UNIVERSITY OF CALIFORNIA,
IRVINE

Improving The Modeling and Analysis of Tropical Convection and Precipitation through Machine Learning Methods

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Earth System Science

by

Griffin Mooers

Dissertation Committee:
Associate Professor Mike Pritchard, Chair
Associate Professor Stephan Mandt
Assistant Professor Tom Beucler
Professor Jin-Yi Yu
Professor Jim Randerson

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The research in this thesis was supported NSF grants 1633631 (and a special thanks the MAPS Program), OAC-1835863, AGS-1734164, OAC-1835769, as well as DARPA contracts HR001119S0038 and HR001120C0021, Division of Atmospheric and Geospace Sciences grant AGS-1912134, Division of Information and Intelligent Systems grants IIS-2047418, IIS-2003237, IIS-2007719, Division of Social and Economic Sciences grant SES-1928718, and Division of Computer and Network Systems grant CNS-2003237 for funding support and co-funding by the Enabling Aerosol-cloud interactions at GLobal convection-permitting scalES (EAGLES) project (74358), of the U.S. Department of Energy Office of Biological and Environmental Research, Earth System Model Development program area. Gifts from Intel, Disney, and Qualcomm also need to be acknowledged. We further acknowledge funding from NSF Science and Technology Center LEAP (Launching Early-Career Academic Pathways) award 2019625.

Computational resources were provided by the Extreme Science and Engineering Discovery Environment supported by NSF Division of Advanced Cyberinfrastructure Grant number ACI-1548562 (charge number TG-ATM190002). DYAMOND data management was provided by the German Climate Computing Center (DKRZ) and supported through the projects ESiWACE and ESiWACE2. The projects ESiWACE and ESiWACE2 have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 675191 and 823988. This work used resources of the German Climate Computing Centre (DKRZ) granted by its Scientific Steering Committee (WLA) under project IDs bk1040 and bb1153.

I want to thank my entire thesis committee, it has been a great honor to collaborate with you. Thank you to Jim Randerson and Jin-Yi Yu for your sharp insights and feedback that kept this thesis on track. I thank my advisor Mike Pritchard for accepting me as a graduate many years ago, for his guidance and mentorship, for the flexibility he gave me as a graduate student, and for his generous support for my career since I arrived at UC Irvine. I would also like to thank Stephan Mandt for agreeing to take on the role as my second advisor under the MAPS program and for his active engagement with my projects since; Thank you for the energy and ideas every time we met that really elevated my research in ways that would otherwise not have been possible. Working with Stephan, I learned so much about not only machine learning but how to better communicate ideas and think in different ways. I need to thank Tom Beucler for first taking me under his wing, for his mentorship since including meeting with me every week without fail even after taking a professorship at Lausanne, for teaching me about so many tools and resources, and for always pushing me outside my comfort zone.

I also need to extend my appreciation to my undergraduate research advisor, Dr. Arthur DeGaetano for his mentorship and training. I would not have gone into a Ph.d. program without his support and without his showing me how rewarding research can be.

I'm very grateful to the Neural-GCMs team at Google Accelerated Sciences, and especially Peter Norgaard, Stephan Hoyer, and Dimitri Kochkov for their mentorship. Working with

you changed the way I think both as a programmer and a researcher.

I also want to acknowledge and thank Marc Alessi, Chris Terai, Veronika Eyring, Yibo Yang, Ruihan Yang, Ilan Koren, Tom Dror, Peter Blossey, Peter Caldwell, Claire Monteleoni, David Rolnick, Imme Ebert-Uphoff, and Maike Sonnewald for their advice, time, and contributions to these chapters.

Last but certainly not least I want to thank my friends and family for their love and support throughout the entirety of my Ph.D. To friends old and new who helped me through my twenties. And to my parents who always encouraged me to follow my passions instead of what made the most sense or money.

VITA

Griffin Mooers

EDUCATION

Doctor of Philosophy in Earth System Science — **2023**
University of California Irvine — *Irvine, California*

Bachelor of Science in Atmospheric Sciences — **2018**
Cornell University — *Ithaca, New York*

RESEARCH EXPERIENCE

Graduate Research Assistant — **2018–2023**
University of California, Irvine — *Irvine, California*

Student Researcher — **2022**
Google LLC — *Mountain View, California*

Research Intern — **2018**
Mount Washington Observatory — *North Conway, New Hampshire*

Research Assistant — **2017-2018**
Cornell University — *Ithaca, New York*

Research Intern — **2017**
Northeast Regional Climate Center — *Ithaca, New York*

TEACHING EXPERIENCE

Teaching Assistant — **2021–2022**
University of California, Irvine — *Irvine, CA*

Teaching Assistant — **2020**
California State University — *Los Angelas, CA*

Teaching Assistant — **2017-2018**
Cornell University — *Ithaca, NY*

REFEREED JOURNAL PUBLICATIONS

Comparing Storm Resolving Models and Climates via Unsupervised Machine Learning. (Under Construction) **2023**
To Be Submitted

Understanding Extreme Precipitation Changes through Unsupervised Machine Learning. (In Revision) **2023**
Environmental Data Science

Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. **2021**
Journal of Advances in Modeling Earth Systems 13, e2020MS002385. https://doi.org/10.1029/2020MS002385

Temporal Changes in the Areal Coverage of Daily Extreme Precipitation in the Northeastern United States Using High-Resolution Gridded Data. **2020**
Journal of Applied Meteorology and Climatology, 59(3), 551-565. https://doi.org/10.1175/JAMC-D-19-0210.1.

