Towards Hybrid Physics-Machine Learning Parameterizations: Employing Data Assimilation for Online Learning of Turbulence and Convection Closures in a Unified Scheme

Thesis by
Costa Christopoulos

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended November 8, 2024

Costa Christopoulos
ORCID: 0000-0002-8552-465X

ACKNOWLEDGEMENTS

Endless pages would not suffice to thank everyone who made my PhD years the happiest and most productive of my life. This experience surpassed even the highest expectations I had when I started applying to PhD programs in 2018, largely because of the people who surrounded me and the work I was doing.

First, to my advisor Tapio Schneider. Above all else, Tapio was extremely patient and trusting. He had a lot of confidence in me—sometimes more than I had in myself—which boosted my self-assurance at the times I needed it most. The same goes for the rest of his students. When problems arose, he wanted to face them head-on, and not kick the can down the road. Having spent a career in climate modeling, he knows that unresolved issues reemerge and plague climate models later in complicated ways. In a world where it is tempting to quickly hack together code and stitch together limited solutions for flashy results, Tapio taught me the importance of digging into problems and addressing them one by one. For tackling any large-scale technological effort, like climate modeling, this is essential. This approach is, and will continue to be, the key to CliMA's success. I also want to thank him for being patient with me throughout my journey, especially when code was not working or when I asked rudimentary questions. Tapio always tried to put the happiness and well-being of his students above all else.

I'm also grateful to Tom Beucler, who, despite living in Switzerland, devoted his personal evening hours to meet with me on a weekly basis, for years. His patience and dedication to students are truly exceptional. To Christian Frankenberg, who served as an advisor for my second quals project and an unofficial mentor during my first two years. He guided me through the complexities of satellite inversions and radiation codes, and our shared sense of humor made the journey all the more enjoyable. From Tapio, Christian, and Tom, who served as advisors (but not all on paper) during my time here, I learned distinct and invaluable lessons and approaches.

I want to acknowledge Ignacio Lopez-Gomez, who initially took on the challenge of calibrating single-column models and taught me much of what I know today about working with the EDMF. To the entire early EDMF team—Yair Cohen, Charlie Kawczynski, Anna Jaruga, and Haaakon Ervik—working with you was a foundational experience, and this work would not be possible without each and every one of you. To the rest of the CliMA team, your advice and good company helped me

tackle many challenges, whether by finding solutions through our conversations or simply by lifting my spirits. Even during tough times and tight deadlines at CliMA, laughter often echoed from the conference room, in the kitchen, and around the patio lunch tables. A special mention goes to Zhaoyi Shen for her endless support and deep, broad, and practical knowledge of climate modeling, fluid dynamics, numerics, and everything in between. There is perhaps no person who received more questions from me, and she (almost) always had the answer. And Ollie Dunbar, who patiently taught me and other students how to apply inverse methods to practical problems, always ready to help with issues and answer questions, big and small.

To Daniel Rothenberg, my former manager and professional mentor. I looked up to him immensely and continue to. I joined an early-stage startup because of him, and stayed for the same reason. Together we faced seemingly innumerable challenges—from tight deadlines to impossibly complex weather forecasts for clients. Recognizing my curiosity, he encouraged me to pursue a PhD despite my uncertainty, and was a significant factor in my decision to apply for one, which turned out to be one of the best decisions I have made in life. Daniel has been rooting for me since we met on a forecasting team during my undergrad years at MIT, and he helped me get here. He's an incredible mentor and friend and was always the type of person I aspire to be—a true role model.

On a personal note, I extend heartfelt thanks to my roommates during and just after the COVID period. Sharing some of the best moments of grad school—and my life—with them at the "treatment sanctuary" was unforgettable. And to my partner Alexander Hu, who thoroughly supported me through the most difficult times in graduate school and traveled with me to some incredible places. My friends Leo Calderas and Lisandro Jimenez Leon have kept in touch through the years since MIT, supporting me through highs and lows, virtually and in person.

Over all these years, my family has always been my backbone. My mom nurtured my early interest in weather, taking me to National Weather Service open days and gifting me the best Christmas present ever— the fanciest weather station at the time. My grandmother taught me the value of grit and being steadfast in the face of the challenges of life.

Looking back, it's astonishing to think that the ten-year-old who eagerly clicked through weather maps on the NOAA website would end up at Caltech, working with leading experts like Tapio, on the very types of numerical models I looked at (but did not understand) as a kid. I am profoundly grateful for this incredible journey. Lastly,

to my childhood heroes: Carl Sagan, whose writings convinced me that someone as ordinary as me could pursue physics, and instilling in me an emotional love for science that continues to this day. And James Spann, the local meteorologist whose reporting during severe weather events instilled in me a fascination of the extremes of our atmosphere. His communication of complicated and uncertain forecasts has saved countless lives. His work reminds me of the significant real-world impact atmospheric modeling can have when paired with the right people to communicate their predictions.

ABSTRACT

Despite advances in climate modeling, the spread in equilibrium climate sensitivity estimates has remained largely unchanged over generations of modeling, mainly due to uncertainties in cloud feedback mechanisms arising from subgrid-scale turbulence, convection, clouds, and the resulting cloud-radiation interactions. Misrepresentations of these processes affect both long-term climate projections and the simulation of short-term atmospheric phenomena, such as the diurnal cycle of precipitation. These limitations are most pronounced in regimes like stratocumulus clouds and their transition to cumulus over ocean basins—areas where climate models have the largest cloud biases in the historical record. This thesis aims to constrain the critical subgrid-scale physics of turbulence and convection by developing and calibrating a hybrid physics–machine learning parameterization using the Eddy-diffusivity Mass-flux (EDMF) framework. By integrating machine learning components into the EDMF and employing data assimilation techniques for online learning, we attempt to directly target some of the processes responsible for uncertainties in cloud feedbacks.

In this thesis, we employ ensemble Kalman inversion within a single-column setup to simultaneously perform online calibration of parameters in empirical closures and embedded neural networks, targeting large-eddy simulations as ground truth. The online learning framework ensures stability and physical consistency, as machine learning components are trained within the context of the full model dynamics. By directly targeting poorly constrained processes like lateral entrainment/detrainment and turbulent mixing lengths, we improve the representation of subgrid-scale fluxes and resulting cloud properties across various atmospheric regimes. We uncover limitations of traditional semi-empirical closures, providing insights for future model development. The calibrated hybrid parameterization outperforms existing schemes, particularly in regions where climate models have historically underperformed, and maintains accuracy in out-of-sample forcings from a warmer climate. This work demonstrates that integrating machine learning with physics-based parameterizations through data assimilation offers a systematic and robust approach for reducing biases in climate models and understanding the physics of elusive subgrid-scale closures.

PUBLISHED CONTENT AND CONTRIBUTIONS

Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O. R. A., & Schneider, T. (2024). Online Learning of Entrainment Closures in a Hybrid Machine Learning Parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(11), e2024MS004485. https://doi.org/10.1029/2024MS004485

C.C. led the project, carried out the code development and data analysis, and wrote the manuscript.

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). https://doi.org/10.1029/2022MS003105

C.C. collaborated on code development and reviewed/edited the manuscript.

Christopoulos, C., & Schneider, T. (2021). Assessing Biases and Climate Implications of the Diurnal Precipitation Cycle in Climate Models. *Geophysical Research Letters*, *48*(13), e2021GL093017. https://doi.org/10.1029/2021GL093017

C.C. led the project, carried out the code development and data analysis, and wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Forward

Since Earth's coalescence over 4.5 billion years ago, our planet has experienced a myriad of climates—from global glaciations known as "Snowball Earth," where ice sheets reached the equator, to hothouse periods devoid of polar ice caps. These dramatic variations eventually led to a climate hospitable to human life. Soon after, the most curious among us began meticulously observing the sky. Around 2,400 years ago, figures like Aristotle, one of the earliest meteorologists, began systematically attempting to explain the atmospheric patterns he observed. The term "climate" itself is derived from the Greek word $\kappa\lambda\iota\mu\alpha$ (klima), meaning "inclination" or "slope," reflecting the ancient understanding of how the sun's angle influenced regional weather patterns and created distinct climatic zones.

The same species that once gave names to the systematically varying skies and began to describe them with rudimentary rules have, through rapid technological development, become powerful enough to unnaturally influence the climate at an unprecedented rate. Today, the word "climate" has taken on a completely new meaning. Through technological innovations in computing power and better process understanding, we now have the power to predict aspects of future weather and climate at a global scale, allowing us to take actions to prepare for or avoid potential realities—a scenario unimaginable to the Oracles of Delphi. Today, we live in a world where approximations of entire simulated Earth systems can be generated on computers, playing out atmospheric realities that may or may not happen. The usefulness of these models, however, hinges on their fidelity and ability to generalize. As the complexity of processes in these models surpasses our ability to precisely measure and constrain them with traditional techniques and measurements, we face new challenges. Paradigm shifts are occurring that favor using AI techniques largely as black boxes, but these tools and their evaluation must be approached systematically and carefully, given the complexity of the Earth system and the large-scale decision making that results from the predictions of climate models. While Aristotle's theories of the atmosphere loosely fit what he could observe in the sky above him, many aspects of his broader reasoning of weather and climate turned out

to be incorrect—and failed to predict characteristics of global weather he did not directly observe.

1.2 Overview

A variety of aggregate metrics may be used to evaluate the physical realism of atmospheric simulations and track the accuracy of their predictions. There is no universally agreed upon metric for identifying the quality of a weather or climate model, as stakeholders utilizing their predictions and researchers analyzing their output are interested in accuracy of different aspects. For climate, perhaps among the most consequential and generally informative metric is equilibrium climate sensitivity (ECS), which measures the average, equilibrium increase in global-mean surface temperature in response to a doubling of CO2 (Meehl et al., 2020). The spread in ECS among climate models has remained largely unchanged since initial estimates with simpler models in 1979, despite rapid growth in the complexity of climate models (Charney, J. G. et al., 1979). The uncertainty of climate models in their prediction of ECS, as indicated by the large spread in climate model ensembles, has been largely traced to limitations in how cloud properties respond to their environment, and the changes in these interactions brought about by a changing climate. The term "cloud feedbacks" is used to describe the response of cloud radiative properties to climate change (Stephens, 2005). It is well-established that tropical low clouds dominate uncertainties in Earth's global cloud feedback through shortwave interactions (Bony & Dufresne, Jean-Louis, 2005; Ceppi et al., 2017; Klein et al., 2017). The misrepresentation of cloud feedbacks stems from inaccuracies of subgrid-scale physics, namely turbulence, convection, clouds, radiation, and the interaction between them (Bretherton et al., 2013; Gettelman & Sherwood, 2016).

The objective of this thesis is to constrain the subgrid-scale physics that is known to link short-term atmospheric variability to long-term climate prediction uncertainty, namely the processes responsible for dictating cloud feedbacks. Building upon the initial work of Lopez-Gomez et al. (2022) and utilizing the calibration and uncertainty quantification framework of CliMA (Schneider et al., 2017), in conjunction with model improvements, this thesis aims to inform representations of the subgrid phenomena, namely those dictating turbulence and convection. Perhaps the largest contribution of this thesis is pushing data assimilation frameworks, frequently published with simplified or idealized problems, towards use in more realistic climate modeling settings, where numerous challenges arise. If the promise of these methods is to make a difference in climate models, these challenges must

be addressed systematically and one-by-one. We specifically target regimes where state-of-the-art climate models broadly struggle to capture cloud properties, even in the historical record—stratocumulus clouds and their transition to cumulus regimes in ocean basins (Vignesh et al., 2020).

Numerous emergent constraints have been identified, establishing potential links between observable geophysical variables in the present climate to future climate outcomes (Klein & Hall, 2015). Proposed emergent constraints span various aspects of the present-day climate, including the depth of tropical low clouds (Brient & Schneider, 2016), the strength of the double-ITCZ bias (Tian & Dong, 2020), and the degree of vertical convective mixing (Sherwood et al., 2014), among others. The Coupled Model Intercomparison Project (CMIP) consists of a large ensemble of climate models run by institutions and modeling agencies all around the world. It is noted that many proposed emergent constraints derived for CMIP5 show considerable decrease in their skill when applied to CMIP6 (Schlund et al., 2020). Chapter 2 contains published work outlining an observed correlation between oceanic diurnal precipitation characteristics and ECS. However, the physical mechanism responsible for the correlation remains elusive. This chapter also highlights the shortcomings of the latest generations of climate models, CMIP5 and CMIP6, in simulating the diurnal cycle of precipitation (Christopoulos & Schneider, 2021). It is hypothesized that subgrid-scale physics surrounding cloud microphysics and entrainment that manifests locally and on daily timescales, through the diurnal precipitation cycle, may also impact the mean climate. Stated differently, diurnal cycle biases and ECS variations in CMIP models may have common causes, without the diurnal cycle biases directly causing ECS variations. While several hypotheses are laid out for the relationship surrounding cloud depth and radiative effects, the claims are ultimately unfalsifiable without a controlled modeling study utilizing a subgrid-scale parameterization. Although numerous biases in climate models have been documented and categorized across generations of CMIP models (Wang et al., 2014), the remainder of the thesis aims to tackle these problems at their source.

To that end, Chapters 3 and 4 focus on improving and calibrating a unified subgrid-scale parameterization of turbulence, convection, and clouds, known as the Eddy-diffusivity Mass-flux (EDMF) parameterization. The EDMF represents transport from coherent updrafts and environmental turbulence in a unified manner across a variety of convective and turbulent regimes (Tan et al., 2018; Thuburn et al., 2022). When coupled to a global climate model (GCM), the scheme accounts for

the effect of subgrid-scale processes and predicts vertical subgrid-scale fluxes and cloud properties as a function of large-scale forcings. The primary contribution of these chapters is to utilize data-driven models and data assimilation techniques to target uncertain processes within the EDMF, using data from large-eddy simulations for training. The concept of a "hybrid" EDMF is introduced, where semi-empirical closures are replaced with expressive and data-driven models. In contrast to emerging approaches that replace subgrid-scale parameterizations entirely with machine learning (ML) methods (Rasp et al., 2018; Yuval & O'Gorman, 2020), this research adopts a physics-first approach that incorporates ML methods at a low level while retaining well-established physical equations of motion (Schneider et al., 2024). In this way, learned relationships can be more directly analyzed, distilled, and reasoned about. The approach is especially pertinent for cloud mixing, where closures are numerous and diverse in climate models (de Rooy et al., 2013; Gregory, 2001). Chapter 4 pushes the setup to even further realism, accounting for interactions with radiation and surface fluxes and using coarse vertical resolution typical of climate models. Whereas Chapter 3 focuses on the dynamics of convection, Chapter 4 turns to the representation of turbulence in the EDMF. We find that data-driven mappings for turbulent mixing length can be distilled into relatively simple, interpretable, and predictive expressions.

Parameterizations in Earth System Models have always been fundamental in distilling our physical understanding of subgrid-scale processes. Thus, improving parameterizations refines our physical understanding and sheds light on the limits of parameterizations to represent even presently observable metrics. Where and how to replace physical closures in hybrid subgrid-scale parameterization remains an open and challenging question, with guidance provided by the experiments detailed in Chapters 3 and 4. The result of this thesis is ultimately a framework that can address biases, like those laid out in Chapter 2, at the closure level in a systematic and universal way. While not addressed in this thesis, the framework opens a variety of avenues for addressing limitations of additional critical closures (such as microphysics) and allows for the use of global, space-based, or sparse surface observations to fine-tune models to match measurements. In this way, we may constrain subgrid-scale closures of varying complexities indirectly. As the climate modeling enterprise continues to rapidly adopt ML methods to emulate processes in the Earth System Models, it is important to offer physically-motivated and scientifically-sound paradigms which avoid black boxes. Balancing accuracy with interpretability and generalizability are the core tenets of this work.

References

Bony, S., & Dufresne, Jean-Louis. (2005). Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophysical Research Letters*, *32*. https://doi.org/10.1029/2005GL023851

Bretherton, C. S., Blossey, P. N., & Jones, C. R. (2013). Mechanisms of marine low cloud sensitivity to idealized climate perturbations: A single-LES exploration extending the CGILS cases. *Journal of Advances in Modeling Earth Systems*, *5*(2), 316–337. https://doi.org/10.1002/jame.20019

Brient, F., & Schneider, T. (2016). Constraints on climate sensitivity from space-based measurements of low-cloud reflection. *Journal of Climate*, *29*(16), 5821–5835. https://doi.org/10.1175/JCLI-D-15-0897.1

Ceppi, P., Brient, F., Zelinka, M. D., & Hartmann, D. L. (2017). Cloud feedback mechanisms and their representation in global climate models. *WIREs Climate Change*, *8*(4), e465. https://doi.org/10.1002/wcc.465

Charney, J. G., Arakawa, A., Baker, D. J., Bolin, B., Dickinson, R. E., Goody, R. M., Leith, C. E., Stommel, H. M., & Wunsch, C. I. (1979). *Carbon Dioxide and Climate: A Scientific Assessment* (tech. rep.). U.S. National Academy of Sciences, Washington, DC. https://doi.org/10.17226/12181

Christopoulos, C., & Schneider, T. (2021). Assessing Biases and Climate Implications of the Diurnal Precipitation Cycle in Climate Models. *Geophysical Research Letters*, *48*(13), e2021GL093017. https://doi.org/10.1029/2021GL093017

de Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, D., Pier Siebesma, A., Teixeira, J., & Yano, J.-I. (2013). Entrainment and detrainment in cumulus convection: An overview. *Quarterly Journal of the Royal Meteorological Society*, *139*(670), 1–19. https://doi.org/10.1002/qj.1959

Gettelman, A., & Sherwood, S. C. (2016). Processes Responsible for Cloud Feedback. *Current Climate Change Reports*, *2*(4), 179–189. https://doi.org/10.1007/s40641-016-0052-8

Gregory, D. (2001). Estimation of entrainment rate in simple models of convective clouds. *Quarterly Journal of the Royal Meteorological Society*, *127*(571), 53–72. https://doi.org/10.1002/qj.49712757104

Klein, S. A., & Hall, A. (2015). Emergent Constraints for Cloud Feedbacks. *Current Climate Change Reports*, *1*(4), 276–287. https://doi.org/10.1007/s40641-015-0027-1

Klein, S. A., Hall, A., Norris, J. R., & Pincus, R. (2017). Low-Cloud Feedbacks from Cloud-Controlling Factors: A Review. *Surveys in Geophysics*, *38*(6), 1307–1329. https://doi.org/10.1007/s10712-017-9433-3

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). https://doi.org/10.1029/2022MS003105

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., & Schlund, M. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020). Emergent constraints on equilibrium climate sensitivity in CMIP5: Do they hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. https://doi.org/10.5194/esd-11-1233-2020

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations. *Geophysical Research Letters*, *44*(24). https://doi.org/10.1002/2017GL076101

Schneider, T., Leung, L. R., & Wills, R. C. J. (2024). Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmospheric Chemistry and Physics*, *24*(12), 7041–7062. https://doi.org/10.5194/acp-24-7041-2024

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1038/nature12829

Stephens, G. L. (2005). Cloud Feedbacks in the Climate System: A Critical Review. *Journal of Climate*, *18*(2), 237–273. https://doi.org/10.1175/JCLI-3243.1

Tan, Z., Kaul, C. M., Pressel, K. G., Cohen, Y., Schneider, T., & Teixeira, J. (2018). An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid-scale turbulence and convection. *Journal of Advances in Modeling Earth Systems*, *10*(3), 770–800. https://doi.org/10.1002/2017MS001162

Thuburn, J., Efstathiou, G. A., & McIntyre, W. A. (2022). A two-fluid single-column model of turbulent shallow convection. Part 1: Turbulence equations in the multifluid framework. *Quarterly Journal of the Royal Meteorological Society*, *148*(748), 3366–3387. https://doi.org/10.1002/qj.4366

Tian, B., & Dong, X. (2020). The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation. *Geophysical Research Letters*, *47*(8), 0–3. https://doi.org/10.1029/2020GL087232

Vignesh, P. P., Jiang, J. H., Kishore, P., Su, H., Smay, T., Brighton, N., & Velicogna, I. (2020). Assessment of CMIP6 cloud fraction and comparison with satellite observations. *Earth and Space Science*, *7*(2), e2019EA000975. https://doi.org/10.1029/2019EA000975

Wang, C., Zhang, L., Lee, S.-K., Wu, L., & Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, *4*(3), 201–205. https://doi.org/10.1038/nclimate2118

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

Chapter 2

ASSESSING BIASES AND CLIMATE IMPLICATIONS OF THE DIURNAL PRECIPITATION CYCLE IN CLIMATE MODELS

2.1 Abstract

The diurnal cycle is a common benchmark for evaluating the performance of weather and climate models on short timescales. For decades, capturing the timing of peak precipitation during the day has remained a challenge for climate models. In this study, the phase and amplitude of the diurnal precipitation cycle in Coupled Model Intercomparison Project (CMIP) models are compared to satellite data. While some improvements align CMIP6 models closer to satellite observations, significant biases in the timing of peak precipitation remain, especially over land. Notably, precipitation over land in CMIP6 models still occurs ~5.4 hours too early; the diurnal cycle amplitude is ~0.81 mm day^{-1} too small over the oceans. Further, the diurnal phase of oceanic precipitation correlates weakly with the equilibrium climate sensitivity in CMIP6 models: models with a later precipitation peak over oceans tend to exhibit a higher climate sensitivity. However, it is unclear whether this relationship is robust.

2.2 Introduction

The diurnal cycle of precipitation is among the fastest modes of variability in the climate system. Simulating diurnal variations of fundamental variables such as cloud cover and precipitation has been a long-standing issue for weather and climate models (Dai & Trenberth, 2004). Disagreements at short timescales indicate fundamental processes are misrepresented, even when seasonal and longer model averages agree with observations. Simulating the diurnal cycle of precipitation with fidelity requires correct accounting of surface-atmosphere interactions, cloud-radiative feedbacks, boundary layer dynamics, and cloud microphysics (Bechtold et al., 2004). Diurnal precipitation variability is multifaceted, and many mechanisms operate across a range of scales to control its behavior, making it an important benchmark for atmospheric models.

The diurnal cycle of precipitation has been characterized extensively in surface observations (Dai et al., 1999), satellite observations (Bowman et al., 2005; Dai et al., 2007; Kikuchi & Wang, 2008; Tan et al., 2019; Yang & Slingo, 2001) and in weather and climate models (Bechtold et al., 2004; Covey et al., 2016; Dai, 2006; DeMott et al., 2007; Lee et al., 2008; Pritchard & Somerville, 2009). The physical mechanisms governing the diurnal cycle over land and ocean are distinct (Dai, 2001; Ruppert & Hohenegger, 2018), leading to fundamentally different characteristics between these regions. Oceanic precipitation tends to peak in the early morning hours (Bowman et al., 2005; Sorooshian et al., 2002). Warm-season precipitation often peaks in the late afternoon to early evening over land areas, with the central U.S. and a few other regions peaking around midnight to early morning (Dai, 2001; Dai et al., 1999, 2007; Lin et al., 2000). The diurnal characteristics of precipitation in climate models differ from observations, most notably in terms of diurnal timing. Simulated precipitation tends to peak too early over land (Collier & Bowman, 2004; Covey et al., 2016; Dai, 2006). Over oceans, the diurnal precipitation amplitude has been noted to be weak in some climate models, possibly as a result of weak temperature variations in the ocean boundary layer and low atmosphere-ocean coupling frequency (Dai & Trenberth, 2004; Randall et al., 1991).

Overall, the diurnal phase and amplitude of convection remain a challenge to properly simulate in CMIP6 climate models. We find that only marginal improvements in diurnal precipitation phase and amplitude have been made since CMIP5, and, by some metrics, CMIP6 models on the whole perform worse. We further uncover a tenuous relationship between diurnal cycle metrics and the equilibrium climate

sensitivity (ECS), reinforcing the need to correctly capture the diurnal cycle in global climate models. The relationship is found to be strongest over oceans in CMIP6 models, with a negligible relationship over land. Physical hypotheses are laid out for why such a relationship may emerge in the models, but the relationship should be further explored in future research. Studies aimed at constraining climate sensitivity typically rely on indices and variability computed on seasonal, annual, and decadal timescales. Because the diurnal cycle occurs on sub-daily timescales, statistics can be generated over a relatively short time period relative to other emergent constraints.

While numerous studies have documented and quantified the diurnal precipitation cycle and its biases in global climate models, as outlined above, less research has been devoted to investigating its broader relationship to aggregate measures of climate change such as ECS and TCR. Diurnal variability may affect the mean climate through timescale feedbacks, as demonstrated by idealized cloud-resolving modeling of cumulus clouds (Ruppert, 2015). Literature surrounding diurnal cycle biases in CMIP6 is still sparse. A recent study looked at the diurnal cycle in 3 CMIP6 models (Watters et al., 2021), but the present study considers 21 CMIP5 and 26 CMIP6 models. The present paper complements existing studies of diurnal cycle biases in climate models while also exploring the relationship between the diurnal precipitation cycle and climate sensitivity.

2.3 Methods

To determine the phase and amplitude of the 24-hour precipitation composite, we perform a 2-mode cosine transform fit to the precipitation rate for a given gridcell, season, and CMIP model. More specifically, a function containing a diurnal (24 hour) and semi-diurnal (12 hour) component is fit after Universal Coordinated Time is transformed to local solar time [LST] in each gridcell, and precipitation rate means for each time bin are computed over the analysis period. The daily mean is subtracted such that the amplitude represents a deviation from the daily mean. The periods of the cosine functions are fixed, and the resulting phase [hours] and amplitude [mm day^{-1}] of the 24-hour mode are used for the analysis laid out in later sections. The Levenberg-Marquardt algorithm is employed for performing the non-linear cosine fit. To perform comparisons between model output and satellite observations at different grid resolutions, calculations of diurnal phase and amplitude are performed on the native CMIP model grid before regridding diurnal parameters to a common analysis grid. A common analysis grid is needed to appropriately account for differing grid resolutions, both among models in CMIP and between models and satellite

observations. The analysis grid resolution is chosen as an intermediate resolution between CMIP output and IMERG observations. Nearest-neighbor regridding is employed to interpolate the derived parameters to a $0.5^\circ \times 0.5^\circ$ analysis grid, which avoids issues with the cyclic discontinuity at midnight for phase. Otherwise, circular statistics are used when aggregating phase in space or time. Polar regions are excluded by only including latitudes in the range $[60^\circ\text{S}, 60^\circ\text{N}]$.

