Fig. 2.7). Because transferability across model types is not guaranteed *a priori*, the ability to use scenarios optimized exclusively on the simple climate model to train a skillful emulator for the intermediate complexity model supports the practical utility of the method. Whereas the high computational cost of running thousands of simulations of a full-scale model prohibits us from directly identifying an optimal training scenario, this cross-model application indicates that a simple surrogate model may be sufficient for this purpose; future work can investigate simulating our optimized scenarios using a full-scale ESM. Although using the simple model as a surrogate requires optimizing over all scenarios simultaneously, which inherently introduces information leakage, our independent evaluation using the intermediate complexity model confirms that we are able to successfully isolate salient physical features rather than merely overfitting to the evaluation metric.

Because our optimization procedure strictly requires a differentiable climate model, our work demonstrates the utility of differentiability across emulator training, calibration, and experimental design. First, we show that differentiability enables our approach to generating maximally informative training datasets. While backpropagation would be computationally intensive for a full-scale differentiable climate model[212], a modified version of this method could be used to inform online emulator training as a simulation is running, using the gradient to select the next data point that minimizes the emulator's loss. Second, we utilize the differentiability of our simple model to calibrate it to reproduce the median temperature response of the constrained, calibrated FaIR ensemble[213] without the expert intervention required by standard calibration techniques (e.g., minimizing the loss between observed and modeled climate

[212] Moses et al., 2025

[213] Smith et al., 2024

statistics)[214–216]. Automatic differentiation accelerates this process and provides a systematic approach to calibration[109,217–219]. Finally, we use the model to generate the sulfur injection trajectory necessary to recreate the climate intervention scenario (GeoMIP *G6sulfur*) via automatic differentiation. This allows us to compute the sensitivity of the output temperature to the sulfur trajectory, addressing the lack of consistent emissions protocols for such experiments[203].

[214] Kennedy and O'Hagan, 2001; [215] Schneider et al., 2017; [216] Schneider, Leung, and Wills, 2024

[109] Kochkov et al., 2024; [217] Heimbach, Hill, and Giering, 2005; [218] Forget et al., 2015; [219] Davenport et al., 2026

[203] Kravitz et al., 2015

Although the question of emulator interpretability is always present with nonlinear/black-box methods, our results highlight that the choice of training data plays a large role in an emulator's physical consistency. While not fully interpretable, the improved extrapolative capability of our optimized emulator may support the development of future emulators targeted towards interpretability. By successfully learning individual forcing effects and the full system response using only the scalar GMST output from our optimized scenarios, we demonstrate a rigorous surrogate for ESM emulation where the availability of spatial information would likely simplify the separation of distinct forcing signatures.

However, there are trade-offs in this approach. Training on multiple structurally distinct scenarios can occasionally lead to destructive interference, as seen when the emulator attempts to average the physical behavior triggered by conflicting training regimes on the intermediate model's *CT* scenario. Additionally, scenarios with extreme structural differences may have competing optimization goals, requiring more iterations to achieve high performance. Future work can explore resolving these issues in two ways. Methodologically, ensemble learning concepts like boosting[220] could sequentially generate optimal features missing from prior datasets. Physically, a two-step training procedure could separate system identification from optimization: (1) estimate intrinsic climate timescales through idealized experiments (e.g., *abrupt-4xCO2*[2]); (2) use sinusoids of those frequencies as initial conditions for our methodology, allowing the optimizer to find the remaining salient structures.

[220] Friedman, 2002

[2] Womack et al., 2026

As full-scale ESMs cannot keep pace with the ever-increasing demand for climate projections beyond CMIP, the popularity of climate emulators for scenario assessment continues to grow. While this study demonstrates the foundational theory and approach for generating optimal emulator training scenarios, fully realizing the utility of this method requires operational implementation to scale these results. This involves applying it to a differentiable intermediate complexity model[219] to evaluate how additional variables like precipitation alter the optimal emissions trajectory, and concurrently utilizing the trajectories derived in this work as forcing inputs for a full-scale ESM. Evaluating an emulator on these outputs will enable a direct performance comparison against a baseline emulator trained on standard policy projections. Since previous work has shown that standard scenarios are suboptimal for emulator development, modeling centers should consider dedicating resources to generate simulation data explicitly designed for machine learning. Moreover, the inherent uncertainty of future socio-economic pathways (e.g., CMIP8 and beyond) requires training emulators to capture the broadest possible range of climate dynamics; the optimization presented in this work provides one structured approach to achieve this scenario diversity. Because these trajectories diverge significantly from standard model intercomparison protocols, they effectively force models into regimes outside their typical tuning. These stress tests offer utility beyond emulator training by quantifying model uncertainties under out-of-distribution forcings.

[219] Davenport et al., 2026

Establishing a formal intercomparison project for emulator development would benefit both the climate modeling and impacts communities. Such an initiative would produce robust emulators capable of generating large, impact-relevant ensembles in a fraction of the time, ultimately freeing computational resources to focus full-scale models on frontier Earth system science.

2.3 Materials and methods

1. Training data optimization. We frame the generation of training data as a bi-level optimization problem (Fig. 2.1); we outline our procedure here and include more detail in Appendix 1. Our objective is to find a specific set of training emissions ($\mathbf{U}_{\text{train}}$) that minimizes the error of an emulator trained on that data when tested against a target set. This problem consists of an implicit inner level (training the emulator parameters θ) and an explicit outer level (updating the training emissions). The optimization objective is given mathematically as

$$\arg\min_{\mathbf{U}_{\text{train}}} \mathscr{L}_{\text{test}}(\mathbf{U}_{\text{train}},\ \theta_{\text{train}},\ D_{\text{test}}),$$

where θ_{train} represents the parameters of the emulator after training on the data generated by $\mathbf{U}_{\text{train}}$ and D_{test} is a test dataset held constant during optimization.

1.1 Inner level (emulator training): The inner level consists of training an emulator to map from emissions to temperature anomalies. We construct training features ($\mathbf{X}_{\text{train}}$) from our emissions time series ($\mathbf{U}_{\text{train}}$). We then force the SCM with $\mathbf{U}_{\text{train}}$ to generate the corresponding GMST anomalies ($\mathbf{y}_{\text{train}}$), which serve as ground-truth targets. The emulator is trained via Stochastic Gradient Descent (SGD) to minimize the Mean Squared Error (MSE) between its predictions and $\mathbf{y}_{\text{train}}$, resulting in optimized network weights (θ).

1.2 Outer level (emissions update): The outer level tests the performance of the trained emulator on D_{test}. We quantify test performance using scenario length-weighted NRMSE ($\mathscr{L}_{\text{test}}$); length-weighting prevents short scenarios from being overrepresented during optimization. To update $\mathbf{U}_{\text{train}}$ to minimize $\mathscr{L}_{\text{test}}$, we utilize Automatic Differentiation (AD) to efficiently calculate the gradient $\nabla_{\mathbf{U}_{\text{train}}}\mathscr{L}_{\text{test}}$ by backpropagating through the testing, training, and data generation processes. We then apply these updates via an SGD optimizer with momentum[221]. A complete breakdown of the chain rule expansion of our procedure and the corresponding pseudocode is provided in Appendix B.1.2.

[221] Liu, Gao, and Yin, 2020

2. Simple and intermediate complexity climate models. To enable our optimization procedure, we present a differentiable implementation of an SCM based on the FaIR SCM[54]. Implemented in JAX, this model leverages automatic differentiation for efficient gradient-based calibration while retaining the core structural components of FaIR. We use a three-box impulse response model to calculate GMST anomaly time series based on total effective radiative forcing from five forcing agents: CO_2, CH_4, N_2O, sulfur, and black carbon; a full description of the model and its calibration can be found in Appendix B.2.

[54] Leach et al., 2021

To more rigorously test the suitability of our optimized scenarios to train emulators of more sophisticated models than our SCM, we utilize

MESM[222], an EMIC that includes a two-dimensional, zonally averaged atmospheric model with interactive chemistry coupled to a zonally averaged land model and an anomaly-diffusing ocean model; see Sokolov et al. (2018)[222] for a full description. As MESM is not differentiable, we use outputs from the optimization procedure from our differentiable SCM as inputs to MESM, simulating the zonal temperature response to these emissions. We additionally simulate the scenarios outlined in Appendix B.5; all MESM simulations are run as a thirty-member initial condition ensemble.

[222] Sokolov et al., 2018

[222] Sokolov et al., 'Description and Evaluation of the MIT Earth System Model (MESM)', *Journal of Advances in Modeling Earth Systems*, 2018

3. Neural network emulator. We implement a neural network emulator to predict temperature from emissions for both climate models considered in this work. We emulate GMST from our SCM and ensemble-average zonal temperatures from MESM. As this work focuses on the impact of training data, rather than emulator architecture (i.e., emulator structure and feature design), on predictive skill, we use the simplest possible neural network: a multi-layer perceptron; improvements in predictive skill are likely possible with more advanced architectures. We train several emulator configurations for each climate model: a baseline emulator trained on the proposed Priority 1 scenarios from ScenarioMIP-CMIP7, along with one emulator for each set of optimized training data as described in Appendix B.5. A full description of the emulator architectures used for each climate model can be found in Appendix B.3.

4. AI use disclosure. We used Gemini 3.1 Pro to help refactor and document the Python scripts and Jupyter notebooks underlying our training data optimization and figure generation. The AI was used strictly for code assistance, not for the direct generation of graphic assets. All AI-assisted code was reviewed, tested, and verified by the lead author, who assumes full responsibility for its accuracy. The resulting codebase is made publicly available on GitHub[†], and users assume all responsibility when running or adapting the code for their own applications.

[†] https://github.com/cbwomack/Emulator_Training_Data

Assessing spatially explicit sensitivities to scenario uncertainty through climate emulation

3

It occurs to me that our survival may depend on talking to one another.

— Dan Simmons, *Hyperion*

As the impacts of climate change continue to intensify, the demand for climate projections to inform adaptation and mitigation efforts grows. Earth System Models (ESMs) are our most comprehensive tools for analyzing physical climate uncertainty, but their high computational and temporal costs limit their utility for exploring the vast space of scenarios resulting from different socio-economic development pathways and climate policy assumptions[33,42,43]. Conversely, Integrated Assessment Models (IAMs) efficiently project the coupled socio-economic outcomes of physical and social interactions for different scenarios[62,223–226]. However, translating their outputs to climate impacts has typically entailed running projected emissions either through full ESMs or simpler models like Simple Climate Models (SCMs) (e.g., MAGICC[51] or FaIR[53,54]) and Earth system Models of Intermediate Complexity (EMICs) (e.g., MESM[222]). ESMs give spatially explicit climate outcomes, but are so computationally expensive that they can only be used for a limited number of scenarios. In practice, they are rarely used for custom IAM scenarios outside of the the Coupled Model Intercomparison Project (CMIP). SCMs and EMICs are computationally efficient and capable of exploring many scenarios, but are limited to aggregated global- or regional-scale climate outcomes, which lack the spatial resolution required for spatially explicit risk assessment. To get to that scale, an additional downscaling step is required, making the approach more complex and less accessible. Decision makers need risk assessment tools that more efficiently bridge the IAM and ESM paradigms, providing probabilistic spatially explicit information on actionable timescales without the prohibitive computational overhead of full-scale ESM ensembles.

[33] Flato, 2011; [42] Balaji et al., 2017; [43] Balaji et al., 2022

[62] Monier et al., 2013; [223] Edmonds and Reiley, 1984; [224] Vuuren et al., 2011; [225] Intergovernmental Panel on Climate Change, 2015; [226] Weyant, 2017

[51] Meinshausen, Raper, and Wigley, 2011

[53] Smith et al., 2018; [54] Leach et al., 2021

[222] Sokolov et al., 2018

Traditionally, translating policy decisions into spatially explicit climate outcomes for emissions scenarios beyond those considered in CMIP has employed the second approach mentioned, and has therefore required some form of spatially resolved climate emulation[64]. This process involves mapping coarse variables (e.g., global mean surface temperature) to spatially explicit scales. Pattern scaling—the linear regression of spatial climate variables against global mean or zonally averaged temperature—is one of the most widely used methods for rapidly projecting the impacts of future emissions scenarios, demonstrating success in reproducing mean climate trends for a number of variables[74,76,79,82,127]. More broadly, emulators are increasingly being used for impact assessment[65–68], along with applications such as attribution and net-zero pathway comparisons[69–73].

[64] Tebaldi et al., 2025

[74] Santer et al., 1990; [76] Mitchell, 2003; [79] Tebaldi and Arblaster, 2014; [82] Mathison et al., 2025; [127] Beusch, Gudmundsson, and Seneviratne, 2020

[65] Shiogama, Takakura, and Takahashi, 2022; [66] Munday et al., 2025; [67] Polonik, Burney, and Ricke, 2025; [68] Varney et al., 2026

[69] Beusch et al., 2022; [70] Kitsios, O'Kane, and Newth, 2023; [71] Schwaab et al., 2024; [72] Schöngart et al., 2025; [73] Quilcaille et al., 2025

However, impact assessment requires both an accurate representation of spatially explicit climate processes and the ability to quantify the three primary sources of uncertainty in climate projections[36]: internal variability (inherent chaos of the climate, dominant on annual timescales), model structural uncertainty (representation of the physics and dynam-

[36] Hawkins and Sutton, 2009

ics, dominant on annual-to-decadal timescales), and scenario uncertainty (human factors, dominant on decadal-to-centennial timescales). While scenario uncertainty is typically addressed through a handful of different emissions or concentration scenarios[227], there have been some probabilistic approaches to quantify the socio-economic uncertainty that drives scenarios[227–231]. Efforts such as CMIP have made great strides towards quantifying model structural uncertainty under a common set of anthropogenic forcings[37,38]. Ensemble modeling strategies have also been used to address model structural uncertainty, as well as to assess both present and future ranges of internal variability[31,39,40]. For example, the MIT Integrated Global Systems Model (IGSM) utilizes an ensemble pattern scaling approach to capture model structural uncertainty while incurring much lower computational costs than an ESM[62,78], allowing for greater exploration of scenario uncertainty. However, this approach does not allow for the sampling of individual realizations without first running an EMIC to project zonal temperatures for emulation, limiting its accessibility. Capturing compound climate risks that directly impact societal metrics like labor productivity, such as wet bulb temperature (heat stress) and Vapor Pressure Deficit (VPD) (fire risk)[134,135], requires individual realizations that incorporate internal variability with accurate spatial correlations and cross-correlations between variables.

[227] Morris et al., 2025

[227] Morris et al., 2025; [228] Gillingham et al., 2018; [229] Morris et al., 2022; [230] Rennert et al., 2022; [231] Fyke, Swart, and Huard, 2026

[37] Taylor, Stouffer, and Meehl, 2012; [38] Eyring et al., 2016

[31] Maher et al., 2019; [39] Shiogama et al., 2023; [40] King et al., 2024

[62] Monier et al., 2013; [78] Gao, Sokolov, and Schlosser, 2023

[134] Stull, 2011; [135] Williams et al., 2019

To address the challenge of rapidly projecting compound climate hazards for any custom scenario, we couple the MIT Emissions Prediction and Policy Analysis (EPPA) model[232] and the Finite Amplitude Impulse Response (FaIR) SCM[54] with a generative diffusion-based climate emulator[106]. Traditional pattern-scaling approaches can effectively capture the monotonic mean response of climate variables to increasing Global Mean Surface Temperature (GMST), but are limited in their ability to represent statistical distributions. In contrast, this generative emulator directly models the full joint probability distribution of the parent ESM at the grid-cell level. This allows it to capture spatially explicit correlations that would otherwise be omitted by pattern scaling. We utilize this emulator to generate stochastic realizations—defined here as an individual ensemble member representing a single potential climate state—of spatially explicit climate anomaly fields. This approach combines the accessibility of a Python-based SCM with the spatial fidelity of an ESM, solving the computational bottleneck while preserving the cross-variable correlations necessary for assessing compound hazards. Combining this approach with an IAM enables rapid assessment of spatially explicit climate impacts and sensitivities to scenario uncertainty.

[232] Paltsev et al., 2005

[54] Leach et al., 2021

[106] Bouabid, Souza, and Ferrari, 2026

We first benchmark our emulator against the existing MIT IGSM pattern scaling approach[78], demonstrating that it performs equivalently in reproducing climate fields while offering increased flexibility and accessibility. We then apply this framework to investigate the detectability of statistically significant spatially explicit climate differences between several policy scenarios. Finally, we use the emulator to proactively assess the projected ScenarioMIP-CMIP7 scenarios prior to the release of full-scale ESM datasets[126]. By generating ensembles of wet-bulb temperature and VPD projections, we provide an early assessment of how different emissions pathways may amplify acute climate risks, highlighting implications for disparate impacts informing regional adaptation planning.

[78] Gao, Sokolov, and Schlosser, 2023

[126] Van Vuuren et al., 2026

3.1 Methods

3.1.1 Socio-economic emissions scenarios

To characterize future climate risks under socio-economic uncertainty and varying degrees of policy stringency, we utilize a set of emission pathways generated using the MIT EPPA model. EPPA is a multi-sector, multi-region Computable General Equilibrium (CGE) model of the world economy designed to project economic growth, energy transitions, and anthropogenic emissions of greenhouse gas and air pollutants[232–234].

[232] Paltsev et al., 2005; [233] Chen et al., 2016; [234] Morris et al., 2019

We focus on three EPPA scenarios: a Reference (current policies) scenario, a 2°C stabilization scenario, and a 1.5°C stabilization scenario; see Table 3.1 for descriptions of each scenario considered in this work. For each scenario, we utilize a 400-member ensemble of simulations that survey a broad range of uncertainties in key socio-economic parameters, including labor and capital productivity, population growth, energy technology costs, and fossil fuel resource availability; see Morris et al. (2019)[234] and Morris et al. (2025)[227] for full details. While EPPA generates detailed regional emissions data, we aggregate these regional projections into global total CO_2-equivalent emissions trajectories to drive our climate emulator (Section 3.1.2). We project climate outcomes for the median emissions pathway for each parameter ensemble to represent the central tendency of each policy outcome. These scenarios provide a granular view of socio-economic and policy uncertainty while enabling us to directly benchmark against previous efforts that applied the pattern scaling framework described in Section 3.1.3.

[234] Morris et al., 'Representing the costs of low-carbon power generation in multi-region multi-sector energy-economic models', *International Journal of Greenhouse Gas Control*, 2019

[227] Morris et al., 'Quantifying both socioeconomic and climate uncertainty in coupled human–Earth systems analysis', *Nature Communications*, 2025

In addition to the EPPA ensembles, we also consider the projected emissions pathways of ScenarioMIP for the upcoming CMIP7 exercise[126]. The pathways utilized in this study are preliminary, stylized trajectories produced by the FaIR SCM, rather than the final scenarios being generated by IAMs at the time of writing this manuscript. By emulating these preliminary pathways now, we can assess potential regional impacts months to years before the official ESM outputs become publicly available. As a result, these emulated results represent approximations in four distinct ways: first, the emissions pathways themselves are stylized and subject to change; second, the global temperature outcomes may differ significantly from the FaIR projections once ESMs are run in fully emissions-driven configurations; third, the resulting spatial climate fields are emulated rather than simulated; and fourth, the emulator is trained to reproduce CMIP6 model output (described in the following section), as CMIP7 model output is not yet available for training. Despite these structural limitations, this framework provides an early look at how the next generation of standard climate scenarios might unfold under CMIP6 structural assumptions.

[126] Van Vuuren et al., 2026

3.1.2 Generative climate emulation pipeline

To translate emissions pathways into spatially explicit climate realizations, we employ a two-stage modeling pipeline. First, we utilize the FaIR SCM calibrated to reproduce the MPI-ESM1-2-LR ESM GMST—the same model used to train our emulator—to simulate the deterministic, forced GMST anomalies relative to preindustrial conditions[54]. FaIR is either driven by CO_2-equivalent emissions derived from the EPPA model scenarios, or the full set of emissions (i.e., all greenhouse gases and aerosols) from the approximate ScenarioMIP-CMIP7 scenarios (detailed in Section 3.1.1).

[54] Leach et al., 2021

This step provides the annual-mean GMST trajectories necessary to drive stage two: emulating monthly average climate anomalies at grid-cell resolution.

We use a machine learning-based climate emulator implemented via the *climemu* Python package; see Bouabid, Souza, and Ferrari (2026)[106] for further details on emulator architecture, training, and evaluation. The emulator is a diffusion model conditioned on GMST that simultaneously projects monthly averaged anomalies for near-surface air temperature, precipitation, relative humidity, and near-surface wind speed. It is trained on output from the MPI-ESM1-2-LR ESM for the Shared Socioeconomic Pathways (SSPs), *piControl*, and *historical* scenarios from the sixth phase of the Coupled Model Intercomparison Project (CMIP6)[31,38,235]. The emulator is evaluated against its parent ESM by comparing the distributions of emulated climate variables with those from large ensemble MPI-ESM1-2-LR simulations. Across both pre-industrial regime and warming scenarios not seen during training, the distributions show good agreement in Earth mover's distance (measures similarity between probability distributions), cross-correlation, and tail behavior. Inaccuracies in the emulator are generally small relative to the magnitude of internal variability in the ESM ensemble, or occur at scales close to the numerical grid, for which the ESM can already not be treated as perfectly accurate. An exception

[106] Bouabid, Souza, and Ferrari, 'Score-Based Generative Emulation of Impact-Relevant Earth System Model Outputs', *Journal of Advances in Modeling Earth Systems*, 2026

[31] Maher et al., 2019; [38] Eyring et al., 2016; [235] Schupfner et al., 2021

Table 3.1: Complete list of scenarios used in this work. Scenario descriptions for the EPPA scenarios are taken from Morris et al. (2025)[227], while ScenarioMIP-CMIP7 descriptions are derived from van Vuuren et al. (2026)[126a]. Each ScenarioMIP-CMIP7 scenario additionally has an *Extension* variant, which extends the emissions trajectories out to 2500.

Activity	Scenario	Short Description
EPPA	*Reference*	No Paris Agreement targets, but expansion of renewables policies
	2°C	Paris NDC targets are met by all countries by 2030, after which there is an emissions cap, implemented with a global emissions price, ensuring 2100 GMST does not exceed 2°C above pre-industrial levels with a 66% probability
	1.5°C	Paris NDC targets are met by all countries by 2030, after which there is an emissions cap, implemented with a global emissions price, ensuring 2100 GMST does not exceed 1.5°C above pre-industrial levels with a 50% probability
ScenarioMIP	*H*	**High:** High emission scenario exploring potential high-end impacts.
(Priority 1)	*M*	**Medium:** Medium emission scenario consistent with current policies.
	ML	**Medium-Low:** Delayed mitigation effort, insufficient to meet Paris Agreement goals.
	L	**Low:** Scenario consistent with likely staying below 2°C.
	VLLO	**Very Low with Limited Overshoot:** Consistent with limiting warming to 1.5°C by 2100 with limited overshoot.
	VLHO	**Very Low after High Overshoot:** Scenario with similar end-of-century temperature impact to VLLO, but with delayed near-term mitigation and reliance net-negative emissions, resulting in a higher overshoot.

[a] At the time of performing this investigation and writing this manuscript, the final version of ScenarioMIP-CMIP7 was not yet published. As a result, we use the scenarios outlined in the preprint manuscript, not including the *High-to-Low* scenario added in the final version.

is precipitation, where the emulator struggles to capture pronounced regime shifts for example in regions influenced by the seasonal migration of the Intertropical Convergence Zone. By conditioning on the deterministic FaIR-generated GMST, the diffusion model generates independent stochastic realizations of regional climate anomaly fields. This approach allows us to rapidly generate large ensembles, capturing shifts in internal variability and exploring extremes without the prohibitive computational cost associated with running full-scale ESMs. Unlike other approaches that emulate climate variables individually[79,82,91,92,127,139], this approach preserves both cross-variable correlations at the grid-cell level and spatial correlations within each variable. This capability is crucial for integrating emulators into impact assessment frameworks, as these abilities allow us to capture compound impacts (e.g., wet bulb temperature and vapor pressure deficit, Section 3.1.5) and synoptic-scale structures, such as heat domes or monsoons[106].

[79] Tebaldi and Arblaster, 2014; [82] Mathison et al., 2025; [91] Womack et al., 2025; [92] Sandstad et al., 2025; [127] Beusch, Gudmundsson, and Seneviratne, 2020; [139] Freese et al., 2024

[106] Bouabid, Souza, and Ferrari, 2026

3.1.3 Benchmarking against CMIP6 pattern scaling

To evaluate our methodology, we benchmark our results against the pattern scaling framework used within the MIT IGSM. As described by Gao, Sokolov, and Schlosser (2023)[78], this procedure begins by utilizing 18 CMIP6 models to calculate baseline monthly climatologies for each climate variable of interest. They then derive pattern-scaling kernels and pattern-change kernels that define how each spatially explicit variable scales with temperature. The IGSM then couples the EPPA model with the MIT Earth System Model (MESM) to simulate a 50-member initial condition ensemble for each scenario, giving stochastic, zonal temperatures that are pattern-scaled to grid-cell resolutions. This approach ensures consistency across interactions between key socio-economic drivers (e.g., economic development, energy and land system changes) and physical climate responses, improving assessments of climate impacts across multiple sectors.

[78] Gao, Sokolov, and Schlosser, 'A Large Ensemble Global Dataset for Climate Impact Assessments', *Scientific Data*, 2023

However, this IGSM workflow presents specific limitations for spatially explicit compound risk assessment. Traditional pattern scaling approaches successfully preserve physical consistency between the mean projected climates of different variables; e.g., an increase in mean temperature will appropriately correlate with a decrease in mean relative humidity. They do not, however, capture the physical cross-correlations between variables within an individual realization. Despite the IGSM's ability to capture some notion of internal variability, its zonal resolution when generating realizations limits its representation of processes that exist on finer spatial scales. Consequently, while its pattern-scaled outputs are climatologically useful, they cannot accurately simulate the co-occurrence of spatially explicit extremes required to model compound climate impacts. Furthermore, while the IGSM's use of an intermediate complexity model is more computationally efficient than a full-scale ESM, running a new large ensemble still requires on the order of a day, limiting its utility for real-time applications and online simulation of socio-economic feedbacks.

We benchmark our generative emulator against this established methodology by generating realizations for mid-century (2040-2050) and end-of-century (2090-2100) periods. We utilize the *Reference* policy scenario outlined in Gao, Sokolov, and Schlosser (2023)[78] and Morris et al. (2025)[227] (described in Table 3.1). To ensure an accurate comparison between the two approaches, we calculate the annual GMST from the IGSM projec-

[78] Gao, Sokolov, and Schlosser, 'A Large Ensemble Global Dataset for Climate Impact Assessments', *Scientific Data*, 2023

[227] Morris et al., 'Quantifying both socioeconomic and climate uncertainty in coupled human–Earth systems analysis', *Nature Communications*, 2025

tions over the target decades and use these values as the input to our emulator. This is in contrast to when we project new scenarios, where we use FaIR as described above. While the IGSM pattern-scaling approach includes patterns from 18 CMIP6 models, we restrict our analysis to the Max Planck Institute ESM to maintain consistency with our emulator's training data. Because the IGSM reproduces the high-resolution version of the model[41] while our generative emulator reproduces the low-resolution version[235], we regrid the IGSM output to match the lower resolution. We therefore expect large-scale agreement between the two approaches, with minor deviations corresponding to the structural differences between the high- and low-resolution ESMs configurations.

[41] Müller et al., 2018

[235] Schupfner et al., 2021

3.1.4 Hypothesis testing

A distinct advantage of our generative emulator over traditional pattern-scaling approaches is its ability to sample from a probability distribution that is statistically consistent with the parent ESM[103]. This is particularly relevant over land, where the emulator captures not only the mean response, but also forced shifts in internal variability, enabling human-relevant uncertainty quantification. By repeatedly sampling from the emulator, we generate large ensembles of realizations for any given scenario.

[103] Bouabid, Sejdinovic, and Watson-Parris, 2024

Using these grid-cell-level distributions, we can assess the impact of differing scenarios and their policy assumptions on spatially explicit climate outcomes. We perform grid-cell-wise independent two-sample t-tests on each generated distribution. For a given location, we compare the ensemble of spatially explicit anomalies generated under different policy scenarios (e.g., *Reference* vs. 1.5°C) to determine if their mean responses are statistically distinguishable. Because conducting simultaneous hypothesis tests across an entire spatial grid inherently increases the probability of falsely identifying significant differences, we control the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure[236,237]; implementation details are provided in Appendix C.1.

[236] Wilks, 2006; [237] Wilks, 2016

3.1.5 Impact metrics

We analyze metrics related to human heat stress and wildfire risk as examples of societal and environmental impacts from the approximate ScenarioMIP-CMIP7 scenarios. All metrics considered require accurate reconstruction of cross-variable correlations, a capability provided by our generative emulator framework. We assess broad shifts in baseline human heat stress using Wet-Bulb Degree-Days (WBDD), a cumulative proxy metric that integrates the intensity of emulated heat events that exceed a baseline thermal threshold. Because our emulator generates monthly averaged fields rather than daily data—inherently smoothing the acute diurnal extremes that drive immediate physiological heat stress—we utilize a conservative critical mean threshold of 25°C rather than the 31°C often cited in daily analyses; future work can apply another version of the emulator trained to generate daily statistics to this problem. During the summer months (JJA in the Northern Hemisphere and DJF in the Southern Hemisphere), the standard deviation of daily wet-bulb temperatures ranges from 0.5°C to 3.0° depending on the region (Appendix C.2.1). Although true daily temperature distributions can be non-Gaussian[129,238], this regional variance provides a practical heuristic. As the monthly average wet-bulb temperature crosses the 25°C proxy

[129] Geogdzhayev et al., 2026; [238] Stefanova, Sura, and Griffin, 2013

threshold, there is a high likelihood of experiencing acute daily extremes, such as sustained multi-hour exposures to 28°C or 31°C.

We first generate absolute temperature (T) and relative humidity (RH) fields by combining the emulator's anomaly projections with a baseline climatology calculated from the parent climate model's *piControl* run. We then compute the wet-bulb temperature (T_w) using the approximation described by Stull (2011)[134]. While monthly averaged T_w is highly correlated with monthly averaged T, incorporating humidity changes provides a more physically complete proxy for long-term shifts in regional heat stress. This approximation is valid for $T \in [-20°, 50°\text{C}]$ and $RH \in [5\%, 99\%]$; data outside these ranges are discarded. The excess intensity is calculated as $\max(0, T_w - 25)$, which we multiply by the number of days in the respective month to derive the monthly WBDD. We aggregate these values to compute the total cumulative heat stress and spatially weighted global averages to track the evolution of risk across different emissions pathways.