REFEREED CONFERENCE PUBLICATIONS

An Unsupervised Learning Perspective on the Dynamic Contribution to Extreme Precipitation Changes. **Dec 2022**
NeurIPS Climate Change AI

Analyzing high-resolution clouds and convection using multi-channel VAEs. **Dec 2021**
NeurIPS

Generative Modeling of Atmospheric Convection. **June 2020**
International Conference on Climate Informatics

ABSTRACT OF THE DISSERTATION

Improving The Modeling and Analysis of Tropical Convection and Precipitation through Machine Learning Methods

By

Griffin Mooers

Doctor of Philosophy in Earth System Science

University of California, Irvine, 2023

Associate Professor Mike Pritchard, Chair

Our knowledge of the atmosphere has increased immensely in the last few decades because of high-resolution "storm-resolving" climate models. With these models, we can simulate atmospheric processes including deep, moist convection with detail previously not possible giving us a more accurate representation of storms, precipitation, and atmospheric waves. However, limits continue to constrain our understanding of the dynamics of the atmosphere. We presently lack the ability to run these new storm-resolving models (SRMs) for the durations we need to understand the cloud-climate feedback. Meanwhile, running these SRMs for any amount of time produces very large volumes of data which are difficult to analyze properly. This work leverages disparate machine-learning approaches in an attempt to break through these deadlocks. First, we implement feed-forward neural networks to replace the computationally expensive explicit convection calculations within the "Super-parameterized Community Atmospheric Model" (SPCAM) allowing us to run the model at a fraction of the original computational cost but with the same accuracy even when realistic geographic boundary conditions are included. Second, we use deep generative models to analyze and organize SPCAM output. This allows us to identify unique types of convection as well as convective storm anomalies within the data. A third outcome involves expanding on this unsupervised learning work to compare different SRMs - including uniform resolution global

cloud resolving models - and quantify which have similar representations of the dynamics of the atmosphere. We find that even among high-resolution SRMs there are substantial differences in the type, proportion, and intensity of convection in representations of atmospheric dynamics. Fourth, we leverage these deep, generative machine learning models to make a novel metric of climate change and use it to better understand the physical mechanisms driving changes in extreme precipitation. We capture anticipated signals of global warming with minimal human intervention while showing the importance of the convection regime type to controlling the changing spatial patterns of heavy rainfall.

Chapter 1

Introduction

1.1 Background

1.1.1 Our Cloud-Climate Deadlock

Understanding Earth's atmosphere requires thinking not only about large-scale disturbances but the small scale physical processes behind the waves and structures. Though our observational records are (both spatially and temporally) incomplete, the climate science community has successfully constructed large scale "General Circulation Models" (GCMs) to represent large-scale circulation and its coupling to many subgrid processes. Agreement of these simulators' predictions with many observational constraints has yielded some confidence in our understanding of large-scale trends in mean temperature and precipitation while helping to fill in the gaps in our observational records.

While climate models are integral tools in understanding and preparing for the effects of climate change, the size and complexity of the Earth's atmosphere necessitate parameterizations (approximations) embedded in the models. For the past several decades, these imprecise

parameterizations of cloud processes have been linked to large biases and error bars on model prediction of key processes such as the intensity of extreme precipitation and tropical cyclones, and the frequency and distribution of low-level clouds, particularly over the equatorial regions [139, 37]. Thankfully, alternatives to the inaccuracy of parameterization of cloud processes are becoming available with the deployment of modern climate models that make fewer assumptions about unresolved convective processes [76, 82, 8].

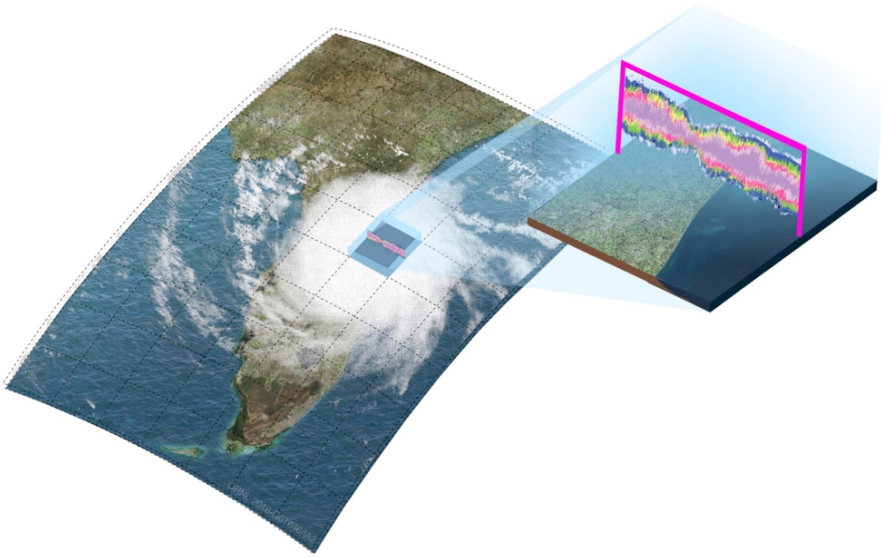

Figure 1.1: Visual of the design of a modern super parameterized climate model with embedded 2D fields in the host General Circulation Model to explicitly resolve deep, moist convection. Figure from the Energy Exascale Earth System Model Project.