A standard practice in the diurnal cycle literature is to mask out grid cells where the diurnal cycle is weak and ill-defined, which reduces noise and improves the validity of satellite-model comparisons. The masking is often performed by removing cells with either low precipitation or low diurnal cycle amplitude ratio, as determined by the ratio of diurnal amplitude to mean precipitation (Covey et al., 2016). A more objective approach, employed here, is to find a parameter distribution of diurnal amplitude and exclude grid cells that contain 0 mm day^{-1} within 3 standard deviations of the estimated amplitude (i.e., cells that include 0 in the 99.7% confidence interval when the number of degrees of freedom in the estimate is large and statistics are normal). The remaining cells thus have a robustly detectable diurnal cycle amplitude. To build the parameter distribution, we use stationary bootstrapping (Politis & Romano, 1994). Briefly, the method entails continually sampling, with replacement, blocks of variable length from the full timeseries to build an ensemble of bootstrap samples, each representing a resampled version of the full dataset. Diurnal analysis is performed separately on each bootstrap sample, and the set of derived diurnal amplitudes forms a parameter distribution that quantifies uncertainty. For this analysis, 200 bootstrap samples with a size of 3 years are generated from the full IMERG timeseries. Block sizes are an integer number of days in length and follow a geometric distribution with a mean length of 10 days. The mask obtained by excluding cells outside 3 standard deviations is then applied to model output, such that only regions with a robust diurnal cycle in observations are analyzed. However, our results are insensitive to whether and how precisely this masking is carried out.

2.4 Data

Observational data

To assess the fidelity of simulated diurnal precipitation cycles, parameters estimated using an identical methodology are computed for NASA's Integrated Multi-satellite Retrievals for GPM (IMERG) V06B data product (Huffman et al., 2019). Briefly, the IMERG dataset combines estimates of precipitation from several passive mi-

crowave sounders aboard satellites in the GPM constellation. The "final" satellite product is inter-calibrated and regridded to a 0.1° grid before undergoing a series of advanced interpolation, re-calibration, assimilation, and correction procedures. The aforementioned processing steps are performed by NASA to produce the IMERG data product. Satellite-based rainfall products infer surface precipitation indirectly from emitted cloud top infrared radiation or the detection of hydrometers with microwave sounders, leading to diurnal phase biases of 2–4 hours with respect to rain gauge observations (Dai et al., 2007). However, IMERG V06 has been shown to reliably capture details of the diurnal phase with a smaller bias of around +0.6 hours compared to surface-based estimates, albeit the validation was only done for the Southeast U.S. (Tan et al., 2019). Estimates of precipitation rate are provided at 30-minute time intervals for the IMERG product. The IMERG V06 satellite product is reliably available starting June 2000. For this study, diurnal parameters are computed using data in the 15-year period spanning June 2000 to May 2015 for latitudes between 60°S and 60°N.

Climate Simulations

The analysis includes all historical CMIP runs with 3-hourly precipitation flux output available on the Earth System Grid Federation (ESGF) data server (https://esgf-node.llnl.gov/), including 26 models for CMIP6 and 21 models for CMIP5 (Eyring et al., 2016; Taylor et al., 2012). Because model realization and initialization are not expected to affect the fundamental representation of the diurnal cycle, a single ensemble member is used for each model. The latest 30-year period for each CMIP iteration is used. The analysis period spans 1976–2005 for CMIP5 and 1985–2014 for CMIP6. Estimates of the transient climate response (TCR) and equilibrium climate sensitivity come from Meehl et al. (2020), which uses the Gregory method to compute climate sensitivity. It is noted the Gregory method has been recently shown to underestimate the true ECS by 10% on average, and up to 25% for models with an ECS over 3 K (Dai et al., 2020). Not all models from ESGF with 3-hourly precipitation output have a reported climate sensitivity by Meehl and colleagues, so comparisons between diurnal parameters and climate sensitivity are made using the 21 overlapping CMIP6 models and 17 CMIP5 models.

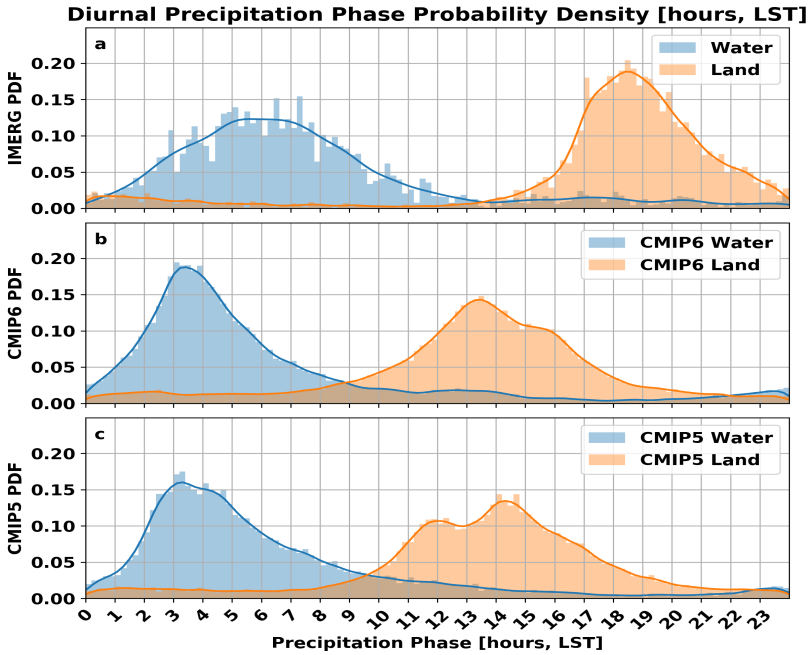

Figure 2.1: Probability density functions of diurnal precipitation phase (local solar time in hours of maximum) in IMERG observations (top, for 6/2000–5/2015 mean) and all-forcing historical simulations from 26 CMIP6 (middle, for 1985–2014 mean) and 21 CMIP5 (bottom, for 1976–2005 mean) models for land (orange) and water (blue) grid boxes between 60°S and 60°N on a common analysis grid. The diurnal phase is estimated using the diurnal component of a sinusoidal fit with diurnal (24 hour) and semi-diurnal (12 hour) modes. Grid cells for which 0 mm day^{-1} lies within 3 standard deviations of the diurnal amplitude are masked out.

2.5 Results

Diurnal Precipitation Cycle Biases

We compare models to satellite observations using probability density functions (PDF), spatial maps of phase and amplitude, and radial plots. PDFs highlight how characteristics of the phase distribution across space, time, and models differ from the phase distribution across space and time in satellite estimates. Figure 1 shows annual-mean PDFs of the diurnal precipitation phase. The GLDAS land mask, which includes large inland lakes, is used to identify grid cells over land and water. The spatial variance of precipitation phase over water is much larger in IMERG observations relative to both iterations of CMIP, although the model phase distribution includes variations across models. While spatial variance of the precipitation phase over land is comparable to satellite estimates, mean phase

remains ~ 5.4 hours too early in CMIP6 and ~ 5.2 hours too early hours in CMIP5. The CMIP5 biases are largely in line with the 6-10 hour phases biases found in (Covey et al., 2016) relative to TRMM satellite observations, although that study uses a different masking method and looks at warm-season diurnal cycles. Using the mode of the PDF instead of the mean, the land phase is ~ 5.1 hours too early in CMIP6 and ~ 4.5 hours too early in CMIP5. A notable outlier in land phase is FGOALs, which has a peak around 1.6 LST (FGOALS-g3 in CMIP6) and 0.8 LST (FGOALS-g2 in CMIP5).

The regional manifestations of the aforementioned biases over land and water become clearer in spatial plots of diurnal amplitude and phase. Figure 2 shows the annual-mean phase and amplitude, where the CMIP simulations are averaged across all models in each experiment after regridding to a common grid. Grid cells for which 0 mm day^{-1} lies within 3 standard deviations of the diurnal amplitude in satellite observations are masked out. Using 2 standard deviations retains much of the noisy signals in the extratropics (Fig. S1). The most striking difference globally is the early triggering of precipitation over land in climate models, a well known problem that remains an issue in CMIP6. The diurnal phase over extratropical continents has shifted earlier from CMIP5 to CMIP6, further from observations, notably over northern Asia and North America. Previous studies have noted issues with simulating nocturnal precipitation peaks associated with eastward-propagating mesoscale convective systems during summer (Liang et al., 2004; Trenberth et al., 2003), especially over the central U.S. (Jiang et al., 2006). The characteristic signature of these systems is a convective phase that smoothly transitions from early morning just east of the Rockies to late afternoon towards the southeastern U.S. CMIP6 models that robustly demonstrate this signal in northern hemisphere summer include MRI-ESM2-0, EC-Earth3, and EC-Earth3-Veg-LR. Diurnal phase over the rainiest ocean regions in the Intertropical Convergence Zone (ITCZ) is systematically a couple hours too early in both CMIP iterations.

While significant issues in phase remain, several improvements in the diurnal amplitude are noted. A more realistic diurnal amplitude (~ 7 mm day^{-1}) is observed over land in the Maritime Continent and South America for CMIP6. This may result from higher-resolution outputs in CMIP6 models, which better capture the localized nature of convection instead of spreading out the signal across a larger gridcell. Studies employing cloud-resolving models have revealed better agreement of diurnal amplitude and phase with satellite estimates with increases in horizontal

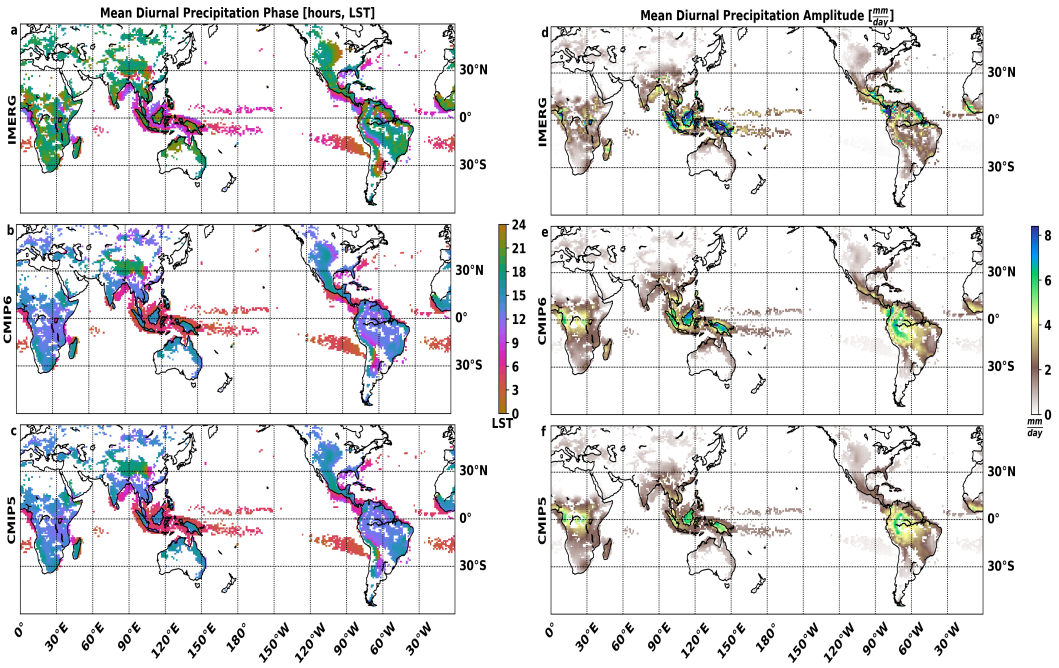

Figure 2.2: Annual-mean diurnal precipitation amplitude (deviation from daily mean in mm day^{-1}, right column) and phase (local solar time in hours of the maximum, left column) in IMERG observations (top row, for 6/2000–5/2015 mean) and averaged across 26 CMIP6 models (middle row, for 1985–2014 mean) and 21 CMIP5 models (bottom row, for 1976–2005 mean). The amplitude and phase are estimated using the diurnal component of a sinusoidal fit with diurnal (24 hour) and semi-diurnal (12 hour) modes. Grid cells for which 0 mm day^{-1} lies within 3 standard deviations of the diurnal amplitude are masked out.

resolution (Dirmeyer et al., 2012; Sato et al., 2009). A slight improvement in the diurnal cycle amplitude over the South Pacific Convergence Zone (SPCZ) brings the models more in line with observations, but the double-ITCZ bias still exists in CMIP6 (Tian & Dong, 2020).

To quantify the ability of models to simulate the spatial structure of the phase, correlations between model and satellite parameters are computed for each model across gridcells and are averaged in space. The sine of phase is used because local solar time is a circular quantity. The phase correlation is slightly higher in CMIP6 over oceans (0.30 and 0.35 for CMIP5 and CMIP6, respectively) and land (0.23 and 0.29). The slight, but insignificant, improvement in the spatial correlations over land from CMIP5 to CMIP6 is largely attributable to regions influenced by topography, notably east of the Andes mountains in South America, the periphery of the Tibetan Plateau, and over the Central Plains of the U.S. In CMIP6, the phase

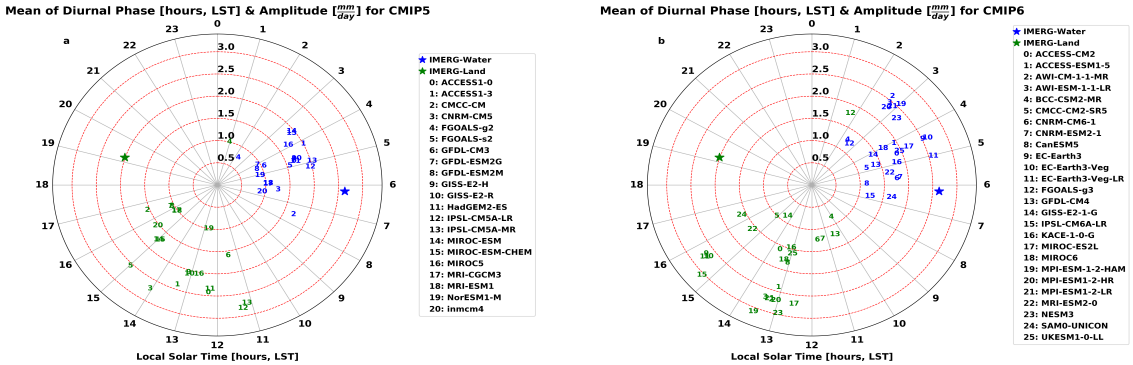

Figure 2.3: Mean diurnal phase and amplitude averaged over 60°S–60°N land (green) and water (blue) for 21 CMIP5 (left) and 26 CMIP6 (right) models and satellite-derived estimates (IMERG as stars). Grid cells for which 0 mm day^{-1} lies within 3 standard deviations of the diurnal amplitude are masked out. The radius from the center represents the mean diurnal amplitude (deviation from daily mean) and the angular position represents the mean phase (LST in hours of the maximum). The dashed concentric circles (representing diurnal amplitude) are spaced at 0.5 mm day^{-1}.

in these regions shifts earlier by 1–2 hours. Over the oceans, no discernible regional pattern is noticeable outside the ITCZ; in the ITCZ, the phase of precipitation is shifted 0–1 hours earlier in CMIP6 models.

In addition to model-satellite correlations, spatially-averaged phase and amplitude over land and water are used as a summary metric to systematically and objectively assess changes between CMIP iterations. Figure 3 demonstrates the spatially-averaged phase and amplitude on a clock-like radial plot, where the distance from the center corresponds to diurnal amplitude and the azimuth angle to diurnal phase. The spread in mean diurnal amplitude and phase among models is larger over land in both iterations of CMIP. The spread over land, as measured by the standard deviation, decreases slightly from 2.4 in CMIP5 to 2.2 hours in CMIP6. The spread in amplitude increases from 0.64 to 0.76 mm day^{-1}, although there are more models in CMIP6 . Over water, the standard deviation of phase increases slightly from 1.2 in CMIP5 to 1.3 hours in CMIP6, and the spread in amplitude increases from 0.44 to 0.51 mm day^{-1}. When spatially averaged, CMIP6 models have a larger bias than CMIP5 over both land and water for diurnal phase but a reduced bias for diurnal amplitude. The spread in diurnal parameters remains large in both CMIP iterations.

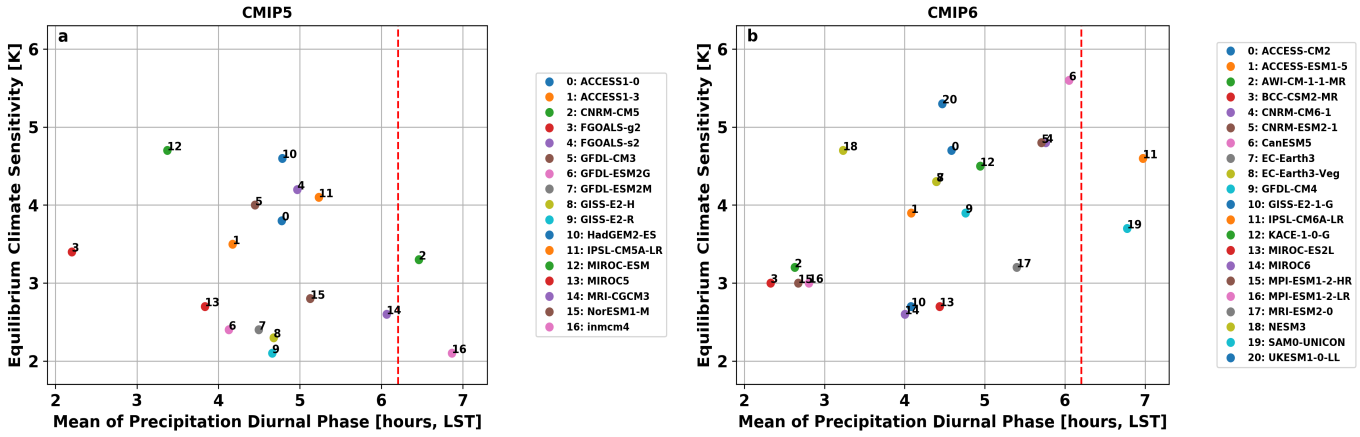

Figure 2.4: Scatter plot of ECS against mean diurnal cycle phase over global oceans between 60S and 60N for 26 CMIP6 (Corr = 0.51) and 21 CMIP5 (Corr = -0.26) models. Grid cells for which 0 mm day^{-1} lies within 3 standard deviations of the diurnal amplitude are masked out. Satellite-based estimate of diurnal phase from IMERG shown as red dashed line.

ECS Relationship and Potential Physical Mechanisms

To assess broader climate implications of the diurnal precipitation cycle, mean phase and amplitude over both land and water are regressed against ECS and TCR. We also broke down the regressions by season and hemisphere (Table S1); the robust relationships that emerged are summarized in what follows. A weak but persistent relationship between precipitation phase over oceans and ECS is found across hemispheres in the annual mean and in individual seasons. The strongest relationship exists in northern hemisphere winter over the oceans (Corr = 0.63), with a comparable correlation in southern hemisphere winter (Corr = 0.62). In the global and annual mean, the ECS-phase correlation over oceans in CIMP6 is 0.51, while the correlation is only −0.26 in CMIP5. To illustrate the relationship, scatter plots of the ECS–phase relationships are shown in Figure 4, together with the observed oceanic phase. Weighting the ocean phase by annual-mean precipitation or subselecting regions by mean precipitation has a minor impact on this relationship. Correlations between ECS/TCR and diurnal parameters over land are negligible. The correlation between ECS and TCR for the available CMIP6 models is 0.7, which is comparable to the ECS–phase relationship over oceans in winter months.

Previous studies have pointed to the potential for using short-term variability such as diurnal and seasonal cycles to characterize a climate model's sensitivity (Brient & Schneider, 2016; Covey et al., 2000; Williams et al., 2020). However, it is important

to understand the physical mechanism that accounts for any such relationship. These are several possible mechanisms for why such a relationship between diurnal phase and ECS may exist in CMIP6 models:

- **A reflection of entrainment** ECS is strongly sensitive to the selected cumulus parameterization, specifically to the bulk detrainment efficiency and details of how cumulus cloud droplets are converted to precipitation (Zhao et al., 2016). The diurnal cycle is also sensitive to cumulus parameterization, as noted by Liang et al. (2004). For instance, diurnal studies employing the Community Atmosphere Model (CAM) have revealed that the entrainment rate in the cumulus parameterization affects both the diurnal timing and intensity of precipitation significantly (DeMott et al., 2007). Similarly, modifying cumulus mixing through entrainment/detrainment rates in GCMs has also been shown to influence the phase and amplitude of oceanic precipitation, with little impact on mean precipitation amounts (Hohenegger & Stevens, 2013). That is, subgrid-scale physics surrounding cloud microphysics and mixing that manifests itself locally and on daily timescales through the diurnal precipitation cycle may also impact the mean climate. In other words, the diurnal cycle biases and ECS variations may have common causes, without the diurnal cycle biases directly causing ECS variations.

- **A proxy for tropical low cloud amount or depth** The depth of tropical low clouds in subsidence regions has been identified as correlating with climate sensitivity, owing to competing effects of how convective drying and turbulent moistening are parameterized in models (Brient et al., 2016). The diurnal phase may also depend on cloud depth or cloud amount and reflect the well-known uncertainties associated with low clouds in GCMs. If we posit a relationship between cloud depth and the diurnal precipitation cycle, models with deeper clouds may have a more robust diurnal cycle.

- **A proxy for cloud radiative effects** Presuming a later precipitation peak in the early morning hours corresponds to a later minimum of precipitation in the afternoon, the observed relationship may reflect how clouds interact with shortwave radiation during the day. We expect shortwave reflection to depend both on cloud properties as well as on the solar zenith angle. For instance, maximum cloudiness that occurs at midday (small solar zenith angle) would result in more shortwave reflection than maximum cloudiness at night or early

morning, even for the same daily mean cloud cover. Nevertheless, such an effect does not fully explain a relationship to ECS.

In an attempt to falsify some of the mechanisms suggested above, we assess whether either deep convective or shallow cloud regions are contributing disproportionately to the ECS-phase relationship. The analysis is repeated by correlating the mean diurnal phase in both low precipitation ($< 1.5\text{mm day}^{-1}$) and high precipitation ($> 5\text{mm day}^{-1}$) regions against ECS, in both cases without the observational mask (Table S2). Low precipitation regions are found to have a marginally higher correlation (Corr = 0.52) than high precipitation regions (Corr = 0.47), meaning the source of the relationship may involve mechanisms operating in both regions or involve a combination of the hypotheses listed above. However, further work is needed to elucidate the mechanisms involved—if the relation between diurnal cycle phase and ECS in CMIP6 in fact turns out to be significant.

2.6 Conclusions

This study quantified diurnal precipitation biases in a consistent manner in CMIP5 and CMIP6 and highlights that biases in diurnal parameters improve marginally between these CMIP iterations. In particular, the mean diurnal precipitation phase remains ~5.4 hours too early over land, and the diurnal amplitude remains ~0.81 mm day^{-1} too small over the oceans. While comparisons of aggregate statistics such a spatial means and correlations with satellite-based observations reveal no significant improvements, more realistic characteristics of the diurnal cycle are noted in CMIP6. Improvements include the more robust simulation in several CMIP6 models of diurnal cycle characteristics that appear to be shaped by nocturnal mesoscale convective systems, and a more realistic diurnal amplitude over the Maritime Continent.

A secondary aim of this study was to assess the broader importance of the diurnal precipitation cycle by regressing diurnal-cycle parameters against ECS. Climate models with a later precipitation phase over the oceans tend to have a higher climate sensitivity in CMIP6; however, this relationship is not evident in CMIP5, calling into question its robustness.

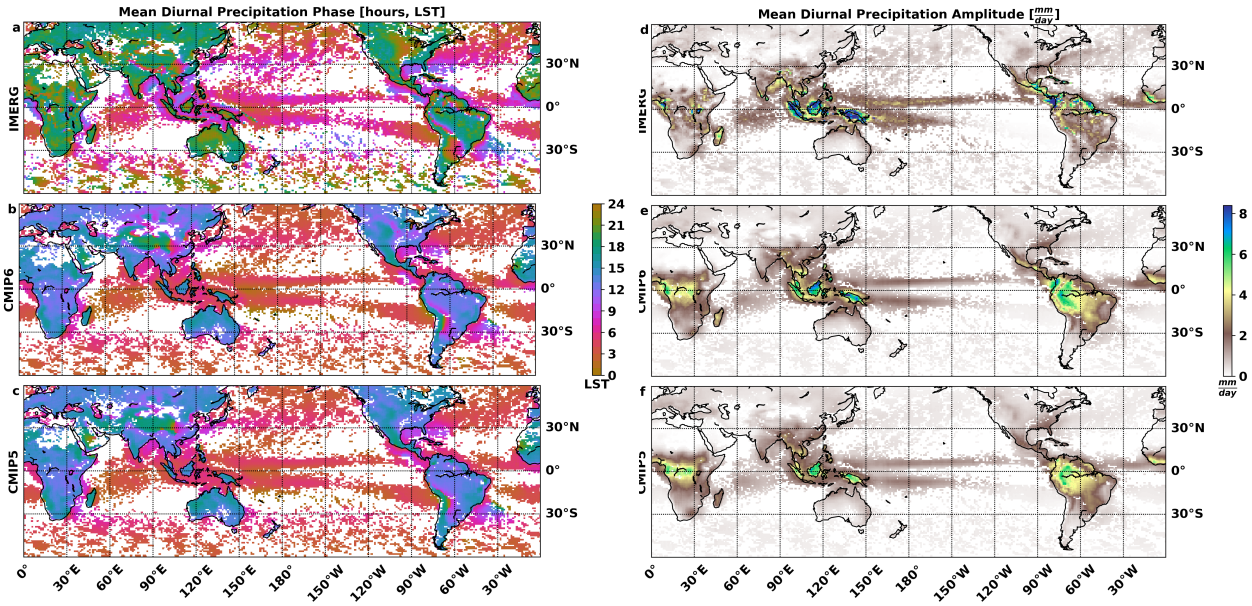

Figure 2.5: Annual-mean diurnal precipitation amplitude (deviation from daily mean in mm day^{-1}, right column) and phase (local solar time in hours of the maximum, left column) in IMERG observations (top row, for 6/2000 – 5/2015 mean) and averaged across 26 CMIP6 models (middle row, for 1985–2014 mean) and 21 CMIP5 models (bottom row, for 1976–2005 mean). The amplitude and phase are estimated using the diurnal component of a sinusoidal fit with diurnal (24 hour) and semi-diurnal (12 hour) modes. Grid cells for which 0 mm day $^{-1}$ lies within 2 standard deviations of the diurnal amplitude are masked out.