[134] Stull, 'Wet-Bulb Temperature from Relative Humidity and Air Temperature', *Journal of Applied Meteorology and Climatology*, 2011

To evaluate changes in fire danger, we calculate VPD, which serves as a primary indicator of atmospheric water demand and represents a reliable predictor of dead fuel moisture content[239,240]. Sustained elevated VPD induces severe stress on living vegetation, driving plant mortality and increasing the availability of highly flammable dead fuel. Because of this mechanistic link to fuel moisture content, VPD has been shown to be strongly correlated with fire activity across boreal, temperate, Mediterranean and tropical ecosystems[135,240–246]. We calculate VPD as the difference between the saturation vapor pressure (e_s, derived via the Clausius-Clapeyron relation) and the actual vapor pressure (e_a):

[239] Resco de Dios et al., 2015; [240] Clarke et al., 2022

[135] Williams et al., 2019; [240] Clarke et al., 2022; [241] Sedano and Randerson, 2014; [242] Williams et al., 2014; [243] Abatzoglou and Williams, 2016; [244] Higuera and Abatzoglou, 2021; [245] Li and Banerjee, 2021; [246] Resco de Dios et al., 2021

$$\text{VPD} = e_s(T) - e_a(T, RH). \tag{3.1}$$

We focus our fire weather analysis on two regions in the Continental United States (CONUS): the Pacific Northwest and Southwestern United States, regions where intense fire seasons are becoming increasingly frequent due to climate change[243,247,248]. We calculate the average VPD over the summer months (June, July, August; JJA) across an ensemble of realizations under future ScenarioMIP-CMIP7 projections. We additionally calculate the 95% confidence interval alongside the mean for all projections. While monthly average VPD cannot resolve acute daily fire-weather events, we use it as a proxy for overall seasonal risk. The 95% confidence interval quantifies the internal variability of the climate system learned by the generative emulator. By bounding the natural fluctuations of the monthly mean, this uncertainty quantification provides decision makers with a probabilistic understanding of compound risks that pattern-scaling approaches cannot reliably capture.

[243] Abatzoglou and Williams, 2016; [247] Balch et al., 2017; [248] Keyser and LeRoy Westerling, 2017

3.2 Results

3.2.1 Benchmarking against MIT IGSM pattern scaling

We emulate a 50 member ensemble for the *Reference* scenario to benchmark the generative emulator against the IGSM ensemble pattern scaling approach. Overall, the emulator is consistent with the baseline technique in reproducing mean climatological states in mid-century (2040-2050, Figure 3.1). However, structural differences arise due to the training data

sources. The pattern scaling approach is trained from the high-resolution MPI model, and we regrid its outputs to match the coarser resolution of the generative emulator trained on the low-resolution MPI model. As a result, the pattern scaled data retain signatures of small-scale features, such as orographic effects over the Andes, Himalayas, and Rockies, that the high-resolution model resolves explicitly. While these fine-scale features represent physically plausible responses to regional topography, the added value of high-resolution climate simulations remains a subject of ongoing debate; increased spatial resolution can sometimes introduce false precision due to imperfections in sub-grid physical parameterizations, and internal variability may mask fine-scale response patterns[249–251]. Our emulator, trained on the low-resolution model, does not reproduce these sharp features. Additionally, the diffusion process inherent to the generative model tends to smooth the climate fields slightly relative to the parent model[106]. Despite this smoothing, the emulator captures the dominant large-scale spatial patterns across the evaluated variables.

[249] Racherla, Shindell, and Faluvegi, 2012; [250] Lloyd, Bukovsky, and Mearns, 2021; [251] Lenderink et al., 2023

[106] Bouabid, Souza, and Ferrari, 2026

Differences are more apparent when analyzing higher-order statistics, specifically the standard deviation and 95th percentile fields for variables beyond temperature. While the spatial features remain broadly consistent between the two techniques, the generative emulator generally predicts a lower standard deviation, particularly for relative humidity over South America, Africa, and Northern Asia. This is again likely attributable to the difference in resolution of the two models, as the variability of the high-resolution model is generally considered more realistic in these regions[41]. Conversely, the generative emulator predicts greater extremes in relative humidity over high-latitude regions compared to pattern scaling, while predicting lower extremes over the Sahara and Australian deserts. Precipitation fields are nearly identical, with the exception of the ENSO region off the coast of western South America, where our emulator projects a greater mean, standard deviation, and 95th percentile into the Pacific Ocean. We see a similar pattern for wind fields, where the large-scale circulation is matched, but the 95th percentile shows smoothing in the tropics, along with decreased variability in the Antarctic. Previous work, however, has shown that CMIP6 models tend to show large inter-model variability in Antarctic winds[252], which is likely reflected here as well. Additionally, we note a ringing or wave-like artifact in the spatial fields of certain variables, particularly relative humidity and precipitation. This behavior is likely an artifact of the neural network architecture[106]. Future work can explore whether training the generative model on high-resolution ESM outputs resolves this issue.

[41] Müller et al., 2018

[252] Davrinche et al., 2025

[106] Bouabid, Souza, and Ferrari, 2026

These benchmarking results hold for end-of-century emulation (Appendix C.2.2), despite higher warming projections (2090-2100, Fig. C.2). While temperature and precipitation distributions are functionally equivalent to the mid-century in terms of their relative performance to pattern scaling, we observe slightly more pronounced differences in the standard deviations of relative humidity and wind speeds. The pattern scaling approach projects a larger increase in standard deviation of relative humidity over North America under warming; our emulator captures this signal with a reduced amplitude and smoother gradients. Additionally, the generative emulator predicts greater wind extremes over the Southern Ocean compared to pattern scaling, suggesting a difference in how the generative model (or the low-resolution parent ESM) extrapolates in this region.

In the context of these benchmarks, our emulator operates within a useful envelope of uncertainty, producing results that are statistically comparable to both ESM and pattern-scaled IGSM outputs. However, the generative approach offers the advantage of not requiring an EMIC run prior to emulation while additionally capturing forced shifts in internal variability. By generating physically consistent realizations of cross-correlated variables without this computational bottleneck, our framework enables uncertainty quantification (Section 3.2.2) and compound risk assessment (Section 3.2.3).

3.2.2 Statistical significance and scenario uncertainty

We use the generative emulator to assess whether grid-cell-level changes in climate anomalies are statistically significant at the end of century between the *Reference* and 1.5°C scenarios, along with the 2°C and 1.5°C scenarios; emissions and corresponding FaIR-simulated GMST are shown in Figure 3.2a and b. We focus on the most extreme months of the year, January and July, as these yield the clearest changes in our variables of interest. Figure 3.2c illustrates that the generative emulator reproduces key physical behaviors of the climate system, particularly the magnitude of internal variability relative to the forced anthropogenic signal for variables beyond temperature. At grid-cell scales, internal variability can mask the forced trend between scenarios for precipitation, relative humidity, and wind speeds, leading to statistically insignificant differences between scenarios; this effect is pronounced when comparing the low-warming scenarios. This behavior is physically consistent with other work describing the detection of both future forced climate responses and differences between emissions scenarios, which demonstrate that forced climate trends become detectable first in the tropics due to relatively low internal variability, whereas high-latitude signals are often masked by synoptic-scale noise[253–255].

[253] Mahlstein et al., 2011; [254] Hawkins and Sutton, 2012; [255] Tebaldi and Friedlingstein, 2013

Global spatial analysis

Changes in temperature anomalies between the *Reference* and 1.5°C scenarios are statistically significant across nearly all grid cells in both January (99.55% of global area) and July (99.46%) in 2100. The maximum change in temperature anomaly in January is roughly 6°C greater than in July, however. This aligns with the well-established phenomenon of Arctic amplification, which is most pronounced in winter months due to the release of accumulated heat from the Arctic Ocean[256]. Spatially explicit differences between the 2°C and 1.5°C worlds are less pronounced (50.97% of global area in January and 63.28% in July). Statistical significance over land correlates with the regional summer (e.g., July in North America, January in Central and Southern Africa), though other areas such as South America and Central Asia experience detectable changes regardless of season. In contrast, several oceanic regions, such as the Southern Ocean and the ENSO region, do not exhibit statistically significant differences between the two stricter policy scenarios. Broadly, our results show that significant temperature differences between scenarios are more likely to be detectable in the tropics than in extratropical regions. The tropics exhibit low internal variability due to consistent solar insolation and high thermal inertia[253,254], whereas extratropical regions are characterized by high internal variability driven by baroclinic wave activity and atmospheric teleconnections[257].

[256] Taylor et al., 2022

[253] Mahlstein et al., 2011; [254] Hawkins and Sutton, 2012

[257] Zhang, Liu, and Dong, 2024

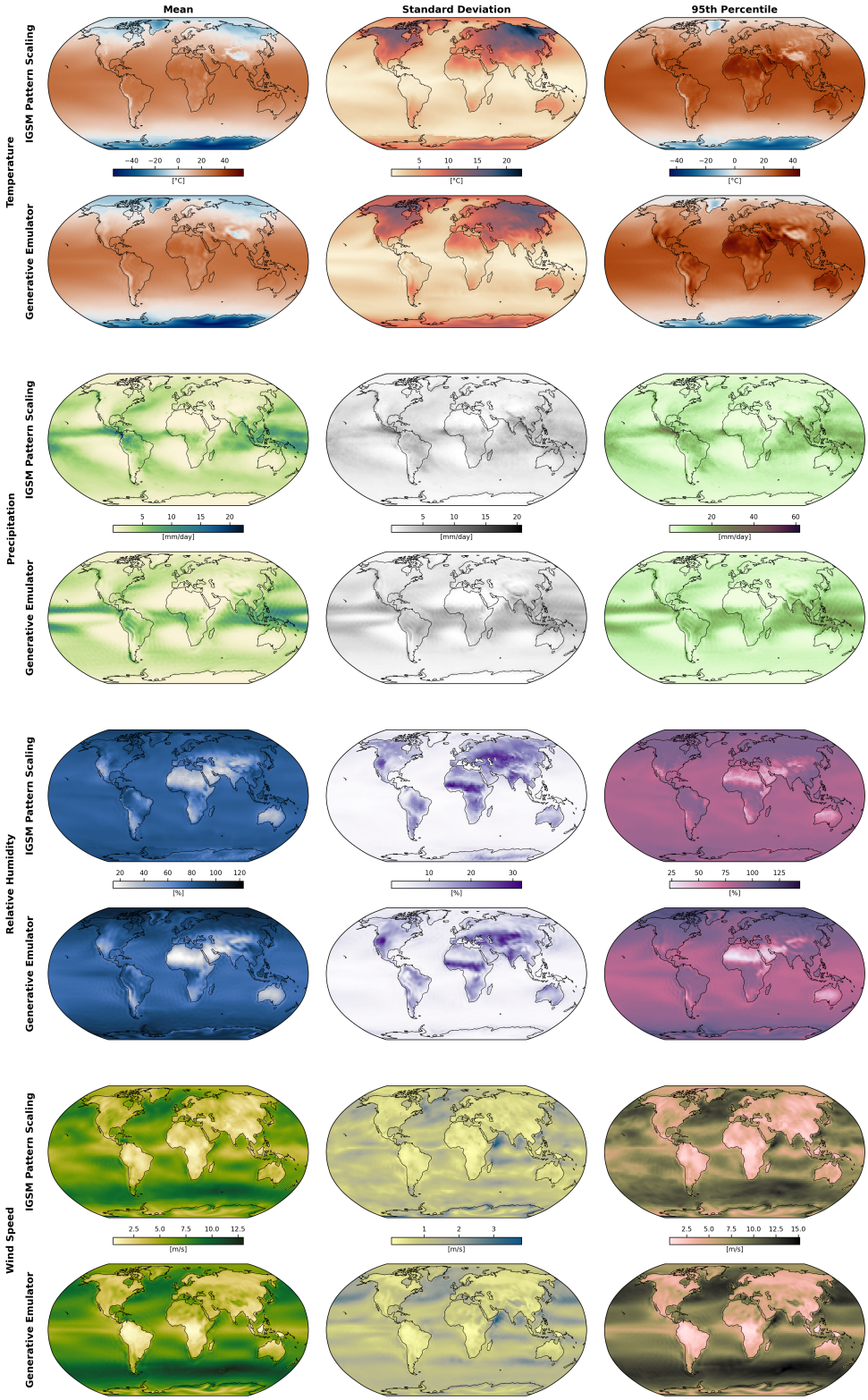

Figure 3.1: Mean, standard deviation, and 95th percentile of emulated climate fields with 50 ensemble members for the EPPA *Reference* scenario between 2040-2050 emulated with the MIT IGSM + pattern scaling (upper half of each row) and our generative emulator (lower half of each row). Climate fields are given from top to bottom as: near-surface air temperature, precipitation, relative humidity, and near-surface wind speed.

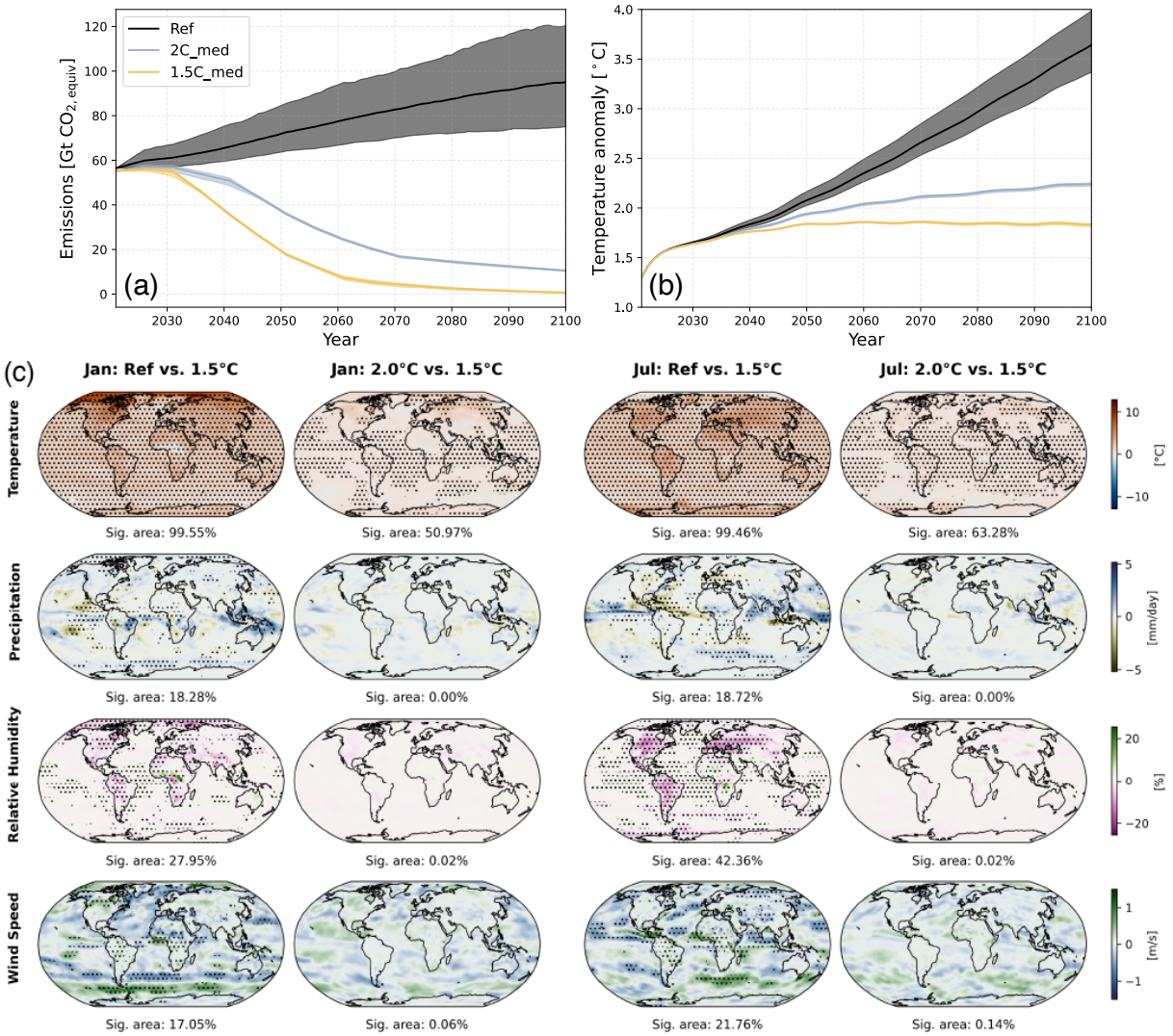

Figure 3.2: Global mean and spatial quantities for EPPA scenarios considered in Section 3.2.2. (a) Emissions (Gt CO_2-equivalent) trajectories for 400-member EPPA parameter ensemble; (b) global mean temperature anomaly (°C) corresponding to (a) as simulated by the FaIR SCM. Shaded regions indicate 95% confidence interval, while the solid line indicates the median trajectory. (c) Difference between mean climate anomalies in 2100 produced by the generative climate emulator over 100 realizations between the EPPA *Reference*, 2°C, and 1.5°C scenarios in January (left) and July (right). The left sub-column compares the *Reference* and 1.5°C scenarios, while the right sub-column compares the *Reference* and 2°C scenarios. Climate fields are given from top-to-bottom as near-surface air temperature, precipitation, relative humidity, and near-surface wind speed. Stippling shows statistically significant regional differences between scenarios and labels under each map indicate the percent of Earth's surface area with statistically significant changes.

For precipitation, relative humidity, and surface winds, statistically significant changes in 2100 are generally restricted to the comparison between the *Reference* and 1.5°C scenarios. For example, roughly 18% of the globe has statistically significant differences for precipitation between these scenarios, with no statistically significant differences between the low-warming scenarios. Similar to temperature, detection of the forced trend is more likely during the regional summer. Relative humidity exhibits the strongest signal among these variables, particularly in the Northern Hemisphere summer. However, when comparing the 1.5°C and 2°C worlds, almost no regions emerge with statistically significant changes for these variables. While minor shifts in the mean exist at the grid-cell level, they are negligible compared to the magnitude of internal variability. This null result aligns with IPCC AR6 findings and other recent work[258–260], which highlight that regional precipitation changes

[258] IPCC, 2023; [259] Tebaldi, O'Neill, and Lamarque, 2015; [260] Dai et al., 2024

are difficult to detect due to internal variability, with emergence often not occurring until late in the century, if at all, for low-warming scenarios[261].

[261] Schuhen et al., 2026

Regional uncertainty analysis

Figure 3.3 displays the distributions of absolute temperatures in 2100 across the *Reference*, 2°C, and 1.5°C scenarios for example locations in four distinct climatological zones: (a) Denali, Alaska; (b) Palau; (c) Bogotá, Colombia; and (d) the Kalahari Desert, Botswana. We derive these distributions by generating 250 realizations for each scenario and averaging the resulting temperature over each region, and constructing the full probability distributions of the monthly temperature outcomes.

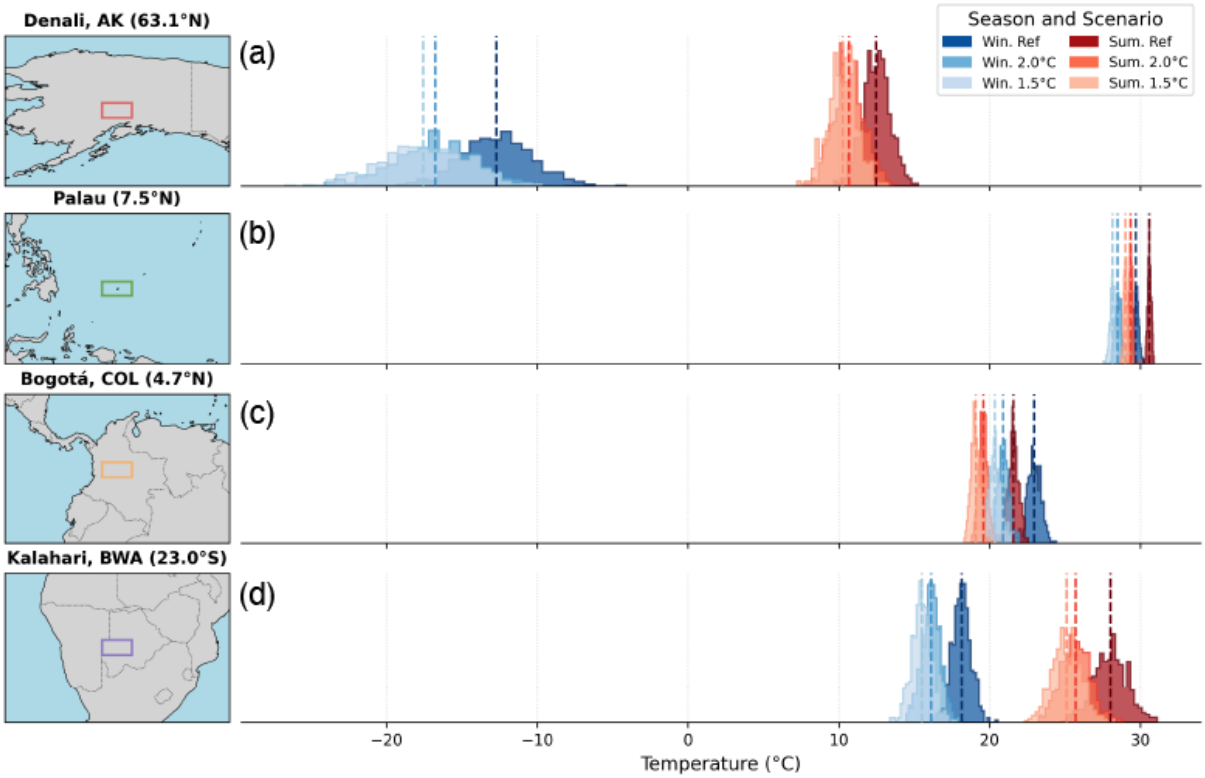

Figure 3.3: Frequency of occurrence of absolute temperature values across (a) Denali, Alaska; (b) Palau; (c) Bogotá, Colombia; and (d) the Kalahari Desert, Botswana. Distributions are given for winter (blue) and summer (red) for the EPPA *Reference*, 2°C, and 1.5°C scenarios; summer (winter) is JJA (DJF) for the Northern Hemisphere and inverted for the Southern Hemisphere. Dashed lines indicate distribution means.

The spatial distributions further highlight the role of internal variability in limiting our ability to detect policy impacts at a spatially explicit level. The two tropical locations, Palau and Bogotá, exhibit narrower distributions with less overlap between scenarios than the extratropical regions. This separation supports the observation that tropical locations see clearer forced signal trends due to lower internal variability[254]. Denali and the Kalahari Desert exhibit much broader distributions, indicating a high intra-seasonal standard deviation. Neither Denali nor the Kalahari Desert shows statistically significant separation in mean temperatures between the low-warming scenarios during their respective winter months. The spread of near-surface temperatures in winter is wider than in summer, obscuring the signal of the different scenarios. This offers an additional validation of the emulator's physical consistency, as previous work has shown that internal variability is maximized in extratropical winters[256].

[254] Hawkins and Sutton, 2012

[256] Taylor et al., 2022

In the Northern Hemisphere, this is driven by modes such as the Pacific-North American (PNA) pattern, which strongly modulates monthly temperatures[262]. In these locations, internal variability can mask spatially explicit differences between globally distinct scenarios.

[262] Leathers, Yarnal, and Palecki, 1991

3.2.3 CMIP7 compound impact assessment

A major use case of emulators is their ability to assess the potential outcomes of emissions scenarios in a fraction of the time required for full-scale ESMs. To that end, we utilize our generative emulator to assess potential outcomes for the approximate ScenarioMIP-CMIP7 scenarios prior to the availability of full-scale ESM simulations. While ESMs remain essential for validation, the emulator provides a first estimate of the spatial distribution of emerging risks. The emulator may exhibit biases in strong overshoot scenarios or when strong non-linear feedbacks that are not fully captured by conditioning solely on GMST occur (e.g., Arctic amplification post-2100). However, the goal of this assessment is not to provide precise forecasts, but to characterize the range of CMIP7 outcomes. Specifically, we focus on compound impact metrics to highlight the emulator's ability to generate spatially and cross-correlated realizations of multiple variables.

Heat stress and wet-bulb temperature

We utilize the Stull (2011)[134] approximation to calculate wet-bulb temperature (T_w) at each grid cell, excluding values that fall outside of the approximation's functional range. We generate 50 samples for each month from 2000-2300, conditioned on the annual GMST anomaly trajectory simulated by FaIR; as the GMST anomaly varies smoothly, we take five year time steps during the analysis period and interpolate the results. We track cumulative heat stress by identifying regions where the monthly average T_w exceeds a proxy baseline of 25°C. This threshold is used, as a sustained monthly mean of T_w implies a severe lack of nighttime thermal relief and acute daily spikes approaching human physiological limits (Appendix C.2.1).

[134] Stull, 'Wet-Bulb Temperature from Relative Humidity and Air Temperature', *Journal of Applied Meteorology and Climatology*, 2011

Figure 3.4 presents cumulative heat stress (measured by Wet-Bulb Degree-Days (WBDD) above the 25°C baseline proxy) for the *Medium-Low Extension* and *Low Extension* ScenarioMIP-CMIP7 scenarios. These scenarios achieve roughly the same GMST in 2300, but *Medium-Low Extension* exhibits a significant increase in temperature before decreasing, while *Low Extension* never achieves those temperatures in the first place. Regionally, both scenarios have the largest impacts in the same zones (Fig. 3.4c), with the largest differences between the two scenarios occurring in highly vulnerable regions such as South Asia and South East Asia, where the *Low Extension* pathway reduces cumulative heat stress by approximately 23% and 59%, respectively. While exceeding the baseline threshold implies threats to human habitability, cumulative regional changes do not indicate prolonged conditions in any single grid cell. These changes reflect the spatial expansion of dangerous conditions across broader geographic areas and their earlier onset during the emulated period.

The largest impacts are felt in South Asia, with the *Medium-Low Extension* scenario leading to an average of roughly 53 WBDD greater than 25°C per year, though this number is likely much higher near the end of the emulated period. Our findings align with recent work highlighting

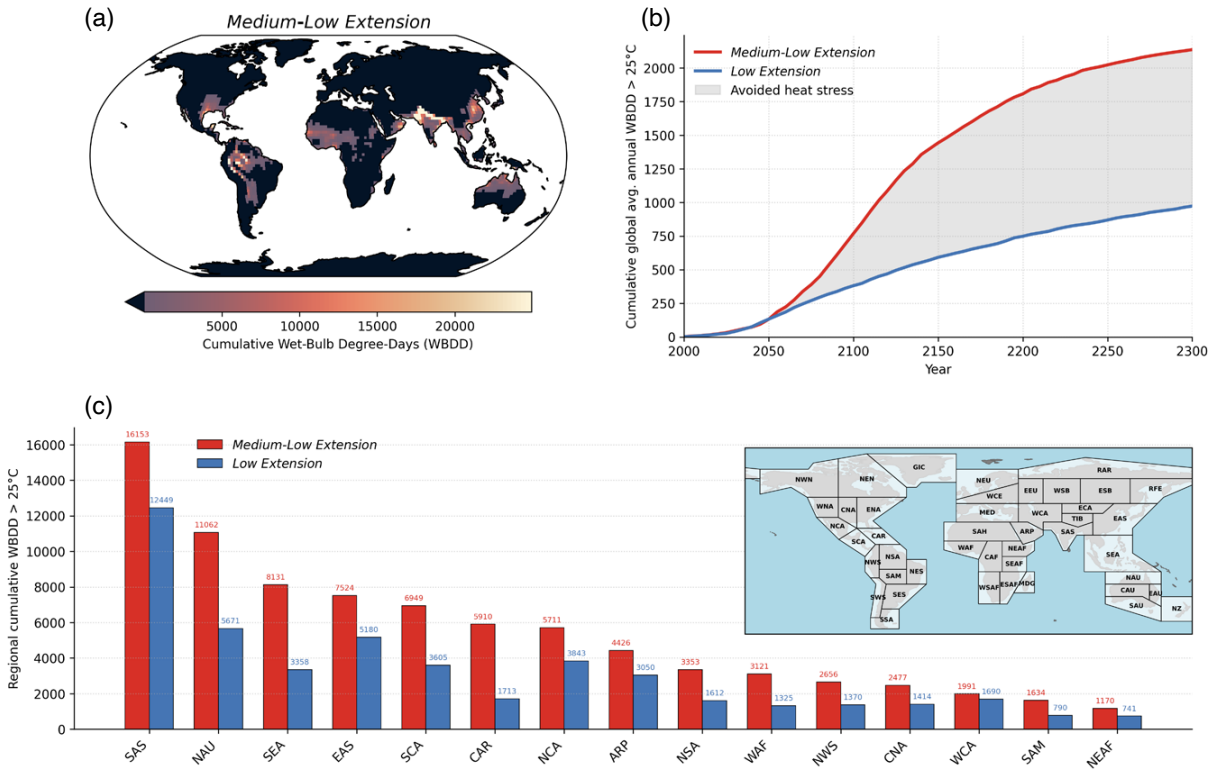

Figure 3.4: Heat stress measured by Wet-Bulb Degree-Days (WBDD) for the approximate *Medium-Low Extension* and *Low Extension* ScenarioMIP-CMIP7 emissions scenarios projected by the generative emulator. (a) Cumulative spatially explicit heat stress for the *Medium-Low Extension* scenario in 2300. (b) Globally averaged cumulative heat stress across both scenarios; shaded area indicates heat stress avoided under the *Low Extension* scenario. (c) Cumulative WBDD in 2300 across both scenarios for the fifteen IPCC AR6 regions for which the generative emulator projects the largest impacts; inset map indicates AR6 regions.

the expansion of heat stress in the Indus River Valley (SAS) and sub-Saharan Africa, e.g., W. Africa (WAF) and N.E. Africa (NEAF), under future warming scenarios[263,264]. Our results (Fig. 3.4b) additionally support findings that suggest that limiting global mean warming by 0.5°C (roughly the difference between these scenarios in 2100) can reduce heat stress exposure by up to a factor of two[265]. While interior continental regions often exhibit a negative correlation between warming and relative humidity due to land surface desiccation (e.g., compare temperature and relative humidity, Figure 3.2c), which partially mitigates the rise in wet-bulb temperature, coastal and river-basin environments experience an increase in specific humidity alongside global temperature increases (following the Clausius-Clapeyron relation)[266].

[263] Coffel, Horton, and Sherbinin, 2018;
[264] Vecellio et al., 2023

[265] Saeed, Schleussner, and Ashfaq, 2021

[266] Willett et al., 2007

Figure 3.5 illustrates distributional differences in wet-bulb temperatures between the three regions projected by the generative emulator to experience the greatest amount of heat stress (SAS, NAU, and SES) across all ScenarioMIP-CMIP7 Priority 1 scenarios in 2100. All three regions clearly reflect the global mean trend of warming, with wet-bulb temperatures increasing the most in the *High* scenario and the least in the *Very Low with Limited Overshoot* scenario. In the *High* scenario, monthly mean wet-bulb temperatures in South Asia can exceed 26°C, while both the *Medium* and *Medium-Low* scenarios exceed 25°C. In contrast, Southeast Asia is projected only to exceed the 25°C threshold by 2100 in the highest warming scenario. This suggests the bulk of the excess heat stress identified in Figure 3.4 for Southeast Asia primarily occurs after 2100, highlighting the delayed emergence of compound risks in certain regions.

Both South Asia and Southeast Asia exhibit a lower standard deviation in wet-bulb temperatures than Northern Australia. The larger range of possible outcomes in Australia is likely due to the region containing both tropical and grassland climates, with the high relative humidity in the tropical areas leading to the higher end of wet-bulb temperature outcomes.