For instance, the implementation of modern "storm-resolving" climate models (SRMs) has begun to transform our understanding of the way the atmosphere works. In SRMs, the use of kilometer-scale horizontal resolution allows for the explicit resolution of deep, moist convection and more accurate representation of storm patterns while simulation biases like the early diurnal onset of deep convection and precipitation over the tropics, which were once ubiquitous, are now reduced [139, 32, 37, 96, 95, 84].

However, there are still problems with these global SRMs that limit their utility for widespread use toward an increasingly comprehensive understanding of our atmosphere. First, the

computational cost of running these simulations for even short amounts of time is exorbitant; for long timescales we will need a 10^{11} increase in computing power [146]. This restricts our ability to run these simulations for the decades we would need to better understand global warming and narrow the uncertainty of the cloud-climate feedback. Meanwhile, even these Storm-Resolving Models still suffer some uncertainties and biases due to their continued need to parameterize sub-km processes including fine-scale turbulence as well as microphysical processes regulating precipitation formation, droplet growth, and descent [146].

But in addition to increased computational costs, a downside of these storm-resolving simulations is the sheer volume of SRM simulations' output required to resolve horizontal grid cells across the globe at this substantially higher spatial resolution. This simulation output from GRSMs over just a few weeks can quickly generate terabytes of data [152], becoming overwhelming to the human eye and for long-term storage. Traditional analysis approaches rely on scale selectivity and linear dimensionality reduction which can fail to capture non-linear relationships we know exist in the atmosphere [19, 168, 165].

1.1.2 The Tropical Atmosphere

Our present understanding of the atmosphere, and in particular the tropics is derived in the shadow of this gridlock between SRMs which resolve major details of the tropical atmosphere but cannot be run for long durations of time and General Circulation Models (GCMs) which fail to accurately represent details of convection and precipitation but can be run for decades [75].

Extracting detail about the tropical atmosphere from such simulations, especially in the era of SRMs, requires methods to reduce dimensionality that have led to helpful insights. Simple linear methods have been shown to be able to reduce coupled models of the atmosphere in a representative manner [19, 168]. Representing 2D fields of atmospheric information

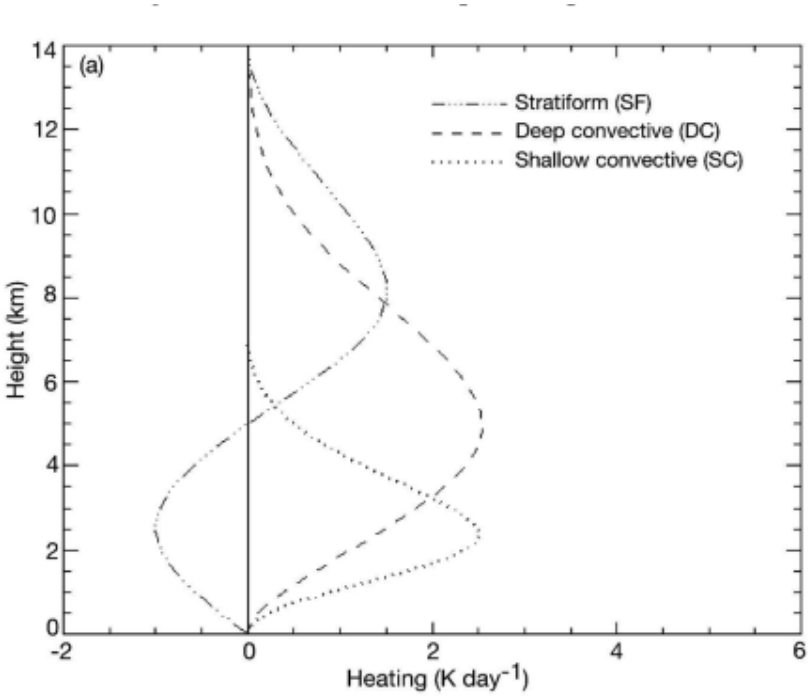

Figure 1.2: Idealized vertical modes of convection as derived in [148].

as just a few statistical moments in a summary profile is a common approach. Clustering convection by these summary statistics is considered a clean way to sidestep data volume in a high resolution model that is being analyzed. Just the basic vertical structure, through one [107, 104] or preferably two baroclinic modes [105, 58] of the troposphere derived from the shallow water equations allows for a multi-cloud partitioning that can produce tropical waves and emulate the observed structure of the atmosphere. Most of these large scale partitioning techniques (though not all [106]) identify three distinct modes of convection [70, 159, 135] that help elucidate the dominant patterns of weather in the tropics (Figure 1.2). While encouraging, the above approximations do not leverage the full benefits of the high resolution simulations available in SRMs and can ignore crucial spatial details needed to understand the details of small scale storm formation, growth, and in particular precipitation shifts that are sources of large uncertainty in climate simulations today.

Rainfall is among the most crucial weather phenomena on the planet for humans. It drives agriculture and development but can cause significant property damage and threats to life.

Its variable amount largely drives regional differences across Earth's climate [131]. Since predicting rainfall in numerical simulations is intimately linked to assumptions drawn about formulating deep convection, it remains highly challenging to predict [84]. Climate models show large uncertainty about changing precipitation patterns, the role of convective memory, precipitation extremes, and even the timing of the diurnal cycle of precipitation [75, 37, 134].