2.7 Supporting Information

Table S1 Correlation coefficient between spatial mean of diurnal precipitation phase over oceans against (**a**) ECS and (**b**) TCR for 26 CMIP6 models. Correlations are broken down by season (row) and hemisphere (column). Grid cells for which 0 mm day $^{-1}$ lies within 3 standard deviations of the diurnal amplitude are masked out.

a

CMIP6 Oceanic Phase–ECS Correlations

Season	Northern Hemisphere	Southern Hemisphere
DJF	0.63	0.49
MAM	0.46	0.55
JJA	0.41	0.62
SON	0.45	0.47

CMIP6 Oceanic Phase–TCR Correlations

	Season	Northern Hemisphere	Southern Hemisphere
b	**DJF**	0.32	0.25
	MAM	0.25	0.25
	JJA	0.23	0.34
	SON	0.24	0.27

Table S2 Correlation coefficient between spatial mean of diurnal precipitation phase over oceans and ECS in low-precipitation areas (< 1.5 mm day^{-1}, top row) and high-precipitation areas (> 5 mm day^{-1}, bottom row) for 26 CMIP6 models and 21 CMIP5 models. Precipitation masks are calculated and applied separately for each model.

CMIP Oceanic Phase–ECS Correlations

Precipitation Regime	CMIP5	CMIP6
Low	-0.15	0.52
High	-0.31	0.47

References

Bechtold, P., Chaboureau, J. P., Beljaars, A., Betts, A. K., Köhler, M., Miller, M., & Redelsperger, J. L. (2004). The simulation of the diurnal cycle of convective precipitation over land in a global model. *Quarterly Journal of the Royal Meteorological Society*, *130 C*(604), 3119–3137. https://doi.org/10.1256/qj.03.103

Bowman, K. P., Collier, J. C., North, G. R., Wu, Q., Ha, E., & Hardin, J. (2005). Diurnal cycle of tropical precipitation in Tropical Rainfall Measuring Mission (TRMM) satellite and ocean buoy rain gauge data. *Journal of Geophysical Research Atmospheres*, *110*(21), 1–14. https://doi.org/10.1029/2005JD005763

Brient, F., & Schneider, T. (2016). Constraints on climate sensitivity from space-based measurements of low-cloud reflection. *Journal of Climate*, *29*(16), 5821–5835. https://doi.org/10.1175/JCLI-D-15-0897.1

Brient, F., Schneider, T., Tan, Z., Bony, S., Qu, X., & Hall, A. (2016). Shallowness of tropical low clouds as a predictor of climate models' response to warming. *Climate Dynamics*, *47*(1-2), 433–449. https://doi.org/10.1007/s00382-015-2846-0

Collier, J. C., & Bowman, K. P. (2004). Diurnal cycle of tropical precipitation in a general circulation model. *Journal of Geophysical Research D: Atmospheres*, *109*(17). https://doi.org/10.1029/2004JD004818

Covey, C., Guilyardi, E., Jiang, X., Johns, T. C., Treut, H. L., Madec, G., Meehl, G. a., Miller, R., Power, S. B., Roeckner, E., & Russell, G. (2000). The seasonal cycle in coupled ocean-atmosphere general circulation models. *Climate Dynamics*, 775–787.

Covey, C., Gleckler, P. J., Doutriaux, C., Williams, D. N., Dai, A., Fasullo, J., Trenberth, K., & Berg, A. (2016). Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models. *Journal of Climate*, *29*(12), 4461–4471. https://doi.org/10.1175/JCLI-D-15-0664.1

Dai, A. (2001). Global precipitation and thunderstorm frequencies. Part II: Diurnal variations. *Journal of Climate*, *14*(6), 1112–1128. https://doi.org/10.1175/1520-0442(2001)014<1112:GPATFP>2.0.CO;2

Dai, A. (2006). Precipitation characteristics in eighteen coupled climate models. *Journal of Climate*, *19*(18), 4605–4630. https://doi.org/10.1175/JCLI3884.1

Dai, A., Giorgi, F., & Trenberth, K. E. (1999). Observed and model-simulated diurnal cycles of precipitation over the contiguous United States. *Journal of Geophysical Research Atmospheres*, *104*(D6), 6377–6402. https://doi.org/10.1029/98JD02720

Dai, A., Huang, D., Rose, B. E., Zhu, J., & Tian, X. (2020). Improved methods for estimating equilibrium climate sensitivity from transient warming simulations. *Climate Dynamics*, *54*(11-12), 4515–4543. https://doi.org/10.1007/s00382-020-05242-1

Dai, A., Lin, X., & Hsu, K. L. (2007). The frequency, intensity, and diurnal cycle of precipitation in surface and satellite observations over low- and mid-latitudes. *Climate Dynamics*, *29*(7-8), 727–744. https://doi.org/10.1007/s00382-007-0260-y

Dai, A., & Trenberth, K. E. (2004). The diurnal cycle and its depiction in the community climate system model. *Journal of Climate*, *17*(5), 930–951. https://doi.org/10.1175/1520-0442(2004)017<0930:TDCAID>2.0.CO;2

DeMott, C. A., Randall, D. A., & Khairoutdinov, M. (2007). Convective precipitation variability as a tool for general circulation model analysis. *Journal of Climate*, *20*(1), 91–112. https://doi.org/10.1175/JCLI3991.1

Dirmeyer, P. A., Cash, B. A., Kinter, J. L., Jung, T., Marx, L., Satoh, M., Stan, C., Tomita, H., Towers, P., Wedi, N., Achuthavarier, D., Adams, J. M., Altshuler, E. L., Huang, B., Jin, E. K., & Manganello, J. (2012). Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization. *Climate Dynamics*, *39*(1-2), 399–418. https://doi.org/10.1007/s00382-011-1127-9

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Hohenegger, C., & Stevens, B. (2013). Controls on and impacts of the diurnal cycle of deep convective precipitation. *Journal of Advances in Modeling Earth Systems*, *5*(4), 801–815. https://doi.org/10.1002/2012ms000216

Huffman, G., Stocker, E., Bolvin, D., Nelkin, E., & Jackson, T. (2019). GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/GPM/IMERG/3B-HH/06

Jiang, X., Lau, N. C., & Klein, S. A. (2006). Role of eastward propagating convection systems in the diurnal cycle and seasonal mean of summertime rainfall over the U.S. Great Plains. *Geophysical Research Letters*, *33*(19), 1–6. https://doi.org/10.1029/2006GL027022

Kikuchi, K., & Wang, B. (2008). Diurnal precipitation regimes in the global tropics. *Journal of Climate*, *21*(11), 2680–2696. https://doi.org/10.1175/2007JCLI2051.1

Lee, M. I., Schubert, S. D., Suarez, M. J., Schemm, J. K. E., Pan, H. L., Han, J., & Yoo, S. H. (2008). Role of convection triggers in the simulation of the diurnal cycle of precipitation over the United States Great Plains in a general circulation model. *Journal of Geophysical Research Atmospheres*, *113*(2), 1–10. https://doi.org/10.1029/2007JD008984

Liang, X. Z., Li, L., Dai, A., & Kunkel, K. E. (2004). Regional climate model simulation of summer precipitation diurnal cycle over the United States. *Geophysical Research Letters*, *31*(24), 1–4. https://doi.org/10.1029/2004GL021054

Lin, X., Randall, D. A., & Fowler, L. D. (2000). Diurnal variability of the hydrologic cycle and radiative fluxes: Comparisons between observations and a GCM. *Journal of Climate*, *13*(23), 4159–4179. https://doi.org/10.1175/1520-0442(2000)013<4159:DVOTHC>2.0.CO;2

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap.

Pritchard, M. S., & Somerville, R. C. J. (2009). Assessing the diurnal cycle of precipitation in a multi-scale climate model. *Journal of Advances in Modeling Earth Systems*, *2*. https://doi.org/10.3894/james.2009.1.12

Randall, D. A., Harshvardhan, & Dazlich, D. A. (1991). Diurnal Variability of the Hydrologic Cycle in a General Circulation Model. *Journal of the Atmospheric Sciences*, *48*.

Ruppert, J. H. (2015). Diurnal timescale feedbacks in the tropical cumulus regime. *Journal of Advances in Modeling Earth Systems*, *7*, 1339–1350. https://doi.org/10.1002/2017MS001065

Ruppert, J. H., & Hohenegger, C. (2018). Diurnal circulation adjustment and organized deep convection. *Journal of Climate*, *31*(12), 4899–4916. https://doi.org/10.1175/JCLI-D-17-0693.1

Sato, T., Miura, H., Satoh, M., Takayabu, Y. N., & Wang, Y. (2009). Diurnal cycle of precipitation in the tropics simulated in a global cloud-resolving model. *Journal of Climate*, *22*(18), 4809–4826. https://doi.org/10.1175/2009JCLI2890.1

Sorooshian, S., Gao, X., Hsu, K., Maddox, R. A., Hong, Y., Gupta, H. V., & Imam, B. (2002). Diurnal variability of tropical rainfall retrieved from combined GOES and TRMM satellite information. *Journal of Climate*, *15*(9), 983–1001. https://doi.org/10.1175/1520-0442(2002)015<0983:DVOTRR>2.0.CO;2

Tan, J., Huffman, G. J., Bolvin, D. T., & Nelkin, E. J. (2019). Diurnal Cycle of IMERG V06 Precipitation. *Geophysical Research Letters*, *46*(22), 13584–13592. https://doi.org/10.1029/2019GL085395

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Tian, B., & Dong, X. (2020). The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation. *Geophysical Research Letters*, *47*(8), 0–3. https://doi.org/10.1029/2020GL087232

Trenberth, K. E., Dai, A., Rasmussen, R. M., & Parsons, D. B. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society*, *84*(9), 1205–1217. https://doi.org/10.1175/BAMS-84-9-1205

Watters, D., Battaglia, A., & Allan, R. P. (2021). The Diurnal Cycle of Precipitation According to Multiple Decades of Global Satellite Observations , Three CMIP6 Models , and the ECMWF Reanalysis. *Journal of Climate*, 1–58. https://doi.org/10.1175/JCLI-D-20-0966.1.

Williams, K. D., Hewitt, A. J., & Bodas-Salcedo, A. (2020). Use of Short-Range Forecasts to Evaluate Fast Physics Processes Relevant for Climate Sensitivity. *Journal of Advances in Modeling Earth Systems*, *12*(4), 1–9. https://doi.org/10.1029/2019MS001986

Yang, G. Y., & Slingo, J. (2001). The diurnal cycle in the tropics. *Monthly Weather Review*, *129*(4), 784–801. https://doi.org/10.1175/1520-0493(2001)129<0784:TDCITT>2.0.CO;2

Zhao, M., Golaz, J. C., Held, I. M., Ramaswamy, V., Lin, S. J., Ming, Y., Ginoux, P., Wyman, B., Donner, L. J., Paynter, D., & Guo, H. (2016). Uncertainty in model climate sensitivity traced to representations of cumulus precipitation microphysics. *Journal of Climate*, *29*(2), 543–560. https://doi.org/10.1175/JCLI-D-15-0191.1

Chapter 3

ONLINE LEARNING OF ENTRAINMENT CLOSURES IN A HYBRID MACHINE LEARNING PARAMETERIZATION

3.1 Abstract

This work integrates machine learning into an atmospheric parameterization to target uncertain mixing processes while maintaining interpretable, predictive, and well-established physical equations. We adopt an eddy-diffusivity mass-flux (EDMF) parameterization for the unified modeling of various convective and turbulent regimes. To avoid drift and instability that plague offline-trained machine learning parameterizations that are subsequently coupled with climate models, we frame learning as an inverse problem: Data-driven models are embedded within the EDMF parameterization and trained online in a one-dimensional vertical GCM column. Training is performed against output from large-eddy simulations (LES) forced with GCM-simulated large-scale conditions in the Pacific. Rather than optimizing subgrid-scale tendencies, our framework directly targets climate variables of interest, such as the vertical profiles of entropy and liquid water path. Specifically, we use ensemble Kalman inversion to simultaneously calibrate both the EDMF parameters and the parameters governing data-driven lateral mixing rates. The calibrated parameterization outperforms existing EDMF schemes, particularly in tropical and subtropical locations of the present climate, and maintains high fidelity in simulating shallow cumulus and stratocumulus regimes under increased sea surface temperatures from AMIP4K experiments. The results showcase the advantage of physically-constraining data-driven models and directly targeting relevant variables through online learning to build robust and stable machine learning parameterizations.

3.2 Introduction

The latest suite of global climate models (GCMs) continues to exhibit a large range of climate sensitivities, the measure of Earth's equilibrium temperature response to a doubling of atmospheric greenhouse gas concentrations (Meehl et al., 2020). Variance in modeled responses has been traced to disparate representations of subgrid-scale (SGS) processes not explicitly resolved by climate models, specifically those controlling the characteristics of cloud feedbacks (Bony et al., 2015; Sherwood et al., 2014; Vial et al., 2013; Zelinka et al., 2020). Furthermore, climate models often fail to reproduce several key statistics from the recent past when run retrospectively (Vignesh et al., 2020). In light of these discrepancies, researchers have launched systematic efforts across the climate modeling enterprise to incorporate machine learning (ML) methods into GCMs, in order to improve the ability of climate model components to learn from high fidelity data. This study specifically uses a training dataset focused on marine low cloud regimes in the central and eastern Pacific—areas that are particularly problematic to model in GCMs (Črnivec et al., 2023; Nam et al., 2012), yet are critical for precise assessments of equilibrium climate sensitivity due to cloud feedbacks (Brient & Schneider, 2016; Myers et al., 2021; Siler et al., 2018).

Initiatives to replace existing physics-based parameterizations in atmospheric models entirely with ML are often marred with challenges surrounding numerical instability and extrapolation performance. Instabilities, such as the generation of unstable gravity wave modes (Brenowitz et al., 2020), largely arise from feedbacks between the learned SGS parameterization and the dynamical core upon integration. Currently, the favored strategy is to train ML models offline via supervised learning to predict SGS tendencies as a function of the resolved atmospheric state, then couple trained models to a dynamical core to perform inferences at each model timestep (Krasnopolsky et al., 2013; Rasp et al., 2018; Yuval & O'Gorman, 2020). As an example of the offline training procedure for atmospheric turbulence, a recent encoder-decoder approach was used to learn vertical turbulent fluxes in dry convective boundary layers on the basis of coarse-grained large-eddy simulations (Shamekh & Gentine, 2023). Although significant progress has been made towards advancing and stabilizing data-driven parameterizations (Brenowitz & Bretherton, 2019; Wang et al., 2022; Watt-Meyer et al., 2023), the conventional offline training strategy precludes learning unobservable processes indirectly from relevant climate statistics. Furthermore, instabilities arising from system feedbacks are not typically incorporated into training, and cannot be easily assessed until ML models are cou-

pled to a dynamical core (Ott et al., 2020; Rasp, 2020). More recently, the advent of differentiable simplified general circulation models (e.g., without phase transitions of water) has enabled spatially three-dimensional (3D) online training of ML-based SGS parameterizations using short-term forecasts (Kochkov et al., 2024). These strategies have not yet overcome the problems of instability and extrapolation to warmer climates and remain difficult to interpret.

We take steps to address these issues by employing ensemble Kalman inversion (EKI) to perform parameter estimation within a SGS parameterization from statistics of atmospheric profiles in a single column setup (Dunbar et al., 2021; Huang, Schneider, & Stuart, 2022; Iglesias et al., 2013). Treating learning as an inverse problem directly enables online learning. Inverse problems are characterized by setups where the dependent variable of some target process is neither directly observable nor explicitly included in the loss function. In this case, it is through secondary causal effects of atmospheric dynamics on observable atmospheric quantities that parameters are optimized. In the field of dynamical systems, theory underpinning the use of inversion techniques to infer parameters is well established (Huang, Huang, et al., 2022; Iglesias et al., 2013), and they have also been shown to be effective for learning neural networks (NNs), especially in chaotic system where the smoothing properties of ensemble methods can be advantageous (Dunbar et al., 2022; Kovachki & Stuart, 2019). In practice, ensemble Kalman methods have been used to learn drift and diffusion terms in the Lorenz '96 model (Schneider et al., 2021), nonlinear eddy viscosity models for turbulence (Zhang et al., 2022), the effects of truncated variables in a quasi-geostrophic ocean-atmosphere model (Brajard et al., 2021), and NN-based parameterizations of the quasi-biennial oscillation and gravity waves (Pahlavan et al., 2024). An alternative approach to online learning relies on differentiable methods to explicitly compute gradients through the physical model to learn data-driven components (Shen et al., 2023; Um et al., 2020). The differentiable learning approach has been used successfully to learn NN-based closures in numerous idealized turbulence setups (Kochkov et al., 2021; List et al., 2022; MacArt et al., 2021; Shankar et al., 2023). In an Earth system modeling setting, differentiable online learning has been used to learn stable turbulence parameterizations in an idealized quasi-geostrophic setup (Frezat et al., 2022) and residual corrections to an upper-ocean convective adjustment scheme (Ramadhan et al., 2023). While promising, differentiable methods preclude computing gradients through physical models with non-differentiable components, such as the physics stemming from water phase changes in cloud parameterizations. Furthermore,

given existing work surrounding differentiable and inverse methods for geophysical fluid dynamics, there remains a lack of literature demonstrating indirect learning of data-driven components in more comprehensive atmospheric parameterizations of convection, turbulence, and clouds. Our contribution is the application of these methods in a more realistic climate modeling setting, a use case which can directly improve operational Earth system models.

We extend a flexible and modular framework that allows for the selective addition of expressive, non-parametric components where physical knowledge is limited, introduced by (Lopez-Gomez et al., 2022). Our approach promotes generalizability and interpretability. Interpretability comes by virtue of targeting specific physical processes, which enables a mechanistic analysis of their effect on climate. Generalizability is a result of both retaining this physical framework and employing an inversion strategy that targets climate statistics. The physical framework includes the partial differential equations in which the closure is embedded, the nondimensionalization of data-driven input variables, and the dimensional scales that modulate learned nondimensional closures. In contrast, a fully data-driven parameterization benefits from expressivity at the expense of sensitivity to training data, leading to difficulties in extrapolating to unobserved climates. Generalizability is verified in our setup by assessing performance on an out-of-distribution climate where SSTs are uniformly increased by 4 K; test error decreases in lockstep with training error from the present climate and overfitting is not observed.

In this study, we will investigate the performance of a single column model containing data-driven lateral mixing closures spanning a range of complexities, from linear regression models to neural networks. In section 2, we describe in detail the data-driven architectures, training data, and online calibration pipeline. Section 3 outlines the performance of the data-driven eddy-diffusivity mass-flux (EDMF) scheme in terms of the root mean squared error of the mean atmospheric state in a current and warmer climate, and representative vertical profiles are presented with physical implications discussed. Relative to the previous work of Lopez-Gomez et al. (2022), modeling improvements are made by both modifying the calibration pipeline and addressing structural biases in the EDMF model itself, namely boundary conditions and the lateral mixing formulation.

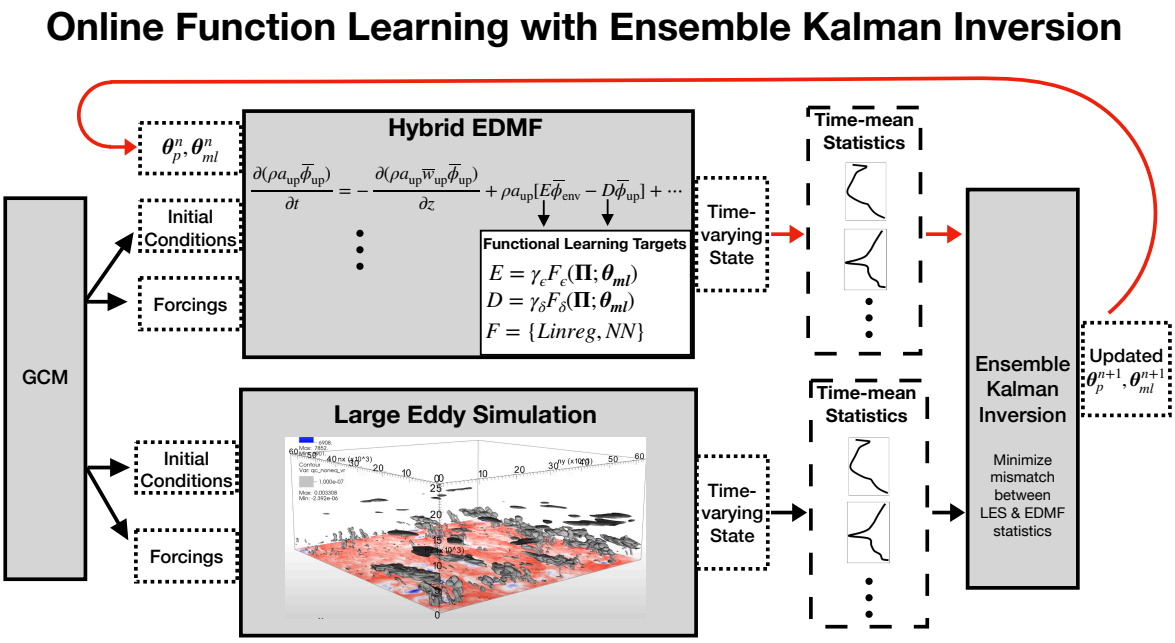

Figure 3.1: Schematic illustrating the ensemble Kalman inversion pipeline used for online training of a one-dimensional (1D) atmospheric model with both physics-based and data-driven components (hybrid EDMF). Black arrows indicate fixed operations between components, and red arrows indicate dynamic information flow on the basis of Kalman updates to EDMF parameters. The training data comprises 176 LES simulations from the AMIP climate, processed in batches of 16 cases for each ensemble Kalman iteration. Lateral mixing rates are formulated as the product of a dimensional scale γ and a data-driven, nondimensional function F.

3.3 Online Training Setup

An overarching goal of SGS modeling is to produce computationally-efficient schemes that emulate expensive high-resolution simulations, given the same large-scale forcings, boundary conditions, and initial conditions. Of primary importance are the prediction of SGS fluxes and cloud properties, which are determined by small-scale processes not resolvable by the GCM dynamical core. In the setup described here, parameters in a full-complexity SGS scheme are systematically optimized through the ensemble Kalman inversion technique to match characteristics of high-resolution simulations, namely time-mean vertical profiles and vertically-integrated liquid water content produced by large-eddy simulations (LES) (Shen et al., 2022). A variant of the SGS scheme is introduced, which imposes fewer assumptions and incorporates more general data-driven functions that can be determined with data. The SGS model is an eddy-diffusivity mass-flux (EDMF) scheme that parameterizes the effects of turbulence, convection, and clouds. The refer-

ence high-resolution simulations are performed with PyCLES (Pressel et al., 2015), which explicitly models convection and turbulent eddies larger than $O(10\text{ m})$. The process diagram in Figure 3.1 illustrates how calibrations are performed using the SGS model. Components of the diagram are detailed in the sections that follow, starting with the EDMF scheme.

Eddy-diffusivity Mass-flux (EDMF) Scheme Overview

EDMF schemes partition GCM grid boxes into two or more subdomains, each characterized by containing either coherent structures (updrafts) or relatively isotropic turbulence (environment). While most SGS schemes use separate parameterizations for the boundary layer, shallow convection, deep convection, and stratocumulus regimes, the extended EDMF scheme we use (herein referred to as EDMF) simulates all regimes in a unified manner by making fewer simplifying assumptions (Thuburn et al., 2018). The scheme includes partial differential equations (PDEs) for prognostic updraft properties (notably temperature, humidity, area fraction, and mass flux), which are coupled to PDEs for environmental variables (temperature, humidity, and turbulent kinetic energy). The physical skeleton of the EDMF consists of these coarse-grained equations of motion and houses a collection of closures, appearing as right-hand-side tendency terms for the prognostic variable equations.

The EDMF scheme we use was initially introduced by Tan et al. (2018). It contains closure functions, for example, for entrainment and detrainment, which capture physics without a known, closed-form expression; specifying them is necessary to fully define the set of EDMF PDEs such that they can be numerically integrated. Closures in the EDMF equations play a role similar to SGS parameterizations in grid-scale prognostic equations. Tendencies from SGS parameterizations appear in dynamical core equations, and, similarly, tendencies from closures appear in the EDMF equations. In the context of GCMs, the EDMF parameterization predicts vertical SGS fluxes and cloud properties due to unresolved processes. The present EDMF parameterization, which is run at 50 m vertical resolution, has been shown to effectively generalize between isotropic and stretched vertical grids (Lopez-Gomez et al., 2022). Its prediction of second-order quantities such as turbulent kinetic energy (TKE), which approach zero as the resolution increases, and its inherent SGS memory endow it with some "scale-aware" properties that become especially important as convection begins to be partially resolved in the "gray zone" (Boutle et al., 2014; Schneider et al., 2024; Tan et al., 2018); however, we have not explicitly tested its resolution dependence yet.

Following domain decomposition, the contributions of EDMF SGS fluxes ($\rho\langle w'\phi'\rangle_{\text{sgs}}$) to the grid-scale equation for a general quantity ϕ are

$$\rho\langle w'\phi'\rangle_{\text{sgs}} = -\rho a_{\text{env}} K_{\phi,\text{env}} \frac{\partial \bar{\phi}_{\text{env}}}{\partial z} + \rho a_{\text{up}} (\bar{w}_{\text{up}} - \langle w\rangle)(\bar{\phi}_{\text{up}} - \langle \phi\rangle). \tag{3.1}$$

Here, $\langle\cdot\rangle$ indicates a grid-mean quantity and $\overline{(\cdot)}$ a subdomain mean. Subscripts "up" and "env" signify updraft and environmental subdomain properties, respectively. We define ρ as the air density, a as the subdomain area fraction, $K_{\phi,\text{env}}$ as the environmental diffusivity of quantity ϕ, and w as the vertical velocity. The first term parameterizes the turbulent flux due to eddy diffusion (ED) in the environment; the second term represents the mass flux (MF) from coherent updrafts. The environmental eddy diffusivity, which governs the diffusive flux, is determined by a mixing length closure (Lopez-Gomez et al., 2020) and environmental TKE, following Mellor and Yamada (1982). In shallow maritime regimes, the turbulent kinetic energy budget is dominated by a balance between shear, buoyancy, and viscous dissipation (Heinze et al., 2015). Thus, lateral mixing primarily affects the updraft mass flux term.