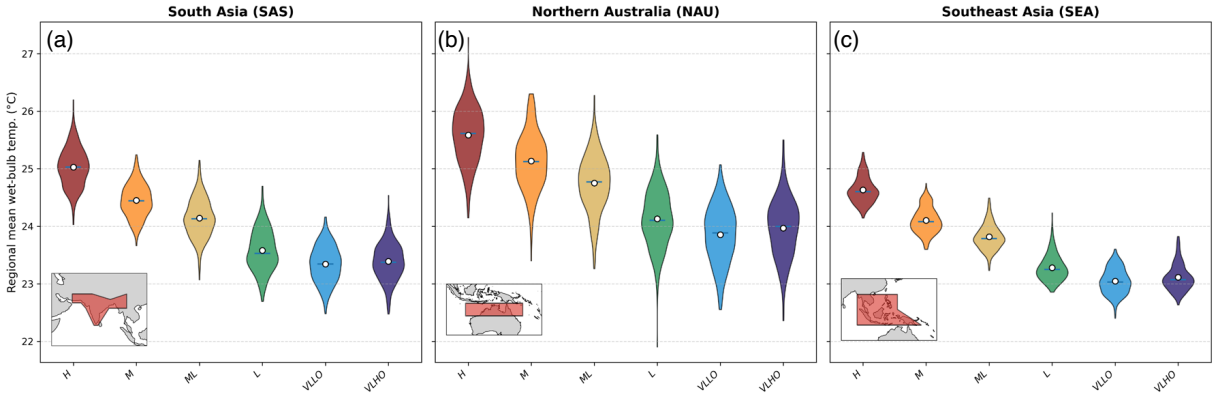

Figure 3.5: Wet-bulb temperature distribution in 2100 for all approximate ScenarioMIP-CMIP7 scenarios across the three IPCC regions most vulnerable to heat stress as projected by the generative emulator (Fig. 3.4): (a) South Asia, (b) Northern Australia, and (c) Southeast Asia. Dots indicate mean wet-bulb temperatures, while lines indicate median.

Vapor pressure deficit

To assess future fire danger, we utilize our spatially and cross-correlated variables to compute VPD, which is highly correlated with wildfire occurrence as it drives transpiration and fuel desiccation[135,267]. Figure 3.6 highlights changes in VPD between 2050 and 2100 under the *High* and *Very Low after High Overshoot* scenarios for the summer (JJA) and fall (SON) seasons in the US. We observe absolute increases of up to 1 kPa between 2050 and 2100 in the *High* emissions scenario, particularly in the Southwest US. However, because many arid ecosystems in this region are fuel-limited rather than climate-limited, increases in atmospheric aridity may not correspond to proportional increases in burned area, as these landscapes often lack the continuous biomass necessary to sustain large fires[268]. Conversely, in forested or climate-limited regions, given the nonlinear relationship between atmospheric water demand and wildfire behavior, a monthly average increase of 1 kPa represents a large amplification in the risk of exceeding dangerous daily fire probability thresholds; biome-specific critical thresholds may range from 1.3 to 2.7 kPa[240]. In contrast, the *Very Low after High Overshoot* scenario effectively mitigates this increase, leading to little change in VPD over time.

[135] Williams et al., 2019; [267] Grossiord et al., 2020

[268] Littell et al., 2009

[240] Clarke et al., 2022

Figure 3.7 highlights the change in VPD over time in the *High Extension* scenario over several North and Central American regions: (a) Western North America, (b) Central North America, (c) Eastern North America, and (d) Northern Central America. We observe a clear, upward trend in VPD across all regions that correlates closely with the increase in GMST between 2000-2300 in this scenario. This aligns with broader CMIP6 projections of intensifying atmospheric aridity, particularly across the western United States[269].

[269] Zhuang et al., 2021

In Western North America (Fig. 3.7a), VPD increases from approximately 1.25 to 2.75 kPa. This shift crosses dangerous thresholds for daily fire probability for both temperate (1.3 kPa) and Mediterranean (2.3 kPa)

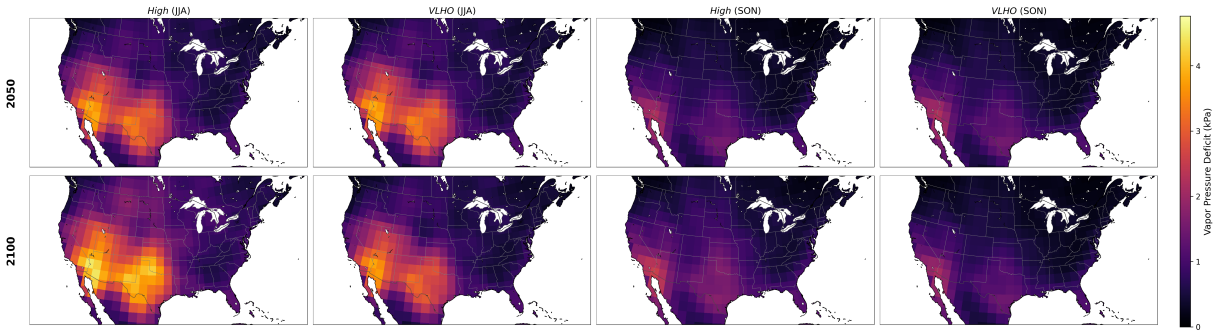

Figure 3.6: Vapor pressure deficit in the summer (JJA) and fall (SON) over the continental United States in 2050 and 2100 compared across the *High* and *Very Low after High Overshoot* approximate ScenarioMIP-CMIP7 scenarios. Increased VPD values correlated with increased fire danger, particularly in the Western United States.

biomes[240]. Because the forested portions of this region are historically climate-limited[270], crossing these thresholds represents a sizable amplification of actual fire risk. Central North America (Fig. 3.7b) displays a similar pattern of increase, rising from roughly 1.0 to 2.75 kPa with the highest magnitude of variability across all four regions. While this region also crosses the same thresholds, much of it is dominated by Great Plains grasslands. As these ecosystems are fuel-limited, fires are constrained by the accumulation of biomass rather than purely by atmospheric aridity[268]; the actual increase in fire danger is limited by fuel availability.

[240] Clarke et al., 2022

[270] Kampf et al., 2025

[268] Littell et al., 2009

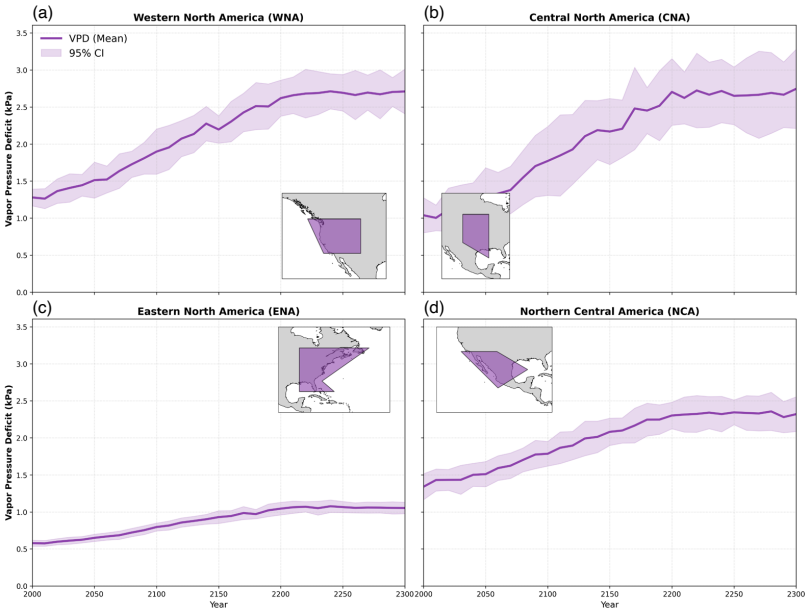

Figure 3.7: Summer (JJA) average vapor pressure deficit time series for the *High* approximate ScenarioMIP-CMIP7 scenario over (a) Western North America, (b) Central North America, (c) Eastern North America, and (d) Northern Central America. Solid line indicates the mean, while the shaded region indicates the 95% confidence interval.

In contrast, Eastern North America and Northern Central America (Fig. 3.7c and d) exhibit lower overall increases and remain constrained by their respective biomes. In Eastern North America, VPD rises from approximately 0.5 to 1.0 kPa, representing the lowest mean increase due to higher regional relative humidity. This trajectory remains below the 1.3 kPa threshold for temperate forests[240]. Because these forests are climate-limited, remaining below this threshold suggests a mitigation of fire risk compared to western forests. Finally, in Northern Central America, VPD rises from roughly 1.4 to 2.25 kPa. While this increase is statistically significant, changes in fire risk are likely to be disparate across the region; Northern Central America has both fuel-limited deserts, along with the climate-limited Temperate Sierra and tropical dry forests.

[240] Clarke et al., 2022

Fire risk is unlikely to increase significantly for the former, but could be highly amplified for the latter. By representing these disparate regional outcomes along with the bounds of internal climate variability, our generative framework can provide decision-makers with the probabilistic, spatially resolved data required to evaluate policies targeting future fire risk.

3.3 Discussion and conclusions

Current approaches to climate impact assessment often face a trade-off between computational cost and resolution. Physics-based Earth System Models (ESMs) provide detailed projections at spatially explicit scales, but are too computationally expensive to run large ensembles or explore a wide variety of policy scenarios. Conversely, Simple Climate Models (SCMs) and Earth system Models of Intermediate Complexity (EMICs) are computationally efficient but lack the spatial resolution required for spatially explicit impact analysis, often necessitating additional downscaling or pattern-scaling emulation to derive spatially explicit climate fields. In this work, we present a climate impact assessment pipeline to bridge this gap by coupling the MIT EPPA model and FaIR with a score-based generative emulator enabling efficient computation of spatially explicit climate variables. This approach differs from traditional mean-state pattern scaling emulation and other statistical emulators by explicitly modeling the full joint probability distribution of several climate variables. The emulator effectively captures the mean, standard deviation, and extremes of climate anomalies at grid-cell resolution, allowing for the rapid generation of spatially and cross-correlated fields. This pipeline enables rapid projection and uncertainty quantification of spatially explicit compound climate impacts for any custom scenario or ensemble of scenarios.

Benchmarking against the established IGSM ensemble pattern scaling approach[78] demonstrates that the generative emulator operates within a useful envelope of uncertainty, successfully reproducing climatological states and large-scale atmospheric structures for both mid- and end-of-century projections (Fig. 3.1). While structural differences arise, particularly around orographic effects, this is primarily due to differences between the high- and low-resolution MPI models used in training the pattern scaling and generative approaches, respectively[41,78,106,235]. The generative framework offers a distinct operational advantage, however, by decoupling the generation of regional fields from the requirement of EMIC simulations. While it requires GMST values generated from tools like FaIR, the generative emulator significantly lowers the technical barrier to entry. Our interface makes it possible for a researcher with no expertise in running ESMs or EMICs to generate physically consistent, spatially correlated climate realizations in minutes. By democratizing access to these data and removing the computational bottlenecks of generating spatially explicit outputs, the emulator facilitates a more comprehensive assessment of uncertainty and compound risks than is feasible with computationally expensive ESM ensembles.

[78] Gao, Sokolov, and Schlosser, 2023

[41] Müller et al., 2018; [78] Gao, Sokolov, and Schlosser, 2023; [106] Bouabid, Souza, and Ferrari, 2026; [235] Schupfner et al., 2021

Distinct global climate policies (e.g., 1.5°C vs. 2°C targets) often produce statistically indistinguishable spatially explicit outcomes in specific locations, exposing the potential for mitigation strategies to be perceived as a failure. While tropical regions with lower internal variability exhibit clearer scenario separation[253,255], higher variability in the extratropics

[253] Mahlstein et al., 2011; [255] Tebaldi and Friedlingstein, 2013

frequently obscures the forced signal of mitigation efforts (Fig. 3.2c), particularly for variables such as precipitation, relative humidity, and wind at monthly frequencies. Moreover, even when statistically significant differences exist in the mean state between large ensembles, the full probability distributions often exhibit substantial overlap. Because we ultimately experience only a single climate realization, a detectable shift in the mean does not guarantee a noticeably different climate outcome for an individual community. This reality presents a significant climate communication challenge: if the benefits of mitigation do not manifest as distinct changes in local, near-term patterns, there is an inherent risk of perceived policy failure and subsequent loss of public support[271,272]. This underscores the necessity of a probabilistic risk-management framework that integrates the full distribution into the decision-making process, rather than relying solely on shifts in the ensemble mean.

[271] Shao et al., 2016; [272] Keys et al., 2022

Applying this generative framework to the approximate ScenarioMIP-CMIP7 scenarios demonstrates that emulators can serve as efficient scenario screening tools, highlighting that end-point temperature targets are insufficient for fully characterizing climate risk. By generating physically consistent, spatially and cross-correlated climate fields, our emulator moves beyond single-variable emulation to the assessment of compound impact-relevant metrics, such as wet-bulb temperature and VPD. We observe that an overshoot scenario leads to drastically different cumulative heat stress than a low-warming scenario, despite achieving similar end-point temperatures. Because heat stress is non-linear, even a transient 0.5°C difference may significantly alter cumulative risk. Consequently, overshoot pathways—which often rely on unproven negative emissions technologies—carry a much higher burden of potential climate damages and adaptation costs than pathways that abate emissions early. Furthermore, our finding that distinct global emissions pathways often yield statistically indistinguishable spatially explicit impacts emphasizes the importance of ongoing discussions within the climate modeling community around how far apart emissions scenarios must be in terms of global radiative forcing to justify the computational expense of an ESM[126,259]. With this generative framework, researchers can proactively evaluate whether proposed policy targets warrant full-scale simulations before committing computational resources.

[126] Van Vuuren et al., 2026; [259] Tebaldi, O'Neill, and Lamarque, 2015

While this generative framework offers several advantages over other emulation techniques, there are several areas for improvement. As the emulator is trained from an ESM, it inherits any structural biases present in the parent model. The emulator being trained on CMIP6 data means that the approximate ScenarioMIP-CMIP7 pathways represent an out-of-sample projection. This limits our analysis, as data-driven emulators may fail to capture shifts in climate dynamics under novel forcing regimes, particularly under the long-term extension scenarios. As a result, our CMIP7 projections should be interpreted as probabilistic screening tools rather than forecasts. As new ESM ensembles become available for CMIP7, future work can validate the generative emulator against these data to better quantify its extrapolative reliability.

Several steps remain to fully integrate emulators into the impact assessment pipeline. First, increasing emulator temporal resolution to daily or hourly intervals, while retaining spatial coherence across variables, is necessary to accurately capture temporally localized compound risks, such as the Fosberg Fire Weather Index[273]. Approaches like DiffESM are promising for generating daily samples from monthly data[105], and future work can explore integrating daily emulation directly into this

[273] Deeming, Burgan, and Cohen, 1977

[105] Bassetti et al., 2024

generative framework. Second, emulators must be able to capture both temporal and spatial system memory. Generating temporally coherent data, rather than relying solely on instantaneous fields, is necessary to assess risks dependent on duration, such as drought propagation. Additionally, recent work demonstrates that the climate system exhibits spatial memory, where spatially explicit temperature responses may differ from the global mean—particularly during overshoot scenarios[2,91]. Conditioning the emulator on rates of GMST change or historical forcing trajectories may be necessary to accurately assess localized risks. In addition, the emulator's outputs must be utilized to quantify the economic impacts of climate damages, for example, through climate damage functions[274–276]. Finally, an important future goal is the fully coupled representation of human-natural system feedbacks, where the emulator is directly coupled to an economic model. For instance, economic growth drives emissions, which increases wet-bulb temperatures; this, in turn, reduces labor productivity and dampens economic output. Similarly, physical feedbacks could limit adaptation efficiency, such as increased VPD and fire danger hindering the mitigating effects of afforestation.

[2] Womack et al., 2026; [91] Womack et al., 2025

[274] Diaz and Moore, 2017; [275] Neumann et al., 2020; [276] Waidelich et al., 2024

Summary and future work

Big things have small beginnings, sir.

— Mr. Dryden, *Lawrence of Arabia*

This work has focused on bridging the gap between basic and applied research in the context of developing efficient computational methods for emulating the statistics of Earth System Models (ESMs). While the design of accurate surrogate models is of great importance to many domains, the Earth system presents a unique challenge in that it is extremely data-limited relative to other fields. Instead of hindering us, this limitation pushes the bounds of our creativity, requiring novel advancements in the theory, operation, and application of climate emulators to fully realize their potential. In the following section, we outline the contributions of this work to the fields of computational and climate science, which provide fundamentally new tools to understand and design these systems.

Contributions

1. Under what conditions do structural assumptions cause emulators to fail, and what trade-offs emerge across different emulation techniques?

- *Establishment of a theoretical framework for emulator methodological comparisons*

The problem of climate emulation is not new, and over time, many efforts have been made to understand the strengths and weaknesses of these methods. However, it is not enough to simply benchmark an emulator's projections against model output, as high skill on one test dataset does not necessarily translate to an emulator that can extrapolate to any possible scenario configuration. We instead need to look deeper, exploring the fundamental assumptions that make up the mathematics and physics behind our emulators to ensure the tools we develop are not only fast, but trustworthy.

To this end, we use principles of statistical mechanics and stochastic calculus to place a number of disparate emulation techniques along a common theoretical spectrum. This marks the first concerted effort to analyze a broad set of climate emulators from a theoretical perspective, and our work critically identifies a missing link in the typical emulator typology: operator-based emulators. Despite overlapping goals with traditional emulation approaches—such as identifying modes of variability within a complex system—operator approaches have yet to be fully explored in the context of emulation. We additionally provide practical implementation details for these approaches that are under-utilized in the geosciences, aiming to bring together communities that otherwise would not have connected despite similar aims.

Developing a theoretical framework for climate emulation additionally enables comprehensive error analyses for existing techniques. We focus on memory effects, hidden variables, system noise, and nonlinearities as potential sources of emulator error, demonstrating that common methods like pattern scaling contain irreducible structural errors. Our framework gives us a systematic method to assess and improve emulation techniques independently of ESM results, ensuring emulators are prepared to train on new results as ESMs continue to improve through CMIP7 and beyond.

2. To what extent do varying scenario structures constrain or enhance an emulator's predictive skill?

– *Development of a training data generation methodology to maximize emulator predictive skill*

One key finding from our first exploration is that ScenarioMIP—the standard set of scenarios typically used to train climate emulators—may not actually be the most efficient choice of data to train an emulator. Exponential forcings can cause the system's response to reduce to a single timescale, effectively obscuring the full set of system behaviors an emulator would need to learn to accurately extrapolate to unseen forcings. This immediately raises the question: *what is the best possible set of training data?* Furthermore, even if we can identify this optimal dataset, we are still fundamentally limited in our ability to run new scenarios; our optimal dataset may not be useful if it requires running an ESM dozens or hundreds of times.

To alleviate this issue, we introduce an algorithm to generate optimal training datasets, using a computationally efficient Simple Climate Model (SCM) to identify those datasets. We treat the training trajectory as a set of tunable parameters rather than assuming the trajectory is fixed (i.e., ScenarioMIP), and leverage a differentiable SCM to backpropagate through the testing, training, and data generation processes. This gives the sensitivity of the test loss with respect to the training data, which we then use to iteratively update the initial emissions trajectory to maximize emulator predictive skill. We demonstrate the utility of this approach on both simple and intermediate-complexity climate models, with emulators trained on just one or two of our optimal scenarios outperforming a baseline emulator trained on six ScenarioMIP-CMIP7 scenarios.

Our results highlight that scenarios characterized by high structural diversity and high-frequency variations are better suited for emulator training and understanding system behavior than baseline scenarios. While not fully interpretable due to the use of a neural network, the improved extrapolative capability of our optimized emulators suggests our optimized scenarios better highlight the physically consistent behavior required to build a robust surrogate model for a physical system. In light of our results, modeling centers could consider dedicating resources to generate simulation data explicitly designed for machine learning. More broadly, establishing a Model Intercomparison Project (MIP) for emulator development would benefit both the climate modeling and impacts communities by producing robust emulators capable of generating large, impact-relevant ensembles in a fraction of the time of traditional ESMs.

3. How can operationalizing emulators enhance the assessment of impact-relevant metrics?

- *Demonstration of the practical utility of a generative emulator for assessing compound climate hazards*

In order to effect real change with our work, we cannot stop at the development of these techniques; uploading a project to GitHub or even publishing in a peer-reviewed journal does not equate to the adoption of our techniques. Our final contribution tackles this issue by developing a lightweight pipeline mapping from an economic model (Emissions Prediction and Policy Analysis (EPPA)) to spatially explicit climate outcomes. By using a generative emulator, we capture the spatial and cross-correlations necessary for assessing future compound climate risks.

We benchmark our approach against one existing method for projecting regional climate outcomes from economic model output: the MIT Integrated Global Systems Model (IGSM). Our approach is statistically consistent with both the parent ESM and IGSM emulator, lending further credibility to its utility in the context of impact assessment. The generative framework offers a distinct operational advantage, however, by decoupling the generation of regional fields from the requirement of Earth system Model of Intermediate Complexity (EMIC) simulations. Our interface additionally makes it possible for a researcher with no expertise in running ESMs or EMICs to generate physically consistent, spatially correlated climate realizations in minutes. By democratizing access to these data and removing the computational bottlenecks of generating spatially explicit outputs, the emulator facilitates a more comprehensive assessment of uncertainty and compound risks than is feasible with computationally expensive ESM ensembles.

Our emulation pipeline moves beyond single-variable emulation to the assessment of compound impact-relevant metrics, such as wet-bulb temperature and Vapor Pressure Deficit (VPD). Applying this framework to the approximate ScenarioMIP-CMIP7 scenarios highlights that end-point temperature targets are insufficient for fully characterizing climate risk. We observe that an overshoot scenario, despite achieving similar end-point temperatures to a low-warming scenario, leads to drastically different cumulative heat stress. This suggests that the approximate ScenarioMIP-CMIP7 overshoot pathways, which often rely on unproven negative emissions technologies, carry a much higher burden of potential climate damages and adaptation costs than pathways that abate emissions early. Furthermore, our results demonstrate that distinct global emissions pathways may produce local climate outcomes that are statistically indistinguishable at monthly frequencies. This highlights ongoing discussions within the climate modeling community around how far apart emissions scenarios must be in terms of global radiative forcing to justify the computational expense of running them through an ESM. By generating full spatial probability distributions, our emulator provides a timely, efficient screening tool. It allows researchers to proactively evaluate whether proposed policy targets will yield statistically distinguishable regional impacts before committing resources to ESM simulations.

Future work

The work is never done, and this work is no exception. With each contribution in this thesis, myriad possibilities for future research have emerged. We outline several of the most promising areas for immediately extending this work here.

Methodological explorations

A logical next step from Chapter 1 is to conduct further research to determine if operator-based methods like Dynamic Mode Decomposition (DMD) and Extended DMD (EDMD) can be practically realized to emulate nonlinear processes in full-scale ESMs directly. This can be explored in tandem with hybrid efforts to improve purely data-driven techniques (e.g., the generative approach applied in Chapter 3) by incorporating physically grounded approaches like the Fluctuation Dissipation Theorem (FDT). In a similar vein, we can also investigate whether operator-based emulators can be used to explicitly learn the true underlying parameters of the climate system, offering utility for system understanding beyond just emulation.

Chapter 2's optimization algorithm is also ripe for additional experimentation, particularly with respect to its performance when applied to other machine learning architectures. We can investigate whether the optimal scenarios derived in Chapter 2 are useful for other architectures, or how significant architectural changes affect the optimization results. It would also be useful to explore how ensemble machine learning techniques like boosting can be used to sequentially generate optimal features and resolve destructive interference when training on scenarios with extreme structural differences.

Operational scaling and experimental design

Prior to a true emulator-development MIP, any efforts to extend the work from Chapter 2 should focus first on scaling the approach to a differentiable EMIC. This will allow us to evaluate how optimizing for performance over additional climate variables (e.g., precipitation and relative humidity) or on different timescales (e.g., daily or monthly) changes the optimal emissions trajectory. Similarly, we can investigate how the operator-based approaches from Chapter 1, along with the FDT, can be applied to EMICs before attempting to scale these computationally expensive techniques to full ESMs. Either simultaneously with these efforts or at a later time, we can use the optimized trajectories derived in Chapter 2 as actual forcing inputs for a full-scale ESM to stress-test the model and evaluate emulator performance against standard protocols.

Emulator-driven impact assessment

To fully integrate emulators into future impact assessment efforts, several clear steps remain. Expanding the generative framework discussed in Chapter 3 to produce daily or even hourly data would enable us to capture more granular, temporally localized compound risks, such as fire weather indices. This would pair well with efforts to move beyond producing instantaneous realizations, incorporating autoregressive structures to model risks that depend on system memory and event duration, such as the propagation of droughts. Because the main benefit of using an emulator over an EMIC/ESM for impact assessment is its computational efficiency, we can additionally explore the development of a fully coupled representation of human and natural system feedbacks. By linking the

emulator directly to an economic model—through damage functions or other methods to link metrics like heat stress to labor productivity—we can progress towards a fully integrated system, wherein physical feedbacks are used online to deepen our understanding of nature-society interactions.

Appendices

A

Appendices for Chapter 1

A.1 Additional derivations

A.1.1 Pattern scaling errors

To understand the potential sources of error in pattern scaling, we start from the linear equation for temperature evolution

$$\frac{\partial}{\partial t}T(\boldsymbol{x},t) = \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')T(\boldsymbol{x}',t) + P(\boldsymbol{x})F(t), \tag{A.1}$$

where $T(\boldsymbol{x},t)$ is the spatially explicit temperature, $\mathcal{K}(\boldsymbol{x},\boldsymbol{x}')$ is the Koopman operator that governs the autonomous system dynamics, $P(\boldsymbol{x})$ is the spatial forcing pattern, and $F(t)$ is the time series of the forcing.

We can examine errors in pattern scaling by considering the case in which the pattern scaled temperature, $T_{PS}(\boldsymbol{x},t)$, is trained using an exponential forcing, $F(t) = e^{t/\tau}$, where τ indicates the growth rate of the exponential. Forcing our governing equation with this yields

$$T_{PS}(\boldsymbol{x},t) = \left[\frac{1}{\tau}\delta(\boldsymbol{x}-\boldsymbol{x}') - \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')\right]^{-1} P(\boldsymbol{x}')e^{t/\tau}. \tag{A.2}$$

Here $\delta(\boldsymbol{x}-\boldsymbol{x}')$ is the Dirac delta, so $\frac{1}{\tau}\delta - \mathcal{K}$ plays the role of $\frac{1}{\tau}I - \mathcal{K}$ in discretized form; we assume τ lies outside the spectrum of $\mathcal{K}$ so the inverse exists. Factoring out the exponential from this expression leaves us with

$$a_1(\boldsymbol{x}) = \left[\frac{1}{\tau}\delta(\boldsymbol{x}-\boldsymbol{x}') - \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')\right]^{-1} P(\boldsymbol{x}'). \tag{A.3}$$

$a_1(\boldsymbol{x})$ is therefore the spatial scaling pattern used as our emulator. Inserting $T(\boldsymbol{x},t) = a_1(\boldsymbol{x})F(t)$ into the governing equation with the same exponential forcing leaves us with

$$\frac{1}{\tau}a_1(\boldsymbol{x}) = \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')a_1(\boldsymbol{x}') + P(\boldsymbol{x}). \tag{A.4}$$

This identity expresses how the pattern, $a_1(\boldsymbol{x})$, balances internal dynamics with an external forcing.

We now consider an alternate scenario with an arbitrary forcing, F_{alt}, that is not the exponential forcing used for training. We denote the error between the true solution and our emulator as

$$T'(\boldsymbol{x},t) = T_{alt}(\boldsymbol{x},t) - a_1(\boldsymbol{x})F_{alt}(t). \tag{A.5}$$

We then recognize that $T_{alt}(\boldsymbol{x},t) = T'(\boldsymbol{x},t) + a_1(\boldsymbol{x})F_{alt}(t)$. Inserting this into our governing equation and using the identity from Equation A.4

gives an equation describing the evolution of errors over time

$$\frac{\partial}{\partial t}T'(\boldsymbol{x},t) = \mathcal{K}(\boldsymbol{x},\boldsymbol{x}')T'(\boldsymbol{x}',t) + \frac{1}{\tau}a_1(\boldsymbol{x})F_{alt}(t) - a_1(\boldsymbol{x})\frac{\partial}{\partial t}F_{alt}(t) \quad \text{(A.6)}$$

From this expression, we see that there are two distinct sources of error in pattern scaling when trained on an exponential (ScenarioMIP-like forcing). The first corresponds to an equilibrium-offset. If $F_{alt}(t)$ asymptotes to a constant F_f, the time derivative in Equation A.6 vanishes, leaving us with

$$\lim_{t\to\infty} T'(\boldsymbol{x},t) = -\frac{1}{\tau}\mathcal{K}^{-1}(\boldsymbol{x}',\boldsymbol{x})a_1(\boldsymbol{x})F_f. \quad \text{(A.7)}$$

Since we assume $\mathcal{K}^{-1}$ exists, there does not exist a non-zero vector such that $\mathcal{K}^{-1}(\boldsymbol{x}',\boldsymbol{x})a_1(\boldsymbol{x}) = 0$. Therefore the temperature produced by pattern scaling does not perfectly match the true equilibrium pattern.

The second source of error occurs in the transient case. When $F_{alt}(t)$ varies in time, the final term in Equation A.6 does not go to zero. If $F_{alt}(t)$ changes more quickly than the training growth rate (i.e., $\frac{\partial F_{alt}(t)}{\partial t} > \frac{1}{\tau}F_{alt}(t)$), then pattern scaling under-predicts the true temperature change. Conversely, very slow changes in $F_{alt}(t)$ lead to an over-prediction of the true temperature change. A non-negligible rate of change term signals that system memory will be significant in that scenario.

Physically, the first error arises because the system's equilibrium pattern depends on its slow internal modes, whereas the second arises because those modes cannot keep pace with forcing that accelerates faster (or slower) than the training rate τ.

A.1.2 Deconvolution instabilities

Deconvolution can amplify noise or in the worst case, cause the response function to blow up entirely. Here we identify where those instabilities arise. While issues with deconvolution are apparent in the time domain, they are easier to diagnose in frequency space. We use the Fourier transform (denoted by $\mathcal{F}$) to rewrite convolution as multiplication:

$$\mathcal{F}\left[g(w_t)\right] = \mathcal{F}\left[\int_{-\infty}^{\infty} d\tau\, R(\boldsymbol{x},\tau)F(t-\tau)\right] \quad \text{(A.8)}$$

$$\hat{g}(w_\omega) = \hat{R}(\boldsymbol{x},\omega)\hat{F}(\omega), \quad \text{(A.9)}$$

where $g(w_t)$ is our statistical quantity of interest, $R(\boldsymbol{x},t)$ is the response function, $F(t)$ is the forcing, the hat denotes the (continuous-time) Fourier transform, and ω is the angular frequency. Recovering the response function therefore becomes division:

$$\hat{R}(\omega) = \frac{\hat{g}(w_\omega)}{\hat{F}(\omega)}, \quad \text{(A.10)}$$

$$R(t) = \mathcal{F}^{-1}\left[\frac{\hat{g}(w_\omega)}{\hat{F}(\omega)}\right], \quad \text{(A.11)}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. In discrete space, we use the fast Fourier transform.