When we consider the scale of tropical convection and precipitation in relation to the size of the Earth's atmosphere as a whole, these modeling problems shouldn't be surprising. Rainfall is generated and driven by a myriad of factors including latent heat transport, convection, and transport of heating and moisture, microphysics, and land-sea interactions. Many of these processes occur over meters to just a few kilometers, while the typical climate model has a horizontal resolution of 50 to 100 kilometers, meaning that the physics behind rainfall and convective processes must be parameterized. We hence require tools to better model and understand these crucial atmospheric processes that don't require infeasible amounts of computational resources.

1.1.3 Machine Learning

Given the present (and anticipated future [146]) limitations on understanding the atmosphere via traditional GCMs and SRMs, this work investigates the utility of Deep Learning (multi-layer neural networks), to enhance climate model performance, analyze large simulation outputs, and contribute to our understanding of the tropical atmosphere all within the computational limits we are governed by today.

The idea of using neural networks to advance climate science is not in and of itself a novel idea but and rather a topic of debate for decades. Since the 1990's, attempts to use primitive neural networks and supervised learning frameworks for cloud classification and data-driven exploration of key climate information have been attempted [40, 91]. However, much of this

early work was unsuccessful, with limited training data and neural network architectures failing to properly extract features from the weather data.

At the same time, there have been efforts to embed neural networks into climate models themselves. The concept of replacing sub-grid convective parameterizations in GCMs with much (computationally) cheaper neural networks has long been tantalizing [50]. This would unlock the potential to run high-resolution climate models which explicitly resolve key atmospheric processes for decades and narrow the cloud climate feedback. But previous attempts at this engineering challenge also have hit roadblocks. Early neural networks were often not optimized to the task at hand in their design, lacking the necessary complexity and ability to generalize across climate data [87, 66, 86].

However, more recently there have been significant strides toward the successful deployment of artificial intelligence in climate and atmospheric sciences. In seasonless, aqua-planet climate simulations, hybrid-climate modeling powered by machine learning was proven possible by manually tuned Deep Neural Networks (DNNs), even in online "prognostic" modes in the Community Atmospheric Model (CAM) [50, 140]. Simultaneously, rapid advances in machine learning architectures, and in particular generative approaches [52] have set the stage for more successful feature extraction of high-resolution climate model output and observational data as well as a framework to replace stochastic parameterizations [40, 91].

These advances are particularly timely given the recent increase in the number of high-resolution simulations and observational data sets of the Earth. With these massive simulations, we have treasure-troves of information about the atmosphere, but lack techniques to analyze it to its fullest potential, particularly through more objective, data-driven methods. A common approach up to this point has been Principle Component Analysis (PCA or Empirical Orthogonal Functions (EOF)) – this method has the advantage of reducing the data to just a few, easy to visualize dimensions [165]. Recent applications of PCA to high-resolution satellite data yield distinct cloud patterns in just a couple of interpretable

dimensions [41, 67]. Basic unsupervised clustering algorithms, particularly K-Means Clustering, have also shown novel analysis potential for geo-spatial data recently [151]. While convenient, these simpler approaches can sacrifice crucial non-linear relationships resulting in important information loss critical in the full representation of stochastic processes like convection [165]. We envision the potential for deep, generative machine learning models to provide novel dynamical analysis of high resolution, multi-variate climate simulations not tractable by the more conventional approaches outlined above. Rudimentary Autoencoders (AEs) used on Geostationary Operational Environment Satellite (GOES-16) imagery have already demonstrated the ability of machine learning methods to find new structures and regimes in climate data [40, 91]. But we believe that other unsupervised machine learning models, in particular, Variational Autoencoders (VAEs) remain under-exploited for use on SRMs and large observational data.

With the hope of building on the climate science and machine learning advances described above, we developed the following research questions:

- (Chapter 2) To what extent can feed-forward neural networks skillfully replace sub-grid convective parameterizations when trained beyond aquaplanets in situations of Earth-like complexity?
- (Chapter 3) Can the generative machinery of Variational Autoencoders organize a high-resolution convection simulation's output in an unbiased manner helpful for objective analysis?
- (Chapter 4) If so, do the resulting learned embeddings provide a novel way to measure differences in dynamics of convection across the massive simulation outputs of various SRMs with minimal assumptions?
- (Chapter 5) Building on this foundation, is it possible to gain new insight into how extreme precipitation will change in the future, including the complexity of changing

convective regime structures?

1.1.4 Outline

To address our objectives, we introduce several models, techniques, and simulation datasets. More concretely:

- Simulation output from the Super Parameterized Community Atmospheric Model (SPCAM).
  - CAM version 3 (Chapter 2, baseline model output for comparison).
  - SPCAM version 3 (Chapter 2, aquaplanet simulation).
  - SPCAM version 5 (Chapters 2, 3, 4, 5 realistic Earth-like simulations).
  - SPCAM version 5 with sea surface temperatures warmed by +4 Kelvin (Chapter 5, approximated climate change).
- 8 SRM simulations from the DYAMOND Initiative including SAM, ICON, NICAM, UM, GEM, IFS, ARPEGE, and SHIELD (Chapter 4).
- Feed-forward neural networks composed of densely connected stacks (Chapter 2).
- Fully convolutional Variational Autoencoders (Chapters 3, 4, 5).
- Statistically constrained Variational Autoencoders (Chapter 3).
- Dimensionality reduction techniques and distribution shift measurements (Chapters 4 and 5).