Baseline EDMF: EDMF-20

We compare a hybrid EDMF, detailed in the next section, to a baseline version we call the EDMF-20. The EDMF-20 model includes physically motivated closures for eddy diffusivity (Lopez-Gomez et al., 2020), entrainment/detrainment (Cohen et al., 2020), and perturbation pressure. The physically motivated closure functions were manually tuned so that the simulated EDMF profiles closely match field campaigns. Parameters in EDMF-20 were tuned to match field campaigns representing a spectrum of convective and turbulent regimes, including Bomex (marine shallow convection) (Holland & Rasmusson, 1973), TRMM (deep convection) (Grabowski et al., 2006), a dry convective boundary layer (Soares et al., 2004), ARM-SGP (continental shallow convection) (Brown et al., 2002), RICO (precipitating shallow cumulus) (vanZanten et al., 2011), and DYCOMS (drizzling stratocumulus) (Ackerman et al., 2009; Stevens et al., 2003).

Hybrid EDMF

Building on the baseline EDMF-20, two notable modifications have been implemented since to improve the realism and relax assumptions imposed by previous bottom boundary specifications. Firstly, the surface Dirichlet boundary condition

on area fraction, a free parameter found in previous work (Lopez-Gomez et al., 2022) to be correlated with numerous other EDMF parameters, is modified to be a free boundary condition (Appendix A1). The modification allows updrafts to be generated directly by entrainment and detrainment source terms, rather than being "pinned" to the surface, and eliminates the dependence on lower boundary specification of mass flux and area fraction required by most mass-flux schemes. Secondly, the surface Dirichlet boundary condition on TKE in previous versions is replaced by a TKE flux boundary condition that depends on surface conditions and turbulence parameters (Appendix A2).

The key distinction between the hybrid EDMF and EDMF-20 lies in the formulation of data-driven entrainment closures. We consider an EDMF scheme that uses linear regression to determine entrainment rates, designated EDMF-Linreg, and an EDMF scheme that uses a neural network for entrainment rates, designated EDMF-NN. These data-driven closures take the place of the semi-empirical but physically motivated closures implemented in EDMF-20 (Cohen et al., 2020).

Functional Learning for Entrainment and Detrainment

Functional Learning Targets

Entrainment and detrainment are two forms of cloud mixing, which describe the exchange of mass, momentum, and tracers between coherent updrafts and their turbulent environment (de Rooy et al., 2013). Entrainment is the process whereby environmental properties are incorporated into updrafts, whereas detrainment describes the ejection of updraft properties into the environment. Entrainment and detrainment appear as rates (units of s^{-1}) in the EDMF tendency equations. These processes are often decomposed into the sum of turbulent and dynamical contributions, which represent cloud mixing driven by horizontal turbulent mixing from eddies and exchange due to more organized cloud-scale flows, respectively (de Rooy & Siebesma, 2010). The closures learned for this study combine the contributions into a single function. Inputs for data-driven closures are chosen to be nondimensional variables $\mathbf{\Pi}$. For the closure formulation, we adopt the approach of learning a nondimensional function, which modulates a dimensional scale of the same units as the entrainment/detrainment rates:

$$E = \gamma_\epsilon F_\epsilon(\mathbf{\Pi}; \mathbf{\Theta_{ml}}), \tag{3.2a}$$

$$D = \gamma_\delta F_\delta(\mathbf{\Pi}; \mathbf{\Theta_{ml}}). \tag{3.2b}$$

Here, γ_ϵ and γ_δ are inverse time scales while F_ϵ and F_δ are nondimensional functions for entrainment and detrainment, respectively. The data-driven functions F parameterize the relationship between nondimensional groups $\mathbf{\Pi}$ and nondimensional mixing rates, given a vector of learnable parameters $\boldsymbol{\Theta_{ml}}$. We note that F_ϵ and F_δ are vertically local functions, and thus map $\mathbf{\Pi}$ groups defined from local quantities at some level to a single lateral mixing rate at that level. Thus, applying the local function to every level yields a vertical profile of mixing rates that varies with height.

The entrainment dimensional scale is chosen as the ratio of updraft-environment vertical velocity difference $\Delta\overline{w}$ to height z:

$$\gamma_\epsilon(z) = \frac{\Delta\overline{w}}{z}. \tag{3.3a}$$

We denote the difference between subdomains with the symbol Δ. Thus, the difference between the mean updraft and environmental vertical velocity is $\Delta\overline{w} = \overline{w}_{\text{up}} - \overline{w}_{\text{env}}$. The inverse height scaling is chosen here as an easy-to-diagnose proxy of the inverse updraft radius or eddy size at a given height (Siebesma et al., 2007). Thus, γ_ϵ defines a horizontal shear that gives rise to entrainment (Griewank et al., 2022). For detrainment, γ_δ is chosen as a dimensional scale that corresponds to the rate needed to sustain mass flux profiles in steady-state. Taking the EDMF continuity equation (Equation 3.10) as steady and assuming no horizontal convergence or entrainment yields the detrainment expression

$$\gamma_\delta(z) = \frac{1}{\rho a_{\text{up}}}\text{ReLU}\left(-\frac{\partial M}{\partial z}\right). \tag{3.3b}$$

Here, a_{up} is the updraft area fraction, ρ is the air density, and $M = \rho a_{\text{up}}\overline{w}_{\text{up}}$ is the updraft mass flux, where $\overline{w}_{\text{up}}$ is the updraft vertical velocity. ReLU is the rectified linear function, which ensures detrainment only occurs when the mass flux divergence is negative.

Nondimensionalization of Input Variables

A consequential step in designing ML problems is the choice of input variables and their preprocessing, including normalization, transformation, and feature engineering. Effective training of data-driven closures requires inputs of similar magnitude so that disproportionate importance is not assigned to variables with larger magnitudes. The online training approach complicates variable normalization since

the input variables and their associated distributions are strongly dependent on entrainment mixing, and thus will vary as parameters change through the calibration process. A natural and physically motivated approach to transform input variables is to form nondimensional groups by combining dimensional variables in a manner that removes physical units. An additional advantage of doing this is that it increases the likelihood of obtaining climate-invariant closures that generalize well out of distribution (Beucler et al., 2024), in much the same way that Monin-Obukhov similarity theory is fairly generally applicable (Schneider et al., 2024).

In principle, nondimensional functions may depend on any nondimensional groups associated with lateral mixing processes. Here, nondimensional groups are found on the basis of Buckingham's Pi Theorem, which states: given N variables containing M primary dimensions, the nondimensionalized equations relating all the variables will have $(N - M)$ dimensionless groups (Buckingham, 1914). We consider a set $\mathbf{D}$ of $N = 7$ primary variables, containing some already nondimensional quantities, namely, relative humidity (RH) and updraft area fraction (a_{up}), in addition to other variables deemed relevant for SGS turbulence and convection:

$$\mathbf{D} = \left\{ \Delta\overline{b}, \Delta\overline{w}, \overline{\text{TKE}}_{\text{env}}, z, H_{\text{scale}}, \Delta\overline{\text{RH}}, \sqrt{a_{\text{up}}} \right\}. \tag{3.4}$$

The set contains two length scales: the height coordinate z and the standard atmospheric scale height $H_{\text{scale}} = R_d T_{\text{ref}}/g$; $\overline{\text{TKE}}_{\text{env}}$ denotes environmental turbulent kinetic energy. Note that we use $\sqrt{a_{\text{up}}}$ instead of a_{up} because it represents a nondimensionalized length scale. Because entrainment mixing transports properties between subdomains, we defined dimensional variables as differences between the updraft and environmental properties. Using subdomain differences also ensures Galilean invariance, such that the diagnosed entrainment rates are independent of the reference frame. Given that these variables contain $M = 2$ primary dimensions (length and time), this leaves $N - M = 5$ dimensionless groups.

We use the nondimensional $\mathbf{\Pi}$ groups

$$\mathbf{\Pi} = \left\{ \frac{z\Delta\overline{b}}{\Delta\overline{w}^2}, \frac{\overline{\text{TKE}}_{\text{env}}}{\Delta\overline{w}^2}, \sqrt{a_{\text{up}}}, \Delta\overline{\text{RH}}, \frac{gz}{R_d T_{\text{ref}}} \right\}, \tag{3.5}$$

and refer to group i as Π_i. These $\mathbf{\Pi}$ groups, defined locally at each level of the atmosphere, serve as inputs to data-driven models that return continuous, non-negative outputs. Π_1 and Π_2 are unbounded and typically have magnitudes larger than 1, so they are normalized by characteristic values of 10^2 for Π_1 and 2 for Π_2,

such that they typically lie in the range $[-1, 1]$. Π_1 resembles the classic $\Delta\overline{b}/\Delta\overline{w}^2$ scaling introduced by Gregory (2001), and may be interpreted as a proxy for the ratio between updraft buoyancy and the updraft-environment shear. Π_2 is indicative of whether turbulent or convective kinetic energy dominate. Π_3 and Π_4, which are already dimensionless, allow for explicitly learning the dependence of lateral mixing on updraft area and relatively humidity, respectively. Finally, Π_5 serves as an easy-to-compute measure of geometric height, nondimensionalized by the density scale height.

Data-driven Entrainment Architectures

The data-driven models considered for this study are linear regression and a fully-connected neural network. The linear closure is a linear mapping between $\mathbf{\Pi}$ groups and the nondimensional mixing rate. A separate regression model is used for entrainment and detrainment, totaling 12 trainable mixing parameters, including bias terms. Linear regression outputs are passed through a ReLU function to ensure positivity of mixing rates. The fully-connected NN contains 237 parameters with three hidden layers containing 10, 10, and 5 neurons, respectively. Neurons in all layers have ReLU activation functions. We confine ourselves here to relatively shallow network architectures, as they already yielded substantial gains in accuracy of the EDMF scheme; exploration of whether deeper networks can yield additional gains is left for future work.

GCM-driven Simulations

We aim to learn compact representations of directly-simulated, SGS processes as a function of large-scale forcings. Forcings are taken from Cloud Feedback Model Intercomparison Project sites (cfSites), which correspond to locations where high-frequency GCM output is saved for systematically diagnosing cloud feedbacks (Webb et al., 2017). To generate spread in forcings, one model from CMIP6 (CNRM-CM6) and two models from CMIP5 (HadGEM2-A and CNRM-CM5) are used, the latter two representing the upper and lower end of tropical low-cloud reflection response. The LES and EDMF scheme are driven with the same large-scale forcings from the corresponding GCM dynamical core. LES simulations are forced with GCM-prescribed tendencies for large-scale subsidence, horizontal advection, and vertical eddy advection. Additionally, entropy and total water specific humidity profiles are relaxed to the initial background GCM state with a 24 hour relaxation timescale above 3.5 km, where convective and turbulent activity cease. Momentum

profiles are relaxed on a 6 hour timescale throughout the column to prevent drift. Radiation is computed interactively with RRTMG. The EDMF scheme is forced in the same manner, with the exception that radiative cooling tendencies obtained from RRTMG are prescribed from LES. LES simulations are run for 6 days; a steady state response to large-scale forcings is often observed after a couple of simulation days. Single column model simulations are ran for 3 days and more readily reach steady state. For calibration, we consider a total of 176 LES simulations across the east Pacific stratocumulus-to-cumulus transition regions. The setup discussed here is described in Shen et al. (2022).

Ensemble Kalman Inversion

For calibration we employ ensemble Kalman inversion (EKI), an iterative data assimilation technique that blends Bayesian inference with stochastic ensemble sampling to efficiently find optimal parameters (Iglesias et al., 2013; Schillings & Stuart, 2017). Starting with a prior distribution over parameters, the method iteratively updates and narrows the parameter distribution by minimizing the EDMF–LES mismatch without explicitly computing gradients. After a sufficient number of iterations, the spread of the ensemble tightens around the ensemble mean, a phenomenon referred to as ensemble collapse. The method is built into a framework that optimizes EDMF parameters on the basis of LES simulations forced in the same manner. The EDMF calibration framework described here was first introduced in Lopez-Gomez et al. (2022), where further details can be found.

The Kalman update equation estimates parameters iteratively following

$$\boldsymbol{\Theta}_{n+1} = \boldsymbol{\Theta}_n + \mathrm{Cov}\left(\boldsymbol{\Theta}_n, \boldsymbol{\mathcal{G}}_n\right)\left[\mathrm{Cov}\left(\boldsymbol{\mathcal{G}}_n, \boldsymbol{\mathcal{G}}_n\right) + \Delta t^{-1}\boldsymbol{\Gamma}\right]^{-1}\left(\boldsymbol{y} - \boldsymbol{\mathcal{G}}_n\right), \tag{3.6}$$

where $\boldsymbol{\Theta}$ is a vector containing EDMF parameters, $\boldsymbol{\mathcal{G}}$ are EDMF statistics evaluated with parameters $\boldsymbol{\Theta}$, $\boldsymbol{y}$ is a vector of the reference LES statistics, and $\boldsymbol{\Gamma}$ is a noise covariance matrix. Subscripts denote iteration number. The sample covariance matrices $\mathrm{Cov}\left(\boldsymbol{\Theta}_n, \boldsymbol{\mathcal{G}}_n\right)$ and $\mathrm{Cov}\left(\boldsymbol{\mathcal{G}}_n, \boldsymbol{\mathcal{G}}_n\right)$ are computed from the ensemble members, reflecting the covariance between parameters and EDMF statistics, and within the EDMF statistics themselves. The artificial timestep is denoted Δt, and represents an EKI hyperparameter analogous to the learning rate in the gradient descent algorithm. The quantities $\boldsymbol{\Gamma}$, $\boldsymbol{y}$, $\boldsymbol{\mathcal{G}}$, and $\mathrm{Cov}\left(\boldsymbol{\mathcal{G}}_n, \boldsymbol{\mathcal{G}}_n\right)$ are formed by concatenating operations over all cases in a given iteration. Statistics in $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{y}$ are computed with the following sequence of operations for each LES configuration. First, state variables are individually normalized by their respective time-variance over the simulation

period. A time-mean is then computed over the final 12 simulation hours before a low-dimensional encoding that preserves 99% of the variance is applied through principal component projection. The projection reduces the dimensionality of each case from 401 to 8–40. Finally, the resulting statistics are concatenated over cases to form $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{y}$. The six variables whose statistics appear in the loss function are:

1. $\bar{s}$: entropy
2. $\bar{q}_t$: total water specific humidity
3. $\overline{w's'}$: vertical entropy flux
4. $\overline{w'q_t'}$: vertical total water specific humidity flux
5. $\bar{q}_l$: liquid water specific humidity
6. LWP: Liquid Water Path

The overbar denotes a temporal and horizontal average and primes deviations therefrom. The first five variables are vertical profiles, whereas liquid water path is a vertically integrated quantity. The pooled LES time variance, used to estimate observation noise $\boldsymbol{\Gamma}$, is scaled by 0.1 for the vertical flux and liquid water specific humidity variables. We found that noise estimated from LES time variances over the full simulation results in uncertainty bands that overwhelm important details about the vertical structure of these variables. Stated differently, the temporal variability in LES simulations, used as a proxy for observation noise, likely overestimates the noise relevant for calibration for these variables. The artificial timestep Δt is determined adaptively by a Data Misfit Controller (DMC) learning rate scheduler, and generally increases with iteration number (Iglesias & Yang, 2021). The DMC scheduler has no hyperparameters, as timestep is computed as a function of observation noise, data misfit, and integrated timestep. The calibrations are terminated after a specified number of iterations, which are quantified below.

In the Kalman update equation, parameters encoding functional relationships of lateral mixing are denoted $\boldsymbol{\Theta}_{ml}$ (machine learning parameters), and are calibrated alongside parameters $\boldsymbol{\Theta}_p$ appearing in eddy diffusivity and perturbation pressure closures with imposed functional forms, which we denote physical parameters.

$$\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_{ml}\}. \tag{3.7}$$

Many parameter combinations lead to unstable simulations, an issue addressed by sampling from regions of the parameter space with successfully completed simulations. For a given iteration, only the subset of ensemble members with stable simulations are used to approximate the parameter distribution for the subsequent iteration, an approach detailed more in Section 3.1.1 of Lopez-Gomez et al. (2022). Model failure rates are typically 50% - 80% in the initial few iterations and diminish to zero after $\sim$ 10 iterations. To further promote stability and determine robust initial priors, we employ a 2-stage calibration process where the initial phase contains only a subset of the full LES library. The first calibration, which we denote precalibration, is performed on 5 cases using the linear regression closure and 300 ensemble members for 20 iterations. The 5 precalibration cases are representative, and span cloud regimes along the stratocumulus-to-cumulus transition. Priors for the precalibration stage are chosen from Lopez-Gomez et al. (2022) for physical parameters. Linear regression prior means are randomly drawn from a uniform distribution on the interval [0.75, 1.25] with a prior uncertainty of 5. Following this step, the neural network model is independently optimized via gradient descent to reproduce the linear regression mapping learned from EKI in the precalibration stage. For the linear closure, the second phase is initialized directly with prior means from the precalibration phase. The NN calibration is initialized with parameter means learned from gradient descent. The second phase contains all 176 LES cases and a batch size of 16 cases per iteration. Rather than evaluating the full LES dataset in each iteration, 16 cases are drawn from the full dataset without replacement until the entire dataset is processed. A complete pass through the dataset is referred to as an epoch. The final calibrations are run for 50 iterations, or ~4.5 epochs. The need for batching is two-fold: computational efficiency and generation of noise in the training loss. Using the full dataset of 176 cases in each iteration is expensive given the runtime and memory requirements of single model runs. Additionally, variability in the forcing and cloud regimes between batches translates to variability in the evaluated loss and root mean square errors. The noise generated by the batching process inhibits convergence to local minima and is commonly used in data assimilation and machine learning (Houtekamer & Mitchell, 2001).

3.4 Calibration Results

Calibration Characteristics and Performance Comparison

To characterize the EKI training process, we consider the evolution of root mean squared error (rmse) separately for each of the six variables in the loss function,

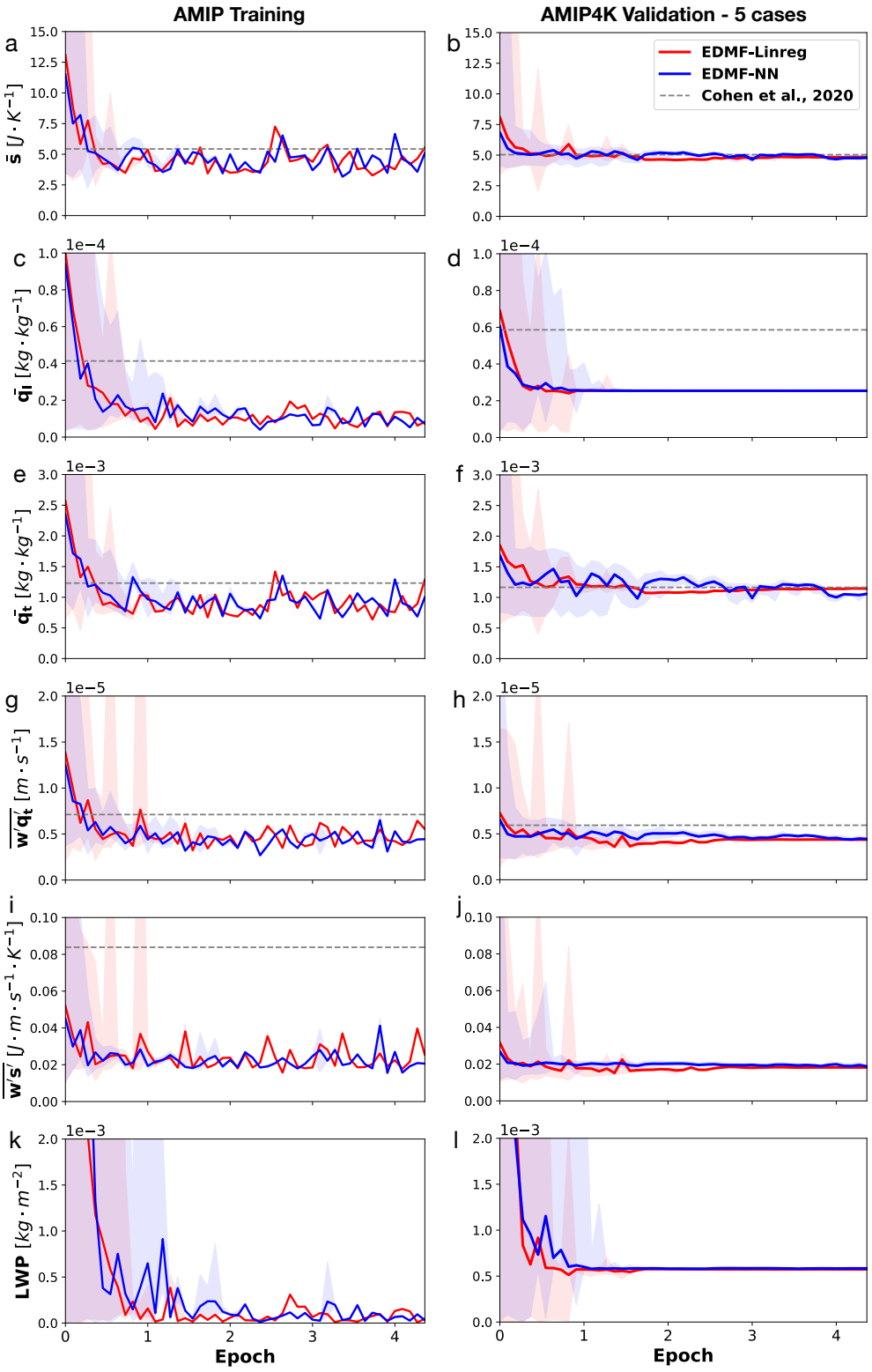

Figure 3.2: Root mean squared error (rmse) by variable for (left) training set from AMIP experiment and (right) validation set with five cases from the AMIP4K experiment. Shaded regions indicate min/max rmse across ensemble members for a given iteration, demonstrating ensemble spread. Dashed horizontal lines indicate baseline simulations from the EDMF-20 version described in Cohen et al. (2020). A summary of rmse comparisons can be found in 3.6.

tracked through the final calibration and following the precalibration step. Figure 3.2 displays the evolution of rmse for the AMIP training set (left column) and a fixed set of 5 LES cases from the AMIP4K climate (right column). The AMIP4K validation cases are a representative set spanning the stratocumulus-to-cumulus transition using HadGEM2-A as the forcing model. Shading indicates the maximum and minimum rmse over ensemble members for a given iteration, as each member is associated with a unique set of parameters. A summary of rmse comparisons between the EDMF variants can be found in 3.6. We note that the training rmse curves are noisier than the validation curves due to the batching processes. During training, the rmse for a given iteration is calculated for the 16 sampled LES cases that vary in location, season, and regime iteration-to-iteration. The validation set is intended to track generalization performance though the calibration process.

The rmse evolution represents an improvement over the precalibration posterior (full calibration prior), constrained initially by the 5 precalibration cases in the AMIP climate. Variables with larger rmse differences between the initial and final iterations benefit more from additional cases from the full AMIP training set, and vice versa. The largest differences are for $\bar{q}_l$ and LWP, where error decreases by an order of magnitude, consistent with the sensitive and multi-scale dynamics needed to simulate cloud variables with fidelity. We note that LWP is the density weighted integral of $\bar{q}_l$, so the rmse values are correlated. Remaining variables, including state variables ($\bar{s}$, $\bar{q}_t$) and flux variables ($\overline{w's'}$, $\overline{w'q_t'}$), demonstrate rmse improvements of roughly 50 – 75% with respect to the prior. The differences in rmse improvement may stem from observation noise differences, but these are scaled to have roughly comparable relative magnitudes, such that they hold similar weight with respect to each other in the loss. This analysis reveals that the accuracy in simulating cloud properties, through parameters that constrain $\bar{q}_l$, is greatly improved by expanding the number of training cases from 5 to 176.

Significant improvements of the hybrid EDMF over EDMF-20 are observed, particularly for cloud-related variables and $\overline{w's'}$. Coplotted are variable-by-variable rmse baselines evaluated with EDMF-20 over the entire AMIP dataset for the training plots and the 5 AMIP4K cases in the validation plots. The most significant improvements of the hybrid EDMF over EDMF-20 are observed for $\bar{q}_l$, LWP, and $\overline{w's'}$. The sizable reduction of entropy flux error likely stems from the modified boundary conditions and larger entrainment rates learned near the surface. Earlier assessments of EDMF-20 demonstrated integrated entropy fluxes that were systematically biased

too large, even after calibration (Lopez-Gomez, 2023). Overly warm and buoyant updrafts in EDMF-20 are likely contributors to the systematically large entropy fluxes. The updraft warm bias has been largely mitigated in the hybrid EDMF, coincident with enhanced surface entrainment that mixes cooler environmental air into the updraft and larger TKE at the surface. Less consequential improvements are identified for state variables $\bar{q}_t$ and $\bar{s}$. In the validation curves, greater differences are observed between the hybrid EDMF schemes and EDMF-20, owing to data-driven closures, structural model improvements, and the larger training dataset.

The comparable performance of EDMF-NN and EDMF-Linreg in training and validation metrics has several potential explanations. Differences in the learned entrainment functions are detailed further in section 3.3. While the NN is pretrained on the linear regression model, significant prior uncertainty is introduced in the NN weights to ensure large regions of parameter space are explored beyond the linear, low-dimensional manifold. Further, given the physical structure surrounding the data-driven mixing closures, including the dimensional scale multipliers and derivation of $\mathbf{\Pi}$ groups for input, expressive and non-linear ML architectures do not appear necessary for learning the optimal mapping. The success of simple nondimensional functions may also be a consequence of simplifications made in the setup. A limitation of the training data is the use of steady large-scale forcings and LES-prescribed radiation tendencies. These preclude the simulation of high-frequency climate variability, such as the diurnal cycle of precipitation and clouds, which is more sensitive to details of entrainment (Del Genio & Wu, 2010). Nonsteady forcings with interactive radiation and deep convection cases may be needed to gain predictive benefits from more expressive mixing closures. A final contributing factor, discussed in section 3.4, is the presence of remaining structural errors in the EDMF formulation itself, which may not be rectified through modifying the cloud mixing process.