If $\hat{F}(\omega)$ has any near-zero frequencies, dividing by it causes $\hat{R}(\omega) \to \infty$ at those frequencies. The corresponding time-domain process requires

an explicit matrix inverse, where small eigenvalues translate into an ill-conditioned matrix. Additionally, if $|\hat{F}(\omega)|$ spans several orders of magnitude, the ratio $\hat{g}(w_\omega)/\hat{F}(\omega)$ amplifies high-frequency measurement noise and round-off error. The condition number of the corresponding matrix becomes very large, yielding an unstable estimate of $R(\boldsymbol{x}, t)$.

These issues are also encountered in signal processing, where a system is said to lack a spectral inverse (i.e., zeros in the frequency domain) if it exhibits the above issues[277,278]. Even in the absence of noise, the relatively flat spectrum of a true impulse response makes it difficult to recover directly. A dominant eigenvalue can obscure the weaker ones.

[277] Yeung and Kong, 1986; [278] Zazula and Gyergyek, 1993

A.1.3 Distinction between Green's and response functions

A scalar field, $w(t)$, governed by the linear time-invariant equation

$$\frac{\partial}{\partial t} w(t) = \mathscr{L} w(t) + F(t), \tag{A.12}$$

has a corresponding Green's function, $G(t)$, that solves

$$\frac{\partial}{\partial t} G(t) = \mathscr{L} G(t) + \delta(t), \quad G(t < 0) = 0. \tag{A.13}$$

For a linear operator, $\mathscr{L}$, the solution is

$$G(t) = H(t) e^{\mathscr{L} t}, \tag{A.14}$$

where $H(t)$ is the Heaviside step function and $e^{\mathscr{L} t}$ is a matrix exponential. From this, any general forcing produces a response given by

$$w(t) = \int_0^t G(\tau) F(t - \tau)\, d\tau. \tag{A.15}$$

A response function, on the other hand, is either an empirical or equation-driven function that reproduces the system's linearized output but is not required to satisfy Equation A.13. When the underlying dynamics are nonlinear, as is the case in climate models, a true Green's function does not exist. In practice however, the success of techniques such as pattern scaling illustrates that temperature response is very nearly linear for most of the globe, suggesting that data-derived response functions may closely approximate Green's functions for certain variables.

A.1.4 Transitioning from $\mathscr{K}$ to $\mathscr{L}$

We begin from a vector form of Equation 1.6, the expectation of a statistical field $g(\boldsymbol{w})$, where bold symbols are used to explicitly denote vectors. The vector $\boldsymbol{w}$ represents a set of state variables, w_i, at discrete points in space. The evolution of $\langle g(\boldsymbol{w}) \rangle$ is

$$\frac{\partial}{\partial t} \langle g(\boldsymbol{w}) \rangle = \left\langle \left[\mathscr{N}_i(\boldsymbol{w}, t) + F_i(t) \right] \frac{\partial}{\partial w_i} g(\boldsymbol{w}) \right\rangle + D \left\langle \frac{\partial^2}{\partial w_i^2} g(\boldsymbol{w}) \right\rangle. \tag{A.16}$$

We consider the case where $g(\boldsymbol{w}) = w_i$ to find the evolution of the mean of the state variables themselves. Substituting this gives

$$\frac{\partial}{\partial t}\langle w_i \rangle = \langle \mathcal{N}_i(\boldsymbol{w}, t) \rangle + F_i(t). \tag{A.17}$$

We then define a steady baseline state, $\overline{\boldsymbol{w}}$, as

$$\langle \mathcal{N}_i(\overline{\boldsymbol{w}}, t) \rangle = -\overline{F}_i, \tag{A.18}$$

where $\overline{F}_i$ is constant in time. Deviations from the baseline satisfy

$$\frac{\partial}{\partial t}\langle w_i' \rangle = \langle \mathcal{N}_i(\overline{\boldsymbol{w}} + \boldsymbol{w}', t) - \mathcal{N}_i(\overline{\boldsymbol{w}}, t) \rangle + F_i'(t), \tag{A.19}$$

where $F_i'(t)$ is the time-varying component of the forcing. We then use a first-order Taylor expansion around $\overline{\boldsymbol{w}}$ to write

$$\frac{\partial}{\partial t}\langle w_i \rangle \simeq \left.\frac{\partial \mathcal{N}_i}{\partial w_j}\right|_{\overline{\boldsymbol{w}}} \langle w_j' \rangle + F_i'(t) = \mathcal{L}_{ij}\langle w_j' \rangle + F_i'(t), \tag{A.20}$$

where the derivative term, $\frac{\partial \mathcal{N}_i}{\partial w_j}$, can be pulled out of the expectation because the baseline state is not stochastic. To conclude, we rewrite this with $\langle w_i' \rangle = T(x_i, t)$ and drop the discrete notation for space

$$\frac{\partial}{\partial t}T(\boldsymbol{x}, t) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{x}')T(\boldsymbol{x}, t) + F(\boldsymbol{x}, t). \tag{A.21}$$

A.1.5 FDT relationship to Fokker-Planck and Koopman

Here we show how the Fluctuation Dissipation Theorem (FDT) relates to the Fokker-Planck operator. The result shows that a linear response function can be computed directly from the forward operator of the unperturbed system.

Let $\mathbf{w}$ represent our full system state. Consider an equation of the form

$$\frac{\partial \mathbf{w}}{\partial t} = \mathbf{f}_0(\mathbf{w}, t) + \mathbf{f}_1(\mathbf{w}, t) + \varepsilon\xi(t), \tag{A.22}$$

where $\mathbf{f}_0$ and $\mathbf{f}_1$ are vectors that govern the unperturbed and perturbation changes to the system dynamics, respectively. The Fokker-Planck equation corresponding to this is

$$\partial_t p + \nabla \cdot \left[(\mathbf{f}_0 + \mathbf{f}_1)p - \frac{\varepsilon^2}{2}\nabla p \right] = 0. \tag{A.23}$$

Without loss of generality, we decompose $p = p_0 + p_1$, where p_0 satisfies

$$\partial_t p_0 + \nabla \cdot \left(\mathbf{f}_0 p_0 - \frac{\varepsilon^2}{2}\nabla p_0 \right) = 0. \tag{A.24}$$

Then p_1 must exactly satisfy,

$$\partial_t p_1 + \nabla \cdot \left(\mathbf{f}_0 p_1 + \mathbf{f}_1 p_0 + \mathbf{f}_1 p_1 - \frac{\varepsilon^2}{2}\nabla p_1 \right) = 0 \tag{A.25}$$

The perturbation variables ($\mathbf{f}_1$ and p_1) form a higher-order term that we

neglect, giving

$$\partial_t p_1 + \nabla \cdot \left(\mathbf{f}_0 p_1 - \frac{\varepsilon^2}{2} \nabla p_1 \right) \approx -\nabla \cdot (\mathbf{f}_1 p_0). \tag{A.26}$$

The solution to this is

$$p_1(\mathbf{w}, t) = -\int_0^t e^{\mathcal{F}_0 (t-t')} \nabla \cdot (\mathbf{f}_1(\mathbf{w}, t') p_0) \, dt', \tag{A.27}$$

assuming that $p_1(\mathbf{w}, 0) = 0$, i.e., there is no perturbation at $t = 0$, and $\mathcal{F}_0$ is the unperturbed (time-independent) Fokker-Planck operator. Multiplying through by an arbitrary statistical quantity of the state, $g(\mathbf{w})$, and integrating with respect to $\mathbf{w}$ then yields the first-order perturbation in $g(\mathbf{w})$

$$\int g(\mathbf{w}) p_1(\mathbf{w}, t) \, d\mathbf{w} = \int g(\mathbf{w}) \left[\int_0^t e^{\mathcal{F}_0 (t-t')} \nabla \cdot (\mathbf{f}_1(\mathbf{w}, t') p_0) \, dt' \right] d\mathbf{w}. \tag{A.28}$$

The quantity on the left hand side is the expected value of the perturbed statistical quantity as a function of time. The right hand side is the cross correlation of the statistical quantity, g, with $h \equiv \nabla \cdot (\mathbf{f}_1 p_0)/p_0$ with respect to the unperturbed system. Noting the Koopman operator is the adjoint of the Fokker-Planck operator gives

$$\left(e^{-\mathcal{F}_0 t} \right)^* = e^{-\mathcal{K}_0 t}, \tag{A.29}$$

where * indicates the adjoint (conjugate transpose in finite dimensions) and $\mathcal{F}^* = \mathcal{K}$, giving an expression for the response function in terms of the Koopman operator.

Alternatively, we can connect the Fokker-Planck operator to the FDT through the score function. Consider the score function of the state given by

$$\mathbf{s}(\mathbf{w}) = \nabla_{\mathbf{w}} \ln p_0(\mathbf{w}). \tag{A.30}$$

For a small, instantaneous perturbation applied at $t = 0$, the linear response of the mean field at a lag t is given by

$$\mathbf{R}(t) = -\langle g(\mathbf{w}_t) \mathbf{s}(\mathbf{w}_0) \rangle_{p_0}, \tag{A.31}$$

where the angle brackets denote an average over the stationary ensemble. We express this correlation with a joint probability density as

$$\mathbf{R}(t) = -\iint p(\mathbf{w}_0, \mathbf{w}_t) g(\mathbf{w}_t) \mathbf{s}(\mathbf{w}_0) \, d\mathbf{w}_t \, d\mathbf{w}_0, \tag{A.32}$$

Using the definition of conditional probability, we factor the joint probability density as

$$p(\mathbf{w}_0, \mathbf{w}_t) = p_0(\mathbf{w}_0) p(\mathbf{w}_t | \mathbf{w}_0), \tag{A.33}$$

where $p(\mathbf{w}_t | \mathbf{w}_0)$ is the conditional probability from $\mathbf{w}_0$ to $\mathbf{w}_t$. For dynamics governed by the Fokker-Planck operator, $\mathcal{F}$, we have

$$p(\mathbf{w}_t | \mathbf{w}_0) = e^{\mathcal{F} t} \delta(\mathbf{w}_0 - \mathbf{w}_t). \tag{A.34}$$

We then insert this expression into Equation A.32 and integrate over $\mathbf{w}_t$:

$$\mathbf{R}(t) = -\int p_0(\mathbf{w}_0) e^{\mathcal{F} t} g(\mathbf{w}_0) \mathbf{s}(\mathbf{w}_0) \, d\mathbf{w}_0. \tag{A.35}$$

Therefore, the linear response function can be obtained by propagating the unperturbed field with $e^{\mathcal{F}t}$ and correlating the result with the stationary score function.

A.2 Regularization for response functions

Estimating a response function from noisy data requires using deconvolution to invert an often ill-conditioned matrix. We choose to model the noise in our field of interest, $g(\mathbf{W})$, with a Gaussian noise term: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Rather than applying an ad-hoc smoothing algorithm, we cast the problem in a Bayesian framework, placing a Gaussian prior on the response matrix: $\mathbf{R} \sim \mathcal{N}(0, \lambda^2\mathbf{I})$. Our measurement model is therefore

$$g(\mathbf{W}) = \mathbf{F}\mathbf{R} + \boldsymbol{\varepsilon}. \tag{A.36}$$

We have dropped Δt and the spatial pattern for conciseness, but this analysis can easily be repeated including those terms.

Under this probabilistic model, we frame the task of estimating $\mathbf{R}$ as finding the vector that maximizes the response function probability given the observable data we have collected, i.e., $p(\mathbf{R}|g(\mathbf{W}))$. This term is called the *maximum a posteriori* (MAP). As it is more convenient to work with log probabilities, we recast this problem as

$$\max_{\mathbf{R}} \log p(\mathbf{R} \mid g(\mathbf{W})). \tag{A.37}$$

Using Bayes theorem, maximizing the log-posterior,

$$\log p(\mathbf{R} \mid g(\mathbf{W})) = -\frac{1}{2\sigma^2}\|g(\mathbf{W}) - \mathbf{F}\mathbf{R}\|^2 - \frac{1}{2\lambda^2}\|\mathbf{R}\|^2 + \text{const}, \tag{A.38}$$

is equivalent to solving

$$\min_{\mathbf{R}} \|g(\mathbf{W}) - \mathbf{F}\mathbf{R}\|^2 + \alpha\|\mathbf{R}\|^2, \quad \alpha = \sigma^2/\lambda^2. \tag{A.39}$$

Thus ridge regression is equivalent to placing a Gaussian prior on the response function and assuming that the data we collect are corrupted by Gaussian noise.

To avoid making an arbitrary choice for our noise and prior variance hyperparameters, σ^2 and λ^2, we propose to compute their maximum likelihood estimates under the distribution of the field of interest. We maximize the marginal likelihood evidence,

$$p(g(\mathbf{W}) \mid \sigma^2, \lambda^2) = \int p(g(\mathbf{W}) \mid \mathbf{R}, \sigma^2) p(\mathbf{R}, \lambda^2)\, d\mathbf{R} \tag{A.40}$$

$$= \mathcal{N}\left(g(\mathbf{W} \mid 0, \boldsymbol{\Sigma}\right), \tag{A.41}$$

with covariance $\boldsymbol{\Sigma} = \sigma^2\mathbf{I} + \lambda^2\mathbf{F}\mathbf{F}^T$. Maximizing the log-evidence,

$$-\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| + g(\mathbf{W})^T\boldsymbol{\Sigma}^{-1}g(\mathbf{W})\right) + \text{const}, \tag{A.42}$$

has no closed-form solution for a general $\mathbf{F}$, so we determine σ^2 and λ^2 numerically.

A.3 Analytic examples

In this appendix, we use a 1D Ornstein-Uhlenbeck (OU) process to analytically derive the Fokker-Planck operator, the Koopman operator, the eigenpairs of both operators, and the linear response function for the system obtained in two ways: (1) by directly solving the forced stochastic differential equation (SDE) and (2) by correlation with the score function.

A.3.1 Fokker-Planck and Koopman operator derivation

We define the OU SDE as

$$dw_t = -w_t dt + \sqrt{2} dW_t, \tag{A.43}$$

where w_t is the statistical field of interest and W_t is a Wiener process. The drift coefficient, $-w_t$, relaxes the state toward zero, while the diffusion coefficient, $\sqrt{2}$, gives a unit variance.

We write the Fokker-Planck equation corresponding to this OU process directly:

$$\frac{\partial}{\partial t} p(w,t) = \frac{\partial}{\partial w}\left(wp\right) + \frac{\partial^2}{\partial w^2} p, \tag{A.44}$$

where $p(w,t)$ is the probability density function of the field. The stationary solution of this expression is the standard normal probability density:

$$p_0(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}. \tag{A.45}$$

From the previous result, we explicitly write the Fokker-Planck operator governing the evolution of the probability density as

$$\mathcal{F}(\cdot) = \frac{\partial}{\partial w}\left[w(\cdot) + \frac{\partial}{\partial w}(\cdot)\right]. \tag{A.46}$$

To find the eigenfunctions, $\phi(w)$, with $\mathcal{F}\phi(w) = \lambda\phi(w)$, we introduce the ansatz $\phi(w) = h(w)e^{-\frac{w^2}{2}}$, giving

$$h''(w) - wh'(w) - \lambda h(w) = 0, \tag{A.47}$$

whose solutions are Hermite polynomials, $H_n(w)$, with eigenvalues $\lambda = -n$ for $n = 0, 1, 2, \ldots$.

A.3.2 Response function via direct diagnosis

Adding a deterministic forcing, $F(t)$, to our OU process gives

$$dy_t = (-y_t + F(t))dt + \sqrt{2} dW_t. \tag{A.48}$$

Taking the expected value of this and assuming $\langle y(0)\rangle = 0$ gives

$$\frac{d}{dt}\langle y\rangle = -\langle y\rangle + F(t), \tag{A.49}$$

whose solution is given by

$$\langle y(t)\rangle = \int_0^t e^{-\tau} F(t-\tau)\, d\tau, \tag{A.50}$$

where the response function is $R(t) = e^{-t}$ for $t \geq 0$.

A.3.3 Response function via correlation with score function

The stationary score function for the Gaussian PDF in Equation A.45 is given by

$$s(w) = \frac{\partial}{\partial w} \ln p_0(w) = \frac{-w\, p_0(w)}{p_0(w)} = -w, \tag{A.51}$$

where we can make this simplification since the stationary probability distribution is given by a standard normal.

The Fluctuation Dissipation Theorem predicts

$$R(t) = \frac{-\langle w(t)s(w(0))\rangle}{\langle w^2(0)\rangle} = \frac{\langle w(t)w(0)\rangle}{\langle w^2(0)\rangle} = e^{-t}, \tag{A.52}$$

which agrees exactly with the direct solution above.

B

Appendices for Chapter 2

In Section B.1, we describe our methodology for optimizing emulator training data by backpropagating through a differentiable model, beginning with a conceptual example to motivate our procedure. We then outline the development of a Python-based differentiable Simple Climate Model (SCM) using the JAX numerical library in Section B.2, including descriptions of the model's structure and gradient-based calibration procedure. In Section B.3, we present a neural network emulator of our SCM, highlighting its architecture, feature design, and training procedure. In Section B.4, we describe how we extend the training data optimization process and emulator creation to the MIT Earth System Model (MESM), followed by an outline of the scenarios and metrics used during evaluation in Section B.5. We conclude with sensitivity analyses for our optimization procedure (Section B.6) and extended results from the main manuscript (Section B.7).

B.1 Training data optimization

B.1.1 Conceptual overview

We use the following problem of estimating unknown linear system parameters as a conceptual example to motivate our optimization procedure. It does not extend directly to our full system because the full system is nonlinear, state dependent, and our full system objective is predictive skill over some set of metrics, rather than parameter estimation. Despite this, the linear system is highly useful because it is interpretable, enabling us to better understand our optimization requirements.

Consider emulating a discrete, linear system via explicit parameter estimation. For example, estimating climate sensitivity and carbon uptake terms for a simple climate model to predict global mean temperature anomaly from CO_2 emissions. Our goal is to determine the set of training data that maximizes the accuracy of our parameter estimates. The system of interest is given by

$$\mathbf{T}_{n+1} = \mathbf{N}\mathbf{T}_n + \mathbf{u}_n, \tag{B.1}$$

where $\mathbf{T}$ is the temperature, $\mathbf{N}$ is the linear operator that evolves the temperature forward in time, and $\mathbf{u}$ is a known forcing (e.g., emissions). We can emulate this system by estimating the parameters of $\mathbf{N}$ and using the recovered operator to step our system forward in time from some initial condition. We use standard dynamic mode decomposition[165] to estimate $\mathbf{N}$ via least squares as

[165] Schmid, 2010

$$\tilde{\mathbf{N}} \approx [\mathbf{T}_{n+1} - \mathbf{u}_n]\,\mathbf{T}_n^{\dagger}, \tag{B.2}$$

where † indicates the Moore-Penrose pseudoinverse.

The accuracy of the estimate of $\mathbf{N}$ is controlled by the conditioning of the training data: the forcing $\mathbf{u}_n$ and the resulting temperature trajectory $\mathbf{T}_n(\mathbf{u}_n)$. If the chosen forcing is uninformative (e.g., nearly constant or exponential, as in Giani et al. (2024)[279] and Womack et al. (2026)[2]), then the columns of the data matrix constructed from $\mathbf{T}_n$ become nearly collinear. In the more realistic case where we only observe noisy states $\tilde{\mathbf{T}} = \mathbf{T} + \varepsilon$, where ε corresponds to some measurement error or stochastic noise (e.g., internal variability), the error in our estimated operator $\tilde{\mathbf{N}}$ is bounded by the condition number, κ, of the data matrix:

[279] Giani et al., *Origin and Limits of Invariant Warming Patterns in Climate Models*, 2024

[2] Womack et al., 'A theoretical framework to understand sources of error in Earth System Model emulation', *Earth System Dynamics*, 2026

$$\frac{\|\tilde{\mathbf{N}} - \mathbf{N}\|}{\|\mathbf{N}\|} \lesssim \kappa(\mathbf{T}_n) \frac{\|\varepsilon\|}{\|\mathbf{T}_n\|}. \tag{B.3}$$

Here, $\kappa(\mathbf{T}_n)$ acts as an amplification factor. If the data are ill-conditioned ($\kappa \gg 1$), even a small amount of noise leads to large errors in the learned dynamics. As a result, designing the forcing $\mathbf{u}_n$ becomes a problem of optimal experimental design[200]. We aim to find the forcing $\mathbf{u}^*$ that minimizes this error bound:

[200] Fedorov, 2010

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \kappa(\mathbf{T}_n(\mathbf{u})). \tag{B.4}$$

If the map $\mathbf{u} \mapsto \mathbf{T}(\mathbf{u})$ is differentiable, we can, in principle, compute $\nabla_{\mathbf{u}}\kappa$ and use gradient descent to iteratively update $\mathbf{u}$ so that the resulting temperature response is maximally informative for parameter estimation.

While minimizing the condition number is optimal for linear parameter recovery, this approach breaks down for our full-scale system. In general, manually deriving and implementing an adjoint model of a system of interest to calculate gradients is intractable due to its complexity (e.g., Earth System Models (ESMs) written in Fortran with millions of lines of code). To address this, we generalize the logic above to leverage backpropagation to calculate gradients.

B.1.2 Framework for training data optimization

We frame the generation of training data as a bi-level optimization problem. Rather than designing an explicit adjoint model, we utilize Automatic Differentiation (AD). AD allows us to accurately and efficiency compute derivatives of complex functions by leveraging the chain rule through the computational graph. This technique is preferable to traditional numerical methods (e.g., finite differences) as it incurs a lower computational cost and computes exact derivatives.

Our objective is to find a specific set of training emissions $\mathbf{U}_{\text{train}}$ that minimizes the error of an emulator trained on that data when tested against a held-out target set. This problem consists of an implicit inner level (training the emulator parameters θ using $\mathbf{U}_{\text{train}}$) and an explicit outer level (updating $\mathbf{U}_{\text{train}}$ to minimize the test loss of the trained emulator). The optimization objective is given mathematically as

$$\arg\min_{\mathbf{U}_{\text{train}}} \mathscr{L}_{\text{test}}(\mathbf{U}_{\text{train}},\, \theta_{\text{train}},\, D_{\text{test}}), \tag{B.5}$$

where θ_{train} represents the parameters of the emulator after training on the data generated by $\mathbf{U}_{\text{train}}$. While this methodology is generalizable to models of other physical systems, we apply it here to an SCM. The procedure is detailed below, with Algorithm 1 providing a summary;

also see Figs. 1 and 2 in the main text for an overview of the optimization process and illustrative example, respectively.

Inner level. The inner level of the optimization consists of training an emulator to map from emissions to temperature anomalies. The process is defined by the following:

1. Emissions training data ($\mathbf{U}_{\text{train}} \in \mathbb{R}^{n_{\text{agents}} \times n_t}$): The trainable parameters are a collection of emission time series for n_{agents} forcing agents (e.g., CO_2, CH_4, etc.) over n_t time steps.
2. Training features ($\mathbf{X}_{\text{train}} \in \mathbb{R}^{n_t \times d}$): We construct features from $\mathbf{U}_{\text{train}}$ using instantaneous emissions, cumulative emissions, and exponential moving averages for each forcing agent. The resulting features are dimension $d = n_{\text{agents}} \times n_{\text{feat.}}$, where $n_{\text{feat.}}$ is equal to the number of features per agent.
3. Training targets ($\overline{\Delta T}(t) = \mathbf{y}_{\text{train}} \in \mathbb{R}^{n_t}$): We force the SCM with $\mathbf{U}_{\text{train}}$ to generate the corresponding Global Mean Surface Temperature (GMST) anomalies, which serve as ground-truth targets.
4. Inner optimization: The emulator (a neural network with parameters θ) is trained to minimize the Mean Squared Error (MSE) between its predictions and $\mathbf{y}_{\text{train}}$. Starting from initial weights θ_0, we perform k steps of Stochastic Gradient Descent (SGD):

$$\theta_k = \text{SGD}(\theta_0;\, \mathbf{X}_{\text{train}},\, \mathbf{y}_{\text{train}},\, k). \tag{B.6}$$

Outer level. The outer level tests the performance of the trained emulator with parameters θ_k on a dataset held constant during optimization, and backpropagates the error through the test, training, and data generation steps to update $\mathbf{U}_{\text{train}}$. The process is defined by the following:

1. Test loss ($\mathscr{L}_{\text{test}}$): The trained emulator is tested on a fixed set of scenarios ($\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}$), constructed from prescribed emissions independent of $\mathbf{U}_{\text{train}}$. However, to ensure consistency, both train and test features are normalized using summary statistics (mean and variance) computed from the current training features $\mathbf{X}_{\text{train}}$. We quantify test performance using Normalized Root Mean Square Error (NRMSE), weighted by scenario length and averaged across N_{scen} test scenarios:

$$\mathscr{L}_{\text{test}} = \frac{1}{N_{\text{scen}}} \sum_{i=1}^{N_{\text{scen}}} w_i \frac{\text{RMSE}\left(f\left(\mathbf{X}_{\text{test}}^{(i)};\, \theta_k\right),\, \mathbf{y}_{\text{test}}^{(i)}\right)}{\max_t |\mathbf{y}_{\text{test}}^{(i)}(t)|}, \tag{B.7}$$

   where f denotes the emulator and w_i accounts for the relative length of scenario i.
2. Gradient calculation: We update $\mathbf{U}_{\text{train}}$ to minimize $\mathscr{L}_{\text{test}}$. By the chain rule, the gradient $\partial\mathscr{L}_{\text{test}}/\partial\mathbf{U}_{\text{train}}$ is decomposed as

$$\frac{\partial\mathscr{L}_{\text{test}}}{\partial\mathbf{U}_{\text{train}}} = \underbrace{\frac{\partial\mathscr{L}_{\text{test}}}{\partial\theta_k} \cdot \frac{\partial\theta_k}{\partial\mathbf{U}_{\text{train}}}}_{\text{Parameter sensitivity}} + \underbrace{\frac{\partial\mathscr{L}_{\text{test}}}{\partial\mathbf{X}_{\text{test}}} \cdot \frac{\partial\mathbf{X}_{\text{test}}}{\partial\mathbf{U}_{\text{train}}}}_{\text{Normalization sensitivity}}. \tag{B.8}$$

   The first term captures how changing emissions alters the trained model parameters. The second term accounts for the dependence of the feature normalization statistics (mean and variance) on the training data $\mathbf{U}_{\text{train}}$. The parameter sensitivity is then expanded

further:

$$\frac{\partial \theta_k}{\partial \mathbf{U}_{\text{train}}} = \underbrace{\frac{\partial \theta_k}{\partial \mathbf{X}_{\text{train}}} \cdot \frac{\partial \mathbf{X}_{\text{train}}}{\partial \mathbf{U}_{\text{train}}}}_{\text{Feature sensitivity}} + \underbrace{\frac{\partial \theta_k}{\partial \mathbf{y}_{\text{train}}} \cdot \frac{\partial \mathbf{y}_{\text{train}}}{\partial \mathbf{U}_{\text{train}}}}_{\text{Physics sensitivity}}. \tag{B.9}$$

Here, $\partial \mathbf{y}_{\text{train}} / \partial \mathbf{U}_{\text{train}}$ requires differentiating through the SCM physics, while $\partial \mathbf{X}_{\text{train}} / \partial \mathbf{U}_{\text{train}}$ involves differentiating through the feature engineering operations (e.g., moving averages). We rely on AD to propagate these gradients through the full pipeline.

3. Emissions update: At iteration n, we update the emissions using the computed gradient:

$$\mathbf{U}_{n+1} = \mathbf{U}_n - \eta \nabla_{\mathbf{U}_{\text{train}}} \mathscr{L}_{\text{test}}, \tag{B.10}$$

where η is the learning rate. In practice, a different learning rate is applied to each forcing agent, as the magnitude of the gradient with respect to each agent can vary by several orders of magnitude. The final updates are applied via an SGD optimizer with momentum[221] to yield a locally optimal emissions trajectory $\mathbf{U}^*$.

[221] Liu, Gao, and Yin, 2020

Algorithm 1: Bi-level training data optimization procedure. The inner loop trains the emulator parameters θ while the outer loop tests performance and updates the training emissions $\mathbf{U}_{\text{train}}$.

Require: $\mathbf{U}_{\text{train}}$ (initial emissions), $\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}$ (test set), θ_0 (initial weights)

while *not converged* **do**
  # 1. Data generation & feature engineering
  $\mathbf{y}_{\text{train}} = \text{SCM}(\mathbf{U}_{\text{train}})$
  $\mathbf{X}_{\text{train}} = \text{Featurize}(\mathbf{U}_{\text{train}})$
  $\mu_{\text{train}}, \sigma_{\text{train}} = \text{get\_stats}(\mathbf{X}_{\text{train}})$ `// Compute normalization statistics from training data`
  $\mathbf{X}^{\text{norm}}_{\text{train}} = (\mathbf{X}_{\text{train}} - \mu_{\text{train}})/\sigma_{\text{train}}$

  # 2. Inner loop: Emulator training
  $\theta = \theta_0$ `// Reset weights completely before training`
  for k in range(K) ***do*** `// K = number of training gradient descent steps`
    $\hat{\mathbf{y}} = f(\mathbf{X}^{\text{norm}}_{\text{train}}; \theta)$
    $\mathscr{L}_{\text{train}} = \text{MSE}(\hat{\mathbf{y}}, \mathbf{y}_{\text{train}})$
    $\theta = \text{SGD}(\theta, \nabla_\theta \mathscr{L}_{\text{train}})$ `// Update weights`

  # 3. Outer loop: Test & update
  $\mathbf{X}^{\text{norm}}_{\text{test}} = (\mathbf{X}_{\text{test}} - \mu_{\text{train}})/\sigma_{\text{train}}$ `// Normalize test data using training stats`
  $\hat{\mathbf{y}}_{\text{test}} = f(\mathbf{X}^{\text{norm}}_{\text{test}}; \theta)$
  $\mathscr{L}_{\text{test}} = \text{NRMSE}(\hat{\mathbf{y}}_{\text{test}}, \mathbf{y}_{\text{test}})$ `// Compute weighted test loss (Eq. B.7)`

  grads $= \nabla_{\mathbf{U}_{\text{train}}} \mathscr{L}_{\text{test}}$ `// Backpropagate through testing, training, and physics via AD`
  $\mathbf{U}_{\text{train}} = \text{Optimizer}(\mathbf{U}_{\text{train}}, \text{grads})$ `// Update emissions via SGD with momentum`
return $\mathbf{U}_{\text{train}}$

B.2 Differentiable simple climate model

To enable the optimization procedure outlined in Section B.1, we present a differentiable simple climate model that calculates annual-average GMST anomalies based on the Finite Amplitude Impulse Response (FaIR) framework[54]. Implemented in JAX, this model leverages automatic

[54] Leach et al., 2021

differentiation to facilitate efficient gradient-based calibration.