Each chapter covers a different way machine learning can be utilized to contribute to climate and atmospheric sciences. Much of the work is intentionally explorative of emerging machine

learning algorithms with breakthrough potential, reflecting an era in which accelerating developments in computer sciences are disrupting multiple fields of science. But the ultimate intent of the work is to seek new scientific foundations for unbiased analysis and novel physical discovery.

We start with Chapter 2 from the basic idea that a hybrid modeling framework can be expanded beyond simple aquaplanets [50, 140] to a more Earth-like atmosphere with the influence of continents and seasons. In the physical host climate model, the sub-grid scale convective parameterizations we wish to emulate with a DNN are cast locally in space and time. This means that even though the learning task is more difficult, we must rely on a simple feed-forward neural network to replace the sub-grid parameterizations. We train this neural network on large volumes of SPCAM5 data and tune the network with a sophisticated, semi-automated sweep to find the optimal hyper-parameters for the best fit. We find that with this sufficient training data and tuning that the neural network is able to emulate the sub-grid parameterizations even with the added complexity of the diurnal heating cycle, land-sea interactions, and seasonal oscillations. Convective Memory does not seem to be a necessary input to the neural network in order to emulate deep convection and storms.

Chapter 3 begins to move beyond the use of neural networks as engineering tools for improving climate models, by beginning to deploy machine learning methods for novel analysis. To accomplish this, we take a different output from SPCAM; the embedded high-resolution 2D vertical velocity fields. This information serves as a proxy for convective dynamics, for machine learning-assisted dynamical analysis. Instead of simple neural networks, we rely on unsupervised learning and Deep Generative Models for analysis of this high-resolution convection data. We train the first-ever Variational Autoencoders (VAEs) to be exposed to high-resolution convection data from a convection permitting climate model. The purpose is scientific interpretability. VAEs are widely used for density estimation and non-linear dimensionality reduction [80]. In a low-dimensional "latent" representation, we will show

VAEs can disentangle confounding aspects of convection by differentiating intensity, vertical structure, turbulent length scale, and more. The latent space can also be used to track the diurnal cycle of deep convection. The VAE, through its density estimation properties, can also assist in anomaly detection by detecting unusual storms and convective formations among large volumes of vertical velocity fields. The next goal is to quantify and focus this data-driven analysis for both model improvement and physical discovery.

Chapter 4 addresses the first goal raised by the analysis in Chapter 3 by building an objective framework to inter-compare different SRM simulation outputs. In addition to the SPCAM data previously discussed, we introduce 8 different SRMs – fully global storm resolving models - from a modern intercomparison archive and train VAEs for all simulations. We design a method to compare the low dimensional latent representations both qualitatively through visual inspection and more formally through vector quantization to measure "Distribution Shift" differences. We find that across the pool of available high dimensional SRMs there is only a subset that exhibits convergent properties, agreeing on the type, intensity, and the proportion of convection properties simulated in the tropical atmosphere; a secondary group is in disagreement. We hope these findings can help elucidate the consequences of parameterization choices in SRMs and help to inform how these models are designed for the future.

Finally, we turn to the possibility of physical discovery through machine learning in Chapter 5. We again return to the use of SPCAM5 data, but now introduce an additional dataset warmed by +4K as a proxy of global warming. The same distribution shift approach innovated in Chapter 4 is applied to studying changing climate, and reveals spatial shifts in regimes of convection that are able to explain the changes in Extreme Precipitation between the "Control" and the "Warmed" climates. Changes in dynamical regimes dominate over changes linked to other variables when it comes to the spatial variability of extreme precipitation. The Variational Autoencoder was able to leverage the rich spatial information provided by

the simulation output to derive new insights into the behavior of extreme precipitation with global warming on both local and regional scales.

The results from the chapters are summarized at the end of the thesis in Chapter 6 along with ideas for future research and reflections on the future of machine learning in atmospheric sciences. Chapter 2 has been published in the *Journal of Advances in Modeling Earth Systems*. Chapter 3 was part of the *2020 Climate Informatics Workshop* before being accepted for publication in the *Association of Computing Machinery*. Chapter 5 was accepted for publication and is currently in revision at the *Journal of Environmental Data Sciences*. Chapter 4 is on the verge of submission.

Chapter 2

Neural-Network Emulation of Sub-grid Parameterizations

2.1 Abstract

We explore the potential of feed-forward deep neural networks (DNNs) for emulating cloud superparameterization in realistic geography, using offline fits for data from the Super Parameterized Community Atmospheric Model. To identify the network architecture of greatest skill, we formally optimize hyperparameters using ∼250 trials. Our DNN explains over 70 percent of the temporal variance at the 15-minute sampling scale throughout the mid-to-upper troposphere. Autocorrelation timescale analysis compared against DNN skill suggests the less good fit in the tropical, marine boundary layer is driven by neural network difficulty emulating fast, stochastic signals in convection. However, spectral analysis in the temporal domain indicates skillful emulation of signals on diurnal to synoptic scales. A close look at the diurnal cycle reveals correct emulation of land-sea contrasts and vertical structure in the heating and moistening fields, but some distortion of precipitation. Sensitivity tests

targeting precipitation skill reveal complementary effects of adding positive constraints vs. hyperparameter tuning, motivating the use of both in the future. A first attempt to force an offline land model with DNN emulated atmospheric fields produces reassuring results further supporting neural network emulation viability in real-geography settings. Overall, the fit skill is competitive with recent attempts by sophisticated Residual and Convolutional Neural Network architectures trained on added information, including memory of past states. Our results confirm the parameterizability of superparameterized convection with continents through machine learning and we highlight the advantages of casting this problem locally in space and time for accurate emulation and hopefully quick implementation of hybrid climate models.