Generalization Performance in AMIP4K Climates

The full library of LES simulations is divided into a training and validation set on the basis of the forcing climate; the hybrid EDMF is calibrated on 176 present-day AMIP simulations and performance is evaluated on simulations from a warmer AMIP4K climate. The AMIP4K climate contains out-of-distribution large-scale forcings and surface heat fluxes. Five AMIP4K cases are chosen to track extrapolation performance through the calibration process, illustrated in the right column of Figure 3.2. For the chosen AMIP4K validation set, consequential performance

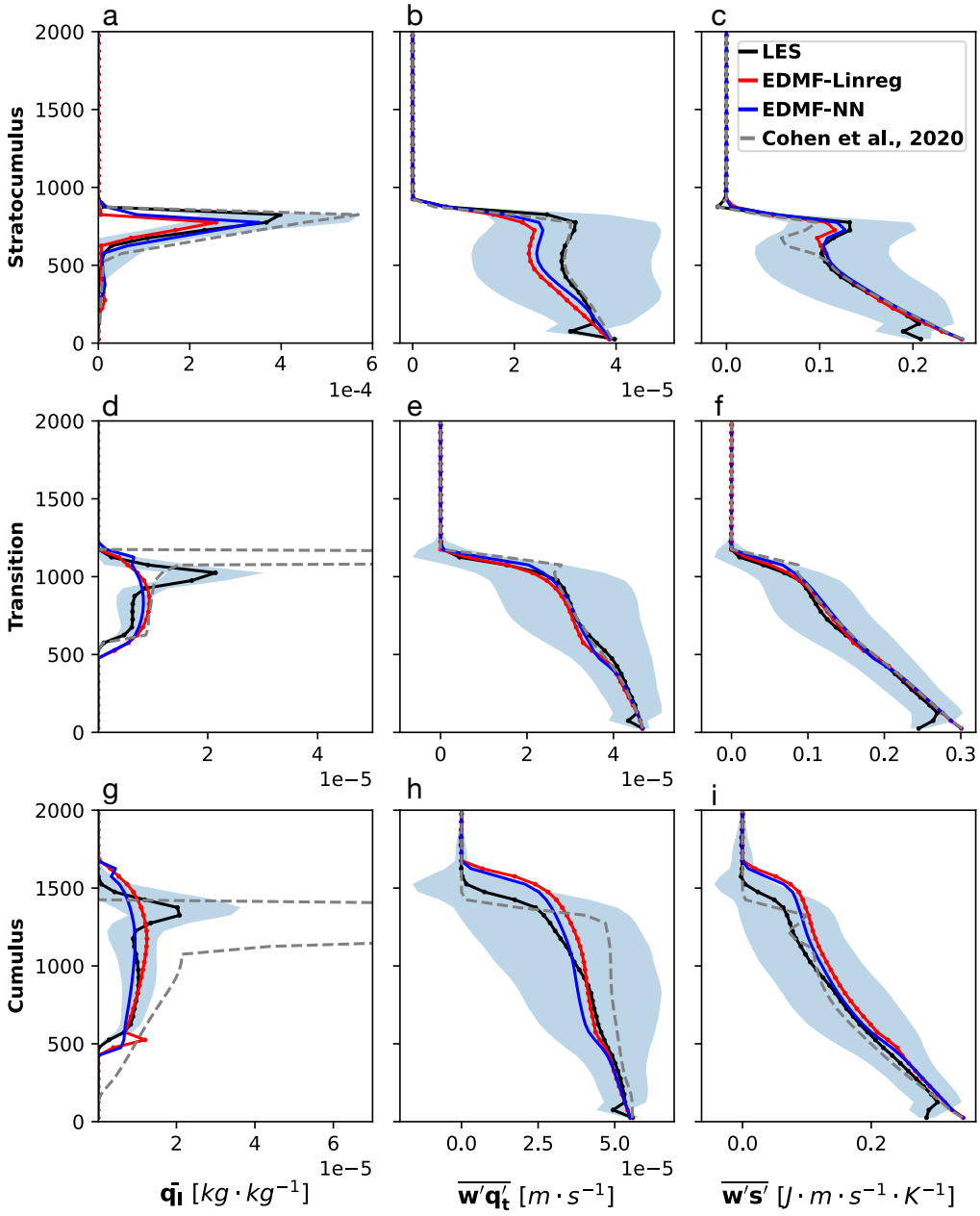

Figure 3.3: AMIP4K, time-mean vertical profiles of liquid water specific humidity ($\bar{q}_l$, left), total water specific humidity flux ($\overline{w'q'_t}$, middle), and entropy flux ($\overline{w's'}$, right) from hybrid EDMF models across a sampling of climate models, seasons, geographic locations, and cloud regimes. Top row: stratocumulus case (cfSite17) in July forced with CNRM-CM5; middle row: transition case (cfSite6) in April forced with CNRM-CM6; bottom row: cumulus case (cfSite22) in July forced with HadGEM2-A. Baseline simulations from Cohen et al. (2020) are plotted in gray dashed lines. Large-eddy simulation (LES) time-mean profiles from Shen et al. (2022) are plotted in black. Calibrated EDMF simulations using a linear regression-based mixing closure (EDMF-Linreg) are depicted in red, while those with a NN-based mixing closure (EDMF-NN) are shown in blue. Light blue shading indicates the 2σ time variance, by level, from LES simulations.

improvements diminish after ~ 1 epoch, consistent with the training rmse. Validation rmse is noted to roughly track training rmse, with rmse for cloud-related variables $\bar{q}_l$ and LWP containing larger extrapolation errors of $2.54 \times 10^{-5}\,\mathrm{kg} \cdot \mathrm{kg}^{-1}$ and $5.84 \times 10^{-4}\,\mathrm{kg} \cdot \mathrm{m}^{-2}$ for EDMF-Linreg, respectively. Nevertheless, it is found that the validation set does not enter the overfitting regime, which is characterized by a u-shaped validation curve.

Robust extrapolation performance is noted in data space as well, where key features learned in training are persistent in a simulated warmer climate. Figure 3.3 depicts a sampling of profiles from the AMIP4K climate across climate models, seasons, location, and cloud regimes. Optimal parameters are chosen from the ensemble member nearest to the ensemble mean at the end of the final training epoch, as the mean itself is not directly evaluated. For a given cfSite, the AMIP4K LES simulations feature changes in boundary layer depth, cloud water content, cloud depth, and vertical fluxes in response to larger surface heat fluxes and changes in local forgings due to large-scale circulation responses. Given these changes, we find hybrid EDMF simulations, trained in a cooler climate, capture these characteristics well. EDMF-20 is noted to have a large bias in $\bar{q}_l$ near the cloud top, particularly for cumulus and transition cases. Remaining biases observed in these profiles are detailed in section 3.4.

Learned Entrainment and Detrainment Profiles

This section turns to the assessment of learned entrainment profiles following the calibration procedure outlined above. To reiterate, the precalibration data-driven cloud mixing priors are initialized with random numbers, and closure learning is indirectly guided by the time-mean profiles alone. Focus is placed on cumulus cases, where cloud mixing is most relevant for determining the formation and behavior of clouds reliant on updraft dynamics. Figure 3.4 illustrates time-mean vertical profiles of the $\mathbf{\Pi}$ groups (left), nondimensional entrainment rates (middle), and total entrainment rates (right). Nonzero liquid water specific humidity ($\bar{q}_l$) is shaded in gray to highlight the cloud layer. The optimal parameters are chosen from the ensemble member nearest to the ensemble mean at the end of the final training epoch, as in Figure 3.3. The first observation to emphasize is the realism of calibrated simulations on the basis of nondimensional input groups (Figure 3.4a, d). Both EDMF-Linreg and EDMF-NN exhibit canonical characteristics of shallow convection. Notably, updraft area (Π_3) begins to shrink considerably above the cloud base due to net detrainment of mass into the environment. Near the cloud

top, the updraft-environment relative humidity difference (Π_4) intensifies, where buoyant and saturated updrafts begin to penetrate into the dry, stable inversion layer. Additionally, the sub-cloud boundary layer is dominated by mixing from turbulent eddies, while the cloud layer is dominated by updraft dynamics, as indicated by the ratio of TKE to vertical velocity squared (Π_2).

The learned cloud mixing profiles themselves further demonstrate realistic and physically robust characteristics, consistent with theory surrounding lateral cloud mixing for shallow convection. Several well-established qualities of entrainment and detrainment in shallow convection include (de Rooy et al., 2013):

- A local maximum of entrainment where updrafts form;
- Net detrainment ($E - D < 0$) through much of the cloud layer;
- Strong detrainment near the cloud top, in the vicinity of a capping inversion layer.

These are consistent with theoretical work and diagnostics of lateral mixing in LES (Savre, 2022).

These key characteristics are observed in lateral mixing profiles (Figure 3.4c, f) for both EDMF-Linreg and EDMF-NN. Many SGS parameterizations feature distinct turbulent surface layer and mass-flux schemes, with the latter typically prescribing a boundary condition closure for the cloud base mass flux. Consequently, this configuration precludes both entrainment below the cloud base and strong entrainment at the cloud base. Because the EDMF scheme employed for this study is unified, updrafts may be either saturated or dry, and extended from the surface where they are generated by strong net entrainment. Coincident with near-surface updraft formation, large entrainment rates are observed in Figure 3.4c, f. Both closures accurately predict net detrainment above the cloud base, where entrainment rates tend to small values and detrainment grows. Finally, a global maximum in detrainment rate is observed near the cloud top.

Several core similarities and differences are discussed for the linear and NN-based entrainment closures on the basis of nondimensional rates, or the components targeted with data-driven closures. The nondimensional functions may be viewed as a multiplicative modulations of dimensional rates introduced in Eqs. 3.3a, 3.3b. Deviations far from unity suggest that the dimensional mixing rate does not accurately

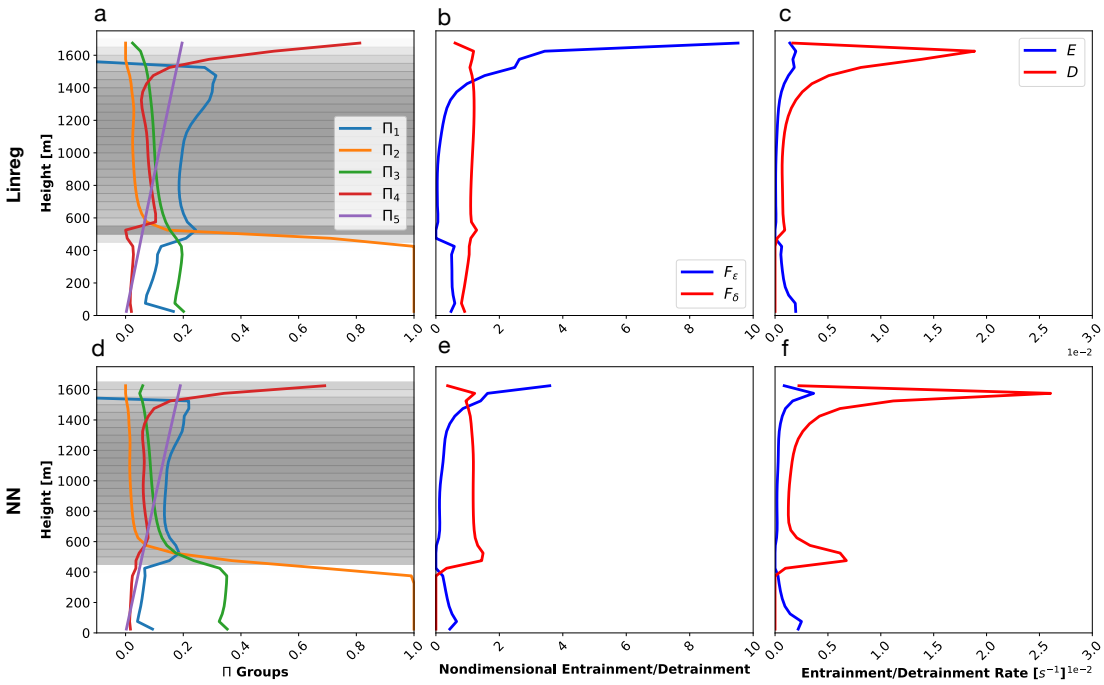

Figure 3.4: Time-mean vertical profiles of lateral mixing variables for cfSite22 with AMIP4K forcings, depicting shallow convection near Hawaii in July. a,d): Nondimensional **Π** groups, with liquid water specific humidity ($\bar{q}_l$) shaded in gray. b,e): nondimensional entrainment and detrainment (data-driven model output). c,f): Total entrainment and detrainment rates.

capture dynamics consistent with LES time-mean profiles. In contrast, nondimensional rates close to unity indicate that the dimensional component effectively approximates cloud mixing without need for modification. Turning to the nondimensional rates (Figure 3.4b, e), we note more consequential differences between the hybrid EDMF schemes in the detrainment rates. Notably, EDMF-NN features a secondary maximum of detrainment near the cloud base, around $\sim$ 500 m above the surface. Such secondary local detrainment maxima are often observed in LES-diagnosed detrainment rates (Romps, 2010). Generally larger detrainment rates are also observed for EDMF-NN through the cloud layer. Alternatively, EDMF-Linreg maintains a less variable nondimensional rate with height, with slight enhancement in the updraft. Focusing on nondimensional entrainment, we find stronger modulation of the dimensional scale than for detrainment. In particular, both closures demonstrate increasing modulation of the dimensional scale with height in the upper cloud levels. This indicates the $\Delta\overline{w}/z$ dimensional scale significantly underpredicts entrainment rates near the updraft top. The behavior driving this learned enhance-

ment may surround the physical mechanisms governing cessation of updrafts, where updraft area fraction or mass flux tend to zero. Updrafts vanish by a combination of strong detrainment, which serves as a sink for area fraction, and entrainment, which diminishes upward mass flux by both reducing updraft buoyancy and entraining environmental parcels with negligible vertical momentum. Despite the two competing effects, studies point to strong net detrainment at the cloud top, as alluded to previously, which is consistent with our simulations. In the sub-cloud layer, the dimensional scale overpredicts entrainment, as indicated by nondimensional values less than unity in both schemes.

The closed-form linear expression for entrainment following the full calibration is

$$E = \frac{\Delta\overline{w}}{z} \times 6\Big[-0.05 + 0.8\left(\frac{z\Delta\overline{b}}{\Delta\overline{w}^2}\right) + 0.6\left(\frac{\overline{\mathrm{TKE}}_{\mathrm{env}}}{\Delta\overline{w}^2}\right) + \\ -3\sqrt{a_{\mathrm{up}}} + 3\left(\Delta\overline{\mathrm{RH}}\right) + 0.2\left(\frac{gz}{R_d T_{\mathrm{ref}}}\right)\Big], \tag{3.8}$$

and that for detrainment is

$$D = \frac{1}{\rho a_u} ReLU(-\frac{\partial M}{\partial z}) \times 8\Big[0.04 - 0.07\left(\frac{z\Delta\overline{b}}{\Delta\overline{w}^2}\right) - 0.07\left(\frac{\overline{\mathrm{TKE}}_{\mathrm{env}}}{\Delta\overline{w}^2}\right) + \\ 0.8\sqrt{a_{\mathrm{up}}} - 0.2\left(\Delta\overline{\mathrm{RH}}\right) + 0.5\left(\frac{gz}{R_d T_{\mathrm{ref}}}\right)\Big]. \tag{3.9}$$

These are determined from the ensemble member nearest to the mean in the final training epoch. These functional relationships may be used to understand the vertical structure of nondimensional mixing in the context of Figure 3.4. In the sub-cloud surface layer, where a local entrainment maximum is observed (Figure 3.4c, f), the linear model has strong contributions from Π_2 as a consequence of large TKE. Above the surface layer, the increase of nondimensional entrainment with height has large contributions from gradually decreasing area fraction (Π_3) through the cloud layer and sharply increasing updraft-environment relative humidity difference (Π_4) near the cloud top (Figure 3.4a, d). The linear nondimensional detrainment rates demonstrate weaker variation with height. The $\mathbf{\Pi}$ groups themselves contain covariances, so variable importance cannot not be read off explicitly from Eq. 3.8 and Eq. 3.9. Because the full calibration is initialized with parameter means from precalibration, differences in the final parameter values indicate sensitivity to number of training cases, particularly when going from 5 to 176 cases. We find the training

data sensitivity to be parameter-dependent. The entrainment weights for Π_3 and Π_4, in particular, demonstrate the most sensitivity. For entrainment, the full calibration modified the Π_3 weight by a factor of ~ 2 and the Π_4 weight by a factor of ~ 3. Alternatively, the detrainment parameters for Π_3 and the bias have little sensitivity beyond 5 cases, and are modified by $< 10\%$ in the full calibration. Remaining parameters exhibit intermediate sensitivities. In 3.6, we provide a comparison of the final linear regression weights following each experiments, as well as precalibration uncertainty estimates from the Calibrate, Emulate, Sample framework (Cleary et al., 2021).

Beyond Calibration: Addressing Structural Errors

Post-calibration, persisting discrepancies between the LES and EDMF may be attributed to three primary contributions: the EKI optimizer, the inverse problem setup, and inherent biases in the underlying physical forward model or data, in this case, the structure and assumptions of the EDMF scheme. The performance of the EKI optimizer, as determined by its convergence, may be sensitive to EKI settings and hyperparameters. Among the most consequential choices are the EKI artificial timestepper and the batch size. Sensitivity to constant artificial timestep values in previous work (Lopez-Gomez et al., 2022) is addressed here by using a hyperparameter-free adaptive timestep (DMC) that increases through the calibration process. For batching, we chose the largest batch size feasible given computational limitations. It is found that batch sizes smaller than ~ 10 generate excessive noise in the loss, preventing descent of the ensemble mean to lower values and convergence of the EKI algorithm. Additional biases may persist as a result of the problem setup, such as the input variables selected for data-driven closures and the choice of priors. In addition to addressing instabilities, the precalibration procedure reduces sensitivities to the priors. Precalibration is initialized with large prior uncertainties over parameters with a relatively large number of ensemble members (300), allowing broad exploration of the parameter space and narrowing of the posterior on the basis of a small but representative dataset. While these approaches curtail EDMF-LES discrepancies and mitigate convergence to local minima, it is possible that more advanced strategies are needed to initialize, pretrain, and calibrate the NN-based EDMF. Attempts to initiate the EDMF-NN calibrations directly with Xavier initialization (Glorot & Bengio, 2010) produced EKI calibrations that exhibited high ensemble failure rates and minimal convergence of the loss function across a range of prior uncertainties.

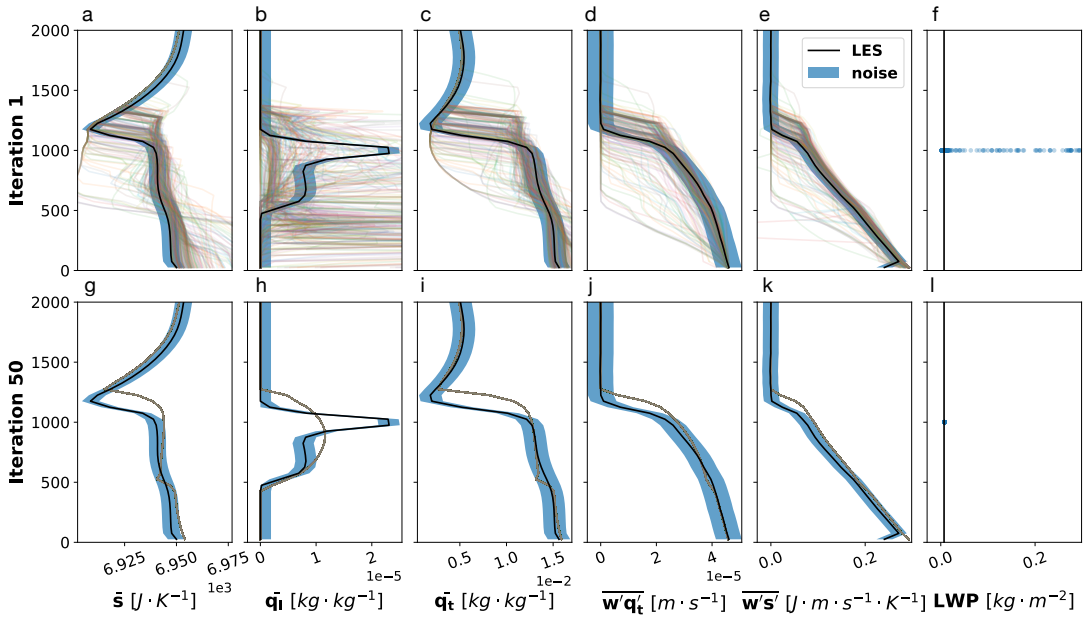

Figure 3.5: Ensemble spread of EDMF-Linreg for all loss function variables in (top) first iteration and (bottom) final iteration. Large-eddy simulation (LES) time-mean profiles are plotted in black (Shen et al., 2022), and each colored lines represents the evaluation from an ensemble member. Blue shading indicates the 2σ observation noise used by EKI, calculated from the pooled variance across levels in LES simulations.

Structural error denotes errors arising from the design of the EDMF scheme itself, including but not limited to the formulation of other closures, boundary conditions, and assumptions made in deriving the EDMF equations. Such limitations may not be corrected by calibration, but must be addressed by modifying the anatomy of the EDMF scheme or adding structural error models within the governing EDMF equations. Relative to Lopez-Gomez et al. (2022), this study addressed three structural errors by modifying the EDMF equations and boundary conditions:

1. A strong warm bias near the surface, resulting from a TKE minimum in the bottom cell center, addressed by implementing a bottom flux boundary condition for the TKE equation;

2. Calibrations with near-zero entrainment throughout the vertical profile, addressed by implementing a free boundary condition on updraft area in the bottom cell center;

3. Divergence of area fraction to values close to 1, addressed by choosing a

dimensional scale for detrainment that ensures area fraction gradually tends to zero when the mass flux gradient is negative.

These modifications led to both improved training and validation errors as well as more realistic cloud mixing profiles following calibration.

Remaining structural errors primarily involve biases in the depth of the mixed layer and cloud-top $\bar{q}_l$ maxima. First, we note an underestimation of capping stratocumulus clouds in stratocumulus-topped cumulus forcing regimes, as demonstrated by $\bar{q}_l$ profiles in the Figure 3.3d and Figure 3.5h. While relatively low $\bar{q}_l$ errors are observed for layers composed of cumulus clouds in these regimes, below roughly 1000 m in Figure 3.3d and 800 m in Figure 3.5h, the grid-mean $\bar{q}_l$ is biased systematically low at cloud tops. Transition cases demonstrating this bias contain saturated updrafts in the cloud layer, but fail to saturate the environment at the level stratocumulus clouds are observed in LES simulations. Because stratocumulus dynamics are dominated by environmental mixing, rather than updraft dynamics, this likely indicates a bias in the TKE equations or other environmental factors. This hypothesis is further supported by the initial spread of $\bar{q}_l$ profiles across ensemble members in data space, illustrated in Figure 3.5b. The initial iteration contains sizeable spread in parameter values, consistent with the prior, and is indicative of the data space subsequent iterations will explore. Characteristics, such as capping stratocumulus clouds, not loosely demonstrated by ensemble members during the initial iterations are unlikely to be developed in later iterations, implying a systematic bias in the model or prior means that are far removed from the optimal solution for a given case. We found the bias to be persistent across many calibration in offline experiments varying the precalibration set and EKI settings. The bias is further demonstrated by systematic collapse of ensembles in the final iteration far beyond the envelope of observation noise (Figure 3.5h). Cloud top maxima of $\bar{q}_l$ are also observed for LES simulations of pure shallow convection, but these features may be an artifact of microphysics in LES simulations. Anvil-like structures in the LES shallow convection cases are coincident with vertical maxima of cloud fraction, and may not be desirable to fit to.

Secondly, we note a bias in mixed layer depth for some cases, resulting in biases across variables near the cloud top. This is evident in the shallow cumulus case illustrated in Figure 3.3, where the mixed layer becomes $\sim$ 100 m too deep, as evidenced by the vertical fluxes in panels h, i. As a consequence, the cloud also develops too deeply (Figure 3.3g). While most cases capture the depth of the

mixed layer with high fidelity, cases with the most prominent bias in cloud-top stratocumulus structures tend to coincide with a bias in the mixed layer depth. Remaining structural errors may be rectified in future work by replacing additional closures with data-driven models or learning structural error models as additional additive terms that modify EDMF tendency equations (Wu et al., 2024). With the latter strategy, care must be taken to ensure conservation of mass, momentum, and energy. Given biases in the depth of the mixed layer and cloud top stratocumulus structures in transition cases, we believe adding data-driven closures or error models to the TKE equation would help address these issues.

3.5 Concluding Remarks

In this study, our aim was to develop realistic hybrid SGS models that combine generalizability with interpretability, targeting the challenging Pacific stratocumulus-to-cumulus transition—a region notorious for being particularly error-prone in state-of-the-art climate models. The primary contribution of this paper is the demonstration of online learning of a 1D hybrid model in realistic climate settings, a step needed to eventually apply such methods in operational GCMs. Application in realistic setups may require pretraining more expressive data-driven components (NNs) to obtain sensible priors, failure handling mechanisms to address numerically unstable simulations in the training process, and procedures or guidelines for identifying remaining structural biases. Development of hybrid models benefits from a bidirectional workflow, where online learning is informative about where structural model biases might lie, and calibrations of data-driven components help improve the predictive power of hybrid models. Finally, and critical in the development of hybrid SGS models, is the assessment of physical validity alongside predictive power. Success of the hybrid EDMF is particularly evident in the realism of cloud mixing closures, which were learned indirectly from extensive LES data with no direct prior information about entrainment and detrainment. The learned closures align closely with existing theoretical understanding and LES-diagnosed characteristics of lateral cloud mixing as it relates to convective and cloud dynamics, reinforcing the model's scientific validity. Furthermore, our results highlight the hybrid model's predictive power, with substantial improvements over a baseline EDMF tuned to match field campaigns. We observe that performance improvements translate to an out-of-distribution AMIP4K climate, as assessed by rmse and qualitative analysis of physical profiles. This generalizability is crucial for the model's application to prediction of future climate scenarios in GCMs.

The online learning approach for hybrid modeling presents several advantages over offline, fully-data driven alternatives. The EKI framework allows for indirectly training SGS model components on the basis of observable statistics or quantities appropriate for long-term climate model projections. While the study focused on high-resolution simulations for training, this may be extended to include sparse observations in the loss function. Numerical instabilities resulting from unstable parameter combinations are directly addressed in the training process, reducing the likelihood of instabilities when the parameterization is incorporated in operational GCMs. Additionally, data-driven components of a hybrid model can be more easily isolated and reasoned about, giving stronger confidence in out-of-distribution predictions of future climate states and promoting physical process understanding.