B.2.1 Model structure

Our model retains the core structural components of FaIR; see Leach et al. (2021)[54] Section 2 for a full model description. We represent the carbon cycle with a four-reservoir model, where each reservoir has an uptake fraction and a decay timescale. These reservoirs are mathematical abstractions representing the different timescales of carbon removal—ranging from rapid biospheric uptake to slow geological weathering—rather than distinct physical stores (e.g., the deep ocean). The decay time constants are scaled by a state-dependent feedback parameter, α. This factor incorporates nonlinear feedbacks, such as the saturation of carbon sinks, based on cumulative carbon uptake and GMST anomalies. We calculate temperature anomalies using a three-box impulse response model based on total effective radiative forcing from forcing agent concentrations. This component accounts for the thermal inertia of the climate system, capturing the delay between radiative forcing and warming caused by heat uptake in the upper and deep ocean.

[54] Leach et al., 'FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration', *Geoscientific Model Development*, 2021

We consider a subset of the forcing agents from FaIR: CO_2, CH_4, N_2O, sulfur and Black Carbon (BC). We exclude minor anthropogenic gases (e.g., CFCs, HFCs) and natural forcings (solar irradiance and volcanic aerosols) to focus on the dominant drivers of future warming while retaining a tractable parameter space. For CH_4, and N_2O, we use the same governing decay equations as CO_2, but model them with a single reservoir. This single-reservoir approach is sufficient to capture their atmospheric residence times without the complex multi-timescale dynamics required for CO_2. CH_4 retains a state-dependent lifetime calculation similar to CO_2 (dependent on temperature and atmospheric burden), while N_2O is modeled with a constant lifetime. Finally, we assume sulfur and black carbon emissions directly impact effective radiative forcing through aerosol-radiation and aerosol-cloud interactions, following the parameterization in Leach et al. (2021)[54].

[54] Leach et al., 'FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration', *Geoscientific Model Development*, 2021

B.2.2 Model calibration

We use automatic differentiation to enable gradient-based calibration, calibrating our model to reproduce the temperature response of the FaIR model. To establish a ground-truth target, we configure FaIR using the median parameters values from the probabilistic ensemble derived in Smith et al. (2024)[213], which was constrained to reproduce climate responses based on ESMs from the sixth phase of the Coupled Model Intercomparison Project (CMIP6) or Intergovernmental Panel on Climate Change (IPCC)-assessed ranges. To isolate the response of each forcing agent for calibration, we use single-forcing experiments generated from the the seventh phase of the Coupled Model Intercomparison Project (CMIP7) ScenarioMIP and DECK protocols[126,280]; for forcing agents without single-forcing experiments in the DECK, we prescribe protocols equivalent to *abrupt-4xCO2* and *1pctCO2* (e.g., *abrupt-4xCH4* and *1pctN2O*). A full list of calibration experiments can be found in Table B.1.

[213] Smith et al., 'fair-calibrate v1.4.1: calibration, constraining, and validation of the FaIR simple climate model for reliable future climate projections', *Geoscientific Model Development*, 2024

[126] Van Vuuren et al., 2026; [280] Dunne et al., 2025

Unlike the standard FaIR calibration methodology, which employs a Bayesian framework to generate a posterior distribution of parameters[213], we perform deterministic parameter estimation via gradient-based optimization. Using the Adam optimizer[281], we minimize NRMSE between

[213] Smith et al., 2024

[281] Kingma and Ba, 2017

our model's GMST and the FaIR reference outputs over the simulation period. We use NRMSE to handle the disparate scales of temperature anomalies across scenarios (e.g., 6°C in *abrupt-4xCO2* vs. < 2°C in *VLLO*), normalizing by the range of the reference temperature trajectory (Equation B.7). This normalization ensures equal weighting across scenarios, preventing high-warming trajectories from dominating the loss function. We conduct calibration sequentially by agent and component (e.g., separating carbon cycle parameters from thermal response parameters), utilizing gradient masking to freeze non-target parameters during each stage.

B.3 Neural network emulator

We implement a neural network emulator for the differentiable SCM outlined in Section B.2 that predicts GMST from emissions. This component serves as a proof-of-concept to demonstrate that the differentiable framework functions as intended to optimize training data (Section B.1).

We employ a lightweight architecture (emulator structure) designed to minimize computational burden while maintaining sufficient fidelity to map emissions to temperature response. While more complex deep learning architectures (e.g., LSTMs or Transformers) could yield higher predictive skill, our objective is to isolate the effect of training data composition rather than model complexity. As a result, we prioritize training speed; all training reported here was completed on a standard laptop CPU (MacBook Pro, M1 Chip).

B.3.1 Feature design and architecture

To capture the temporal dynamics and inertia of the climate system without the computational cost and complexity of an autoregressive model, we construct a feature vector that implicitly encodes the atmospheric state and memory of past forcings. For a given simulation year t, the input vector $\mathbf{X}_t$ summarizes the history of emissions up to year $t-1$. For each forcing agent, we calculate five scalar features (1) the emissions at the previous timestep; (2) the cumulative emissions to date; and (3) Exponential Moving Averages (EMAs) of the emissions calculated with decay timescales of five, thirty, and one hundred years. These timescales were selected to approximate the multiple response timescales of the carbon cycle and thermal response.

These features are fed into a standard Multi-Layer Perceptron (MLP). The network consists of a single hidden layer with hyperbolic tangent (tanh) activations, followed by a linear output layer that predicts the scalar GMST anomaly for year t. Prior to training, all input features are standardized to zero mean and unit variance.

B.3.2 Training

We train the emulator to predict GMST by minimizing the MSE between the emulator predictions and the true SCM output. Note this training loss (MSE) is distinct from the NRMSE used in the outer optimization level. We train distinct emulator configurations to compare the impact of training data on predictive skill. To ensure fair comparison, we use a con-

sistent number of gradient training steps for all emulator configurations, regardless of the size or composition of the training dataset.

Baseline emulator. We train the baseline emulator on Priority 1 scenarios from CMIP7 ScenarioMIP protocol[126]. We selected this baseline as it represents the standard set of scenarios that all modeling centers are expected to generate, ensuring broad accessibility.

[126] Van Vuuren et al., 2026

Optimized emulator. The optimized emulator utilizes an identical architecture, feature set, and training hyperparameters (e.g., learning rate) as the baseline, but is trained on the synthetic datasets generated via the optimization process described in Section B.1.

B.4 Extension to the MIT Earth System Model

We apply our framework to MESM to demonstrate its scalability and utility for generating informative training datasets even when the target model is not differentiable. MESM is a zonally averaged Earth system Model of Intermediate Complexity (EMIC) that includes a two-dimensional, zonally averaged atmospheric model with interactive chemistry coupled to a zonally averaged land model and an anomaly-diffusing ocean model; see Sokolov et al. (2018)[222] for a full description. As MESM is not differentiable, we recalibrate the differentiable SCM to act as a surrogate for MESM.

[222] Sokolov et al., 'Description and Evaluation of the MIT Earth System Model (MESM)', *Journal of Advances in Modeling Earth Systems*, 2018

We recalibrate the SCM to approximate the MESM temperature response in two stages, using CO_2-only scenarios. First, we constrain the climate sensitivity parameters (thermal response) by minimizing the loss between SCM and MESM GMST anomalies under scenarios with prescribed CO_2 concentrations (e.g., *1pctCO2*). Then, we calibrate the carbon cycle parameters using emissions-driven scenarios, minimizing the error in simulated atmospheric CO_2 concentrations. To filter MESM's internal variability, we target the mean of a thirty-member initial condition ensemble for all components of calibration, optimization, and emulation.

We assume that the training data optimized for this MESM-tuned SCM will be highly informative for training an emulator of the actual MESM. This approach relies on the assumption that loss landscape topology of the recalibrated SCM is sufficiently similar to that of MESM, allowing gradients computed through the SCM to guide data selection for the more complex model.

Following recalibration, we utilize the bi-level optimization procedure using the SCM to generate optimized emissions trajectories. We then evaluate these datasets by training a new emulator for MESM. Unlike the global-mean emulator outlined in the previous section, this emulator is modified to predict zonal temperature anomalies by modifying the output layer of the neural network to produce a vector (predictions at each latitude band), rather than a scalar; the feature generation remains the same. We benchmark the performance of the emulators trained on our optimized datasets against a baseline emulator trained on ScenarioMIP Priority 1 scenarios, evaluating all emulators on the remaining scenarios. As the area of each latitude band is non-uniform and decreasing towards the poles, we use area-weighted error metrics during training

and evaluation:

$$\mathscr{L}_{\text{zonal}}(\phi) = \cos(\phi)\langle E \rangle_t, \tag{B.11}$$

$$\mathscr{L}_{\text{global}} = \frac{\sum_\phi \mathscr{L}_{\text{zonal}}(\phi)}{\sum_\phi \cos(\phi)}, \tag{B.12}$$

where ϕ denotes the latitude, E denotes the error metric of interest, and $\langle \cdot \rangle_t$ denotes the temporal average over a scenario. The error metric is MSE during training and NRMSE during evaluation, and must be normalized by the magnitude of the weights to ensure the global metric remains in the same units and scale as the zonal errors.

B.5 Scenario descriptions and evaluation protocol

We use NRMSE (Equation B.7) as our primary evaluation metric. By normalizing RMSE by the maximum absolute magnitude of the SCM- or EMIC-projected temperature trajectory, NRMSE weights all scenarios with equal importance regardless of their warming magnitude. This design choice ensures the emulator is optimized to perform well across a wide range of future pathways, rather than prioritizing high-warming scenarios where absolute errors would otherwise dominate the loss function. Table B.1 details the complete set of experiments used for calibration, optimization, and evaluation.

We compare the performance of several optimized emulator configurations against a baseline emulator, which is trained exclusively on ScenarioMIP-CMIP7 Priority 1 scenarios described in Table B.1. For the optimized emulator, we initialize the optimization with a constant emissions trajectory and optimize for predictive skill (minimize NRMSE) over a specific test set. We optimize over up to seven different test sets, depending on if we are evaluating single-forcing or multi-forcing performance: Priority 1, Priority 2, DECK, CS3, DAMIP (multi-forcing only), GeoMIP (multi-forcing only), and All (the union of all sets). Following optimization, we evaluate the emulator's predictive skill on all other scenario sets.

To assess the emulator's generalization capability and adherence to physical principles, we include structurally distinct, out-of-distribution scenarios in our evaluation. We utilize scenarios analogous to those considered within the Detection and Attribution MIP (DAMIP)[195,202] and Geoengineering MIP (GeoMIP)[203] protocols. The DAMIP-like scenarios (*Medium-GHG* and *Medium-aer*) isolate the contributions of specific forcing agents by extending the *historical* period into the future using the *Medium* ScenarioMIP-CMIP7 scenario forcing for a subset of agents (e.g., GHGs only), while holding others constant. Similarly, the *G6sulfur* scenario from GeoMIP introduces a stratospheric sulfate injection trajectory significantly larger than any found in standard training data, stress-testing the emulator's response to extreme aerosol forcing. These scenarios allow us to test emulator skill in reproducing the contribution of individual forcing agents. The *DECK* scenarios similarly require separation of individual forcing agents, as each experiment in the *DECK* is a single-forcing experiment.

[195] Gillett et al., 2016; [202] Gillett et al., 2025

[203] Kravitz et al., 2015

Table B.1: Complete list of scenarios used for training, optimization, and evaluation. Scenario descriptions for ScenarioMIP are derived from the CMIP7 protocol[282], while CS3 scenarios are taken from the MIT Center for Sustainability Science and Strategy (CS3)'s 2025 Global Change Outlook[201]. *abrupt-4xX* and *1pctX* refer to idealized single-forcing experiments performed for each agent. Scenarios containing *-ext* refer to scenario extensions that end in 2500.

Activity	Scenario	Short Description
ScenarioMIP-CMIP7 (Priority 1)	*H-ext*	**High:** High emission scenario exploring potential high-end impacts.
	M	**Medium:** Medium emission scenario consistent with current policies.
	ML	**Medium-Low:** Delayed mitigation effort, insufficient to meet Paris Agreement goals.
	L	**Low:** Scenario consistent with likely staying below 2°C.
	VLLO-ext	**Very Low with Limited Overshoot:** Consistent with limiting warming to 1.5°C by 2100 with limited overshoot.
	VLHO	**Very Low after High Overshoot:** Scenario with similar end-of-century temperature impact to VLLO, but with delayed near-term mitigation and reliance net-negative emissions, resulting in a higher overshoot.
ScenarioMIP-CMIP7 (Priority 2)	*H-ext-OS*	**High Overshoot:** Radical emissions reductions after 2100 with net zero in 2160.
	M-ext	Extension of the Medium scenario.
	ML-ext	Extension of the Medium-Low scenario.
	L-ext	Extension of the Low scenario.
	VLHO-ext	Extension of the Very Low with High Overshoot scenario.
DECK	*historical*	Simulation of the historical period (1850–2014) using observed forcing.
	abrupt-4xX	Instantaneous quadrupling of agent X (e.g., CO_2, CH_4) concentrations or emissions from pre-industrial levels.
	1pctX	Concentrations or emissions of agent X increase by 1% per year until quadrupling.
CS3	*CT*	**Current Trends:** Current measures for reducing greenhouse gas emissions.
	AA	**Accelerated Actions:** Aggressive reductions which aim to limit and stabilize human-induced global climate warming to 1.5°C by 2100 with a 50% probability.
DAMIP	*Medium-GHG*	Historical + Future *Medium* scenario forcing for Well-Mixed GHGs only (CO_2, CH_4, N_2O); all other forcers held constant.
	Medium-aer	Historical + Future *Medium* scenario forcing for Aerosols only (Sulfur, BC); all other forcers held constant.
GeoMIP	*G6sulfur*	Stratospheric sulfur injection used to reduce GMST from the *High* (H) scenario to match that of the *Medium* (M) scenario.

B.6 Sensitivity analyses

In this section, we examine the sensitivity of the optimization procedure to changes in the emissions Initial Condition (IC), neural network emulator architecture, and training features. Results in this section correspond to CO_2-only experiments; accompanying figures show the average performance for both the baseline and optimized emulators tested across all scenarios in the ScenarioMIP, DECK, and CS3 evaluation sets. We tune our optimization hyperparameters (e.g., gradient step size and momentum decay) to the sixteen-neuron, single-layer architecture optimized from a constant initial condition with EMAs of five, thirty, and one hundred years; this is the default configuration shown in the results of the main text. We do not necessarily expect these hyperparameters to be well-tuned for other architectures. We address the sensitivity of these hyperparameters to architectural changes where relevant.

B.6.1 Sensitivity to initial condition

Figure B.1 illustrates the sensitivity of the optimization convergence rate and resulting emissions time series to the choice of constant, Gaussian, and sinusoidal ICs. The final structure of the optimized emissions time series depends heavily on its initialization; all three ICs yield distinct final trajectories. Despite this, every optimized emulator outperforms the baseline, suggesting that there exists a set of scenarios that are better suited for training than the baseline. This set is likely to be functionally infinite in size, as stochastic perturbations during optimization generate marginally different pathways for identical ICs (not shown). Although the large-scale features of the optimized constant and Gaussian time series differ, we observe similar small-scale features, such as rapid changes in concavity. The sinusoidal emulator does not exhibit this behavior, likely because such variations are inherent to the initialization structure. Instead, the optimization targets the magnitude of each peak, breaking the symmetry of the sinusoid. While all three ICs produce more skillful emulators than the baseline, their relative convergence rates vary. The constant IC converges most rapidly; the sinusoidal IC starts slowly, but approaches the skill of the constant IC by the 1000th iteration. The Gaussian IC yields a more skillful emulator initially, with performance on par with the baseline, but it exhibits slower overall convergence. This rate reduction is likely a consequence of the higher initial skill that results in smaller gradients.

B.6.2 Sensitivity to architecture changes

Figures B.2 and B.3 show the sensitivity of both the optimization convergence rate and resulting emissions time series to the choice of neural network architecture; the former is initialized from a constant IC and the latter from a sinusoidal IC. Architectural modifications alter the skill of the baseline emulator, with single-layer configurations preferable to double-layer ones. Furthermore, increasing the number of neurons in single-layer configurations improves baseline predictive skill. The constant IC exhibits high sensitivity to architectural changes, which results in a lack of convergence relative to the primary architecture (a single hidden layer with sixteen neurons). Specifically, the optimization algorithm fails to converge for configurations with eight neurons or two layers. This failure likely stems from the gradient's sensitivity to architectural choices. Because the gradient descent hyperparameters remain

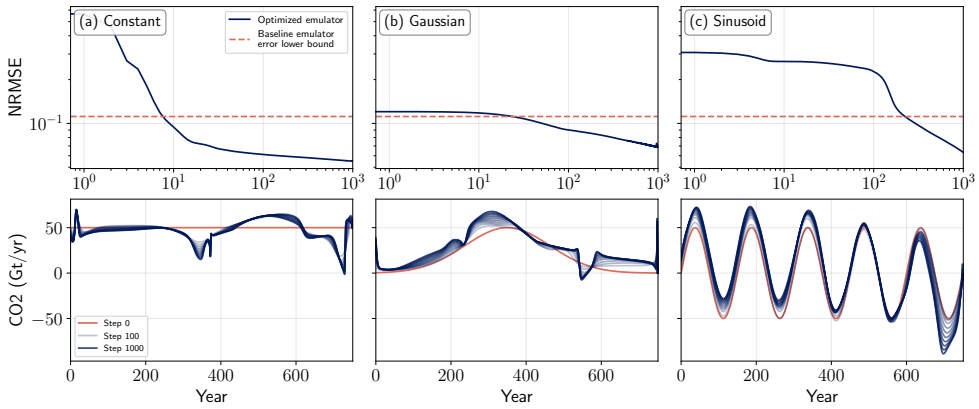

Figure B.1: Top: trajectory of evaluation loss (NRMSE) during optimization compared across three ICs for the emissions time series: (a) constant; (b) Gaussian (normally distributed in time); (c) sinusoidal with a period of 175 years. Emulators are evaluated on their performance in reproducing SCM-projected GMST anomalies caused by CO_2-only across all scenarios included in the ScenarioMIP, DECK, and CS3 activities; see Table B.1 for scenario descriptions. The solid, dark blue line tracks emulator performance throughout the optimization process, while the dashed, red line marks the lower bound of the baseline emulator error (evaluating performance on its own training data). Bottom: evolution of emissions time series over 1000 iterations, corresponding to the ICs listed above.

fixed across architectures, they are ill-tuned for alternate configurations, leading to slow or nonexistent convergence. Conversely, the emissions time series derived from the sinusoidal IC displays less sensitivity to architectural changes, suggesting that the gradients corresponding to this IC are more robust to a wider range of conditions. Updates modify the magnitude of the sinusoid's peak while leaving the period nearly unchanged across all four configurations. Larger networks require additional optimization iterations to converge; this observation indicates that gradients may shrink as network size increases. The two-layer case may suffer from the converse issue of exploding gradients, evidenced by rapid error oscillations near the 100th update. This instability, likely driven by stochasticity, stabilizes quickly. Future work can incorporate results from machine learning literature regarding the origins and mitigation of such gradient issues[283,284].

[283] Hanin, 2018; [284] Philipp, Song, and Carbonell, 2018

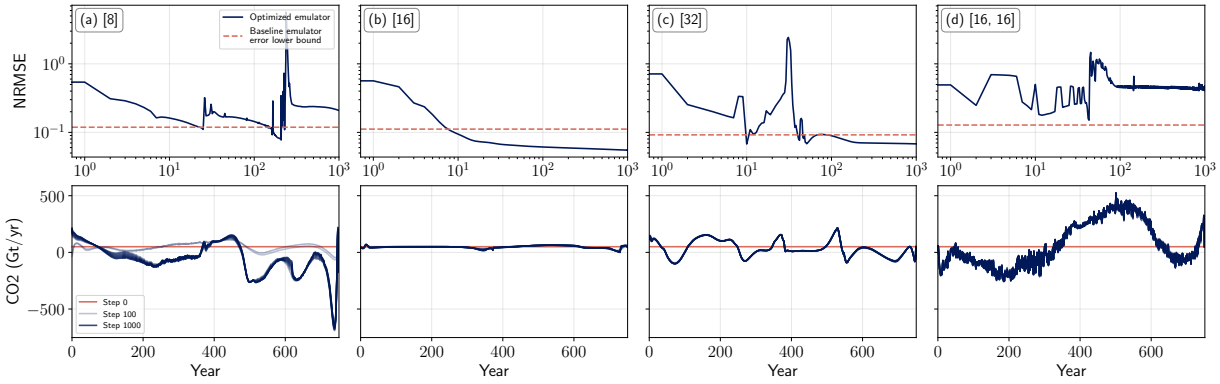

Figure B.2: Top: trajectory of evaluation loss (NRMSE) during optimization compared across four architectures for the neural network emulator initialized from a constant initial condition: (a) a single hidden layer with eight neurons; (b) a single hidden layer with sixteen neurons; (c) a single hidden layer with thirty-two neurons; (d) two hidden layers with sixteen neurons each. Emulators are evaluated on their performance in reproducing SCM-projected GMST anomalies caused by CO_2-only across all scenarios included in the ScenarioMIP, DECK, and CS3 activities; see Table B.1 for scenario descriptions. The solid, dark blue line tracks emulator performance throughout the optimization process, while the dashed, red line marks the lower bound of the baseline emulator error (evaluating performance on its own training data). Bottom: evolution of emissions time series over 1000 iterations, corresponding to the architectures listed above.

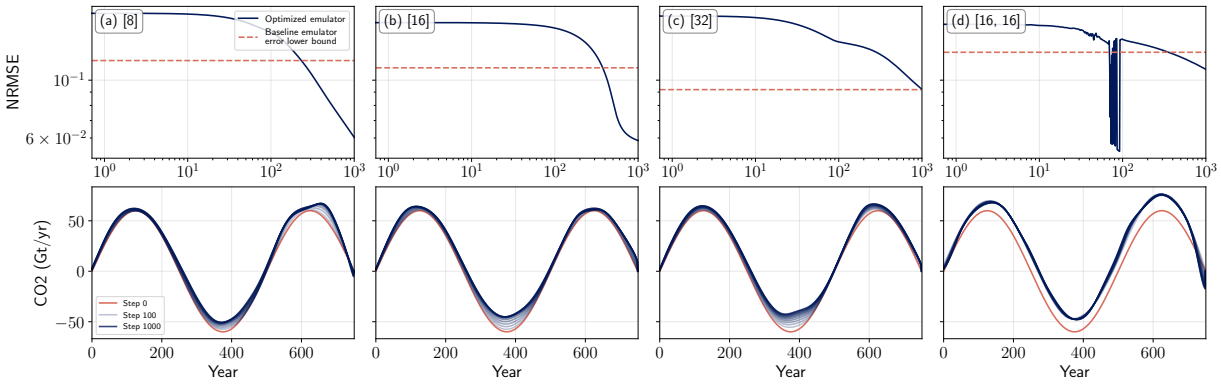

Figure B.3: Top: trajectory of evaluation loss (NRMSE) during optimization compared across four architectures for the neural network emulator initialized from a sinusoidal initial condition: (a) a single hidden layer with eight neurons; (b) a single hidden layer with sixteen neurons; (c) a single hidden layer with thirty-two neurons; (d) two hidden layers with sixteen neurons each. Emulators are evaluated on their performance in reproducing SCM-projected GMST anomalies caused by CO_2-only across all scenarios included in the ScenarioMIP, DECK, and CS3 activities; see Table B.1 for scenario descriptions. The solid, dark blue line tracks emulator performance throughout the optimization process, while the dashed, red line marks the lower bound of the baseline emulator error (evaluating performance on its own training data). Bottom: evolution of emissions time series over 1000 iterations, corresponding to the architectures listed above.

B.6.3 Sensitivity to features

Figures B.4 and B.5 illustrate the sensitivity of optimization convergence rate and resulting emissions time series to the choice of features; the former is initialized from a constant IC and the latter from a sinusoidal IC. Medium features (EMAs of thirty, fifty, and seventy years) yield superior performance for the baseline emulator, although improvement relative to long features (EMAs of fifty, one hundred, and two hundred years) is negligible. Conversely, short features (EMAs of one, five, and ten years) lead to the fastest convergence and highest skill for the constant IC emulator. Both medium and long features exhibit slow initial convergence, suggesting that this feature-IC combination produces small gradients. These small gradients eventually amplify to facilitate optimization convergence; the short and medium features surpass the skill of the baseline emulator for the constant IC. As with the architectural sensitivity analysis, the sinusoidal IC exhibits markedly lower sensitivity to the choice of features and achieves consistent convergence patterns across all three configurations. However, the initial error among the three sinusoidal emulators varies; this suggests the optimal feature set correlates with the period of the sinusoid. The structure of the resulting emissions time series varies for both ICs across the different feature configurations. For the constant IC, both short and medium-length features yield time series with a combination of high- and low-frequency variations. In contrast, long features result in a time series with minimal variation accompanied by the lowest rate of convergence. This behavior indicates that these features are less informative for the constant IC. The sinusoidal case displays the opposite behavior; long features induce high-frequency changes in the optimized time series, whereas short features result in low-frequency changes. Medium features drive more consistent and substantial changes throughout the optimization process, which implies the presence of larger gradients relative to this feature set.

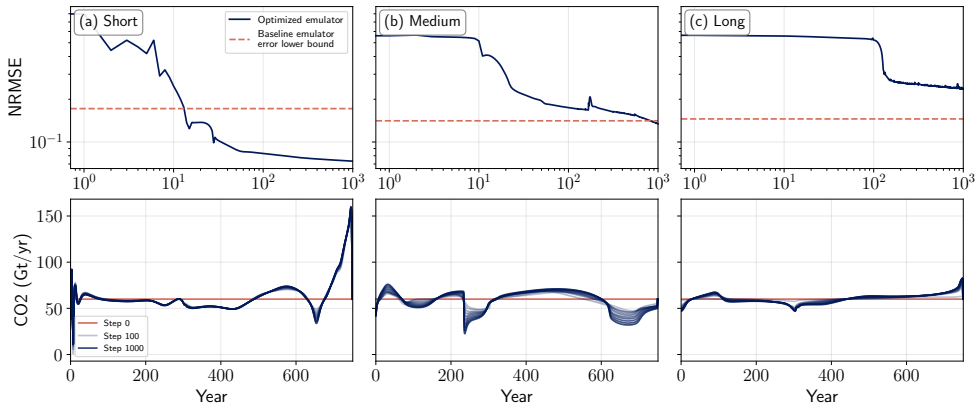

Figure B.4: Top: trajectory of evaluation loss (NRMSE) during optimization compared across three sets of features for the neural network emulator initialized from a constant initial condition: (a) short - EMAs of one, five, and ten years; (b) medium - EMAs of thirty, fifty, and seventy years; (c) long - EMAs of fifty, one hundred, and two hundred years. Emulators are evaluated on their performance in reproducing SCM-projected GMST anomalies caused by CO_2-only across all scenarios included in the ScenarioMIP, DECK, and CS3 activities; see Table B.1 for scenario descriptions. The solid, dark blue line tracks emulator performance throughout the optimization process, while the dashed, red line marks the lower bound of the baseline emulator error (evaluating performance on its own training data). Bottom: evolution of emissions time series over 1000 iterations, corresponding to the features listed above.

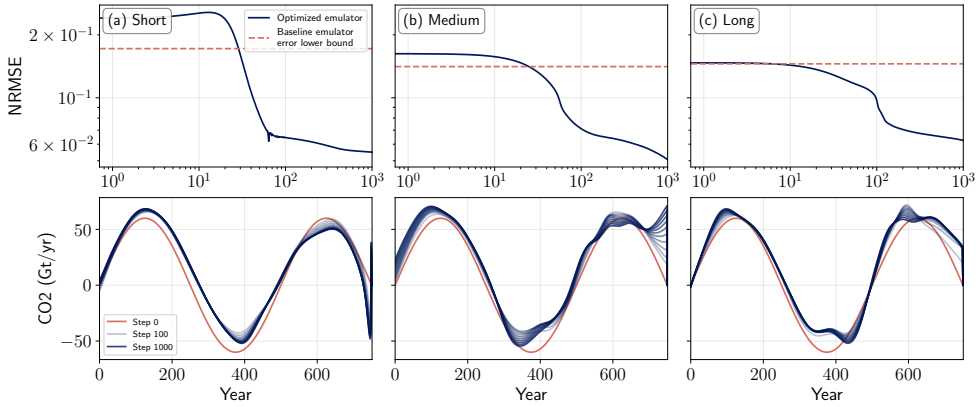

Figure B.5: Top: trajectory of evaluation loss (NRMSE) during optimization compared across three sets of features for the neural network emulator initialized from a sinusoidal initial condition: (a) short - EMAs of one, five, and ten years; (b) medium - EMAs of thirty, fifty, and seventy years; (c) long - EMAs of fifty, one hundred, and two hundred years. Emulators are evaluated on their performance in reproducing SCM-projected GMST anomalies caused by CO_2-only across all scenarios included in the ScenarioMIP, DECK, and CS3 activities; see Table B.1 for scenario descriptions. The solid, dark blue line tracks emulator performance throughout the optimization process, while the dashed, red line marks the lower bound of the baseline emulator error (evaluating performance on its own training data). Bottom: evolution of emissions time series over 1000 iterations, corresponding to the features listed above.

B.7 Extended results

Figure B.6 summarizes the full results for the individual forcing experiments (e.g., CO_2-only, CH_4-only) introduced in the main text. While the main text provides an in-depth discussion of the CO_2-only results, this section focuses on the other forcing agents. The figure illustrates the difference in emulator performance between the baseline and optimized configurations; positive values indicate an increase in performance relative to the baseline, while negative values indicate a decrease. Each configuration corresponds to a different optimization target (e.g., Opt. Priority 1 corresponds to optimizing over all of ScenarioMIP-CMIP7 Priority 1).

As in the CO_2-only case, optimizing over all scenarios leads to increased performance over all scenarios simultaneously, rather than overfitting to a specific scenario type. However, the increase in average performance varies across forcing agents; BC exhibits the largest improvements,

whereas sulfur shows the smallest. The ability of the optimized emulators to generalize across the full set of scenario structures suggests that the training data are more informative overall. In all cases except sulfur, the optimization improves skill across Priority 1; a separate discussion of sulfur follows below. Optimizing for performance over the DECK decreases predictive skill when emulating the other scenario sets. This result is expected because the DECK scenarios are structurally dissimilar to the others. Furthermore, the small sample size (two scenarios) fails to provide the optimizer with sufficient information regarding the most generally informative features. A similar phenomenon occurs when optimizing over CS3, wherein the small number of scenarios (two) leads to overfitting and a subsequent loss of extrapolative skill for CH_4, N_2O, and sulfur. In contrast, black carbon-only scenarios show improvement in all cases. This suggests that the baseline emulator configuration is ill-suited for capturing BC behavior. Changes to baseline emulator features and/or architecture may decrease the performance gap to the optimized emulator in this instance.

One experiment is notable within this suite: optimizing over Priority 2 in the CH_4-only case. Here, performance improvements relative to the baseline were unattainable on any evaluation set, a result similar to the sulfur-only experiments. Further investigation revealed two primary factors causing this result. First, the baseline emulator is well-tuned for these specific methane scenarios. Second, this optimization target is ill-conditioned for methane and exhibits high sensitivity to changes in all optimization hyperparameters. The state-dependence of atmospheric lifetime of methane may be the source of this ill-conditioning, as Priority 2 contains more long-duration scenarios than any other evaluation set. The nonlinear lifetime of methane therefore plays a larger role, and small changes in the optimized time series may lead to vastly different representations of this behavior. As a result, the optimizer oscillates between solutions and fails to find the global minimum. Despite this, adding more scenarios (i.e., optimizing over all datasets) resolves the issue, highlighting the importance of scenario diversity in the optimization process.