2.2 Introduction

Although global atmospheric model simulations are increasingly high-resolution, even under optimistic scenarios of enhanced computing performance, physically resolving the atmospheric turbulence controlling clouds will likely not be feasible for decades. Current climate model horizontal grid cells are typically 50-100 kilometers wide but the turbulent updrafts governing cloud formation occur on scales of just tens to hundreds of meters and the microphysical processes regulating convection occur down at the micro-meter scale [146, 18, 110]. This discrepancy creates large uncertainties about the precise details of deep convection on cloud feedbacks and climate change [20]. Multi-scale methods such as embedding two-dimensional Cloud Resolving Models (CRMs) into General Circulation Model (GCM) grid cells (superpararmeterization) have been used to directly resolve the spatial and temporal progression of moist convection. More recently, explicit kilometer-scale simulation of moist convection has improved the representation of deep convective clouds and the hydrological cycle [139, 146, 95, 32, 37]. These advancements allow models to simulate historically

challenging atmospheric modes of variability like the observed afternoon maxima of deep convection over continents and a more realistic probability distribution of precipitation that captures extremes on the tail-end [84, 96]. However, even the highest resolution global CRM simulations today require some assumption-prone parameterization for microphysics and sub-km turbulence, among other cloud processes [150, 29], although multi-scale algorithms still hold promise for making some of this explicitly tractable [68].

Given these dual physical and computational hurdles, using machine learning emulators to replace sub-grid convective physics in coarse-resolution climate models is an area of rapidly increasing interest. Following seminal works including [85, 88, 81], recent breakthroughs from global aqua-planet simulations have provided a proof of concept for hybrid climate models powered by machine learning. [50] showed 40-100M samples taken from a zonally symmetric aqua-planet simulation were sufficient to train a five to ten layer DNN to emulate superparameterized convective heating and moistening in a hold-out test dataset, with R^2 greater than 0.7 in the mid-troposphere. Building on these results, [140] demonstrated that a similar DNN could even be run in a prognostic setting, coupled to an advective scheme in the Community Atmosphere Model (CAM), thus generating accurate mean climate states and equatorial wave spectra at as low as five percent of the computational cost of actual superparameterization. Recently, [121, 175] showed that similar prognostic success can occur in idealized aqua-planets trained on coarse-grained three-dimensional output using Random Forests (RFs). These RFs used refined inputs and outputs tailored to the prognostic variables underlying the System for Atmospheric Modeling or SAM, which is the embedded storm-resolving model used in the SPCAM multimodel framework. Whereas all of the above studies have focused on aqua-planets, skillfully replicating convection in more complex, realistic settings is a key step towards building a replacement for traditional sub-grid parameterizations of deep convection in climate models.

Achieving competitive emulation of convection under realistic geography may be a much

more significant hurdle for neural networks. At the time of this writing, a first pioneering attempt has been made to fit superparameterized convection in a realistic operational setting. Results of this Chapter indicate that sophisticated network designs involving the addition of 1D convolutions in the vertical dimension, and 'residual" neural network architecture [60] using state information from previous time steps, appeared critical to achieving reasonable fits [59]. This raises two issues. From a practical perspective, implementing neural networks that rely on prior temporal information (such as the Resnet in [59]) as coupled components of a host climate model is technically challenging since previous timesteps are not typically passed to the physics parameterization. From a philosophical perspective, this questions casting machine learning parameterizations of convection locally in time. If confirmed, these two issues make the full potential of neural network (NN) emulators as a tool to advance scientific understanding beyond aqua-planets substantially harder to utilize.

On this basis, we explore whether feed-forward DNNs are capable of emulating convection with real-geography if sufficient hyperparameter tuning is taken advantage of when training our neural network. The hypothesis that even a feed-forward neural network can emulate superparameterized convection with realistic geography is based in part off the results of [123] on aqua-planets in which formal, expansive hyperparameter tuning was identified as essential for the performance of a DNN to emulate convection. We focus on emulating superparameterization to avoid ambiguities due to coarse-grained uniformly-resolved CRM output [23]. We readily acknowledge that other methods such as RFs [176] also show success and promise in sub-grid convection emulation. Indeed, RFs have some advantages, including automatically respecting physical constraints that are linear in their outputs, such as energy conservation, and positive-definite precipitation, which are not guaranteed in DNNs [176]. Furthermore, [163] has demonstrated that RFs can be used to correct, or "nudge" parameterizations and reduce simulation errors even for realistic convection. However, there are also ways to enforce such constraints in DNNs [16, 90], and RFs also come with disadvantages. To cite a few, RFs with deep trees quickly become computationally expensive

for large datasets, requiring large storage capacity which could prevent taking full advantage of Graphics Processing Unit (GPU) infrastructure [176]. RFs may struggle to capture local patterns in the atmosphere as well [163]. For these reasons, we leave RFs for future work.