Despite these promising developments, there are remaining avenues for improving the hybrid EDMF scheme. The paper highlights that the reliance on steady large-scale forcings and prescribed radiation tendencies in the training data limits the ability to learn phenomena important for capturing high-frequency climate variability, such as the diurnal cycle. Additional datasets of high-resolution simulations, such as those introduced by Chammas et al. (2023) and Yu et al. (2023), would likely improve performance over a broader range of forcings and atmospheric regimes. Additionally, some errors in the structure of the model persist after calibration, resulting in a form of underfitting. Remaining structural errors may be remedied in future work by replacing additional closures with expressive, data-driven components or learning structural error corrections as additional additive terms that modify EDMF tendency equations. One avenue is to target closures in the environmental TKE equation, as the data-driven lateral mixing closures presented here primarily affect updraft characteristics and mass flux. Because our EDMF scheme uses a 1.5-order Mellor-Yamada turbulence closure, a natural target is the mixing length function, which determines environmental turbulent diffusivity and viscosity (Mellor & Yamada, 1982). Future work should focus on these aspects, in addition to more expansive training datasets, to ensure that the hybrid modeling approach can be effectively applied in operational Earth system models.

3.6 Appendix

Hybrid EDMF Bottom Boundary Conditions

Updraft Area

The inhomogeneous Dirichlet boundary condition on area in EDMF-20 is replaced by a free boundary condition, where updraft area is generated directly by entrainment

and detrainment source terms at the bottom boundary. Because area is a prognostic variable in the EDMF equations, choices must be made about how the boundary conditions are specified. The EDMF continuity equation for a single updraft reads

$$\frac{\partial(\rho a_{\text{up}})}{\partial t} = -\nabla_h \cdot (\rho a_{\text{up}} \langle u_h \rangle) - \frac{\partial(\rho a_{\text{up}} \bar{w}_{\text{up}})}{\partial z} + \rho a_{\text{up}}(E - D) \tag{3.10}$$

where $\langle u_h \rangle$ is the average grid-scale horizontal velocity, ∇_h is the horizontal divergence, a_{up} is the updraft area fraction, $\bar{w}_{\text{up}}$ is the updraft vertical velocity, ρ is the air density, and E and D are entrainment and detrainment, respectively.

The bottom area fraction was previously specified as an EDMF parameter a_s, typically chosen as 0.1, which remained fixed in all simulations (Cohen et al., 2020; Lopez-Gomez et al., 2022; Tan et al., 2018). The Dirichlet boundary condition on area was defined as

$$\rho a(z_0) = \rho a_s \tag{3.11}$$

where z_0 is the height of the interior point adjacent to the bottom boundary. Removing the surface area parameter and allowing for a free boundary condition permits the generation of surface-based updrafts directly from source terms. The modification allows updrafts to be generated by net entrainment ($E - D > 0$) or grid-scale horizontal convergence near the surface, and thus vary with environmental conditions.

Turbulent Kinetic Energy

We substitute the TKE Dirichlet boundary condition in EDMF-20 by a flux boundary condition at the bottom boundary. The Dirichlet boundary condition was formulated as

$$\overline{\text{TKE}}_{\text{env}}(z_0) = \kappa_\star^2 u_\star^2 \tag{3.12}$$

where $\overline{\text{TKE}}_{\text{env}}$ represents the environmental TKE, $\kappa_\star$ is the ratio of rms turbulent velocity to the friction velocity (an EDMF parameter), $u_\star$ is the friction velocity, and z_0 is the height of the interior point adjacent to the boundary.

We replaced this formulation by a flux boundary condition on the TKE flux at the bottom boundary. To obtain the flux boundary condition, the following simplifying assumptions are made:

1. The mixing length in the surface layer is limited by the distance to the boundary.

2. Storage and mean advection of $\overline{\text{TKE}}_{\text{env}}$ are neglected. This is a good approximation in the surface layer, where TKE is roughly constant.

3. Horizontal derivatives are small compared to the vertical derivatives close to the boundary (the boundary layer approximation).

4. The velocity-pressure gradient correlation term can be neglected. This assumption is consistent with the impenetrability condition for the subdomains and the closure for perturbation pressure in the EDMF model.

These approximations lead to the flux-gradient relation at the surface

$$\rho a_{\text{env}} \overline{w_0' \text{TKE}'}_{\text{env}} \Big|_{z_0} = \rho a_{\text{env}} \left(1 - c_d c_m \kappa_\star^4\right) u_\star^2 \left\| u_{p,\text{int}} \right\|, \tag{3.13}$$

where a_{env} is the environmental area fraction, $u_{p,\text{int}}$ is the near-surface velocity component parallel to the surface, c_d is the turbulent dissipation coefficient, and c_m is the eddy viscosity coefficient (Lopez-Gomez et al., 2022). The modification allows the surface TKE to vary more strongly with environmental conditions.

RMSE Tables

EDMF Version - AMIP	$\bar{\mathbf{s}}$	$\bar{\mathbf{q}}_\mathbf{l}$	$\bar{\mathbf{q}}_\mathbf{t}$	$\overline{\mathbf{w}'\mathbf{q}_\mathbf{t}'}$	$\overline{\mathbf{w}'\mathbf{s}'}$	**LWP**
EDMF-NN	5.55	8.26e-06	1.29e-03	5.54e-06	2.54e-02	4.72e-05
EDMF-Linreg	5.10	7.25e-06	1.00e-03	4.45e-06	2.06e-02	3.14e-05
Cohen et al., 2020	5.43	4.13e-05	1.23e-03	7.12e-06	8.38e-02	1.79e-01

Table 3.1: Table of root mean squared errors for EDMF variants. Reported rmse values for EDMF-NN and EDMF-Linreg are the ensemble-averaged rmse in the final iteration.

EDMF Version - AMIP4K	$\bar{\mathbf{s}}$	$\bar{\mathbf{q}}_\mathbf{l}$	$\bar{\mathbf{q}}_\mathbf{t}$	$\overline{\mathbf{w}'\mathbf{q}_\mathbf{t}'}$	$\overline{\mathbf{w}'\mathbf{s}'}$	**LWP**
EDMF-NN	4.84	2.54e-05	1.14e-03	4.37e-06	1.82e-02	5.73e-04
EDMF-Linreg	4.78	2.54e-05	1.06e-03	4.44e-06	1.88e-02	5.84e-04
Cohen et al., 2020	5.03	5.86e-05	1.16e-03	5.93e-06	7.93e-01	2.13e-01

Table 3.2: Root mean squared errors for EDMF variants on AMIP4K validation set.

Parameter Sensitivities

This analysis compares linear regression parameters for the 5-case precalibration (precal) and 176-case full calibration (full cal), including precalibration uncertainty estimates using the Calibrate, Emulate, Sample (CES) framework (Cleary et al., 2021). The precal prior and posterior distributions help contextualize shifts in the final weights between experiments. Figure 3.6 shows the precal prior and posterior distributions, coplotted with the final parameter values for each experiment. The linear weights multiplying Π_i are labeled C_i^ϵ for entrainment and C_i^δ for detrainment. The corresponding bias terms are labeled bias^ϵ and bias^δ for entrainment and detrainment, respectively. We find the training data sensitivity to be parameter-dependent, as indicated by varying degrees of modification to the final parameter values between experiments. In precalibration, the emulation step consists of training Gaussian processes containing a radial basis function kernel on parameter-to-data pairs from the ensemble. We train the emulator on the first iteration with a failure rate below 50% (iteration 4), and include iterations 8 and 16 to better emulate regions of the parameter space where the ensemble is converging. The sample step probes uncertainty via Markov Chain Monte Carlo (MCMC) sampling of the parameter space around the precal final mean parameter values. We use 100,000 samples for MCMC, with the first 2,000 samples discarded as burn-in to ensure the chain reaches equilibrium and mitigate the impact of initialization bias.

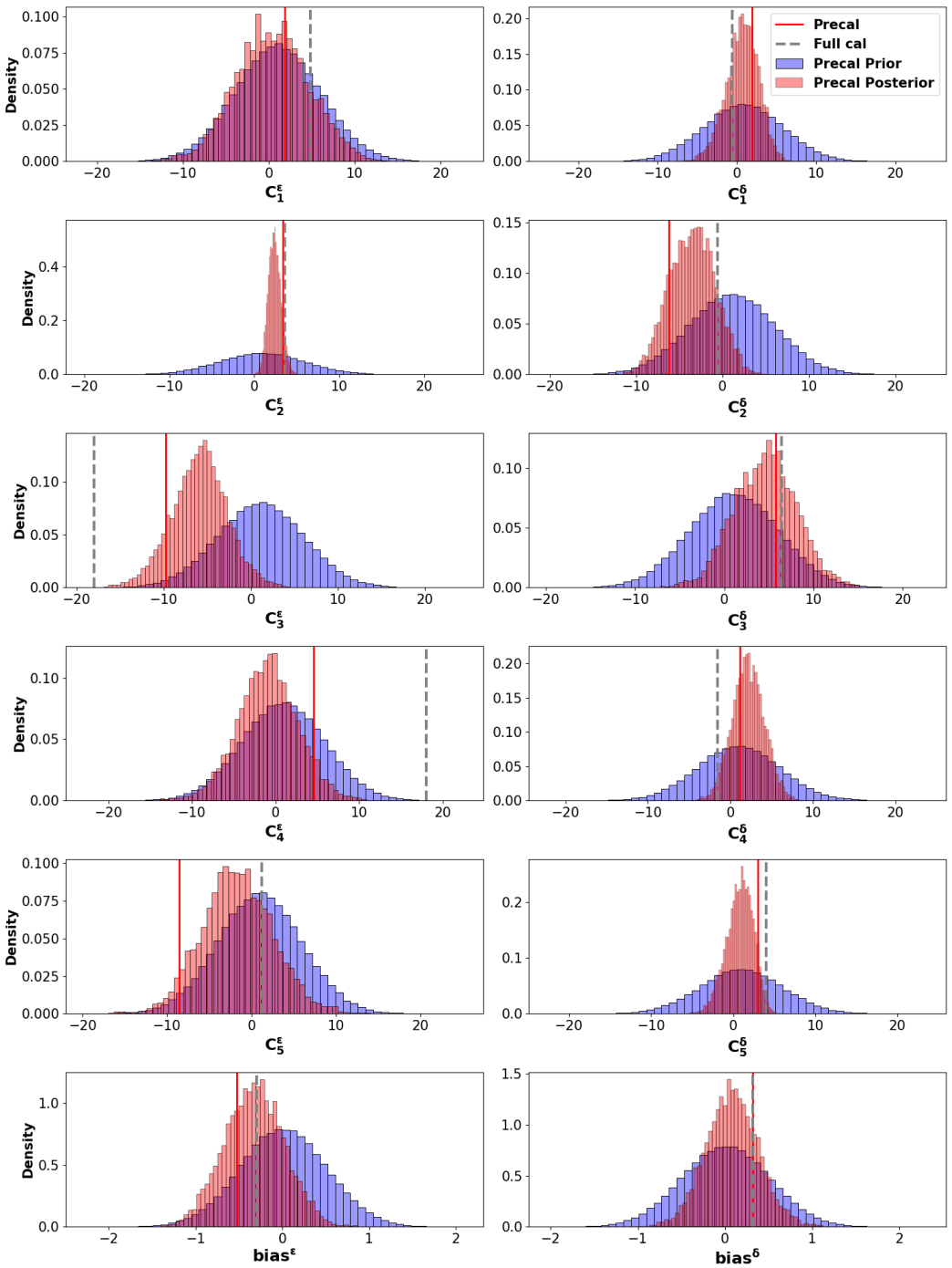

Figure 3.6: Prior and posterior parameter uncertainty estimated by Calibrate, Emulate, Sample (Cleary et al., 2021) for the 5-case precalibration. Blue distributions indicate the prior and red distributions indicate the posterior. Vertical lines mark final parameter values for the precalibration (precal) in solid red and the 176-case full calibration (full cal) in dashed gray, determined by taking the ensemble member nearest to the ensemble mean in the final iteration. Entrainment parameters are in the left column and detrainment parameters are in the right.

References

Ackerman, A. S., VanZanten, M. C., Stevens, B., Savic-Jovcic, V., Bretherton, C. S., Chlond, A., Golaz, J.-C., Jiang, H., Khairoutdinov, M., Krueger, S. K., Lewellen, D. C., Lock, A., Moeng, C.-H., Nakamura, K., Petters, M. D., Snider, J. R., Weinbrecht, S., & Zulauf, M. (2009). Large-eddy simulations of a drizzling, stratocumulus-topped marine boundary layer. *Monthly Weather Review*, *137*(3), 1083–1110. https://doi.org/10.1175/2008MWR2582.1

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., & Pritchard, M. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., Watanabe, M., & Webb, M. J. (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. https://doi.org/10.1038/ngeo2398

Boutle, I. A., Eyre, J. E. J., & Lock, A. P. (2014). Seamless Stratocumulus Simulation across the Turbulent Gray Zone. *Monthly Weather Review*, *142*(4), 1655–1668. https://doi.org/10.1175/MWR-D-13-00229.1

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200086. https://doi.org/10.1098/rsta.2020.0086

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019ms001711

Brient, F., & Schneider, T. (2016). Constraints on climate sensitivity from space-based measurements of low-cloud reflection. *Journal of Climate*, *29*(16), 5821–5835. https://doi.org/10.1175/JCLI-D-15-0897.1

Brown, A. R., Cederwall, R. T., Chlond, A., Duynkerke, P. G., Golaz, J. C., Khairoutdinov, M., Lewellen, D. C., Lock, A. P., MacVean, M. K., Moeng, C. H., Neggers, R. A., Siebesma, A. P., & Stevens, B. (2002). Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Quarterly Journal of the Royal Meteorological Society*, *128*, 1075–1093. https://doi.org/10.1256/003590002320373210

Buckingham, E. (1914). On physically similar systems; illustrations of the use of dimensional equations. *Physical Review*, *4*(4), 345–376. https://doi.org/10.1103/PhysRev.4.345

Chammas, S., Wang, Q., Schneider, T., Ihme, M., Chen, Y.-f., & Anderson, J. (2023). Accelerating large-eddy simulations of clouds with Tensor Processing Units. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003619. https://doi.org/10.1029/2023MS003619

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. https://doi.org/10.1016/j.jcp.2020.109716

Cohen, Y., Lopez-Gomez, I., Jaruga, A., He, J., Kaul, C. M., & Schneider, T. (2020). Unified entrainment and detrainment closures for extended eddy-diffusivity mass-flux schemes. *Journal of Advances in Modeling Earth Systems*, *12*(9). https://doi.org/10.1029/2020MS002162

Črnivec, N., Cesana, G., & Pincus, R. (2023). Evaluating the representation of tropical stratocumulus and shallow cumulus clouds as well as their radiative effects in CMIP6 models using satellite observations. *Journal of Geophysical Research: Atmospheres*, *128*(23), e2022JD038437. https://doi.org/10.1029/2022JD038437

Del Genio, A. D., & Wu, J. (2010). The role of entrainment in the diurnal cycle of continental convection. *Journal of Climate*, *23*(10), 2722–2738. https://doi.org/10.1175/2009JCLI3340.1

de Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, D., Pier Siebesma, A., Teixeira, J., & Yano, J.-I. (2013). Entrainment and detrainment in cumulus convection: An overview. *Quarterly Journal of the Royal Meteorological Society*, *139*(670), 1–19. https://doi.org/10.1002/qj.1959

de Rooy, W. C., & Siebesma, A. P. (2010). Analytical expressions for entrainment and detrainment in cumulus convection: Analytical Expressions for Entrainment and Detrainment. *Quarterly Journal of the Royal Meteorological Society*, *136*(650), 1216–1227. https://doi.org/10.1002/qj.640

Dunbar, O. R. A., Duncan, A. B., Stuart, A. M., & Wolfram, M.-T. (2022). Ensemble inference methods for models with noisy and expensive likelihoods. *SIAM Journal on Applied Dynamical Systems*, *21*(2), 1539–1572. https://doi.org/10.1137/21M1410853

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, *13*(9). https://doi.org/10.1029/2020MS002454

Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003124. https://doi.org/10.1029/2022MS003124

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, *9*.

Grabowski, W. W., Bechtold, P., Cheng, A., Forbes, R., Halliwell, C., Khairoutdinov, M., Lang, S., Nasuno, T., Petch, J., Tao, W.-K., Wong, R., Wu, X., & Xu, K.-M. (2006). Daytime convective development over land: A model intercomparison based on LBA observations. *Quarterly Journal of the Royal Meteorological Society*, *132*(615), 317–344. https://doi.org/10.1256/qj.04.147

Gregory, D. (2001). Estimation of entrainment rate in simple models of convective clouds. *Quarterly Journal of the Royal Meteorological Society*, *127*(571), 53–72. https://doi.org/10.1002/qj.49712757104

Griewank, P. J., Heus, T., & Neggers, R. A. J. (2022). Size-dependent characteristics of surface-rooted three-dimensional convective objects in continental shallow cumulus simulations. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002612. https://doi.org/10.1029/2021MS002612

Heinze, R., Mironov, D., & Raasch, S. (2015). Second-moment budgets in cloud topped boundary layers: A large-eddy simulation study. *Journal of Advances in Modeling Earth Systems*, *7*(2), 510–536. https://doi.org/10.1002/2014MS000376

Holland, J. Z., & Rasmusson, E. M. (1973). Measurements of the atmospheric mass, energy, and momentum budgets over a 500-kilometer square of tropical ocean. *Monthly Weather Review*, *101*(1), 44–55. https://doi.org/10.1175/1520-0493(1973)101<0044:MOTAME>2.3.CO;2

Houtekamer, P. L., & Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *129*(1), 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2

Huang, D. Z., Huang, J., Reich, S., & Stuart, A. M. (2022). Efficient derivative-free Bayesian inference for large-scale inverse problems [arXiv:2204.04386 [cs, math]]. *Inverse Problems*, *38*(12), 125006. https://doi.org/10.1088/1361-6420/ac99fa

Huang, D. Z., Schneider, T., & Stuart, A. M. (2022). Iterated Kalman methodology for inverse problems. *Journal of Computational Physics*, *463*, 111262. https://doi.org/10.1016/j.jcp.2022.111262

Iglesias, M., & Yang, Y. (2021). Adaptive regularisation for ensemble Kalman inversion. *Inverse Problems*, *37*(2), 025008. https://doi.org/10.1088/1361-6420/abd29b

Iglesias, M. A., Law, K. J. H., & Stuart, A. M. (2013). Ensemble Kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001. https://doi.org/10.1088/0266-5611/29/4/045001

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., & Hoyer, S. (2021). Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, *118*(21), e2101784118. https://doi.org/10.1073/pnas.2101784118

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., & Hoyer, S. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kovachki, N. B., & Stuart, A. M. (2019). Ensemble Kalman inversion: A derivative-free technique for machine learning tasks [arXiv: 1808.03620]. *Inverse Problems*, *35*(9), 095005. https://doi.org/10.1088/1361-6420/ab1c3a

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, *2013*, 1–13. https://doi.org/10.1155/2013/485913

List, B., Chen, L.-W., & Thuerey, N. (2022). Learned turbulence modelling with differentiable fluid solvers: Physics-based loss functions and optimisation horizons. *Journal of Fluid Mechanics*, *949*, A25. https://doi.org/10.1017/jfm.2022.738

Lopez-Gomez, I. (2023). *A Unified Data-Informed Model of Turbulence and Convection for Climate Prediction* [Doctoral dissertation, California Institute of Technology]. https://thesis.library.caltech.edu/15063/

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). https://doi.org/10.1029/2022MS003105

Lopez-Gomez, I., Cohen, Y., He, J., Jaruga, A., & Schneider, T. (2020). A Generalized Mixing Length Closure for Eddy-Diffusivity Mass-Flux Schemes of Turbulence and Convection. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020MS002161

MacArt, J. F., Sirignano, J., & Freund, J. B. (2021). Embedded training of neural-network subgrid-scale turbulence models. *Physical Review Fluids*, *6*(5), 050502. https://doi.org/10.1103/PhysRevFluids.6.050502

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., & Schlund, M. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

Mellor, G. L., & Yamada, T. (1982). Development of a turbulence closure model for geophysical fluid problems. *Reviews of Geophysics*, *20*(4), 851–875. https://doi.org/10.1029/RG020i004p00851

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., & Caldwell, P. M. (2021). Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*, *11*(6), 501–507. https://doi.org/10.1038/s41558-021-01039-0

Nam, C., Bony, S., Dufresne, J.-L., & Chepfer, H. (2012). The 'too few, too bright' tropical low-cloud problem in CMIP5 models. *Geophysical Research Letters*, *39*(21), 2012GL053421. https://doi.org/10.1029/2012GL053421

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras deep learning bridge for scientific computing. *Scientific Programming*, *2020*, 1–13. https://doi.org/10.1155/2020/8888811

Pahlavan, H. A., Hassanzadeh, P., & Alexander, M. J. (2024). Explainable offline-online training of neural networks for parameterizations: A 1D gravity wave-QBO testbed in the small-data regime. *Geophysical Research Letters*, *51*(2), e2023GL106324. https://doi.org/10.1029/2023GL106324

Pressel, K. G., Kaul, C. M., Schneider, T., Tan, Z., & Mishra, S. (2015). Large-eddy simulation in an anelastic framework with closed water and entropy balances. *Journal of Advances in Modeling Earth Systems*, *7*(3), 1425–1456. https://doi.org/10.1002/2015MS000496

Ramadhan, A., Marshall, J., Souza, A., Lee, X. K., Piterbarg, U., Hillier, A., Wagner, G. L., Rackauckas, C., Hill, C., Campin, J.-M., & Ferrari, R. (2023, March). Capturing missing physics in climate model parameterizations using neural differential equations [arXiv:2010.12559 [physics]]. https://doi.org/10.1002/essoar.10512533.1

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: General algorithms and Lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. https://doi.org/10.5194/gmd-13-2185-2020

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Romps, D. M. (2010). A direct measure of entrainment. *Journal of the Atmospheric Sciences*, *67*(6), 1908–1927. https://doi.org/10.1175/2010JAS3371.1

Savre, J. (2022). What controls local entrainment and detrainment rates in simulated shallow convection? *Journal of the Atmospheric Sciences*, *79*(11), 3065–3082. https://doi.org/10.1175/JAS-D-21-0341.1

Schillings, C., & Stuart, A. M. (2017). Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, *55*(3), 1264–1290. https://doi.org/10.1137/16M105959X

Schneider, T., Leung, L. R., & Wills, R. C. J. (2024). Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmospheric Chemistry and Physics*, *24*(12), 7041–7062. https://doi.org/10.5194/acp-24-7041-2024

Schneider, T., Stuart, A. M., & Wu, J.-L. (2021). Learning stochastic closures using ensemble Kalman inversion. *Transactions of Mathematics and Its Applications*, *5*(1), tnab003. https://doi.org/10.1093/imatrm/tnab003

Shamekh, S., & Gentine, P. (2023). Learning atmospheric boundary layer turbulence. *ESS Open Archive*. https://doi.org/10.22541/essoar.168748456.60017486/v1

Shankar, V., Puri, V., Balakrishnan, R., Maulik, R., & Viswanathan, V. (2023). Differentiable physics-enabled closure modeling for Burgers' turbulence. *Machine Learning: Science and Technology*, *4*(1), 015017. https://doi.org/10.1088/2632-2153/acb19c

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., ... Lawson, K. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*. https://doi.org/10.1038/s43017-023-00450-9

Shen, Z., Sridhar, A., Tan, Z., Jaruga, A., & Schneider, T. (2022). A library of large-eddy simulations forced by global climate models. *Journal of Advances in Modeling Earth Systems*, *14*(3). https://doi.org/10.1029/2021MS002631

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1038/nature12829

Siebesma, A. P., Soares, P. M. M., & Teixeira, J. (2007). A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of the Atmospheric Sciences*, *64*(4), 1230–1248. https://doi.org/10.1175/JAS3888.1

Siler, N., Po-Chedley, S., & Bretherton, C. S. (2018). Variability in modeled cloud feedback tied to differences in the climatological spatial pattern of clouds. *Climate Dynamics*, *50*(3-4), 1209–1220. https://doi.org/10.1007/s00382-017-3673-2

Soares, P., Miranda, P., Siebesma, A., & Teixeira, J. (2004). An eddy-diffusivity/mass-flux parametrization for dry and shallow cumulus convection. *Quarterly Journal of the Royal Meteorological Society*, *130*(604), 3365–3383. https://doi.org/10.1256/qj.03.223

Stevens, B., Lenschow, D. H., Vali, G., Gerber, H., Bandy, A., Blomquist, B., Brenguier, J.-L., Bretherton, C. S., Burnet, F., Campos, T., Chai, S., Faloona, I., Friesen, D., Haimov, S., Laursen, K., Lilly, D. K., Loehrer, S. M., Malinowski, S. P., Morley, B., . . . Van Zanten, M. C. (2003). Dynamics and chemistry of marine stratocumulus—DYCOMS-II. *Bulletin of the American Meteorological Society*, *84*(5), 593–593. https://doi.org/10.1175/BAMS-84-5-Stevens

Tan, Z., Kaul, C. M., Pressel, K. G., Cohen, Y., Schneider, T., & Teixeira, J. (2018). An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid-scale turbulence and convection. *Journal of Advances in Modeling Earth Systems*, *10*(3), 770–800. https://doi.org/10.1002/2017MS001162

Thuburn, J., Weller, H., Vallis, G. K., Beare, R. J., & Whitall, M. (2018). A framework for convection and boundary layer parameterization derived from conditional filtering. *Journal of the Atmospheric Sciences*, *75*(3), 965–981. https://doi.org/10.1175/JAS-D-17-0130.1

Um, K., Brand, R., Fei, Yun, Holl, Philipp, & Thuerey, Nils. (2020). Solver-in-the-loop: Learning from differentiable physics to interact with iterative PDE-solvers. *Advances in Neural Information Processing Systems*, *33*, 6111–6122. https://doi.org/arXiv:2007.00016

vanZanten, M. C., Stevens, B., Nuijens, L., Siebesma, A. P., Ackerman, A. S., Burnet, F., Cheng, A., Couvreux, F., Jiang, H., Khairoutdinov, M., Kogan, Y., Lewellen, D. C., Mechem, D., Nakamura, K., Noda, A., Shipway, B. J., Slawinska, J., Wang, S., & Wyszogrodzki, A. (2011). Controls on precipitation and cloudiness in simulations of trade-wind cumulus as observed during RICO. *Journal of Advances in Modeling Earth Systems*, *3*(2). https://doi.org/10.1029/2011MS000056

Vial, J., Dufresne, J.-L., & Bony, S. (2013). On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*, *41*(11-12), 3339–3362. https://doi.org/10.1007/s00382-013-1725-9

Vignesh, P. P., Jiang, J. H., Kishore, P., Su, H., Smay, T., Brighton, N., & Velicogna, I. (2020). Assessment of CMIP6 cloud fraction and comparison with satellite observations. *Earth and Space Science*, *7*(2), e2019EA000975. https://doi.org/10.1029/2019EA000975

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes.