Although performance for the sulfur-only optimized emulator falls below the baseline more frequently than the other forcing agents, this result stems from the high skill of the baseline emulator rather than a failure of the optimization process. Unlike CO_2 or CH_4, which exhibit complex, nonlinear atmospheric residence times dependent on concentrations and temperature, forcing the SCM solely with sulfur yields temperature anomalies that are approximately linear in sulfur and respond almost instantaneously. Because this input-output mapping is structurally simple, the standard ScenarioMIP-CMIP7 baseline data already provides enough information for the emulator to learn the underlying physical relationship. As a result, the baseline emulator is effectively near its performance ceiling, leaving negligible margin for improvement via data optimization.

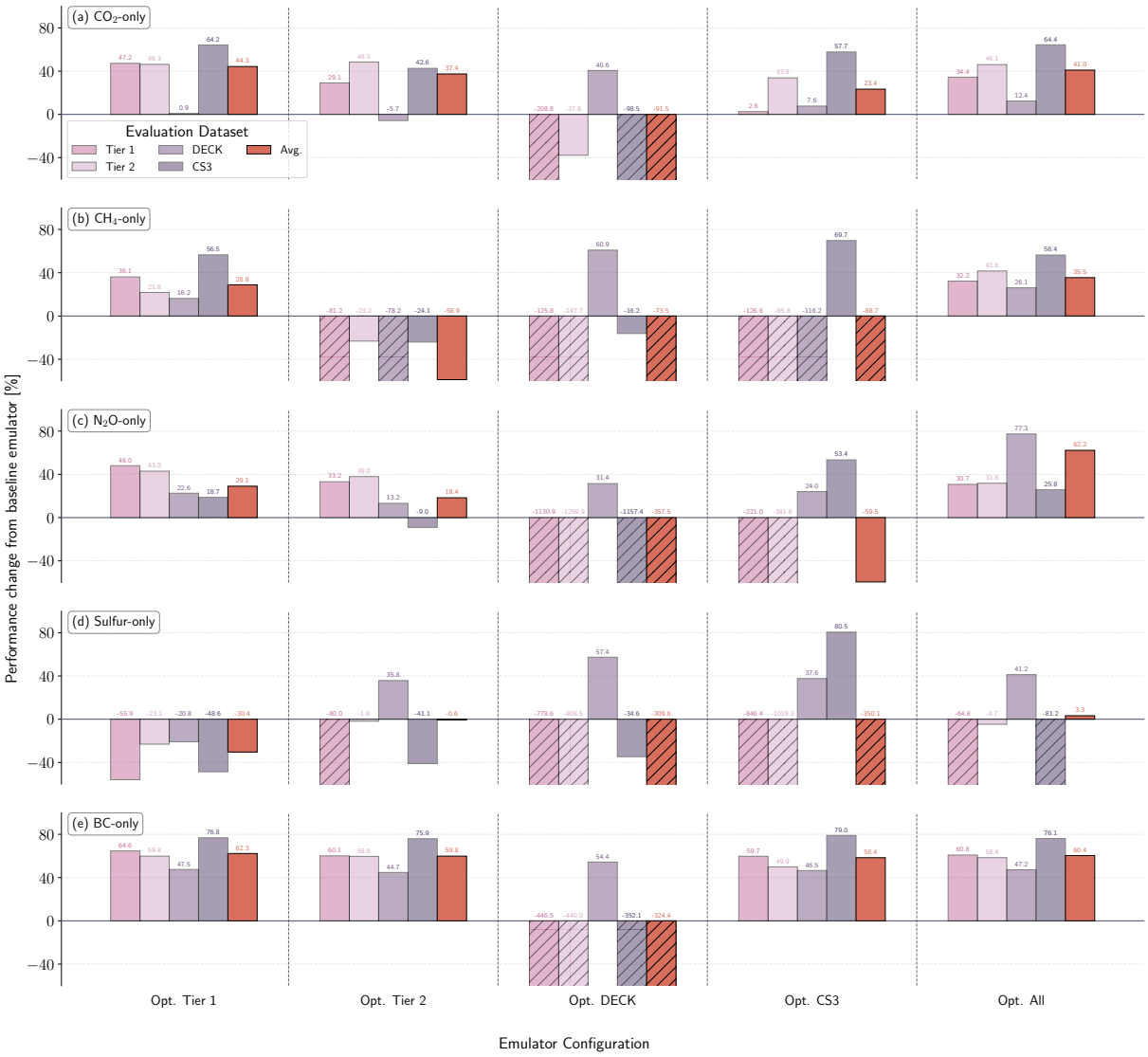

Figure B.6: Relative change in emulator predictive skill (NRMSE) for optimized emulators compared to the baseline configuration. Panels show results for (a) CO_2-only; (b) CH_4-only; (c) N_2O-only; (d) Sulfur-only; (e) BC. Positive values indicate reduced error (increased skill). Bars represent the average performance over all scenarios within a given evaluation dataset. Emulators are categorized by the subset of data used during optimization: ScenarioMIP Priority 1 and 2, DECK, CS3, and the full combined dataset. Hatching indicates a performance decrease that extends beyond the y-axis limits; y-axis limits are chosen for visual clarity.

C

Appendices for Chapter 3

C.1 Statistical significance testing

To formally assess whether climate outcomes between policy scenarios are statistically distinguishable, we evaluate the null hypothesis (H_0) that the mean (μ) of the anomaly distribution at a given spatial location s does not shift between two scenarios (e.g., *Reference* vs. 1.5°C). The alternative hypothesis (H_a) asserts that the means are statistically distinguishable:

$$H_0 : \mu_{Ref,\,s} = \mu_{1.5,\,s} \quad \text{vs.} \quad H_a : \mu_{Ref,\,s} \neq \mu_{1.5,\,s}. \tag{C.1}$$

To account for the multiple hypothesis testing problem inherent in grid-cell-wise comparisons, we apply the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR)[236,237]. We calculate p-values from all m grid cells using an independent two-sample t test for each distribution. We then sort them in ascending order ($p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$) and identify the largest rank k that satisfies the condition:

$$p_{(k)} \leq \frac{k}{m}\alpha_{\text{global}}, \tag{C.2}$$

where α_{global} is the target global significance level (0.05). We reject the null hypothesis for all grid cells corresponding to p-values $p_{(1)}$ through $p_{(k)}$. This ensures that the expected proportion of false rejections among the rejected hypotheses remains below α_{global}.

[236] Wilks, 2006; [237] Wilks, 2016

C.2 Additional results

C.2.1 Daily wet-bulb temperature standard deviation

We calculate daily wet-bulb temperatures using the same approximation from Stull (2011)[134] as the monthly averaged values. However, we take these daily data directly from the MPI-ESM1-2-LR *piControl* experimental output, rather than emulating it. From these data, we calculate the average daily standard deviation for each month. Figure C.1 shows the regional mean summer daily wet-bulb temperature standard deviation against the regional cumulative wet-bulb degree-days above 25°C emulated in Section 3.2.3. We see that some regions have a summer wet-bulb temperature standard deviation as high as 3°C, though these regions are typically not exhibiting monthly-average wet-bulb temperatures above 25°C. The most dangerous regions are those that both exhibit high monthly-average wet-bulb temperature and a high daily summer standard deviation (e.g., CNA, ENA, WCA, ARP, EAS). In these regions, we expect daily wet-bulb temperatures to frequently exceed safe thresholds (e.g., 28°C or potentially 31°C).

[134] Stull, 'Wet-Bulb Temperature from Relative Humidity and Air Temperature', *Journal of Applied Meteorology and Climatology*, 2011

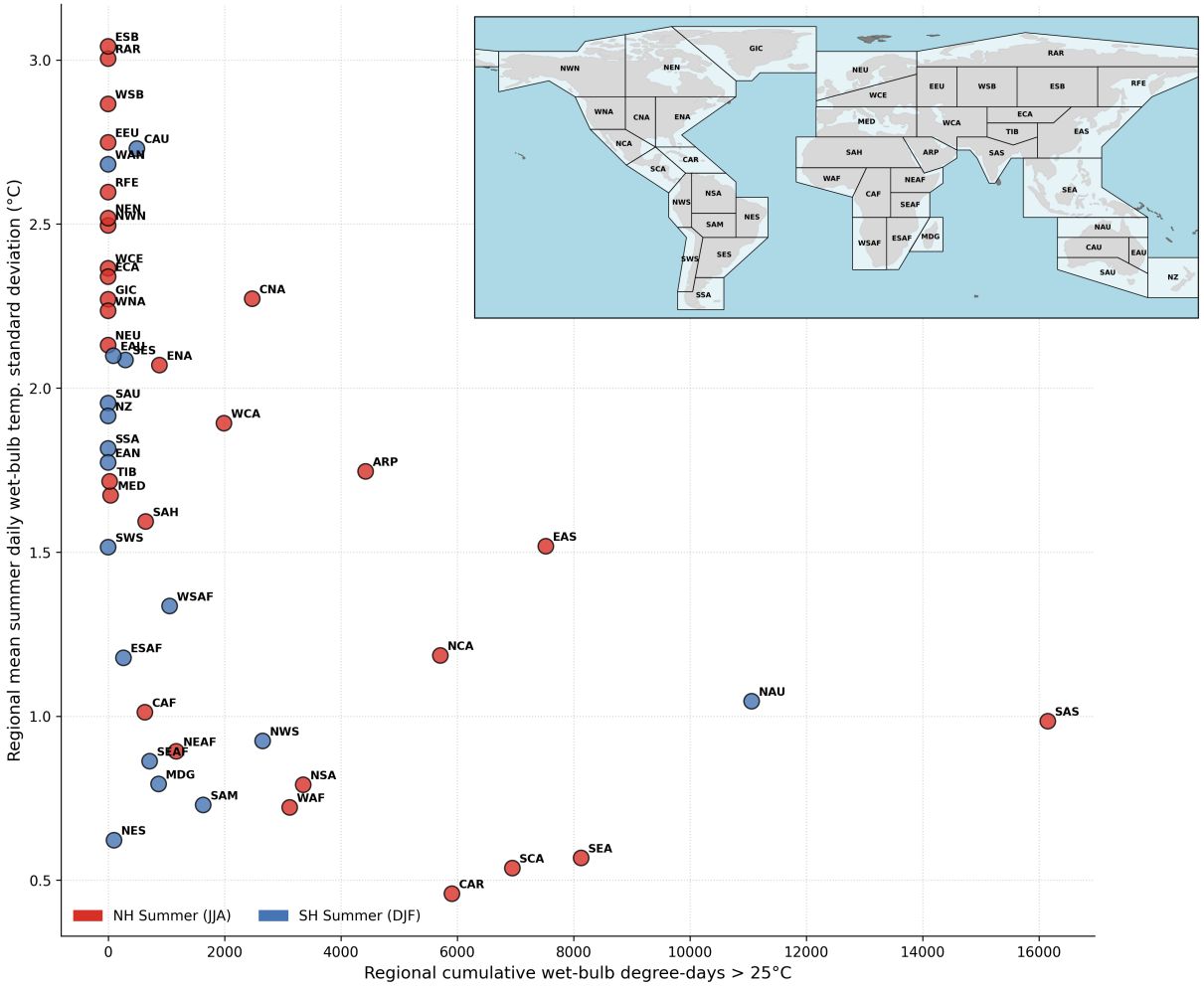

Figure C.1: Mean summer daily wet-bulb temperature standard deviation plotted against cumulative wet-bulb degree-days over 25°C for all IPCC AR6 regions. Inset map provides a visual reference for IPCC AR6 regions. Northern Hemisphere (NH) summer (JJA) is given in red and Southern Hemisphere (SH) summer (DJF) is given in blue.

C.2.2 Additional benchmarking results

Fig. C.2 illustrates results for benchmarking the generative emulator against the baseline IGSM pattern scaling approach for the *Reference* scenario at the end of the century (2090-2100). As was the case with the mid-century results presented in the main text, the generative emulator is consistent with the baseline technique in reproducing mean climatological states. Discrepancies between the two approaches primarily stem from the baseline method reproducing high-resolution ESM data, features of which are retained even when regridding to match the lower resolution of the generative emulator.

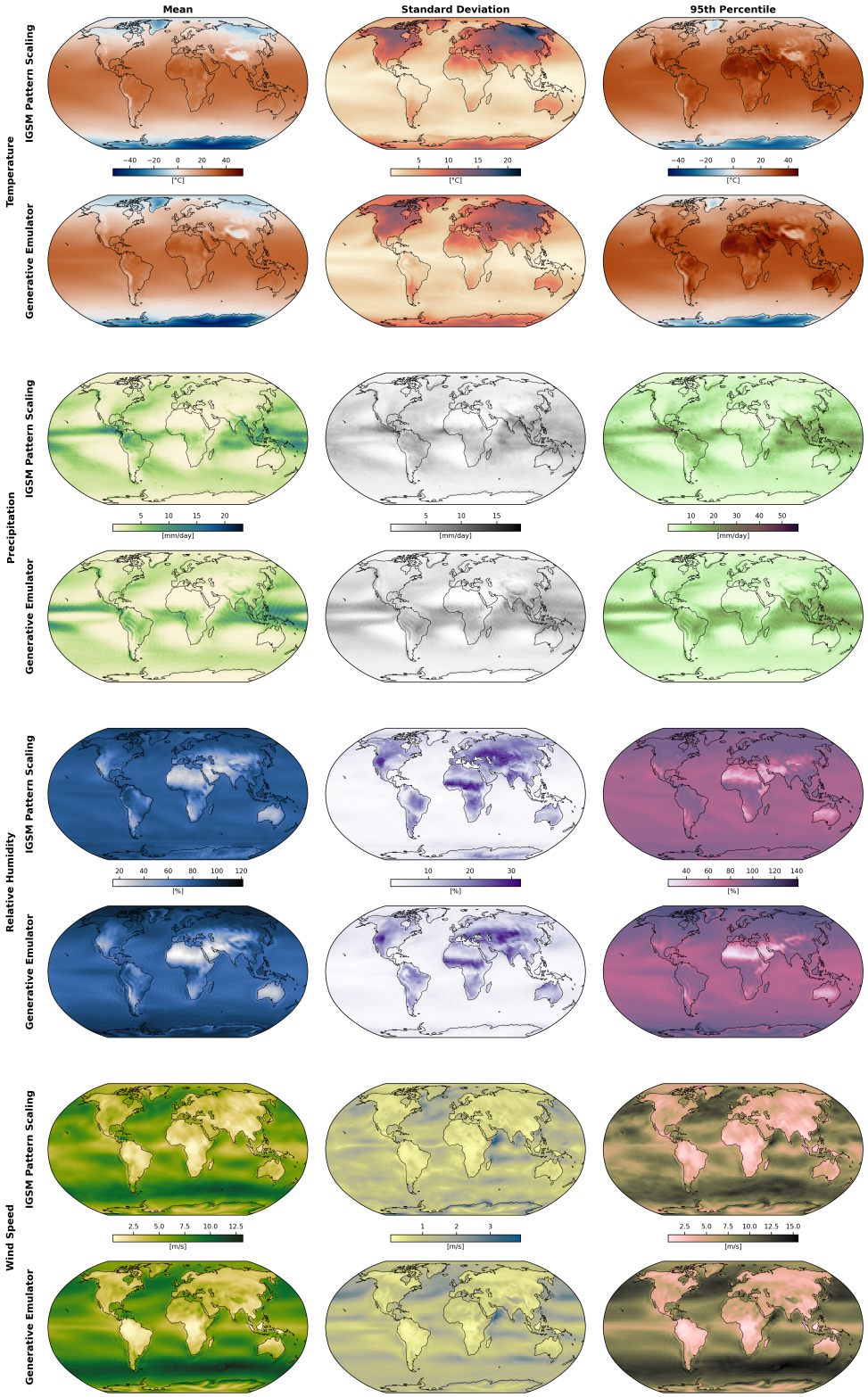

Figure C.2: Mean, standard deviation, and 95th percentile of emulated climate fields for the Emissions Prediction and Policy Analysis (EPPA) *Reference* scenario between 2090-2100 emulated with the MIT Integrated Global Systems Model (IGSM) + pattern scaling (upper half of each row) and our generative emulator (lower half of each row). Climate fields are given from top to bottom as: near-surface air temperature, precipitation, relative humidity, and near-surface wind speed.

List of terms

A
AD Automatic Differentiation.

B
BC Black Carbon.

C
CGE Computable General Equilibrium.
CMIP the Coupled Model Intercomparison Project.
CMIP6 the sixth phase of the Coupled Model Intercomparison Project.
CMIP7 the seventh phase of the Coupled Model Intercomparison Project.
CONUS Continental United States.
CS3 MIT Center for Sustainability Science and Strategy.

D
DAMIP Detection and Attribution MIP.
DMD Dynamic Mode Decomposition.

E
EDMD Extended DMD.
EMA Exponential Moving Average.
EMIC Earth system Model of Intermediate Complexity.
EPPA Emissions Prediction and Policy Analysis.
ESM Earth System Model.

F
FaIR Finite Amplitude Impulse Response.
FDR False Discovery Rate.
FDT Fluctuation Dissipation Theorem.

G
GeoMIP Geoengineering MIP.
GMST Global Mean Surface Temperature.

I
IAM Integrated Assessment Model.
IC Initial Condition.
IGSM Integrated Global Systems Model.
IPCC Intergovernmental Panel on Climate Change.

M
MESM MIT Earth System Model.
MIP Model Intercomparison Project.
ML Machine Learning.
MLP Multi-Layer Perceptron.
MSE Mean Squared Error.

N
NRMSE Normalized Root Mean Square Error.

P
PNA Pacific-North American.

S
SCM Simple Climate Model.
SGD Stochastic Gradient Descent.
SSP Shared Socioeconomic Pathway.

V
VPD Vapor Pressure Deficit.

W
WBDD Wet-Bulb Degree-Days.

Bibliography

[1] Langdon Winner. 'Do Artifacts Have Politics?' In: *Daedalus* 109.1 (1980), pp. 121–136 (cited on page 8).

[2] Christopher B. Womack et al. 'A theoretical framework to understand sources of error in Earth System Model emulation'. English. In: *Earth System Dynamics* 17.1 (Jan. 2026), pp. 107–139. DOI: 10.5194/esd-17-107-2026 (cited on pages 11, 16, 17, 19, 57, 61, 67, 68, 89, 105).

[3] 'Summary for Policymakers'. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Intergovernmental Panel on Climate Change (IPCC). Cambridge: Cambridge University Press, 2023, pp. 3–32. DOI: 10.1017/9781009157896.001 (cited on page 14).

[4] S. Mark Howden et al. 'Adapting agriculture to climate change'. In: *Proceedings of the National Academy of Sciences* 104.50 (Dec. 2007), pp. 19691–19696. DOI: 10.1073/pnas.0701890104 (cited on pages 14, 17).

[5] Drury B. Crawley. 'Estimating the impacts of climate change and urbanization on building performance'. In: *Journal of Building Performance Simulation* 1.2 (June 2008), pp. 91–115. DOI: 10.1080/19401490802182079 (cited on pages 14, 17, 56).

[6] Cédric Clastres. 'Smart grids: Another step towards competition, energy security and climate change objectives'. In: *Energy Policy* 39.9 (Sept. 2011), pp. 5399–5408. DOI: 10.1016/j.enpol.2011.05.024 (cited on pages 14, 17).

[7] Hari Bansha Dulal, Gernot Brodnig, and Charity G. Onoriose. 'Climate change mitigation in the transport sector through urban planning: A review'. In: *Habitat International* 35.3 (July 2011), pp. 494–500. DOI: 10.1016/j.habitatint.2011.02.001 (cited on pages 14, 17).

[8] Muhuddin Rajin Anwar et al. 'Adapting agriculture to climate change: a review'. en. In: *Theoretical and Applied Climatology* 113.1 (July 2013), pp. 225–245. DOI: 10.1007/s00704-012-0780-1 (cited on pages 14, 17).

[9] C. Adam Schlosser et al. 'The future of global water stress: An integrated assessment'. en. In: *Earth's Future* 2.8 (2014), pp. 341–361. DOI: 10.1002/2014EF000238 (cited on pages 14, 15).

[10] P. Döll et al. 'Integrating risks of climate change into water management'. In: *Hydrological Sciences Journal* 60.1 (Jan. 2015), pp. 4–13. DOI: 10.1080/02626667.2014.967250 (cited on page 14).

[11] Yunfang Jiang et al. 'A Review of Urban Planning Research for Climate Change'. en. In: *Sustainability* 9.12 (Dec. 2017), p. 2224. DOI: 10.3390/su9122224 (cited on page 14).

[12] A. T. D. Perera et al. 'Quantifying the impacts of climate change and extreme climate events on energy systems'. en. In: *Nature Energy* 5.2 (Feb. 2020), pp. 150–159. DOI: 10.1038/s41560-020-0558-0 (cited on pages 14, 17).

[13] Seleshi G. Yalew et al. 'Impacts of climate change on energy systems in global and regional scenarios'. en. In: *Nature Energy* 5.10 (Aug. 2020), pp. 794–802. DOI: 10.1038/s41560-020-0664-z (cited on pages 14, 17, 56).

[14] Andrew Hultgren et al. 'Impacts of climate change on global agriculture accounting for adaptation'. en. In: *Nature* 642.8068 (June 2025), pp. 644–652. DOI: 10.1038/s41586-025-09085-w (cited on pages 14, 17, 56).

[15] Etienne Piguet, Antoine Pécoud, and Paul de Guchteneire. 'Migration and Climate Change: An Overview'. In: *Refugee Survey Quarterly* 30.3 (Sept. 2011), pp. 1–23. DOI: 10.1093/rsq/hdr006 (cited on page 14).

[16] Céline Bellard et al. 'Impacts of climate change on the future of biodiversity'. en. In: *Ecology Letters* 15.4 (2012), pp. 365–377. DOI: 10.1111/j.1461-0248.2011.01736.x (cited on page 14).

[17] Scott C. Doney et al. 'Climate Change Impacts on Marine Ecosystems'. In: *Annual Review of Marine Science* 4.1 (2012), pp. 11–37. DOI: 10.1146/annurev-marine-041911-111611 (cited on page 14).

[18] Gennaro D'Amato et al. 'Climate change and respiratory diseases'. EN. In: *European Respiratory Review* 23.132 (May 2014), pp. 161–169. DOI: 10.1183/09059180.00001714 (cited on page 14).

[19] Tomoko Hasegawa et al. 'Economic implications of climate change impacts on human health through undernourishment'. en. In: *Climatic Change* 136.2 (May 2016), pp. 189–202. DOI: 10.1007/s10584-016-1606-4 (cited on page 14).

[20] David J. Kaczan and Jennifer Orgill-Meyer. 'The impact of climate change on migration: a synthesis of recent empirical insights'. en. In: *Climatic Change* 158.3 (Feb. 2020), pp. 281–300. DOI: 10.1007/s10584-019-02560-0 (cited on page 14).

[21] Muzafar Shah Habibullah et al. 'Impact of climate change on biodiversity loss: global evidence'. en. In: *Environmental Science and Pollution Research* 29.1 (Jan. 2022), pp. 1073–1086. DOI: 10.1007/s11356-021-15702-8 (cited on page 14).

[22] Jan C. Semenza, Joacim Rocklöv, and Kristie L. Ebi. 'Climate Change and Cascading Risks from Infectious Disease'. en. In: *Infectious Diseases and Therapy* 11.4 (Aug. 2022), pp. 1371–1390. DOI: 10.1007/s40121-022-00647-3 (cited on page 14).

[23] Jonathan A. Patz et al. 'Impact of regional climate change on human health'. en. In: *Nature* 438.7066 (Nov. 2005), pp. 310–317. DOI: 10.1038/nature04188 (cited on page 14).

[24] J. Samson et al. 'Geographic disparities and moral hazards in the predicted impacts of climate change on human populations'. en. In: *Global Ecology and Biogeography* 20.4 (2011), pp. 532–544. DOI: 10.1111/j.1466-8238.2010.00632.x (cited on pages 14, 15).

[25] Richard S. J. Tol. 'The Economic Impacts of Climate Change'. In: *Review of Environmental Economics and Policy* 12.1 (Jan. 2018), pp. 4–25. DOI: 10.1093/reep/rex027 (cited on pages 14, 15).

[26] Rob Dellink, Elisa Lanzi, and Jean Chateau. 'The Sectoral and Regional Economic Consequences of Climate Change to 2060'. en. In: *Environmental and Resource Economics* 72.2 (Feb. 2019), pp. 309–363. DOI: 10.1007/s10640-017-0197-5 (cited on page 14).

[27] Karen L. O'Brien and Robin M. Leichenko. 'Winners and Losers in the Context of Global Change'. In: *Annals of the Association of American Geographers* 93.1 (Mar. 2003), pp. 89–103. DOI: 10.1111/1467-8306.93107 (cited on page 14).

[28] Kian Mintz-Woo and Justin Leroux. 'What do climate change winners owe, and to whom?' en. In: *Economics & Philosophy* 37.3 (Nov. 2021), pp. 462–483. DOI: 10.1017/S0266267120000449 (cited on page 14).

[29] Edward N Lorenz. 'Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?' en. In: (1972) (cited on pages 14, 19, 21).

[30] J. M. Gregory et al. 'A new method for diagnosing radiative forcing and climate sensitivity'. en. In: *Geophysical Research Letters* 31.3 (2004). DOI: 10.1029/2003GL018747 (cited on page 14).

[31] Nicola Maher et al. 'The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability'. en. In: *Journal of Advances in Modeling Earth Systems* 11.7 (2019), pp. 2050–2069. DOI: 10.1029/2019MS001639 (cited on pages 14, 15, 22, 72, 74).

[32] Valerio Lembo, Valerio Lucarini, and Francesco Ragone. 'Beyond Forcing Scenarios: Predicting Climate Change through Response Operators in a Coupled General Circulation Model'. en. In: *Scientific Reports* 10.1 (May 2020), p. 8668. DOI: 10.1038/s41598-020-65297-2 (cited on pages 14, 16, 20, 28, 29, 31, 32, 55).

[33] Gregory M. Flato. 'Earth system models: an overview'. en. In: *WIREs Climate Change* 2.6 (2011), pp. 783–800. DOI: 10.1002/wcc.148 (cited on pages 14, 19, 71).

[34] G. Flato et al. 'Evaluation of climate models'. eng. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, June 2014, pp. 741–866. DOI: 10.1017/CBO9781107415324.020 (cited on page 14).

[35] Nadir Jeevanjee et al. 'A perspective on climate model hierarchies'. en. In: *Journal of Advances in Modeling Earth Systems* 9.4 (2017), pp. 1760–1771. DOI: 10.1002/2017MS001038 (cited on page 14).

[36] Ed Hawkins and Rowan Sutton. 'The Potential to Narrow Uncertainty in Regional Climate Predictions'. en. In: *Bulletin of the American Meteorological Society* 90.8 (Aug. 2009), pp. 1095–1108. DOI: 10.1175/2009BAMS2607.1 (cited on pages 15, 71).

[37] Karl E. Taylor, Ronald J. Stouffer, and Gerald A. Meehl. 'An Overview of CMIP5 and the Experiment Design'. en. In: (Apr. 2012). DOI: 10.1175/BAMS-D-11-00094.1 (cited on pages 15, 72).

[38] Veronika Eyring et al. 'Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization'. English. In: *Geoscientific Model Development* 9.5 (May 2016), pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016 (cited on pages 15, 19, 56, 58, 72, 74).

[39] Hideo Shiogama et al. 'MIROC6 Large Ensemble (MIROC6-LE): experimental design and initial analyses'. English. In: *Earth System Dynamics* 14.6 (Nov. 2023), pp. 1107–1124. DOI: 10.5194/esd-14-1107-2023 (cited on pages 15, 72).

[40] Andrew D. King et al. 'Exploring climate stabilisation at different global warming levels in ACCESS-ESM-1.5'. English. In: *Earth System Dynamics* 15.5 (Oct. 2024), pp. 1353–1383. DOI: 10.5194/esd-15-1353-2024 (cited on pages 15, 72).

[41] W. A. Müller et al. 'A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR)'. en. In: *Journal of Advances in Modeling Earth Systems* 10.7 (2018), pp. 1383–1413. DOI: 10.1029/2017MS001217 (cited on pages 15, 19, 76, 78, 87).

[42] Venkatramani Balaji et al. 'CPMIP: measurements of real computational performance of Earth system models in CMIP6'. English. In: *Geoscientific Model Development* 10.1 (Jan. 2017), pp. 19–34. DOI: 10.5194/gmd-10-19-2017 (cited on pages 15, 56, 71).

[43] V. Balaji et al. 'Are general circulation models obsolete?' In: *Proceedings of the National Academy of Sciences* 119.47 (Nov. 2022), e2202075119. DOI: 10.1073/pnas.2202075119 (cited on pages 15, 71).

[44] J. A. Edmonds, M. A. Wise, and C. N. MacCracken. *Advanced Energy Technologies and Climate Change an Analysis Using the Global Change Assessment Model (GCAM)*. English. Tech. rep. PNL-9798. Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), May 1994. DOI: 10.2172/1127203 (cited on page 15).

[45] Keywan Riahi, Arnulf Grübler, and Nebojsa Nakicenovic. 'Scenarios of long-term socio-economic and environmental development under climate stabilization'. In: *Technological Forecasting and Social Change*. Greenhouse Gases - Integrated Assessment 74.7 (Sept. 2007), pp. 887–935. DOI: 10.1016/j.techfore.2006.05.026 (cited on page 15).

[46] Juan-Carlos Ciscar et al. 'Physical and economic consequences of climate change in Europe'. In: *Proceedings of the National Academy of Sciences* 108.7 (Feb. 2011), pp. 2678–2683. DOI: 10.1073/pnas.1011612108 (cited on page 15).

[47] Katherine Calvin et al. 'GCAM v5.1: representing the linkages between energy, water, land, climate, and economic systems'. English. In: *Geoscientific Model Development* 12.2 (Feb. 2019), pp. 677–698. DOI: 10.5194/gmd-12-677-2019 (cited on page 15).

[48] C. Adam Schlosser et al. 'Assessing compounding risks across multiple systems and sectors: a socio-environmental systems risk-triage approach'. English. In: *Frontiers in Climate* 5 (Apr. 2023). DOI: 10.3389/fclim.2023.1100600 (cited on page 15).

[49] Gerald R. North. 'Analytical Solution to a Simple Climate Model with Diffusive Heat Transport'. In: *Journal of the Atmospheric Sciences* 32.7 (Mar. 1975), pp. 1301–1307 (cited on page 15).

[50] Gerald R. North. 'Multiple solutions in energy balance climate models'. In: *Global and Planetary Change* 2.3 (Aug. 1990), pp. 225–235. DOI: 10.1016/0921-8181(90)90003-U (cited on page 15).

[51] M. Meinshausen, S. C. B. Raper, and T. M. L. Wigley. 'Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration'. English. In: *Atmospheric Chemistry and Physics* 11.4 (Feb. 2011), pp. 1417–1456. DOI: 10.5194/acp-11-1417-2011 (cited on pages 15, 17, 19, 56, 71).