Here, our task is to understand what convective patterns, cycles, and modes of variation in a realistic setup of superparameterized convection can be fitted with a feed-forward DNN. We additionally aim to establish a set of post-processing metrics to benchmark our own neural network's performance and transparently compare different neural network emulators trained on similar data. Section 2.3 outlines the details of our simulation dataset, introduces the design of a neural network, and describes our automated hyperparameter tuning algorithm, capable of finding a reasonable fit. Then, in Section 2.3.3, we lay out our test benchmarks. In Section 2.4 we present the spatial and temporal breakdown of our neural network predictions for parameterized convective tendencies. We also analyze the plausibility of the neural network emulated hydrological cycle in detail. The last part of Sections 2.4 examines the potential to couple an aqua-planet trained neural network to a land model as another credibility test towards a hybrid climate model. Section 2.5 includes a summary of our work, its limitations, and potential directions for future research.

2.3 Methods

2.3.1 Climate Simulation Data

We leverage three different datasets to train, test, and benchmark DNNs emulating convection with real-geography. The data are based on the Super Parameterized Community Atmospheric Model (SPCAM), a global climate model that nearly explicitly resolves atmospheric moist convection by using idealized embedded CRMs [139, 54]. Each of the host GCM's grid cells embed two-dimensional CRMs of optional horizontal resolution and physical extent, thus

avoiding heuristic parameterization of sub-grid moist convective processes [139, 12].

For a point of reference, we first use outputs from SPCAM v.3 (SPCAM3) at T42 spectral truncation (i.e. 8,192 horizontal grid cells) driven with boundary conditions of a zonally symmetric aqua-planet; as in [140]. We then build beyond previous aqua-planet emulation studies by generating a new dataset from a more modern version (v.5) of SPCAM (SPCAM5) that includes higher horizontal resolution (1.9x2.5 degree finite volume dynamical core, i.e. 13,824 grid cells) and in which we incorporate realistic boundary conditions, including a land surface model, seasonality, and a zonally asymmetric annual climatology of sea surface temperatures (SSTs). The dataset is similar to one recently used in [59] but with a few differences. The simulation itself is 10 years long, but selectively sub-sampled to every 10 days to avoid temporally autocorrelated training samples. We also rely on a shorter GCM timestep (15 minutes as opposed to 20) (Table 2.1). As in [50, 140], but unlike [59], we make the further simplification of using a reduced (32-km) CRM horizontal extent, i.e. CRMs with only 8-columns apiece, instead of the 128-km / 32-column CRM configuration (Table 2.1). This decision, based on [137]'s finding that (for deep convection) small CRM domains do not corrupt the representation of tropical wave dynamics in SPCAM. This also has the advantage of simplifying the comparison of our results to [50, 140]. Meanwhile, there are reasons to think it may facilitate DNN emulation [22, 123]. The codebase for running the "SPCAM5" simulations is the same employed by [129], which is archived at `https://github.com/mspritch/UltraCAM-spcam2_0_cesm1_1_1`; this code was in turn forked from a development version of the CESM1.1.1 located on the NCAR central subversion repository under tag `spcam_cam5_2_00_forCESM1_1_1Rel_V09`, which dates to February 25, 2013. Finally, for reference, we analyze output from the conventionally parameterized version of CAM5; this helps assess the emulation of Superparameterization compared to conventional parameterization.

Simulation Datasets			
Details	CAM5	SPCAM3	SPCAM5
Spatio-temporal resolution	1.9°×2.5°×15 min	2.8°×2.8°×30 min	1.9°×2.5°×15 min
Total Number of Days Simulated	93	93	3,650
Total Number of atmospheric columns Simulated	123,420,672	36,569,088	4,843,929,600

Table 2.1: Details of the three datasets used for benchmarking the results of our DNN trained on real-geography data.

DNN Setup				
Input	Size	Output	Scaling factor	size
Temperature (K)	30	Heating Tendency (K/s)	c_p	30
Specific Humidity (kg/kg)	30	Moistening Tendency ($kg/kg/s$)	L_s	30
Surface Pressure (hPa)	1	TOA LW Flux (W/m^2)	-1e-3	1
Solar Insolation (W/m^2)	1	Surface LW Flux (W/m^2)	1e-3	1
Sensible Heat Flux (W/m^2)	1	TOA SW Flux (W/m^2)	-1e-3	1
Latent Heat Flux (W/m^2)	1	Surface SW Flux (W/m^2)	1e-3	1
		Precipitation (m/s)	1728000	1

Table 2.2: Details of the input and output vectors to the DNN. c_p refers to the specific heat capacity of air at a constant pressure and is assumed to be 1.00464e3 ($J/kg/K$) and L_s is the latent heat of sublimination of water in standard atmospheric conditions calculated by adding the latent heat of vaporization 2.501e6 (J/kg) and the latent heat of freezing 3.337e5 (J/kg). Precipitation is weighted by the same prefactor, 1728000, also used in [140] to ensure it is felt in the loss function of the DNN.