Geoscientific Model Development, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., Brenowitz, N. D., Kashinath, K., Pritchard, M. S., Bonev, B., Peters, M. E., & Bretherton, C. S. (2023, December). ACE: A fast, skillful learned global atmospheric model for climate prediction [arXiv:2310.02074 [physics]].

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., Chepfer, H., Douville, H., Good, P., Kay, J. E., Klein, S. A., Marchand, R., Medeiros, B., Siebesma, A. P., Skinner, C. B., Stevens, B., Tselioudis, G., Tsushima, Y., & Watanabe, M. (2017). The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384. https://doi.org/10.5194/gmd-10-359-2017

Wu, J.-L., Levine, M. E., Schneider, T., & Stuart, A. (2024). Learning about structural errors in models of complex dynamical systems. *Journal of Computational Physics*, *513*, 113157. https://doi.org/10.1016/j.jcp.2024.113157

Yu, S., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J., Loose, N., Stern, C. I., Beucler, T., Harrop, B., Hillman, B. R., Jenney, A., Ferretti, S., Liu, N., Anandkumar, A., ... Pritchard, M. (2023). ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation. *Advances in Neural Information Processing Systems*, *36*, 22070–22084. https://proceedings.neurips.cc/paper_files/paper/2023/file/45fbcc01349292f5e059a0b8b02c8c3f-Paper-Datasets_and_Benchmarks.pdf

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1). https://doi.org/10.1029/2019GL085782

Zhang, X.-L., Xiao, H., Luo, X., & He, G. (2022). Ensemble Kalman method for learning turbulence models from indirect observation data. *Journal of Fluid Mechanics*, *949*, A26. https://doi.org/10.1017/jfm.2022.744

Chapter 4

LEARNING AND DISTILLING TARGETED MIXING LENGTH CLOSURES

4.1 Introduction

Turbulence is pivotal in governing the vertical transport of heat, moisture, momentum, and various tracers in the atmosphere (Garratt, 1994). The treatment of turbulence has been shown to influence the ability of climate models to accurately produce the stratocumulus-to-cumulus transition (Bogenschutz et al., 2013), an important consideration for this thesis. The concept of mixing length, initially introduced by Taylor and further developed by Ludwig Prandtl in the early 1900s, serves as a foundational concept representing the characteristic distance over which turbulent anomalies maintain their initial properties before dissipating. In the Eddy-diffusivity Mass-flux (EDMF) scheme, the diagnostically-determined mixing length fundamentally controls the behavior of turbulence in the environment, in addition to vertical diffusive fluxes and cloud properties. Over the decades, numerous mixing length closures have been proposed for Large Eddy Simulations (LES) and atmospheric parameterizations, ranging in complexity (Boutle et al., 2014; Grisogono & Belušić, 2008; Mellor & Yamada, 1982; Umlauf & Burchard, 2005). Traditional mixing length closures are often semi-empirical, derived from theoretical regime limits and assumptions that may not fully capture the complexities of atmospheric turbulence in coarse resolution models used to predict weather and climate (Honnert et al., 2021). This is especially true in transitional regimes or under varying stratification and shear conditions. The problem is compounded by the fact that parameterizations vary in how they account for subgrid-scale processes and their interactions, complicating the search for a universally valid mixing length.

In this study, we replace the empirical mixing length closure described by Lopez-Gomez et al. (2020) with a neural network (NN) embedded into the EDMF parameterization framework. Both mixing length closures are trained online using the ensemble Kalman inversion (EKI) framework to match statistics of large-eddy simulation (LES) data, as in Chapter 3. Training is performed to simultaneously optimize parameters controlling turbulence, convection, and cloud microphysics. Following online calibration, the learned NN relationship is distilled using sym-

bolic regression in an offline setting in order to interpret the NN relationship. We compare the performance and functional form of our data-driven closure with the existing semi-empirical closure, demonstrating improvements in turbulent mixing representation, particularly in stratocumulus regions where the empirical model tends to under-predict mixing. This work highlights the potential of integrating machine learning techniques with physically-based parameterizations to both improve the fidelity of mixing length parameterizations against LES statistics, while also pinpointing physical limitations of semi-empirical mixing length closures in a manner that can inform model development.

4.2 Overview

EDMF Version 3

Building on previous variants of the previous EDMF setup, specifically those used in Chapter 3, the EDMF used for this study contains consequential improvements, namely

- Quadratures are used to compute environmental cloud properties, rather than the grid-mean state.
- Prognostic EDMF equations are in advective rather than flux form.
- The updraft prognostic variable is changed to moist static energy.
- Clippings on most prognostic and diagnostic variables are removed.
- The use of a uniform grid is replaced with a stretch-grid, which extends to 40 km, well into the stratosphere.
- Interactive surface fluxes are used, based of Monin-Obukhov theory, where only the sea surface temperature needs be specified. Previously, surface sensible and latent heat fluxes were prescribed to match the LES simulations.
- Fully interactive radiation with RRTMGP is used, rather than prescribing radiative tendencies from the LES.

EDMF Turbulence Scheme

In the EDMF, diffusivity acts to vertically smooth grid-mean quantities and accounts for the influence of stratification, shear, and dissipation (among other source and transport terms) on both turbulent kinetic energy (TKE) and the characteristic eddy

sizes in the environment. Local diffusive fluxes, characterized by chaotic and relatively short-lived environmental eddies, operate in tandem with non-local transport by coherent updrafts to distribute properties vertically and modify the grid-mean state. We focus on the eddy-diffusive contribution to grid-mean fluxes, $\rho\langle w'\phi'\rangle_{\mathrm{ED}}$, defined as

$$\rho\langle w'\phi'\rangle_{\mathrm{ED}} = -\rho a_{\mathrm{env}} K_{\phi,\mathrm{env}} \frac{\partial \bar{\phi}_{\mathrm{env}}}{\partial z}, \tag{4.1}$$

where a_{env} is the environmental area fraction, ϕ is a general tracer, $K_{\phi,\mathrm{env}}$ is the environmental diffusivity for scalar ϕ, and ρ is density.

TKE is formally defined as

$$\overline{\mathrm{TKE}}_{\mathrm{env}} = \frac{1}{2}\left\|\mathbf{u}'_{\mathrm{env}}\right\|^2, \tag{4.2}$$

where $\overline{\mathrm{TKE}}_{\mathrm{env}}$ represents the subdomain average of TKE in the EDMF environment and $\mathbf{u}'_{\mathrm{env}}$ denotes fluctuations of the environmental velocity field. For this study, we use the abbreviation TKE to broadly refer to turbulent kinetic energy, with $\overline{\mathrm{TKE}}_{\mathrm{env}}$ specifically indicating its subdomain average in the EDMF environment.

The EDMF makes use of a 1.5-order Mellor-Yamada turbulence scheme (Mellor & Yamada, 1982), where a single second-moment prognostic equation for vertical velocity variance is retained, representing TKE, and turbulent diffusivity is closed by a diagnostic expression for mixing length:

$$K_{\phi,\mathrm{env}} = c_d l (\overline{\mathrm{TKE}}_{\mathrm{env}})^{1/2}, \tag{4.3}$$

where c_{d} is an empirical nondimensional parameter, l is the mixing length, and $\overline{\mathrm{TKE}}_{\mathrm{env}}$ is the environmental turbulent kinetic energy. This formulation effectively parameterizes diffusivity as the product of the characteristic eddy velocity scale and a length scale, referred to as the mixing length. For the purposes of this study, l is used to denote a general mixing length, with subscripts corresponding to specific variants, introduced later.

The governing prognostic equation for $\overline{\mathrm{TKE}}_{\mathrm{env}}$ in the EDMF is

$$\frac{\partial(\rho a_{\text{env}}\overline{\text{TKE}}_{\text{env}})}{\partial t} + \nabla\cdot\left(\rho a_{\text{env}}\overline{\text{TKE}}_{\text{env}}\mathbf{u}\right) = \underbrace{\nabla\cdot\left(\rho a_{\text{env}}K_{m,\text{env}}\nabla\overline{\text{TKE}}_{\text{env}}\right)}_{\text{Turbulent transport}}$$
$$+ \underbrace{\rho a_{\text{env}}K_{m,\text{env}}\left[\left(\frac{\partial\overline{u}}{\partial z}\right)^2 + \left(\frac{\partial\overline{v}}{\partial z}\right)^2 + \left(\frac{\partial\overline{w}_{\text{env}}}{\partial z}\right)^2\right]}_{\text{Shear production}}$$
$$- \underbrace{\rho a_{\text{env}}\left(K_{h,\text{env}}\frac{\partial\overline{b}_{\text{env}}}{\partial z}\right)}_{\text{Buoyancy production}}$$
$$+ \underbrace{\rho a_{\text{env}}\mathcal{E}_p}_{\text{Entrainment/Detrainment}}$$
$$- \underbrace{\rho a_{\text{env}}\mathcal{P}}_{\text{Pressure diffusion}} - \underbrace{\rho a_{\text{env}}\mathcal{D}}_{\text{Dissipation}}. \tag{4.4}$$

In the prognostic TKE equation, $K_{m,\text{env}}$ is the environmental turbulent viscosity and $K_{h,\text{env}}$ is the thermal diffusivity (related to each other by the Prandtl number). The turbulent transport term represents the diffusion of TKE by environmental eddies themselves. Shear production generates TKE through the differential motion of fluid layers, driven by vertical gradients in the horizontal and vertical wind components $\overline{u}$, $\overline{v}$, and $\overline{w}_{\text{env}}$. Note that the lack of subscripts in the $\overline{u}$ and $\overline{v}$ components denotes grid-mean velocity components. The buoyancy production term captures the generation or suppression of TKE due to density stratification, where $\frac{\partial\overline{b}_{\text{env}}}{\partial z}$ is the environmental buoyancy gradient. $\mathcal{E}_p$ is the entrainment/detrainment term which generates TKE if updrafts detrain into the environment and removes TKE if updrafts entrain air from the environment. This term, alongside $\mathcal{P}$, directly couples the environmental TKE budget to updraft dynamics. Finally, dissipation, represented by $\mathcal{D}$, parameterizes the irreversible conversion of TKE to thermal energy due to molecular viscosity, closing the TKE budget by removing energy from the turbulent flow. The TKE dissipation $\mathcal{D}$ is parameterized using Taylor's surrogate:

$$\mathcal{D} = c_d\frac{\overline{\text{TKE}}_{\text{env}}^{\,3/2}}{l}. \tag{4.5}$$

We note that mixing length affects several terms in the prognostic TKE equation, specifically turbulent transport, shear production, and buoyancy production via eddy

diffusivity and viscosity (Eq. 4.3). Mixing length is also inversely related to the dissipation rate (Eq. 4.5).

The Semi-Empirical Mixing Length Closure

The empirical mixing length closure, described extensively in Lopez-Gomez et al. (2020), is derived from theoretical limits of TKE balance in various regimes. The minimum dissipation assumption dictates that turbulence self-organizes in a way that dissipates TKE as efficiently as possible. Following the minimum dissipation assumption, the empirical mixing length closure is written as a smooth minimum between three candidate mixing lengths

$$l_{\mathrm{p}} = s_{\min}(l_{\mathrm{tke}}, l_{\mathrm{w}}, l_{\mathrm{b}}), \tag{4.6}$$

where l_{b} is determined by environmental stratification, l_{w} by interactions with the surface boundary, and l_{tke} from minimum-dissipation limits of the TKE budget equations. The physical mixing length described by Lopez-Gomez et al. (2020) is denoted l_{p} for the purposes of this paper. The foundations for each candidate mixing length are briefly described in the subsequent sections.

Stratification

In the presence of vertically stratified (stable) atmospheric layers, the strong effects of buoyancy rapidly suppresses eddy extent in the vertical. A characteristic length scale that decreases with increasing stability is defined,

$$l_b = c_b \frac{\overline{\mathrm{TKE}_{\mathrm{env}}}^{\frac{1}{2}}}{N_e}, \tag{4.7}$$

where N_e is the effective stability measured by the Brunt-Väisälä frequency.

Wall Constraint

Monin-Obukhov similarity theory provides a framework for describing the behavior of flows in the atmospheric surface layer. Derived from approximations of the Navier-Stokes equations near a solid boundary, the theory fundamentally relates surface turbulence under varying stability conditions, linking turbulence statistics to mean flow profiles. Monin-Obukhov theory also imposes an upper bound on eddy size in the presence of a boundary. This length scale, l_w, takes the form

$$l_w = \frac{\kappa}{c_m \kappa^* \Phi_m(\xi)} z, \tag{4.8}$$

where κ is the von Karman constant and ξ is a stability parameter defined as $\frac{z}{L_{\text{mo}}}$. Φ_m is an empirical stability function. L_{mo} is the Monin-Obukhov length, defined as

$$L_{\text{mo}} = -\frac{u_*^3 \overline{\theta}_v}{\kappa g (\overline{w'\theta'_v})_s}, \tag{4.9}$$

where u_* is the friction velocity, $(\overline{w'\theta'_v})_s$ is the surface virtual potential temperature flux, $\overline{\theta}_v$ is the virtual potential temperature, and g is the gravitational acceleration on Earth.

Minimum TKE Dissipation

The final candidate mixing length, l_{tke}, is derived from a complex mathematical formulation involving numerous diagnostic terms. For a comprehensive explanation and derivation, readers are referred to Lopez-Gomez et al. (2020). The mixing length l_{tke} is based on the principle of minimum TKE dissipation. By positing that TKE is dissipated at a rate equal to or greater than its production at small scales within the environment, a balance in the TKE budget is achieved. This balance results in a mixing length that incorporates entrainment processes and convective motions. Specifically, l_{tke} depends on local environmental conditions and the vertical velocity differences between subdomains. It is most effective in neutral and slightly stable flows. Under certain conditions, l_{tke} is shown to simplify to align with other established mixing length closures.

Limitations of Minimum Mixing Length Closure

A key limitation of the smooth minimum closure lies in its tendency to cause excessive dissipation of TKE when one of the length scales becomes small, particularly in regions where eddy sizes are constrained by stratification or near the surface layer. This over-dissipation can lead to significant underestimations of turbulent fluxes, as the semi-empirical length scales may not fully encapsulate the complexity of turbulent processes. Also, the application of a uniform smooth minimum across different turbulent regimes can result in overly diffusive transitions, making it challenging to capture sharp gradients accurately. Perhaps a regime-dependent smooth minimum might better address the nuances between different turbulent states. Additionally, the buoyancy length scale employed may fail to account for phenomena such as

intermittent turbulence. By assuming that one length scale dominates through the smooth minimum, the model may oversimplify interactions between multiple physical processes, especially in scenarios where several processes contribute equally to the production and dissipation of turbulence. Finally, parameterizing numerous terms (higher-order moments) in the TKE equation and deriving mixing length limits based on these assumptions may introduce inaccuracies, particularly during transitions between regimes.

4.3 Ensemble Kalman Inversion Setup

Precalibration

Following Christopoulos et al. (2024), a precalibration step is performed to get sensible priors before exposing the EKI pipeline to the full library of LES cases. The default, semi-empirically determined priors for the turbulence, convection, and cloud closures are originally taken from Christopoulos et al. (2024) and Lopez-Gomez et al. (2022) and slightly modified to improve stability for AMIP-like global atmospheric simulations. Because the parameters were largely designed for a previous variant of the EDMF and modified in an ad-hoc manner for stability in a global simulation, we perform a precalibration on six representative cases representing stratocumulus, cumulus, and deep convection to give more robust initial priors. The precalibration set consists of LES simulations forced with the HadGEM2-A model for the month of July, and includes cfSites 17, 18, 22, 23, 30, and 94. Precalibration is run using the empirical mixing length closure for 10 iterations. The optimal precalibration parameter set is chosen from the ensemble member nearest to the mean in the final iteration, and referred to as $\mathbf{\Theta}'_{\text{pre}}$.

Full Calibration

Relative to the previous work of Christopoulos et al. (2024), we briefly summarize modifications to the observation map in the EKI pipeline. The low-dimensional encoding via principal component analysis has been removed. The PCA components, computed from the time-variability of domain-average LES profiles, are not guaranteed to encode data in a manner that is optimally needed to explore state space by EKI. Additionally, retaining the full profiles preserves the full covariance structure of the system, likely facilitating faster convergence. Further, we find that memory constraints for even relatively large batches (30) are not prohibitive. In previous work, we found the EKI convergence characteristics were strongly sensitive to the EKI noise covariance level, empirically estimated from the LES. In general, the

LES noise either over or underestimates noise levels relative to the time-mean signal, necessitating separate variable-dependent, empirical scaling factors for each noise covariance matrix. To reduce this sensitivity, we simplify the noise by specifying a diagonal, constant noise of $\sim 5\%$. A log transform is applied to q_l since the quantity generally follows a log-normal distribution and so that constant noise can be applied. All variables are empirically scaled to have zero mean and unit variance on the basis of distributions collected from all LES simulations forced with HadGEM2-A in the AMIP configuration. Following these updates, we find the EKI convergence properties to be generally smoother and less sensitive to hyperparameters relative to previous work. The loss function variables included in EKI are

1. $\bar{\theta}$: potential temperature
2. $\bar{q}_t$: total water specific humidity
3. $\bar{q}_l$: liquid water specific humidity.

Beyond these differences, the setup is comparable to the one used in Christopoulos et al. (2024) and Lopez-Gomez et al. (2022).

The full calibration is initialized using prior means equivalent to $\mathbf{\Theta}'_{\text{pre}}$, and calibrated to all 60 HadGEM2-A cases with a batch size of 30. After three epochs, the final optimal parameter set is chosen from the ensemble member nearest to the mean in the final iteration, in the same manner as precalibration. The optimal parameter set for the NN-based EDMF is denoted $\left(\mathbf{\Theta}'_{\text{full}}\right)^{\text{NN}}$ and that for the physically-based EDMF $\left(\mathbf{\Theta}'_{\text{full}}\right)^{\text{phys}}$.

Pretraining Neural Network

To obtain priors for weights and biases of the NN, pretraining is performed by running all simulations in the training set using $\mathbf{\Theta}'_{\text{pre}}$. Ten timesteps from every simulation in the 60 HadGEM2-A AMIP runs are randomly sampled and concatenated to form a collection of input-output pairs for NN training. The NN is trained using an Adam optimizer and a learning rate of 1e$-$3. The batch size is set to 50000 and the network is trained for 1000 epochs.

Data-driven Inputs

We consider eight input variables for the NN deemed relevant for describing TKE dynamics:

$$X_1 = \frac{\text{shear}^2}{\left(\frac{db}{dz}\right)_{\text{env}}}, \tag{4.10a}$$

$$X_2 = \frac{\overline{\text{TKE}}_{\text{env}}}{\left(\frac{db}{dz}\right)_{\text{env}} \cdot z^2}, \tag{4.10b}$$

$$X_3 = \frac{\overline{\text{TKE}}_{\text{env}}}{\Delta\overline{w}^2}, \tag{4.10c}$$

$$X_4 = \text{shear}^2, \tag{4.10d}$$

$$X_5 = \left(\frac{db}{dz}\right)_{\text{env}}, \tag{4.10e}$$

$$X_6 = \overline{\text{TKE}}_{\text{env}}, \tag{4.10f}$$

$$X_7 = \frac{z}{\text{L}_{\text{mo}}}, \tag{4.10g}$$

$$X_8 = \frac{\Delta z}{\text{L}_{\text{mo}}}, \tag{4.10h}$$

where $\left(\frac{db}{dz}\right)_{\text{env}}$ is the environmental buoyancy gradient, L_{mo} is the Monin-Obukhov length, Δz is the vertical model resolution, and z is the geometric height. The initial input X_1 is the gradient Richardson number, a proxy measure for indicating thermally- or mechanically-driven turbulence. The variable X_3 may be considered an indicator of the relative importance of turbulence versus convection in driving EDMF fluxes. In X_3, the difference between the mean updraft and environmental vertical velocity is $\Delta\overline{w}$, equivalent to $\overline{w}_{\text{up}} - \overline{w}_{\text{env}}$. We note that X_4, X_5, and X_6 are dimensional variables and not Buckingham Pi groups. Quantities with the subdomain subscript "up" or "env" omitted are grid mean quantities.

Empirical scaling was applied to the input variables to encourage relative consistency in their magnitude and range. Each variable was centered by removing its empirical mean and was subsequently scaled by ten times the empirical standard deviation,

leading to values that generally lie in the range (-1, 1). The transformation is expressed mathematically as

$$\tilde{X}_i = \frac{X_i - \mu_i}{10 \cdot \sigma_i}, \tag{4.11}$$

where μ_i and σ_i are the empirical mean and standard deviation of each variable, respectively. The scaling makes X_4, X_5, and X_6 nondimensional due to division by the empirical standard deviation.

Finally, the mixing length l is normalized in the same manner:

$$\tilde{l} = \frac{l - \mu_l}{10 \cdot \sigma_l}. \tag{4.12}$$

Neural Network Architecture

The data-driven model selected for learning mixing length is a fully connected neural network with 666 parameters. The number of parameters was selected by training a series of eight networks with varying parameters and architectures to match the physical mixing closure. The considered number of parameters ranged from $O(10)$ to $O(1000)$. Consequential performance improvements diminish after ~ 500 parameters, indicating networks of this size are expressive enough to capture the sorts of mixing length relationships we hope to learn through EKI. The NN takes the normalized, nondimensional inputs and predicts the normalized mixing length, denoted as $\tilde{l}_{ml}$, where

$$\tilde{l}_{ml} = NN(\tilde{X}_1, \tilde{X}_2, ...\tilde{X}_N). \tag{4.13}$$

Hard Constraints on l

As in Lopez-Gomez et al. (2020), we enforce the following hard constraints on l in the parameterization code following prediction of l with the closure:

1. $l < z$: Enforces eddy length scales which are smaller than the distance to the boundary (ocean surface, in this case).

2. $l = \max(l_{\text{smag}}, l)$, where l_{smag} is the Smagorinsky-Lilly length scale.

The Smagorinsky-Lilly closure, commonly used in LES simulations, parametrizes eddy diffusivity as a function of the resolved-scale strain rate of the flow and the characteristic grid spacing (Smagorinsky, 1963). The limiter properly encourages resolution-dependent mixing lengths, ensuring that subgrid-scale dissipation adjusts

to grid resolution, thus improving the model's ability to capture energy transfer across unresolved scales at different resolution.

Additionally, to ensure no spurious turbulent length scales in non-turbulent regions, l is set to zero when TKE is zero:

$$l = \begin{cases} 0 & \text{if } \overline{\text{TKE}}_{\text{env}} = 0 \\ l & \text{otherwise.} \end{cases}$$

This constraint is also essential for numerical stability when NNs are coupled to the simulation.

Symbolic Regression

To extract interpretable mathematical expressions that approximate the learned NN mapping, we employ symbolic regression using the PySR package (Cranmer, 2023). In atmospheric and ocean modeling, equation discovery and symbolic regression techniques have been effectively used to learn eddy parameterizations in ocean models (Zanna & Bolton, 2020) and cloud cover parameterizations in a global storm-resolving model (Grundner et al., 2024) with success, among other examples in fluid dynamics more generally (Brunton et al., 2016). Symbolic regression searches the space of mathematical formulas to find equations that best fit the data while balancing complexity and accuracy. PySR utilizes a genetic algorithm that evolves a population of candidate equations through operations analogous to natural selection, crossover, and mutation. By iteratively modifying the population, the algorithm converges toward parsimonious expressions that approximate the underlying relationships learned by the NN. From them, we can begin to physically interpret relationships between the mixing length predictors and mixing length. The training dataset for this procedure consists of all 60 training cases in the HadGEM2-A AMIP configuration, run with the optimal parameters $\left(\boldsymbol{\Theta}'_{\text{full}}\right)^{\text{NN}}$.

4.4 Calibration Results

We compare the performance and physical behaviors of both EDMF variants, following full calibration of EDMF-NN_mix (NN-based mixing closure) and EDMF-Physical_mix (empirical mixing closure). We use root mean squared error (rmse) as an aggregate summary metric to broadly capture performance differences between the schemes. Figure 4.1 displays box-and-whiskers plots comparing the schemes. Boxes indicate the interquartile range and the center line indicates the median rmse. Whiskers

extend to 1.5 times the interquartile range, with outliers shown as diamonds. The rmse is computed over all 60 training cases in the HadGEM2-A AMIP configuration, using the optimal parameters $\left(\boldsymbol{\Theta}'_{\text{full}}\right)^{\text{NN}}$ for EDMF-NN_mix and $\left(\boldsymbol{\Theta}'_{\text{full}}\right)^{\text{phys}}$ for EDMF-Physical_mix.

On the whole, EDMF-NN_mix demonstrates slight improvements in $\bar{\theta}$ with a smaller spread and comparable performance with EDMF-Physical_mix for $\bar{q}_t$. While the median $\bar{q}_l$ error is comparable between the schemes, EDMF-NN_mix demonstrates a marginally higher spread and more outliers. The tendency to produce outliers in $\bar{q}_l$ for both schemes centers around the prevalence of "cloud spikes," where a single model level becomes saturated. This may be a consequence of environmental microphysics or grid resolution, but further study is needed.

Differences in Mixing Characteristics: A Stratocumulus Case Study

While rmse statistics broadly capture performance differences, averaging over height can obscure consequential differences in physical mechanisms, despite both leading to robust statistical fits. We tease out the differing physical mechanisms at play in the EDMF for setting the time-mean statistics, using a case study to identify differences in how the simulations evolve and maintain clouds. The most consequential improvements of EDMF-NN_mix over EDMF-Physical_mix are noted for the evolution and maintenance of the boundary layer height in stratocumulus clouds. In particular, EDMF-Physical_mix significantly under-predicts turbulent mixing in the presence of the strong inversions characteristic of subtropical subsidence regions.