[52] Richard J. Millar et al. 'A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions'. English. In: *Atmospheric Chemistry and Physics* 17.11 (June 2017), pp. 7213–7228. DOI: 10.5194/acp-17-7213-2017 (cited on page 15).

[53] Christopher J. Smith et al. 'FAIR v1.3: a simple emissions-based impulse response and carbon cycle model'. English. In: *Geoscientific Model Development* 11.6 (June 2018), pp. 2273–2297. DOI: 10.5194/gmd-11-2273-2018 (cited on pages 15, 71).

[54] Nicholas J. Leach et al. 'FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration'. English. In: *Geoscientific Model Development* 14.5 (May 2021), pp. 3007–3036. DOI: 10.5194/gmd-14-3007-2021 (cited on pages 15, 19, 57, 58, 69, 71–73, 107, 108).

[55] Kyle C. Armour, Cecilia M. Bitz, and Gerard H. Roe. 'Time-Varying Climate Sensitivity from Regional Feedbacks'. en. In: (July 2013). DOI: 10.1175/JCLI-D-12-00544.1 (cited on pages 15, 38, 41).

[56] Paolo Giani et al. 'Origin and Limits of Invariant Warming Patterns in Climate Models'. en. In: *Journal of Climate* (Dec. 2025). DOI: 10.1175/JCLI-D-24-0683.1 (cited on pages 15, 16, 19, 29, 30, 38, 39, 41, 44, 52, 53, 67).

[57] M. Claussen et al. 'Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models'. en. In: *Climate Dynamics* 18.7 (Mar. 2002), pp. 579–586. DOI: 10.1007/s00382-001-0200-1 (cited on page 15).

[58] Susanne L. Weber. 'The utility of Earth system Models of Intermediate Complexity (EMICs)'. en. In: *WIREs Climate Change* 1.2 (2010), pp. 243–252. DOI: 10.1002/wcc.24 (cited on page 15).

[59] Philip B. Holden et al. 'PLASIM–GENIE v1.0: a new intermediate complexity AOGCM'. English. In: *Geoscientific Model Development* 9.9 (Sept. 2016), pp. 3347–3361. DOI: 10.5194/gmd-9-3347-2016 (cited on page 15).

[60] G. Platov et al. 'A new earth's climate system model of intermediate complexity, PlaSim-ICMMG-1.0: description and performance'. en. In: *IOP Conference Series: Earth and Environmental Science* 96.1 (Nov. 2017), p. 012005. DOI: 10.1088/1755-1315/96/1/012005 (cited on page 15).

[61] Paolo Ruggieri et al. 'SPEEDY-NEMO: performance and applications of a fully-coupled intermediate-complexity climate model'. en. In: *Climate Dynamics* 62.5 (May 2024), pp. 3763–3781. DOI: 10.1007/s00382-023-07097-8 (cited on page 15).

[62] E. Monier et al. 'An integrated assessment modeling framework for uncertainty studies in global and regional climate change: the MIT IGSM-CAM (version 1.0)'. English. In: *Geoscientific Model Development* 6.6 (Dec. 2013), pp. 2063–2085. DOI: 10.5194/gmd-6-2063-2013 (cited on pages 15, 17, 71, 72).

[63] M. Eby et al. 'Historical and idealized climate model experiments: an intercomparison of Earth system models of intermediate complexity'. English. In: *Climate of the Past* 9.3 (May 2013), pp. 1111–1140. DOI: 10.5194/cp-9-1111-2013 (cited on page 15).

[64] C. Tebaldi et al. 'Emulators of Climate Model Output'. en. In: *Annual Review of Environment and Resources* 50.Volume 50, 2025 (Oct. 2025), pp. 709–737. DOI: 10.1146/annurev-environ-012125-085838 (cited on pages 15, 17, 19, 20, 29, 35, 52, 54, 56, 57, 71).

[65] Hideo Shiogama, Jun'ya Takakura, and Kiyoshi Takahashi. 'Uncertainty constraints on economic impact assessments of climate change simulated by an impact emulator'. en. In: *Environmental Research Letters* 17.12 (Dec. 2022), p. 124028. DOI: 10.1088/1748-9326/aca68d (cited on pages 15, 71).

[66] Gregory Munday et al. 'Risks of unavoidable impacts on forests at 1.5 °C with and without overshoot'. en. In: *Nature Climate Change* 15.6 (June 2025), pp. 650–655. DOI: 10.1038/s41558-025-02327-9 (cited on pages 15, 71).

[67] Pascal Polonik, Jennifer Burney, and Katharine Ricke. 'Emulation of the Climate Response to Greenhouse Gas and Aerosol Emissions From High- and Low-Income Nations'. en. In: *Geophysical Research Letters* 52.21 (2025), e2025GL117841. DOI: 10.1029/2025GL117841 (cited on pages 15, 71).

[68] Rebecca M. Varney et al. 'Northern high latitudes could become a net carbon source below 2 °C global warming'. English. In: *EGUsphere* (Jan. 2026), pp. 1–22. DOI: 10.5194/egusphere-2025-6075 (cited on pages 15, 71).

[69] Lea Beusch et al. 'Responsibility of major emitters for country-level warming and extreme hot years'. en. In: *Communications Earth & Environment* 3.1 (Jan. 2022), p. 7. DOI: 10.1038/s43247-021-00320-6 (cited on pages 15, 71).

[70] Vassili Kitsios, Terence John O'Kane, and David Newth. 'A machine learning approach to rapidly project climate responses under a multitude of net-zero emission pathways'. en. In: *Communications Earth & Environment* 4.1 (Oct. 2023), pp. 1–15. DOI: 10.1038/s43247-023-01011-0 (cited on pages 15, 16, 71).

[71] Jonas Schwaab et al. 'Spatially resolved emulated annual temperature projections for overshoot pathways'. en. In: *Scientific Data* 11.1 (Nov. 2024), p. 1262. DOI: 10.1038/s41597-024-04122-1 (cited on pages 15, 71).

[72] Sarah Schöngart et al. 'High-income groups disproportionately contribute to climate extremes worldwide'. en. In: *Nature Climate Change* (May 2025), pp. 1–7. DOI: 10.1038/s41558-025-02325-x (cited on pages 15, 71).

[73] Yann Quilcaille et al. 'Systematic attribution of heatwaves to the emissions of carbon majors'. en. In: *Nature* 645.8080 (Sept. 2025), pp. 392–398. DOI: 10.1038/s41586-025-09450-9 (cited on pages 15, 71).

[74] B. Santer et al. 'Developing climate scenarios from equilibrium GCM results'. In: 1990 (cited on pages 16, 19, 29, 71).

[75] Michael E. Schlesinger et al. 'Geographical Distributions of Temperature Change for Scenarios of Greenhouse Gas and Sulfur Dioxide Emissions'. In: *Technological Forecasting and Social Change* 65.2 (Oct. 2000), pp. 167–193. DOI: 10.1016/S0040-1625(99)00114-6 (cited on pages 16, 19).

[76] Timothy D. Mitchell. 'Pattern Scaling: An Examination of the Accuracy of the Technique for Describing Future Climates'. en. In: *Climatic Change* 60.3 (Oct. 2003), pp. 217–242. DOI: 10.1023/A:1026035305597 (cited on pages 16, 19, 29, 52, 71).

[77] Nadja Herger, Benjamin M. Sanderson, and Reto Knutti. 'Improved pattern scaling approaches for the use in climate impact studies'. en. In: *Geophysical Research Letters* 42.9 (2015), pp. 3486–3494. DOI: 10.1002/2015GL063569 (cited on pages 16, 19, 25, 30).

[78] Xiang Gao, Andrei Sokolov, and C. Adam Schlosser. 'A Large Ensemble Global Dataset for Climate Impact Assessments'. en. In: *Scientific Data* 10.1 (Nov. 2023), p. 801. DOI: 10.1038/s41597-023-02708-9 (cited on pages 16, 17, 72, 75, 87).

[79] Claudia Tebaldi and Julie M. Arblaster. 'Pattern scaling: Its strengths and limitations, and an update on the latest model simulations'. en. In: *Climatic Change* 122.3 (Feb. 2014), pp. 459–471. DOI: 10.1007/s10584-013-1032-9 (cited on pages 16, 19, 29, 52, 71, 75).

[80] Christopher D. Wells et al. 'Understanding pattern scaling errors across a range of emissions pathways'. English. In: *Earth System Dynamics* 14.4 (Aug. 2023), pp. 817–834. DOI: 10.5194/esd-14-817-2023 (cited on pages 16, 19, 25, 29, 44, 52).

[81] Lea Beusch et al. 'From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: coupling of MAGICC (v7.5.1) and MESMER (v0.8.3)'. English. In: *Geoscientific Model Development* 15.5 (Mar. 2022), pp. 2085–2103. DOI: 10.5194/gmd-15-2085-2022 (cited on page 16).

[82] Camilla Mathison et al. 'A rapid-application emissions-to-impacts tool for scenario assessment: Probabilistic Regional Impacts from Model patterns and Emissions (PRIME)'. English. In: *Geoscientific Model Development* 18.5 (Mar. 2025), pp. 1785–1808. DOI: 10.5194/gmd-18-1785-2025 (cited on pages 16, 17, 19, 57, 71, 75).

[83] F. Joos and M. Bruno. 'Pulse response functions are cost-efficient tools to model the link between carbon emissions, atmospheric CO2 and global warming'. In: *Physics and Chemistry of the Earth*. Ocean and Atmosphere 21.5 (Oct. 1996), pp. 471–476. DOI: 10.1016/S0079-1946(97)81144-5 (cited on pages 16, 20, 27, 31).

[84] Clara Orbe et al. 'Large-scale tropospheric transport in the Chemistry–Climate Model Initiative (CCMI) simulations'. English. In: *Atmospheric Chemistry and Physics* 18.10 (May 2018), pp. 7217–7235. DOI: 10.5194/acp-18-7217-2018 (cited on pages 16, 20, 27, 31).

[85] Laura Cimoli et al. 'Annually Resolved Propagation of CFCs and SF6 in the Global Ocean Over Eight Decades'. en. In: *Journal of Geophysical Research: Oceans* 128.3 (2023), e2022JC019337. DOI: 10.1029/2022JC019337 (cited on pages 16, 27, 31).

[86] K. Hasselmann et al. 'Sensitivity Study of Optimal CO2 Emission Paths Using a Simplified Structural Integrated Assessment Model (SIAM)'. en. In: *Climatic Change* 37.2 (Oct. 1997), pp. 345–386. DOI: 10.1023/A:1005339625015 (cited on pages 16, 20, 31).

[87] Klaus Hasselmann. '1.27 - Optimizing Long-Term Climate Management'. In: *Global Biogeochemical Cycles in the Climate System*. Ed. by Ernst-Detlef Schulze et al. San Diego: Academic Press, Jan. 2001, pp. 333–343. DOI: 10.1016/B978-012631260-7/50029-7 (cited on pages 16, 31).

[88] K. Hasselmann et al. 'The Challenge of Long-Term Climate Change'. In: *Science* 302.5652 (Dec. 2003), pp. 1923–1925. DOI: 10.1126/science.1090858 (cited on pages 16, 27).

[89] Hege-Beate Fredriksen, Maria Rugenstein, and Rune Graversen. 'Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response'. en. In: *Journal of Geophysical Research: Atmospheres* 126.24 (2021), e2020JD034145. DOI: 10.1029/2020JD034145 (cited on pages 16, 31, 34, 35).

[90] Hege-Beate Fredriksen et al. '21st Century Scenario Forcing Increases More for CMIP6 Than CMIP5 Models'. en. In: *Geophysical Research Letters* 50.6 (2023), e2023GL102916. DOI: 10.1029/2023GL102916 (cited on pages 16, 31, 34).

[91] Christopher B. Womack et al. 'Rapid Emulation of Spatially Resolved Temperature Response to Effective Radiative Forcing'. en. In: *Journal of Advances in Modeling Earth Systems* 17.1 (2025), e2024MS004523. DOI: 10.1029/2024MS004523 (cited on pages 16, 17, 20, 27, 29, 31–33, 52, 54, 56, 75, 89).

[92] Marit Sandstad et al. 'METEORv1.0.1: a novel framework for emulating multi-timescale regional climate responses'. English. In: *Geoscientific Model Development* 18.21 (Nov. 2025), pp. 8269–8312. DOI: 10.5194/gmd-18-8269-2025 (cited on pages 16, 27, 28, 31, 34, 52, 75).

[93] Valerio Lucarini, Francesco Ragone, and Frank Lunkeit. 'Predicting Climate Change Using Response Theory: Global Averages and Spatial Patterns'. en. In: *Journal of Statistical Physics* 166.3 (Feb. 2017), pp. 1036–1064. DOI: 10.1007/s10955-016-1506-z (cited on pages 16, 20, 28, 31, 32, 52, 53, 55).

[94] Valerio Lucarini and Mickaël D. Chekroun. 'Detecting and Attributing Change in Climate and Complex Systems: Foundations, Green's Functions, and Nonlinear Fingerprints'. en. In: *Physical Review Letters* 133.24 (Dec. 2024), p. 244201. DOI: 10.1103/PhysRevLett.133.244201 (cited on pages 16, 52).

[95] Jakob Zscheischler et al. 'A typology of compound weather and climate events'. en. In: *Nature Reviews Earth & Environment* 1.7 (July 2020), pp. 333–347. DOI: 10.1038/s43017-020-0060-z (cited on page 16).

[96] Camilla Mathison et al. 'Description and evaluation of the JULES-ES set-up for ISIMIP2b'. English. In: *Geoscientific Model Development* 16.14 (July 2023), pp. 4249–4264. DOI: 10.5194/gmd-16-4249-2023 (cited on page 16).

[97] Stefano Castruccio et al. 'Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs'. EN. In: *Journal of Climate* 27.5 (Mar. 2014), pp. 1829–1844. DOI: 10.1175/JCLI-D-13-00099.1 (cited on pages 16, 17, 19, 56).

[98] Annalisa Bracco et al. *Machine Learning for the Physics of Climate*. Aug. 2024. DOI: 10.48550/arXiv.2408.09627 (cited on pages 16, 56).

[99] Björn Lütjens et al. 'The Impact of Internal Variability on Benchmarking Deep Learning Climate Emulators'. en. In: *Journal of Advances in Modeling Earth Systems* 17.8 (2025), e2024MS004619. DOI: 10.1029/2024MS004619 (cited on pages 16, 17, 57).

[100] D. Watson-Parris et al. 'ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections'. en. In: *Journal of Advances in Modeling Earth Systems* 14.10 (2022), e2021MS002954. DOI: 10.1029/2021MS002954 (cited on pages 16, 17, 19, 54, 57).

[101] Katie Christensen et al. *Diffusion-Based Joint Temperature and Precipitation Emulation of Earth System Models*. Apr. 2024. DOI: 10.48550/arXiv.2404.08797 (cited on page 16).

[102] Cynthia Rudin. 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. en. In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x (cited on page 16).

[103] Shahine Bouabid, Dino Sejdinovic, and Duncan Watson-Parris. 'FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation'. en. In: *Journal of Advances in Modeling Earth Systems* 16.6 (2024), e2023MS003926. DOI: 10.1029/2023MS003926 (cited on pages 16, 17, 19, 25, 56, 57, 76).

[104] Alexander J. Winkler and Carlos A. Sierra. 'Towards a New Generation of Impulse-Response Functions for Integrated Earth System Understanding and Climate Change Attribution'. en. In: *Geophysical Research Letters* 52.8 (2025), e2024GL112295. DOI: 10.1029/2024GL112295 (cited on pages 16, 20, 28, 52, 54).

[105] Seth Bassetti et al. 'DiffESM: Conditional Emulation of Temperature and Precipitation in Earth System Models With 3D Diffusion Models'. en. In: *Journal of Advances in Modeling Earth Systems* 16.10 (2024), e2023MS004194. DOI: 10.1029/2023MS004194 (cited on pages 16, 17, 19, 20, 25, 56, 88).

[106] Shahine Bouabid, Andre Nogueira Souza, and Raffaele Ferrari. 'Score-Based Generative Emulation of Impact-Relevant Earth System Model Outputs'. en. In: *Journal of Advances in Modeling Earth Systems* 18.3 (2026), e2025MS005558. DOI: 10.1029/2025MS005558 (cited on pages 16, 17, 20, 54, 56, 72, 74, 75, 78, 87).

[107] Oliver Watt-Meyer et al. *ACE: A fast, skillful learned global atmospheric model for climate prediction*. Dec. 2023. DOI: 10.48550/arXiv.2310.02074 (cited on page 16).

[108] James P. C. Duncan et al. 'Application of the AI2 Climate Emulator to E3SMv2's Global Atmosphere Model, With a Focus on Precipitation Fidelity'. en. In: *Journal of Geophysical Research: Machine Learning and Computation* 1.3 (2024), e2024JH000136. DOI: 10.1029/2024JH000136 (cited on page 16).

[109] Dmitrii Kochkov et al. 'Neural general circulation models for weather and climate'. en. In: *Nature* (July 2024), pp. 1–7. DOI: 10.1038/s41586-024-07744-y (cited on pages 16, 17, 20, 56, 68).

[110] William E. Chapman et al. *CAMulator: Fast Emulation of the Community Atmosphere Model*. Apr. 2025. DOI: 10.48550/arXiv.2504.06007 (cited on page 16).

[111] Nathaniel Cresswell-Clay et al. 'A Deep Learning Earth System Model for Efficient Simulation of the Observed Climate'. en. In: *AGU Advances* 6.4 (2025), e2025AV001706. DOI: 10.1029/2025AV001706 (cited on page 16).

[112] James P. C. Duncan et al. *SamudrACE: Fast and Accurate Coupled Climate Modeling with 3D Ocean and Atmosphere Emulators*. Sept. 2025. DOI: 10.48550/arXiv.2509.12490 (cited on page 16).

[113] Spencer K. Clark et al. 'ACE2-SOM: Coupling an ML Atmospheric Emulator to a Slab Ocean and Learning the Sensitivity of Climate to Changed CO2'. en. In: *Journal of Geophysical Research: Machine Learning and Computation* 2.4 (2025), e2024JH000575. DOI: 10.1029/2024JH000575 (cited on page 16).

[114] Katharine Rucker et al. *Benchmarking Regional Thermodynamic Trends in an AI emulator, ACE2, and a hybrid model, NeuralGCM*. Nov. 2025. DOI: 10.48550/arXiv.2511.00274 (cited on page 16).

[115] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. 'Hamiltonian Neural Networks'. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cited on pages 17, 56).

[116] M. Raissi, P. Perdikaris, and G. E. Karniadakis. 'Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations'. In: *Journal of Computational Physics* 378 (Feb. 2019), pp. 686–707. DOI: 10.1016/j.jcp.2018.10.045 (cited on pages 17, 56).

[117] Arvind T. Mohan et al. *Embedding Hard Physical Constraints in Neural Network Coarse-Graining of 3D Turbulence*. Feb. 2020. DOI: 10.48550/arXiv.2002.00021 (cited on pages 17, 56).

[118] Shengze Cai et al. 'Physics-informed neural networks (PINNs) for fluid mechanics: a review'. en. In: *Acta Mechanica Sinica* 37.12 (Dec. 2021), pp. 1727–1738. DOI: 10.1007/s10409-021-01148-1 (cited on pages 17, 56).

[119] George Em Karniadakis et al. 'Physics-informed machine learning'. en. In: *Nature Reviews Physics* 3.6 (May 2021), pp. 422–440. DOI: 10.1038/s42254-021-00314-5 (cited on pages 17, 56).

[120] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 'E(n) Equivariant Graph Neural Networks'. en. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, July 2021, pp. 9323–9332 (cited on pages 17, 56).

[121] Salvatore Cuomo et al. 'Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next'. en. In: *Journal of Scientific Computing* 92.3 (July 2022), p. 88. DOI: 10.1007/s10915-022-01939-z (cited on pages 17, 56).

[122] Shaghayegh Fazliani, Zachary Frangella, and Madeleine Udell. *Enhancing Physics-Informed Neural Networks Through Feature Engineering*. June 2025. DOI: 10.48550/arXiv.2502.07209 (cited on pages 17, 56).

[123] Ye Li, Yiwen Pang, and Bin Shan. *Physics-guided Data Augmentation for Learning the Solution Operator of Linear Differential Equations*. Dec. 2022. DOI: 10.48550/arXiv.2212.04100 (cited on pages 17, 56).

[124] Michael D. Shields et al. 'Active learning applied to automated physical systems increases the rate of discovery'. en. In: *Scientific Reports* 13.1 (May 2023), p. 8402. DOI: 10.1038/s41598-023-35257-7 (cited on pages 17, 56).

[125] Yulin Guo et al. 'Active learning for adaptive surrogate model improvement in high-dimensional problems'. en. In: *Structural and Multidisciplinary Optimization* 67.7 (July 2024), p. 122. DOI: 10.1007/s00158-024-03816-9 (cited on pages 17, 56).

[126] Detlef P. Van Vuuren et al. 'The Scenario Model Intercomparison Project for CMIP7 (ScenarioMIP-CMIP7)'. English. In: *Geoscientific Model Development* 19.7 (Apr. 2026), pp. 2627–2656. DOI: 10.5194/gmd-19-2627-2026 (cited on pages 17, 54, 56–58, 72–74, 88, 108, 110).

[127] Lea Beusch, Lukas Gudmundsson, and Sonia I. Seneviratne. 'Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land'. English. In: *Earth System Dynamics* 11.1 (Feb. 2020), pp. 139–159. DOI: 10.5194/esd-11-139-2020 (cited on pages 17, 19, 25, 56, 57, 71, 75).

[128] Claudia Tebaldi, Abigail Snyder, and Kalyn Dorheim. 'STITCHES: creating new scenarios of climate model output by stitching together pieces of existing simulations'. English. In: *Earth System Dynamics* 13.4 (Nov. 2022), pp. 1557–1609. DOI: 10.5194/esd-13-1557-2022 (cited on pages 17, 57).

[129] Gosha Geogdzhayev et al. 'An EOF-Based Emulator of Means and Covariances of Monthly Climate Fields'. English. In: *Earth System Dynamics* 17.2 (Mar. 2026), pp. 235–263. DOI: 10.5194/esd-17-235-2026 (cited on pages 17, 19, 30, 57, 76).

[130] Peter Van Katwyk et al. 'Rewiring climate modeling with machine learning emulators'. en. In: *Communications Earth & Environment* 7.1 (Jan. 2026), p. 107. DOI: 10.1038/s43247-026-03238-z (cited on pages 17, 57).

[131] Stephen J. Collier, Rebecca Elliott, and Turo-Kimmo Lehtonen. 'Climate change and insurance'. en. In: *Economy and Society* 50.2 (Apr. 2021), pp. 158–172. DOI: 10.1080/03085147.2021.1903771 (cited on pages 17, 56).

[132] Fujin Zhou, Thijs Endendijk, and W.J. Wouter Botzen. 'A Review of the Financial Sector Impacts of Risks Associated with Climate Change'. en. In: *Annual Review of Resource Economics* 15.1 (Oct. 2023), pp. 233–256. DOI: 10.1146/annurev-resource-101822-105702 (cited on pages 17, 56).

[133] Ivan Sudakow, Michael Pokojovy, and Dmitry Lyakhov. 'Statistical mechanics in climate emulation: Challenges and perspectives'. en. In: *Environmental Data Science* 1 (Jan. 2022), e16. DOI: 10.1017/eds.2022.15 (cited on pages 17, 19, 56).

[134] Roland Stull. 'Wet-Bulb Temperature from Relative Humidity and Air Temperature'. EN. In: *Journal of Applied Meteorology and Climatology* 50.11 (Nov. 2011), pp. 2267–2269. DOI: 10.1175/JAMC-D-11-0143.1 (cited on pages 17, 72, 77, 83, 119).

[135] A. Park Williams et al. 'Observed Impacts of Anthropogenic Climate Change on Wildfire in California'. en. In: *Earth's Future* 7.8 (2019), pp. 892–910. DOI: 10.1029/2019EF001210 (cited on pages 17, 72, 77, 85).

[136] Mengze Wang et al. *Stochastic Emulators of Spatially Resolved Extreme Temperatures of Earth System Models*. 2025 (cited on pages 19, 20, 25, 30).

[137] Claudia Tebaldi and Reto Knutti. 'Evaluating the accuracy of climate change pattern emulation for low warming targets'. en. In: *Environmental Research Letters* 13.5 (May 2018), p. 055006. DOI: 10.1088/1748-9326/aabef2 (cited on pages 19, 29).

[138] Henry Addison et al. *Machine learning emulation of precipitation from km-scale regional climate simulations using a diffusion model*. July 2024. DOI: 10.48550/arXiv.2407.14158 (cited on page 19).

[139] Lyssa M. Freese et al. 'Spatially Resolved Temperature Response Functions to CO2 Emissions'. en. In: *Geophysical Research Letters* 51.15 (2024), e2024GL108788. DOI: 10.1029/2024GL108788 (cited on pages 20, 25, 27, 29, 31, 32, 52, 54, 75).

[140] Ludovico Theo Giorgini et al. *Response Theory via Generative Score Modeling*. July 2024. DOI: 10.48550/arXiv.2402.01029 (cited on pages 20, 28, 32, 53).

[141] C. Huntingford and P. M. Cox. 'An analogue model to derive additional climate change scenarios from existing GCM simulations'. en. In: *Climate Dynamics* 16.8 (Aug. 2000), pp. 575–586. DOI: 10.1007/s003820000067 (cited on pages 20, 30).

[142] Long Cao et al. 'Fast and slow climate responses to CO2 and solar forcing: A linear multivariate regression model characterizing transient climate change'. en. In: *Journal of Geophysical Research: Atmospheres* 120.23 (2015), pp. 12, 037–12, 053. DOI: 10.1002/2015JD023901 (cited on pages 20, 30).

[143] J. Nathan Kutz, Xing Fu, and Steven L. Brunton. 'Multiresolution Dynamic Mode Decomposition'. en. In: *SIAM Journal on Applied Dynamical Systems* 15.2 (Jan. 2016), pp. 713–735. DOI: 10.1137/15M1023543 (cited on pages 20, 35, 36).

[144] Georg A. Gottwald and Federica Gugole. 'Detecting Regime Transitions in Time Series Using Dynamic Mode Decomposition'. en. In: *Journal of Statistical Physics* 179.5 (June 2020), pp. 1028–1045. DOI: 10.1007/s10955-019-02392-3 (cited on pages 20, 35, 36).

[145] Antonio Navarra, Joe Tribbia, and Stefan Klus. 'Estimation of Koopman Transfer Operators for the Equatorial Pacific SST'. en. In: *Journal of the Atmospheric Sciences* 78.4 (Apr. 2021), pp. 1227–1244. DOI: 10.1175/JAS-D-20-0136.1 (cited on pages 20, 25–27, 35, 37).

[146] Nathan Mankovich et al. 'Analyzing climate scenarios using dynamic mode decomposition with control'. en. In: *Environmental Data Science* 4 (2025), e16. DOI: 10.1017/eds.2025.8 (cited on pages 20, 27, 35, 36).

[147] Jaideep Pathak et al. *FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators*. en. Feb. 2022. DOI: 10.48550/arXiv.2202.11214 (cited on page 20).

[148] Jared Lewis et al. 'A method to encapsulate model structural uncertainty in ensemble projections of future climate: EPIC v1.0'. English. In: *Geoscientific Model Development* 10.12 (Dec. 2017), pp. 4563–4575. DOI: 10.5194/gmd-10-4563-2017 (cited on page 20).

[149] Fenwick C. Cooper and Peter H. Haynes. 'Climate Sensitivity via a Nonparametric Fluctuation–Dissipation Theorem'. en. In: *Journal of Atmospheric Sciences* 68.5 (May 2011), pp. 937–953. DOI: 10.1175/2010JAS3633.1 (cited on page 20).

[150] Niccolò Zagli et al. *Bridging the Gap between Koopmanism and Response Theory: Using Natural Variability to Predict Forced Response*. Oct. 2024. DOI: 10.48550/arXiv.2410.01622 (cited on pages 20, 28).

[151] Ludovico T. Giorgini, Fabrizio Falasca, and Andre N. Souza. 'Predicting forced responses of probability distributions via the fluctuation–dissipation theorem and generative modeling'. In: *Proceedings of the National Academy of Sciences* 122.41 (Oct. 2025), e2509578122. DOI: 10.1073/pnas.2509578122 (cited on pages 20, 28, 55).

[152] K. Hasselmann. 'Stochastic climate models Part I. Theory'. en. In: *Tellus* 28.6 (1976), pp. 473–485. DOI: 10.1111/j.2153-3490.1976.tb00696.x (cited on pages 21, 24).

[153] Claude Frankignoul and Klaus Hasselmann. 'Stochastic climate models, Part II Application to sea-surface temperature anomalies and thermocline variability'. In: *Tellus A: Dynamic Meteorology and Oceanography* 29.4 (Jan. 1977), p. 289. DOI: 10.3402/tellusa.v29i4.11362 (cited on page 21).

[154] Brian C. O'Neill et al. 'The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6'. English. In: *Geoscientific Model Development* 9.9 (Sept. 2016), pp. 3461–3482. DOI: 10.5194/gmd-9-3461-2016 (cited on pages 22, 53).

[155] S. I. Denisov, W. Horsthemke, and P. Hänggi. 'Generalized Fokker-Planck equation: Derivation and exact solutions'. en. In: *The European Physical Journal B* 68.4 (Apr. 2009), pp. 567–575. DOI: 10.1140/epjb/e2009-00126-3 (cited on page 25).

[156] Stefan Klus, Péter Koltai, and Christof Schütte. 'On the numerical approximation of the Perron-Frobenius and Koopman operator'. In: *Journal of Computational Dynamics* 3.1 (Sept. 2016), pp. 1–12. DOI: 10.3934/jcd.2016003 (cited on page 25).

[157] Stefan Klus et al. 'Data-Driven Model Reduction and Transfer Operator Approximation'. en. In: *Journal of Nonlinear Science* 28.3 (June 2018), pp. 985–1010. DOI: 10.1007/s00332-017-9437-7 (cited on page 25).

[158] Eurika Kaiser, J. Nathan Kutz, and Steven L. Brunton. *Data-driven approximations of dynamical systems operators for control*. Feb. 2019. DOI: 10.48550/arXiv.1902.10239 (cited on page 25).

[159] Andre N. Souza. 'Representing turbulent statistics with partitions of state space. Part 1. Theory and methodology'. en. In: *Journal of Fluid Mechanics* 997 (Oct. 2024), A1. DOI: 10.1017/jfm.2024.658 (cited on page 25).

[160] Andre N. Souza. 'Representing turbulent statistics with partitions of state space. Part 2. The compressible Euler equations'. en. In: *Journal of Fluid Mechanics* 997 (Oct. 2024), A2. DOI: 10.1017/jfm.2024.657 (cited on page 25).

[161] Andre N. Souza and Simone Silvestri. *A Modified Bisecting K-Means for Approximating Transfer Operators: Application to the Lorenz Equations*. Dec. 2024. DOI: 10.48550/arXiv.2412.03734 (cited on page 25).

[162] Igor Mezić. 'Analysis of Fluid Flows via Spectral Properties of the Koopman Operator'. en. In: *Annual Review of Fluid Mechanics* 45.1 (Jan. 2013), pp. 357–378. DOI: 10.1146/annurev-fluid-011212-140652 (cited on pages 26, 27).