EDMF simulations are initialized from a time-averaged GCM state with shallow boundary layers, leading to initially unsteady dynamics. From the initial state, the boundary layer must grow and eventually form and maintain a relatively thin stratocumulus deck (typically $O(100$ m) or less) to match the statistics of the LES. The dominant mechanism maintaining stratocumulus clouds is top-down turbulent mixing driven by cloud-top radiative cooling, in addition to surface moisture fluxes (Stevens et al., 2003; Wood, 2012). In the EDMF, boundary layers can grow and are maintained by the actions of both diffusive turbulent flux (ED) and convective mass flux (MF). We find that the overwhelming majority of stratocumulus cases using the physical mixing length closure, even after calibration, result in significant under-mixing in the environment. The consequence of this can manifest in several ways, as illustrated by Figure 4.2. Panels (d) and (i) display the learned mixing length l, while (e) and (j) display the resulting turbulent kinetic energy. During

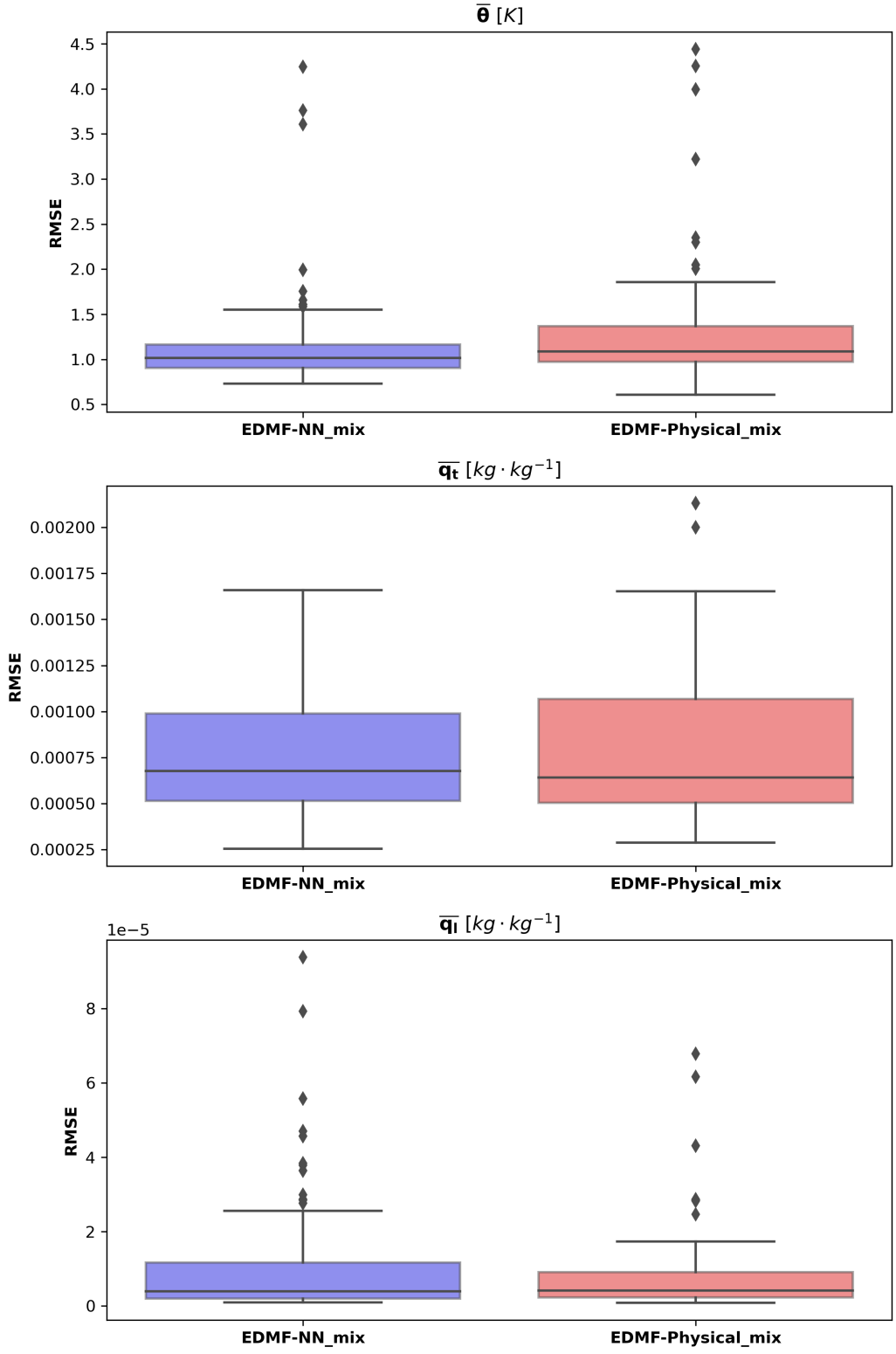

Figure 4.1: Comparison of root mean squared error (rmse) for EDMF-NN_mix (left; blue) and EDMF-Physical_mix (right; red). Boxes indicate the interquartile range and the center line indicates the median rmse. Whiskers extend to 1.5 times the interquartile range, with outliers shown as diamonds. The rmse is computed over all 60 training cases in the HadGEM2-A AMIP configuration, using optimal parameters from the full calibration.

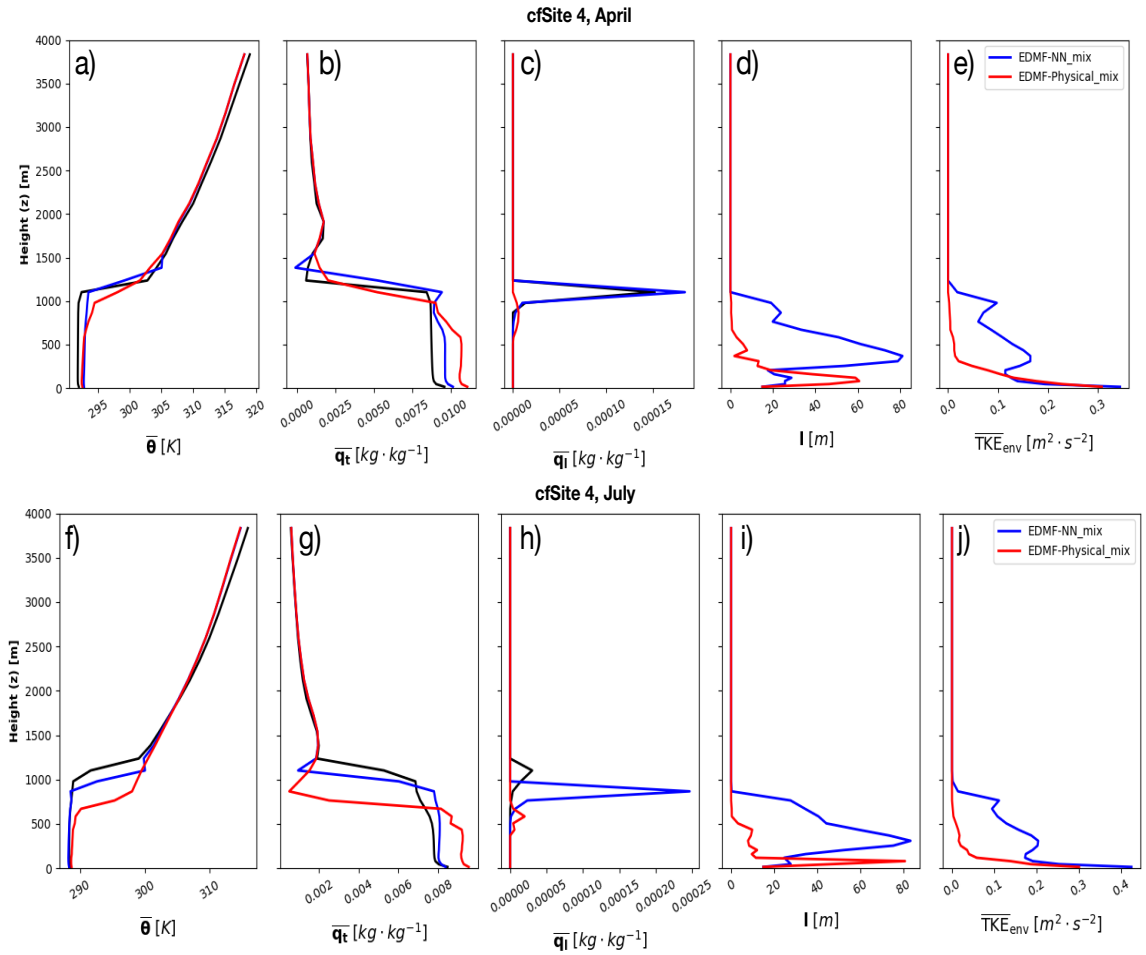

Figure 4.2: Comparison of EDMF-Physical_mix (red), EDMF-NN_mix (blue), and the LES (black) for a characteristic stratocumulus case off the coast of South America in April (top row) and July (bottom row). The first three columns display profiles for variables explicitly included in the EKI loss function.

April (Fig. 4.2, top row), both variants correctly capture the height of the boundary layer ($\sim$ 1000 m), but EDMF-Physical_mix lacks consequential turbulent mixing above $\sim$ 400 m. TKE, alongside l, may be used as a proxy for the diffusive flux. The lack of consequential mixing results from excessively small mixing lengths which lead to over-dissipation of TKE (Eq. 4.5). Given the comparable boundary layer heights despite the large difference in TKE and diffusive flux, this implies EDMF-Physical_mix relies strongly on convective mass fluxes in the upper $\sim$ 600 m to maintain the boundary layer.

For the same site in July, the under-mixing in EDMF-Physical_mix results in a boundary layer that fails to grow far beyond its initial depth, leading to a bound-

ary layer height bias of $\sim$ 500 m and larger errors in $\bar{\theta}$ and $\bar{q}_t$. In contrast, EDMF-NN_mix produces strong mixing in the boundary layer that grows the boundary layer to a height consistent with the LES. At the same time, a stratocumulus cloud forms with excessive $\bar{q}_l$ relative to LES and EDMF-Physical_mix. While the amount of (un)resolved turbulent kinetic energy is resolution dependent, we can broadly compare the properties of the TKE profiles to other studies. In stratocumulus-topped boundary layers TKE magnitudes are typically 0.1 - 0.8 $m^2 \cdot s^{-2}$, maintained to the stratocumulus cloud top, with typical mixing length of (50 -100 m) (Heinze et al., 2015; Lenderink & Holtslag, 2004).

Comparison of TKE Distributions

Figure 4.3 displays the aggregate frequency distributions of mixing length and TKE across all cfSites and vertical grid levels with non-zero TKE, so that they may be systematically compared. The EDMF-NN_mix model shows a more concentrated distribution and limited mixing lengths larger than $\sim$ 400 m. Both models demonstrate a comparable range of TKE, generally ranging below 2.5 $m^2 \cdot s^{-2}$. The physical model notably has a high frequency of large mixing lengths $\sim 200-700$ m for small values of TKE ($<\sim 0.05\ m^2 \cdot s^{-2}$). Furthermore, some counts as high as $\sim$ 1000 m occur, which is comparable to the boundary layer height. While not explicitly limited in the model, mixing lengths in convective and turbulent boundary layers should in general be smaller than the boundary layer height, as turbulent mixing is almost fully suppressed at the top.

Both models demonstrate distinct modes. In particular, both contain a mode characterized by large counts across a large range of TKE values at small mixing lengths. In the space of intermediate TKE values ($\sim 0.1-1.0\ m^2 \cdot s^{-2}$) and mixing lengths $\sim 50-400$ m, EDMF-Physical_mix shows a distinctly broader dispersion, implying a weaker relationship between TKE and mixing length. Alternatively, EDMF-NN_mix demonstrates a roughly linear relationship with modest dispersion.

4.5 Symbolic Regression Results

We build a large set of candidate equations using symbolic regression run with varying hyperparameters and variables. A final list of candidate equations are selected by choosing the simplest expressions with the best fits, while maximizing diversity across the symbolic regression runs. In general, the first candidate equation from each PySR run with a non-dimensional loss $< 10^{-3}$ is selected. Following symbolic regression, constants are added and the equation constants are re-optimized

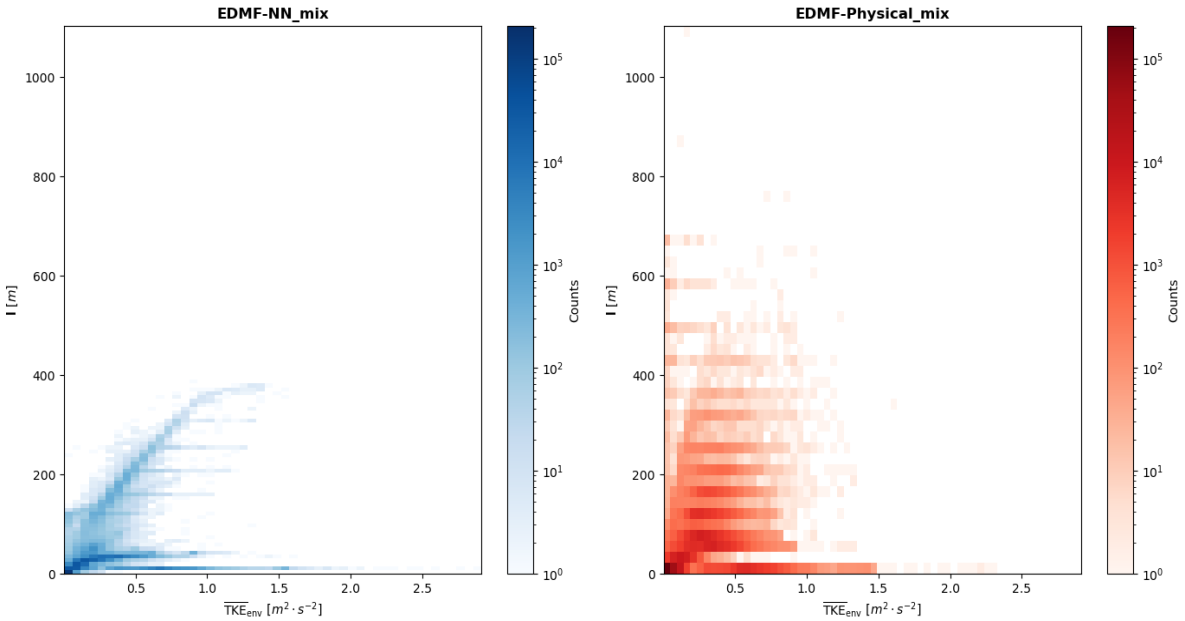

Figure 4.3: Comparison of 2D frequency distributions between EDMF-NN_mix (left) and EDMF-Physical_mix (right) for mixing length (l, y-axis) and environmental turbulent kinetic energy ($\overline{\text{TKE}}_{\text{env}}$, x-axis) across all cfSites and vertical grid levels with non-zero TKE.

to fit the learned NN relationship, as PySR focuses more on functional forms rather than constant optimization. The final candidate equations are enumerated in Table 4.5, alongside their correlation with the NN relationships after optimization.

Among the candidates, $\tilde{l}_1'$ stands out as among the simplest and most predictive expressions, which we will largely focus on here. The exponential decrease of mixing length on shear likely prevents excessive generations of mixing in strongly sheared environments, as shear explicitly appears in the shear production term of the prognostic TKE equation. Furthermore, while shear increases TKE, it often simultaneously inhibits the size and persistence of larger eddies, leading to a reduced mixing length. All expressions have an explicit dependence on TKE in the numerator, which very commonly appears in empirical closures, including l_{tke}. This aligns with the physical expectation that higher turbulence allows eddies to transport momentum and scalar quantities over larger distances. Mixing length closures containing a negative relationship with shear include Grisogono (2010), Huang et al. (2013), Blackadar (1962), and some of the earliest parameterizations of mixing length, including von Karman's in the 1930s. Many parameterizations, including Lopez-Gomez et al. (2020), also include semi-empirical and theoretically based

Candidate Mixing Length Equations	**Correlation**
$\tilde{l}_1' = \left(0.703\,\widetilde{TKE} + 0.0298\right) \exp\left(-14.9\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right) - 8.72\,\widetilde{\mathrm{shear}^2}\right)$	0.936
$\tilde{l}_2' = \dfrac{2.22\,\widetilde{TKE} + 0.0897}{\left(0.00299\,\exp\left(135\left(\widetilde{\frac{z}{\mathrm{L_{mo}}}}\right)\right) + 1\right)^{1.09}}$	0.905
$\tilde{l}_3' = -\left(5.73\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right) - 0.618\,\widetilde{TKE}\right)\left(1.61\,\widetilde{TKE} + 0.0382\right)$ $\times \exp\left(-88.6\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right) - 6.61\left\lvert -0.359\,\widetilde{TKE} + 0.603\left(\widetilde{\frac{z}{\mathrm{L_{mo}}}}\right) + 0.0506\right\rvert\right)$	0.899
$\tilde{l}_4' = \left(-0.556\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right) + 0.417\,\widetilde{TKE}\right) \exp\left(-59.5\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right)\right)$	0.845

Table 4.1: Learned symbolic expressions for non-dimensionalized mixing length functions and their correlations with the NN relationship. Primes indicate candidates and subscripts on l are used to index the different equations. The overbars and subscripts denoting subdomain means have been omitted for easier readability.

relationships that directly depend on the stability (often measured by the Gradient Richardson number X_1). While buoyancy gradients and X_1 do not explicitly appear in l_1', the quantity $\frac{\overline{\mathrm{TKE}}_{\mathrm{env}}}{\Delta\overline{w}^2}$ can implicitly indicate the stability, as the prominence and strength of updrafts, indicated by $\Delta\overline{w}^2$, is influenced by buoyancy differences between the updraft and environment and so too is the amount of environmental turbulent kinetic energy via buoyancy production.

Additionally the factor $\left(c_1\,\widetilde{\mathrm{TKE}} - c_2\left(\widetilde{\frac{\mathrm{TKE}}{(\Delta w)^2}}\right)\right)$ commonly appears across PySR runs with different hyperparameters. The terms couples $\overline{\mathrm{TKE}}_{\mathrm{env}}$ length scales to updraft dynamics through $(\Delta w)^2$. The term increases the mixing length as $\overline{\mathrm{TKE}}_{\mathrm{env}}$ increases,

but modulates the growth by the influence of $\overline{\text{TKE}}_{\text{env}}$ relative to a proxy for mass fluxes due to updrafts. For a fixed $\overline{\text{TKE}}_{\text{env}}$, stronger updrafts would increase the mixing length. This may encode the contribution of updraft detrainment (and, by symmetry, environmental entrainment) to $\overline{\text{TKE}}_{\text{env}}$ via TKE injection. Faster updrafts will detrain more at their tops by virtue of mass continuity, leading to stronger detrainment and more $\overline{\text{TKE}}_{\text{env}}$ production.

4.6 Conclusion and Future Work

Given the trade off between mass fluxes and diffusive fluxes for setting time-averaged statistics of $\bar{\theta}$, $\bar{q}_t$, and $\bar{q}_l$, it may be helpful to add additional variables to the loss function which isolate the contribution of each flux more precisely. In particular, a measure of TKE in the LES may be included or a second-moment statistic that captures spatial variability. Furthermore, nondimensional groups such as X_4, X_5, and X_6 may be turned into proper Pi groups by finding the appropriate dimensional scales, rather than non-dimensionalizing by empirical variances. We would also like to explore and confirm physical limits of the learned symbolic expressions, specifically whether they approach mixing lengths in the surface layer predicted by Monin-Obukhov similarity theory under a range of stability criteria. Finally, the "cloud spike" dilemma may potentially be addressed by increasing the vertical resolution of the model, although it is unclear whether this is the chief cause of these phenomena.

References

Blackadar, A. K. (1962). The vertical distribution of wind and turbulent exchange in a neutral atmosphere. *Journal of Geophysical Research*, *67*. https://doi.org/10.1029/JZ067i008p03095

Bogenschutz, P. A., Gettelman, A., Morrison, H., Larson, V. E., Craig, C., & Schanen, D. P. (2013). Higher-Order Turbulence Closure and Its Impact on Climate Simulations in the Community Atmosphere Model. *Journal of Climate*, *26*, 9655–9676. https://doi.org/10.1175/JCLI-D-13-00075.1

Boutle, I. A., Eyre, J. E. J., & Lock, A. P. (2014). Seamless Stratocumulus Simulation across the Turbulent Gray Zone. *Monthly Weather Review*, *142*(4), 1655–1668. https://doi.org/10.1175/MWR-D-13-00229.1

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113

Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O. R. A., & Schneider, T. (2024). Online Learning of Entrainment Closures in a Hybrid Machine Learning Parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(11), e2024MS004485. https://doi.org/10.1029/2024MS004485

Cranmer, M. (2023, May). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl [arXiv:2305.01582 [astro-ph]]. https://doi.org/10.48550/arXiv.2305.01582

Garratt, J. R. (1994). Review: The atmospheric boundary layer. *Earth-Science Reviews*, *37*, 89–134. https://doi.org/10.1016/0012-8252(94)90026-4

Grisogono, B. (2010). Generalizing 'z-less' mixing length for stable boundary layers.

Grisogono, B., & Belušić, D. (2008). Improving mixing length-scale for stable boundary layers. *Quarterly Journal of the Royal Meteorological Society*, *134*, 2185–2192. https://doi.org/10.1002/qj.347

Grundner, A., Beucler, T., Gentine, P., & Eyring, V. (2024). Data-Driven Equation Discovery of a Cloud Cover Parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(3), e2023MS003763. https://doi.org/10.1029/2023MS003763

Heinze, R., Mironov, D., & Raasch, S. (2015). Second-moment budgets in cloud topped boundary layers: A large-eddy simulation study. *Journal of Advances in Modeling Earth Systems*, *7*(2), 510–536. https://doi.org/10.1002/2014MS000376

Honnert, R., Masson, V., Lac, C., & Nagel, T. (2021). A Theoretical Analysis of Mixing Length for Atmospheric Models From Micro to Large Scales. *Frontiers in Earth Science*, *8*, 582056. https://doi.org/10.3389/feart.2020.582056

Huang, J., Bou-Zeid, E., & Golaz, J.-C. (2013). Turbulence and Vertical Fluxes in the Stable Atmospheric Boundary Layer. Part II: A Novel Mixing-Length Model. *Journal of the Atmospheric Sciences*, *70*, 1528–1542. https://doi.org/10.1175/JAS-D-12-0168.1

Lenderink, G., & Holtslag, A. A. M. (2004). An updated length-scale formulation for turbulent mixing in clear and cloudy boundary layers. *Quarterly Journal of the Royal Meteorological Society*, *130*. https://doi.org/10.1256/qj.03.117

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). https://doi.org/10.1029/2022MS003105

Lopez-Gomez, I., Cohen, Y., He, J., Jaruga, A., & Schneider, T. (2020). A Generalized Mixing Length Closure for Eddy-Diffusivity Mass-Flux Schemes of Turbulence and Convection. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020MS002161

Mellor, G. L., & Yamada, T. (1982). Development of a turbulence closure model for geophysical fluid problems. *Reviews of Geophysics*, *20*(4), 851–875. https://doi.org/10.1029/RG020i004p00851

Smagorinsky, J. (1963). General circulation experiments with the primitive equations. *Monthly Weather Review*, *91*(3), 99–164. https://doi.org/10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2

Stevens, B., Lenschow, D. H., Vali, G., Gerber, H., Bandy, A., Blomquist, B., Brenguier, J.-L., Bretherton, C. S., Burnet, F., Campos, T., Chai, S., Faloona, I., Friesen, D., Haimov, S., Laursen, K., Lilly, D. K., Loehrer, S. M., Malinowski, S. P., Morley, B., ... Van Zanten, M. C. (2003). Dynamics and chemistry of marine stratocumulus—DYCOMS-II. *Bulletin of the American Meteorological Society*, *84*(5), 593–593. https://doi.org/10.1175/BAMS-84-5-Stevens

Umlauf, L., & Burchard, H. (2005). Second-order turbulence closure models for geophysical boundary layers. A review of recent work. *Continental Shelf Research*, *25*(7), 795–827. https://doi.org/10.1016/j.csr.2004.08.004

Wood, R. (2012). Stratocumulus Clouds. *Monthly Weather Review*, *140*(8), 2373–2423. https://doi.org/10.1175/MWR-D-11-00121.1

Zanna, L., & Bolton, T. (2020). Data-Driven Equation Discovery of Ocean Mesoscale Closures. *Geophysical Research Letters*, *47*(17). https://doi.org/10.1029/2020GL088376

Chapter 5

CONCLUSION

This dissertation investigates the representation of subgrid-scale processes, namely turbulence and convection, which are key contributors to biases in climate models and uncertainties in ECS. These issues manifest strongly in stratocumulus-to-cumulus transition regions, where traditional parameterizations struggle to capture key dynamics that dictate cloud properties. Building on the EDMF framework, this research develops a hybrid parameterization that incorporates targeted machine learning to improve representations of the poorly constrained processes of turbulent mixing and lateral entrainment. Ensemble Kalman inversion facilitates calibration of empirical closures alongside data-driven components, ensuring that the hybrid model maintains stability and physical consistency while fitting statistics relevant for modeling on climate timescales. This approach, demonstrated in a single-column setup, improves model fidelity across the stratocumulus-to-cumulus transition and holds promise for addressing systemic biases in climate predictions on the basis of indirect observations.

Looking ahead 5–10 years, the operationalization of ML-based climate prediction systems will require consideration of diverse stakeholders, including government agencies, climate researchers, private sector companies, and non-governmental organizations. Each group has distinct priorities, such as forecasting the frequency of heatwaves for public health planning, predicting changes in crop yields for agriculture, or modeling sea level rise for coastal infrastructure development. The inherently high-dimensional nature of atmospheric predictions poses unique challenges, as stakeholders care about different aspects of the forecasted state. Persistent issues with ML methods, such as numerical instability and limited generalizability, must be addressed, and the development of common benchmarks will be essential to ensure their reliability across diverse applications. Addressing these varied demands will likely require the integration of ML techniques—not only for tasks such as downscaling, uncertainty quantification, and post-processing but also for improving subgrid-scale parameterizations in numerical models.

Operational success will hinge on the ability of ML-based models to produce robust, CMIP-like projections that meet the diverse needs of stakeholders while maintaining

scientific credibility and transparency. Progress will likely evolve in fits and starts, as climate modeling continues to face physical, computational, and data limitations. While advances in computational power bring kilometer-scale horizontal resolution within reach, many small-scale atmospheric processes—including turbulence and cloud microphysics—will remain unresolved. Hybrid approaches should play an increasingly central role, but their effectiveness depends on the quality and availability of training data. Processes like cloud microphysics, occurring at scales too small for consistent and direct observation, will pose persistent challenges due to the lack of robust observational constraints and incomplete physical understanding. As the field evolves, the balance between accuracy, interpretability, and generalizability should remain a critical focus, ensuring that next-generation models can meet the pressing demands of both scientific research and practical decision-making in a changing climate.