[163] Samuel E. Otto and Clarence W. Rowley. 'Koopman Operators for Estimation and Control of Dynamical Systems'. en. In: *Annual Review of Control, Robotics, and Autonomous Systems* 4.1 (May 2021), pp. 59–87. DOI: 10.1146/annurev-control-071020-010108 (cited on pages 26, 27).

[164] Joanna Slawinska, Eniko Szekely, and Dimitrios Giannakis. *Data-Driven Koopman Analysis of Tropical Climate Space-Time Variability*. Nov. 2017. DOI: 10.48550/arXiv.1711.02526 (cited on page 26).

[165] Peter J. Schmid. 'Dynamic mode decomposition of numerical and experimental data'. en. In: *Journal of Fluid Mechanics* 656 (Aug. 2010), pp. 5–28. DOI: 10.1017/S0022112010001217 (cited on pages 27, 35, 36, 53, 104).

[166] Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. 'A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition'. In: *Journal of Nonlinear Science* 25.6 (Dec. 2015), pp. 1307–1346. DOI: 10.1007/s00332-015-9258-5 (cited on pages 27, 35, 37, 52, 53).

[167] Antonio Navarra et al. 'Variability of SST through Koopman Modes'. en. In: *Journal of Climate* 37.16 (July 2024), pp. 4095–4114. DOI: 10.1175/JCLI-D-23-0335.1 (cited on pages 27, 35, 37, 52).

[168] J. Thuburn. 'Climate sensitivities via a Fokker–Planck adjoint approach'. en. In: *Quarterly Journal of the Royal Meteorological Society* 131.605 (2005), pp. 73–92. DOI: 10.1256/qj.04.46 (cited on page 27).

[169] D. K. Henze, A. Hakami, and J. H. Seinfeld. 'Development of the adjoint of GEOS-Chem'. English. In: *Atmospheric Chemistry and Physics* 7.9 (May 2007), pp. 2413–2433. DOI: 10.5194/acp-7-2413-2007 (cited on page 27).

[170] Guokun Lyu et al. 'Adjoint-Based Climate Model Tuning: Application to the Planet Simulator'. en. In: *Journal of Advances in Modeling Earth Systems* 10.1 (2018), pp. 207–222. DOI: 10.1002/2017MS001194 (cited on page 27).

[171] Mackenzie L. Blanusa, Carla J. López-Zurita, and Stephan Rasp. 'Internal variability plays a dominant role in global climate projections of temperature and precipitation extremes'. en. In: *Climate Dynamics* 61.3 (Aug. 2023), pp. 1931–1945. DOI: 10.1007/s00382-023-06664-3 (cited on page 27).

[172] K Caldeira and N P Myhrvold. 'Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration'. en. In: *Environmental Research Letters* 8.3 (Sept. 2013), p. 034039. DOI: 10.1088/1748-9326/8/3/034039 (cited on page 27).

[173] F. Joos et al. 'Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: a multi-model analysis'. English. In: *Atmospheric Chemistry and Physics* 13.5 (Mar. 2013), pp. 2793–2825. DOI: 10.5194/acp-13-2793-2013 (cited on page 27).

[174] Valerio Lucarini et al. 'A general framework for linking free and forced fluctuations via Koopmanism'. In: *Chaos, Solitons & Fractals* 202 (Jan. 2026), p. 117540. DOI: 10.1016/j.chaos.2025.117540 (cited on page 28).

[175] Holger Metzler, Markus Müller, and Carlos A. Sierra. 'Transit-time and age distributions for nonlinear time-dependent compartmental systems'. In: *Proceedings of the National Academy of Sciences* 115.6 (Feb. 2018), pp. 1150–1155. DOI: 10.1073/pnas.1705296115 (cited on page 28).

[176] Ludovico T. Giorgini, Tobias Bischoff, and Andre N. Souza. *Statistical Parameter Calibration with the Generalized Fluctuation Dissipation Theorem and Generative Modeling*. Sept. 2025. DOI: 10.48550/arXiv.2509.19660 (cited on page 28).

[177] Ben Kravitz et al. 'Exploring precipitation pattern scaling methodologies and robustness among CMIP5 models'. English. In: *Geoscientific Model Development* 10.5 (May 2017), pp. 1889–1902. DOI: 10.5194/gmd-10-1889-2017 (cited on page 29).

[178] Andrew D. King et al. 'Transient and Quasi-Equilibrium Climate States at 1.5°C and 2°C Global Warming'. en. In: *Earth's Future* 9.11 (2021), e2021EF002274. DOI: 10.1029/2021EF002274 (cited on page 29).

[179] Bjorn Stevens et al. 'Prospects for narrowing bounds on Earth's equilibrium climate sensitivity'. en. In: *Earth's Future* 4.11 (2016), pp. 512–522. DOI: 10.1002/2016EF000376 (cited on page 30).

[180] Jared Farley et al. 'A Climate Intervention Dynamical Emulator (CIDER) for scenario space exploration'. English. In: *Geoscientific Model Development* 19.5 (Mar. 2026), pp. 1809–1831. DOI: 10.5194/gmd-19-1809-2026 (cited on page 31).

[181] Jonah Bloch-Johnson et al. 'The Green's Function Model Intercomparison Project (GFMIP) Protocol'. In: *Journal of Advances in Modeling Earth Systems* 16.2 (Feb. 2024), e2023MS003700. DOI: 10.1029/2023MS003700 (cited on pages 31, 55).

[182] Umberto Marini Bettolo Marconi et al. 'Fluctuation–dissipation: Response theory in statistical physics'. In: *Physics Reports* 461.4 (June 2008), pp. 111–195. DOI: 10.1016/j.physrep.2008.02.002 (cited on page 31).

[183] Christian L. E. Franzke, Federica Gugole, and Stephan Juricke. 'Systematic multi-scale decomposition of ocean variability using machine learning'. en. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.7 (July 2022). DOI: 10.1063/5.0090064 (cited on pages 35, 36).

[184] Peter J Schmid. 'Dynamic Mode Decomposition and Its Variants'. en. In: (2021) (cited on pages 35, 52).

[185] Masih Haseli and Jorge Cortés. 'Approximating the Koopman Operator using Noisy Data: Noise-Resilient Extended Dynamic Mode Decomposition'. In: *2019 American Control Conference (ACC)*. July 2019, pp. 5499–5504. DOI: 10.23919/ACC.2019.8814684 (cited on page 35).

[186] Marcos Netto et al. 'On analytical construction of observable functions in extended dynamic mode decomposition for nonlinear estimation and prediction'. In: *2021 American Control Conference (ACC)*. May 2021, pp. 4190–4195. DOI: 10.23919/ACC50511.2021.9482747 (cited on page 35).

[187] Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. 'Dynamic Mode Decomposition with Control'. In: *SIAM Journal on Applied Dynamical Systems* 15.1 (Jan. 2016), pp. 142–161. DOI: 10.1137/15M1013857 (cited on page 36).
[188] Steven L. Brunton et al. 'Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control'. en. In: *PLOS ONE* 11.2 (Feb. 2016), e0150171. DOI: 10.1371/journal.pone.0150171 (cited on page 36).
[189] Hiroaki Tatebe and Masahiro Watanabe. *MIROC MIROC6 model output prepared for CMIP6 CMIP piControl*. 2018. DOI: 10.22033/ESGF/CMIP6.5711 (cited on page 39).
[190] Martin Dix et al. *CSIRO-ARCCSS ACCESS-CM2 model output prepared for CMIP6 CMIP piControl*. 2019. DOI: 10.22033/ESGF/CMIP6.4311 (cited on page 39).
[191] Karl-Hermann Wieners et al. *MPI-M MPI-ESM1.2-LR model output prepared for CMIP6 CMIP piControl*. 2019. DOI: 10.22033/ESGF/CMIP6.6675 (cited on page 39).
[192] Edward N. Lorenz. 'Deterministic Nonperiodic Flow'. en. In: *Journal of the Atmospheric Sciences* 20.2 (Mar. 1963), pp. 130–141. DOI: doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2 (cited on page 39).
[193] Andre N. Souza and Charles R. Doering. 'Maximal transport in the Lorenz equations'. In: *Physics Letters A* 379.6 (Mar. 2015), pp. 518–523. DOI: 10.1016/j.physleta.2014.10.050 (cited on page 39).
[194] Jonathan H. Tu et al. 'On Dynamic Mode Decomposition: Theory and Applications'. In: *Journal of Computational Dynamics* 1.2 (2014), pp. 391–421. DOI: 10.3934/jcd.2014.1.391 (cited on page 52).
[195] Nathan P. Gillett et al. 'The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6'. English. In: *Geoscientific Model Development* 9.10 (Oct. 2016), pp. 3685–3697. DOI: 10.5194/gmd-9-3685-2016 (cited on pages 54, 58, 111).
[196] Laura J. Wilcox et al. 'The Regional Aerosol Model Intercomparison Project (RAMIP)'. English. In: *Geoscientific Model Development* 16.15 (Aug. 2023), pp. 4451–4479. DOI: 10.5194/gmd-16-4451-2023 (cited on page 54).
[197] Zongyi Li et al. *Fourier Neural Operator for Parametric Partial Differential Equations*. en. May 2021. DOI: 10.48550/arXiv.2010.08895 (cited on page 56).
[198] Lu Lu et al. 'Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators'. en. In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 218–229. DOI: 10.1038/s42256-021-00302-5 (cited on page 56).
[199] Kai R. Keller, Marta Alerany Solé, and Mario Acosta. 'Replicability in Earth System Models'. English. In: *Geoscientific Model Development* 18.24 (Dec. 2025), pp. 10221–10243. DOI: 10.5194/gmd-18-10221-2025 (cited on page 56).
[200] Valerii Fedorov. 'Optimal experimental design'. en. In: *WIREs Computational Statistics* 2.5 (2010), pp. 581–589. DOI: 10.1002/wics.100 (cited on pages 57, 105).
[201] Sergey Paltsev et al. *2025 Global Change Outlook*. Tech. rep. MIT Center for Sustainability Science and Strategy, Dec. 2025 (cited on pages 58, 112).
[202] Nathan P. Gillett et al. 'The Detection and Attribution Model Intercomparison Project (DAMIP v2.0) contribution to CMIP7'. English. In: *Geoscientific Model Development* 18.14 (July 2025), pp. 4399–4416. DOI: 10.5194/gmd-18-4399-2025 (cited on pages 58, 111).
[203] B. Kravitz et al. 'The Geoengineering Model Intercomparison Project Phase 6 (GeoMIP6): simulation design and preliminary results'. English. In: *Geoscientific Model Development* 8.10 (Oct. 2015), pp. 3379–3392. DOI: 10.5194/gmd-8-3379-2015 (cited on pages 59, 68, 111).
[204] Daniele Visioni et al. *The Geoengineering Model Intercomparison Project (GeoMIP) contribution to CMIP7 – description of new experimental protocols and preliminary results*. en. May 2026. DOI: 10.5194/egusphere-2026-2417 (cited on page 59).
[205] Zhiyuan Li and Sanjeev Arora. *An Exponential Learning Rate Schedule for Deep Learning*. Nov. 2019. DOI: 10.48550/arXiv.1910.07454 (cited on page 61).
[206] Himakar Ganti and Prashant Khare. 'Data-driven surrogate modeling of multiphase flows using machine learning techniques'. en. In: *Computers & Fluids* 211 (Oct. 2020), p. 104626. DOI: 10.1016/j.compfluid.2020.104626 (cited on page 65).
[207] Jincheng Zhang and Xiaowei Zhao. 'Machine-Learning-Based Surrogate Modeling of Aerodynamic Flow Around Distributed Structures'. en. In: *AIAA Journal* 59.3 (Mar. 2021), pp. 868–879. DOI: 10.2514/1.J059877 (cited on page 65).
[208] Tongzhou Wang et al. *Dataset Distillation*. Feb. 2020. DOI: 10.48550/arXiv.1811.10959 (cited on page 65).
[209] Timothy Nguyen et al. 'Dataset Distillation with Infinitely Wide Convolutional Networks'. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 5186–5198 (cited on page 65).
[210] George Cazenavette et al. *Dataset Distillation by Matching Training Trajectories*. Mar. 2022. DOI: 10.48550/arXiv.2203.11932 (cited on page 65).
[211] Ben Kravitz et al. 'Technical note: Simultaneous fully dynamic characterization of multiple input–output relationships in climate models'. English. In: *Atmospheric Chemistry and Physics* 17.4 (Feb. 2017), pp. 2525–2541. DOI: 10.5194/acp-17-2525-2017 (cited on page 67).
[212] William S Moses et al. *DJ4Earth: Differentiable, and Performance-portable Earth System Modeling via Program Transformations*. Nov. 2025. DOI: 10.22541/essoar.176314951.18114616/v1 (cited on page 67).
[213] Chris Smith et al. 'fair-calibrate v1.4.1: calibration, constraining, and validation of the FaIR simple climate model for reliable future climate projections'. English. In: *Geoscientific Model Development* 17.23 (Dec. 2024), pp. 8569–8592. DOI: 10.5194/gmd-17-8569-2024 (cited on pages 67, 108).
[214] Marc C. Kennedy and Anthony O'Hagan. 'Bayesian calibration of computer models'. en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464. DOI: 10.1111/1467-9868.00294 (cited on page 68).
[215] Tapio Schneider et al. 'Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations'. en. In: *Geophysical Research Letters* 44.24 (2017), pp. 12, 396–12, 417. DOI: 10.1002/2017GL076101 (cited on page 68).
[216] Tapio Schneider, L. Ruby Leung, and Robert C. J. Wills. *Opinion: Optimizing climate models with process-knowledge, resolution, and AI*. en. Jan. 2024. DOI: 10.5194/egusphere-2024-20 (cited on page 68).
[217] Patrick Heimbach, Chris Hill, and Ralf Giering. 'An efficient exact adjoint of the parallel MIT General Circulation Model, generated via automatic differentiation'. In: *Future Generation Computer Systems* 21.8 (Oct. 2005), pp. 1356–1371. DOI: 10.1016/j.future.2004.11.010 (cited on page 68).
[218] G. Forget et al. 'ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation'. English. In: *Geoscientific Model Development* 8.10 (Oct. 2015), pp. 3071–3104. DOI: 10.5194/gmd-8-3071-2015 (cited on page 68).
[219] Ellen H. Davenport et al. 'JCM v1.0: A Differentiable, Intermediate-Complexity Atmospheric Model'. English. In: *EGUsphere* (Jan. 2026), pp. 1–20. DOI: 10.5194/egusphere-2025-6266 (cited on page 68).

[220] Jerome H. Friedman. 'Stochastic gradient boosting'. In: *Computational Statistics & Data Analysis*. Nonlinear Methods and Data Mining 38.4 (Feb. 2002), pp. 367–378. DOI: 10.1016/S0167-9473(01)00065-2 (cited on page 68).
[221] Yanli Liu, Yuan Gao, and Wotao Yin. 'An Improved Analysis of Stochastic Gradient Descent with Momentum'. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18261–18271 (cited on pages 69, 107).
[222] Andrei Sokolov et al. 'Description and Evaluation of the MIT Earth System Model (MESM)'. en. In: *Journal of Advances in Modeling Earth Systems* 10.8 (2018), pp. 1759–1789. DOI: 10.1029/2018MS001277 (cited on pages 70, 71, 110).
[223] Jae Edmonds and JM Reiley. 'Global energy-assessing the future'. In: (1984) (cited on page 71).
[224] Detlef P. van Vuuren et al. 'How well do integrated assessment models simulate climate change?' en. In: *Climatic Change* 104.2 (Jan. 2011), pp. 255–285. DOI: 10.1007/s10584-009-9764-2 (cited on page 71).
[225] 'Assessing Transformation Pathways'. In: *Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the IPCC Fifth Assessment Report*. Ed. by Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, 2015, pp. 413–510. DOI: 10.1017/CBO9781107415416.012 (cited on page 71).
[226] John Weyant. 'Some Contributions of Integrated Assessment Models of Global Climate Change'. In: *Review of Environmental Economics and Policy* 11.1 (Jan. 2017), pp. 115–137. DOI: 10.1093/reep/rew018 (cited on page 71).
[227] Jennifer Morris et al. 'Quantifying both socioeconomic and climate uncertainty in coupled human–Earth systems analysis'. en. In: *Nature Communications* 16.1 (Mar. 2025), p. 2703. DOI: 10.1038/s41467-025-57897-1 (cited on pages 72–75).
[228] Kenneth Gillingham et al. 'Modeling Uncertainty in Integrated Assessment of Climate Change: A Multimodel Comparison'. In: *Journal of the Association of Environmental and Resource Economists* 5.4 (Oct. 2018), pp. 791–826. DOI: 10.1086/698910 (cited on page 72).
[229] Jennifer Morris et al. 'Representing Socio-Economic Uncertainty in Human System Models'. en. In: *Earth's Future* 10.4 (2022), e2021EF002239. DOI: 10.1029/2021EF002239 (cited on page 72).
[230] Kevin Rennert et al. 'Comprehensive evidence implies a higher social cost of CO2'. en. In: *Nature* 610.7933 (Oct. 2022), pp. 687–692. DOI: 10.1038/s41586-022-05224-9 (cited on page 72).
[231] Jeremy Fyke, Neil C. Swart, and David Huard. 'An Earth System Model Ensemble Forced With Probabilistic Emissions: Demonstration and Prospects for Climate Risk Assessment'. en. In: *Earth's Future* 14.2 (2026), e2025EF007289. DOI: 10.1029/2025EF007289 (cited on page 72).
[232] Sergey Paltsev et al. *The MIT emissions prediction and policy analysis (EPPA) model: version 4*. Tech. rep. MIT joint program on the science and policy of global change, 2005 (cited on pages 72, 73).
[233] Y.-H. Henry Chen et al. 'Long-term economic modeling for climate change assessment'. en. In: *Economic Modelling* 52 (Jan. 2016), pp. 867–883. DOI: 10.1016/j.econmod.2015.10.023 (cited on page 73).
[234] Jennifer Morris et al. 'Representing the costs of low-carbon power generation in multi-region multi-sector energy-economic models'. In: *International Journal of Greenhouse Gas Control* 87 (Aug. 2019), pp. 170–187. DOI: 10.1016/j.ijggc.2019.05.016 (cited on page 73).
[235] Martin Schupfner et al. *DKRZ MPI-ESM1.2-LR model output prepared for CMIP6 ScenarioMIP*. Jan. 2021 (cited on pages 74, 76, 87).
[236] D. S. Wilks. 'On "Field Significance" and the False Discovery Rate'. en. In: (Sept. 2006). DOI: 10.1175/JAM2404.1 (cited on pages 76, 119).
[237] D. S. Wilks. '"The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It'. EN. In: *Bulletin of the American Meteorological Society* 97.12 (Dec. 2016), pp. 2263–2273. DOI: 10.1175/BAMS-D-15-00267.1 (cited on pages 76, 119).
[238] Lydia Stefanova, Philip Sura, and Melissa Griffin. 'Quantifying the Non-Gaussianity of Wintertime Daily Maximum and Minimum Temperatures in the Southeast'. en. In: (Feb. 2013). DOI: 10.1175/JCLI-D-12-00161.1 (cited on page 76).
[239] Víctor Resco de Dios et al. 'A semi-mechanistic model for predicting the moisture content of fine litter'. In: *Agricultural and Forest Meteorology* 203 (Apr. 2015), pp. 64–73. DOI: 10.1016/j.agrformet.2015.01.002 (cited on page 77).
[240] Hamish Clarke et al. 'Forest fire threatens global carbon sinks and population centres under rising atmospheric water demand'. en. In: *Nature Communications* 13.1 (Nov. 2022), p. 7161. DOI: 10.1038/s41467-022-34966-3 (cited on pages 77, 85, 86).
[241] F. Sedano and J. T. Randerson. 'Multi-scale influence of vapor pressure deficit on fire ignition and spread in boreal forest ecosystems'. English. In: *Biogeosciences* 11.14 (July 2014), pp. 3739–3755. DOI: 10.5194/bg-11-3739-2014 (cited on page 77).
[242] A. Park Williams et al. 'Correlations between components of the water balance and burned area reveal new insights for predicting forest fire area in the southwest United States'. In: *International Journal of Wildland Fire* 24.1 (Nov. 2014), pp. 14–26. DOI: 10.1071/WF14023 (cited on page 77).
[243] John T. Abatzoglou and A. Park Williams. 'Impact of anthropogenic climate change on wildfire across western US forests'. In: *Proceedings of the National Academy of Sciences* 113.42 (Oct. 2016), pp. 11770–11775. DOI: 10.1073/pnas.1607171113 (cited on page 77).
[244] Philip E. Higuera and John T. Abatzoglou. 'Record-setting climate enabled the extraordinary 2020 fire season in the western United States'. en. In: *Global Change Biology* 27.1 (2021), pp. 1–2. DOI: 10.1111/gcb.15388 (cited on page 77).
[245] Shu Li and Tirtha Banerjee. 'Spatial and temporal pattern of wildfires in California from 2000 to 2019'. en. In: *Scientific Reports* 11.1 (Apr. 2021), p. 8779. DOI: 10.1038/s41598-021-88131-9 (cited on page 77).
[246] Víctor Resco de Dios et al. 'Climate change induced declines in fuel moisture may turn currently fire-free Pyrenean mountain forests into fire-prone ecosystems'. In: *Science of The Total Environment* 797 (Nov. 2021), p. 149104. DOI: 10.1016/j.scitotenv.2021.149104 (cited on page 77).
[247] Jennifer K. Balch et al. 'Human-started wildfires expand the fire niche across the United States'. en. In: *Proceedings of the National Academy of Sciences* 114.11 (Mar. 2017), pp. 2946–2951. DOI: 10.1073/pnas.1617394114 (cited on page 77).
[248] Alisa Keyser and Anthony LeRoy Westerling. 'Climate drives inter-annual variability in probability of high severity fire occurrence in the western United States'. en. In: *Environmental Research Letters* 12.6 (May 2017), p. 065003. DOI: 10.1088/1748-9326/aa6b10 (cited on page 77).
[249] P. N. Racherla, D. T. Shindell, and G. S. Faluvegi. 'The added value to global model projections of climate change by dynamical downscaling: A case study over the continental U.S. using the GISS-ModelE2 and WRF models'. en. In: *Journal of Geophysical Research: Atmospheres* 117.D20 (2012). DOI: 10.1029/2012JD018091 (cited on page 78).
[250] Elisabeth A. Lloyd, Melissa Bukovsky, and Linda O. Mearns. 'An analysis of the disagreement about added value by regional climate models'. en. In: *Synthese* 198.12 (Dec. 2021), pp. 11645–11672. DOI: 10.1007/s11229-020-02821-x (cited on page 78).
[251] Geert Lenderink et al. 'A perfect model study on the reliability of the added small-scale information in regional climate change projections'. en. In: *Climate Dynamics* 60.9 (May 2023), pp. 2563–2579. DOI: 10.1007/s00382-022-06451-6 (cited on page 78).

[252] Cécile Davrinche et al. 'Future changes in Antarctic near-surface winds: regional variability and key drivers under a high-emission scenario'. English. In: *The Cryosphere* 19.11 (Nov. 2025), pp. 6023–6042. DOI: 10.5194/tc-19-6023-2025 (cited on page 78).
[253] I Mahlstein et al. 'Early onset of significant local warming in low latitude countries'. en. In: *Environmental Research Letters* 6.3 (July 2011), p. 034009. DOI: 10.1088/1748-9326/6/3/034009 (cited on pages 79, 87).
[254] E. Hawkins and R. Sutton. 'Time of emergence of climate signals'. en. In: *Geophysical Research Letters* 39.1 (2012). DOI: 10.1029/2011GL050087 (cited on pages 79, 82).
[255] Claudia Tebaldi and Pierre Friedlingstein. 'Delayed detection of climate mitigation benefits due to climate inertia and variability'. In: *Proceedings of the National Academy of Sciences* 110.43 (Oct. 2013), pp. 17229–17234. DOI: 10.1073/pnas.1300005110 (cited on pages 79, 87).
[256] Patrick C. Taylor et al. 'Process Drivers, Inter-Model Spread, and the Path Forward: A Review of Amplified Arctic Warming'. English. In: *Frontiers in Earth Science* 9 (Feb. 2022). DOI: 10.3389/feart.2021.758361 (cited on pages 79, 82).
[257] Shulei Zhang, Xiaodong Liu, and Buwen Dong. 'Spatiotemporal characteristics of the time of emergence for anthropogenic tropospheric temperature changes based on the CMIP6 multi-model results'. en. In: *Environmental Research Letters* 19.4 (Apr. 2024), p. 044052. DOI: 10.1088/1748-9326/ad34e6 (cited on page 79).
[258] IPCC. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. en. 1st ed. Cambridge University Press, July 2023 (cited on page 81).
[259] Claudia Tebaldi, Brian O'Neill, and Jean-François Lamarque. 'Sensitivity of regional climate to global temperature and forcing'. en. In: *Environmental Research Letters* 10.7 (July 2015), p. 074001. DOI: 10.1088/1748-9326/10/7/074001 (cited on pages 81, 88).
[260] Panxi Dai et al. 'Constraints on regional projections of mean and extreme precipitation under warming'. In: *Proceedings of the National Academy of Sciences* 121.11 (Mar. 2024), e2312400121. DOI: 10.1073/pnas.2312400121 (cited on page 81).
[261] Nina Schuhen et al. 'Emergence of climate change signal in CMIP6 extreme indices'. English. In: *Natural Hazards and Earth System Sciences* 26.2 (Feb. 2026), pp. 753–773. DOI: 10.5194/nhess-26-753-2026 (cited on page 82).
[262] Daniel J. Leathers, Brent Yarnal, and Michael A. Palecki. 'The Pacific/North American Teleconnection Pattern and United States Climate. Part I: Regional Temperature and Precipitation Associations'. en. In: (May 1991) (cited on page 83).
[263] Ethan D. Coffel, Radley M. Horton, and Alex de Sherbinin. 'Temperature and humidity based projections of a rapid rise in global heat stress exposure during the 21st century'. In: *Environmental research letters : ERL [Web site]* 13.1 (Jan. 2018), p. 014001. DOI: 10.1088/1748-9326/aaa00e (cited on page 84).
[264] Daniel J. Vecellio et al. 'Greatly enhanced risk to humans as a consequence of empirically determined lower moist heat stress tolerance'. In: *Proceedings of the National Academy of Sciences of the United States of America* 120.42 (2023), e2305427120. DOI: 10.1073/pnas.2305427120 (cited on page 84).
[265] Fahad Saeed, Carl-Friedrich Schleussner, and Moetasim Ashfaq. 'Deadly Heat Stress to Become Commonplace Across South Asia Already at 1.5°C of Global Warming'. en. In: *Geophysical Research Letters* 48.7 (Apr. 2021), e2020GL091191. DOI: 10.1029/2020GL091191 (cited on page 84).
[266] Katharine M. Willett et al. 'Attribution of observed surface humidity changes to human influence'. en. In: *Nature* 449.7163 (Oct. 2007), pp. 710–712. DOI: 10.1038/nature06207 (cited on page 84).
[267] Charlotte Grossiord et al. 'Plant responses to rising vapor pressure deficit'. en. In: *New Phytologist* 226.6 (2020), pp. 1550–1566. DOI: 10.1111/nph.16485 (cited on page 85).
[268] Jeremy S. Littell et al. 'Climate and wildfire area burned in western U.S. ecoprovinces, 1916–2003'. en. In: *Ecological Applications* 19.4 (2009), pp. 1003–1021. DOI: 10.1890/07-1183.1 (cited on pages 85, 86).
[269] Yizhou Zhuang et al. 'Quantifying contributions of natural variability and anthropogenic forcings on increased fire weather risk over the western United States'. In: *Proceedings of the National Academy of Sciences* 118.45 (Nov. 2021), e2111875118. DOI: 10.1073/pnas.2111875118 (cited on page 85).
[270] Stephanie K. Kampf et al. 'Fire, Fuel, and Climate Interactions in Temperate Climates'. en. In: *AGU Advances* 6.2 (2025), e2024AV001628. DOI: 10.1029/2024AV001628 (cited on page 86).
[271] Wanyun Shao et al. 'Science, Scientists, and Local Weather: Understanding Mass Perceptions of Global Warming*'. en. In: *Social Science Quarterly* 97.5 (2016), pp. 1023–1057. DOI: 10.1111/ssqu.12317 (cited on page 88).
[272] Patrick W. Keys et al. 'Potential for perceived failure of stratospheric aerosol injection deployment'. In: *Proceedings of the National Academy of Sciences* 119.40 (Oct. 2022), e2210036119. DOI: 10.1073/pnas.2210036119 (cited on page 88).
[273] John E Deeming, Robert E Burgan, and Jack D Cohen. *The national fire-danger rating system, 1978*. Vol. 39. Department of Agriculture, Forest Service, Intermountain Forest and Range . . ., 1977 (cited on page 88).
[274] Delavane Diaz and Frances Moore. 'Quantifying the economic risks of climate change'. en. In: *Nature Climate Change* 7.11 (Nov. 2017), pp. 774–782. DOI: 10.1038/nclimate3411 (cited on page 89).
[275] James E Neumann et al. 'Climate Damage Functions for Estimating the Economic Impacts of Climate Change in the United States'. In: *Review of Environmental Economics and Policy* 14.1 (Jan. 2020), pp. 25–43. DOI: 10.1093/reep/rez021 (cited on page 89).
[276] Paul Waidelich et al. 'Climate damage projections beyond annual temperature'. en. In: *Nature Climate Change* 14.6 (June 2024), pp. 592–599. DOI: 10.1038/s41558-024-01990-8 (cited on page 89).
[277] Wai-Kwong Yeung and Fan-Nian Kong. 'Time domain deconvolution when the kernal has no spectral inverse'. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.4 (Aug. 1986), pp. 912–918. DOI: 10.1109/TASSP.1986.1164882 (cited on page 98).
[278] D. Zazula and L. Gyergyek. 'Direct frequency-domain deconvolution when the signals have no spectral inverse'. In: *IEEE Transactions on Signal Processing* 41.2 (Feb. 1993), pp. 977–981. DOI: 10.1109/78.193238 (cited on page 98).
[279] Paolo Giani et al. *Origin and Limits of Invariant Warming Patterns in Climate Models*. Nov. 2024. DOI: 10.48550/arXiv.2411.14183 (cited on page 105).
[280] John P. Dunne et al. 'An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment'. English. In: *Geoscientific Model Development* 18.19 (Oct. 2025), pp. 6671–6700. DOI: 10.5194/gmd-18-6671-2025 (cited on page 108).
[281] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. en. Jan. 2017. DOI: 10.48550/arXiv.1412.6980 (cited on page 108).
[282] Detlef P. van Vuuren et al. 'The representative concentration pathways: an overview'. en. In: *Climatic Change* 109.1 (Aug. 2011), p. 5. DOI: 10.1007/s10584-011-0148-z (cited on page 112).
[283] Boris Hanin. 'Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?' In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018 (cited on page 114).

[284] George Philipp, Dawn Song, and Jaime G. Carbonell. *The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions*. Apr. 2018. DOI: 10.48550/arXiv.1712.05577 (cited on page 114).