2.3.2 Neural Network Design

We design a DNN that takes the same inputs as standard convection parameterizations in CAM to predict sub-grid scale tendencies at each vertical level and across each timestep globally. The neural network inputs can be thought of as atmospheric thermodynamics components in the eight year SPCAM5 data training simulation including: both temperature (K) and specific humidity (kg/kg) for each of the 30 vertical levels spanning the column, as well as surface latent heat flux (W/m^2), surface sensible heat flux (W/m^2), top of atmosphere (TOA) solar insolation (W/m^2), and surface pressure (hPa). By including surface pressure in the input vector, we allow the neural network to fit horizontal variations in the vertical pressure grid, which is based on a hybrid terrain-following coordinate [116]. Concatenating these state variables creates an input vector to the neural network of length 64. Each of the input variables was pre-normalized before exposure to the neural network by subtracting its respective mean and dividing by its respective range, with these statistics computed and applied separately for each vertical level in the case of the vertically-resolved temperature and humidity profiles (Table 2.2). The reason we divide by the range instead of the more traditional standard deviation, in line with the methods of [140], is to avoid dividing by near-zero numbers, e.g. in the case of stratospheric humidity. Some previous aqua-planet experiments also used the meridional wind vertical profile as part of the input vector to the neural network, but it was omitted in this case as preliminary neural network tests indicate it had an insignificant effect on the skill of the trained network while increasing the input vector length by 30 and thus substantially increasing training time [140, 50]; we note that [59] also deem this an avoidable input.

Our DNN ultimately predicts the sub-grid scale time tendency of temperature (K/s) or heating tendency for short, which includes the sub-grid advection of temperature by convection and fine-scale turbulence, as well as grid average radiative heating throughout the column. It also predicts the sub-grid scale time-tendency of specific humidity throughout the column

($kg/kg/s$) - or moistening tendency for short. A scalar is predicted for precipitation (mm/day) as well as for the long and shortwave net radiative fluxes (W/m^2) at both the surface and the TOA. This fully concatenated output vector is of length 65 (Table 2.2). The state variables that comprise the output vector have different units, making the ultimate Mean Squared Error (MSE) of the neural network devoid of physical meaning. To ensure that all of the predicted variables have comparable magnitude and can be felt in the optimizer, we apply multiplicative prefactors as in [140], recognizing that other choices can also be made such as additionally weighting by the mass of each pressure level [16].

2.3.3 Performance Analysis and Postprocessing

To assess the skill of our DNNs after training, we benchmark them against an offline hold-out test dataset with multiple metrics. This is a first step to determine whether our DNNs could be candidates for online coupling, which we leave for future work. How well a neural network emulator appears to perform is in part a reflection of statistical analysis choices. Multiple conventions have been used and the degree of spatial averaging before applying error statistics has not been sufficiently reported to do inter-study comparison confidently, though spatial averaging is common practice [59, 140, 23]. In some cases, snapshots of unaveraged data [140] have helped reveal issues at the finest resolved scales while zonally averaged temporal standard deviations [59] have helped reveal issues in emulation of convective tendencies at small time intervals and spatial scales in neural network fits. Precipitation time series and Probability Density Functions (PDFs) [140, 59, 121] have also been used to assess neural network performance.

In our case, to examine the magnitude of the error between the neural network prediction and the SPCAM5 target data, which we treat as truth, we will calculate a sum of squared errors (SSE) separately for each longitude and latitude and, in the case of 3D variables, vertical

level (based on the hybrid, terrain following sigma coordinate):

$$\mathrm{SSE} \stackrel{\mathrm{def}}{=} \sum_{\mathrm{j}=1}^{\mathrm{N_t}} (\mathrm{y_j} - \hat{\mathrm{y}}_\mathrm{j})^2 \tag{2.1}$$

where N_t is the length of the time series, y is the target SPCAM5 data, and $\hat{y}$ is the corresponding neural network predicted value based on coarse-grained variables. In this case, we examine the performance of the neural network predicting heating and moistening tendencies. The primary metric for assessing DNN prediction and the associated spatial error structure is the coefficient of determination, R^2, defined as:

$$\mathrm{R}^2 \stackrel{\mathrm{def}}{=} 1 - \frac{\mathrm{SSE}}{\sum_{\mathrm{j}=1}^{\mathrm{N_t}} (\mathrm{y_j} - \bar{\mathrm{y}})^2}, \tag{2.2}$$

where $\bar{y}$ is the temporally-averaged heating or moistening tendency at a given latitude, longitude, and vertical level.

We apply R^2 to data entirely unaveraged in the latitude, longitude, and pressure. When visualizing averaged R^2, we often use a latitude-longitude cross sections at specific vertical levels to reveal error structure at the native 15 minute sampling interval (Figure 2.2). In other portions of the analysis, we use spatial averaging prior to the error calculation, as in [59]. For instance, in pressure-latitude cross-sections (Figures 2.1, 2.3), we first zonally average the predictions and targets, before computing R^2 over the time dimension. Furthermore, we examine R^2 at two different temporal resolutions: the native model timestep (15 minute sampling; Figures 2.1, 2.2, 2.6) since the strong diurnal cycle over land regions could bias the analysis between land and ocean regions, and then visualize with temporal averaging reducing the data to daily means (Figure 2.3).

We also wish to elucidate whether there is any detectable "mode-specific" performance, i.e. certain temporal patterns that are especially predictable such as the diurnal cycle over continents in our moist convection emulation. To that end, we calculate the temporal Power Spectral Density (PSD) for a single month (July):

$$\mathrm{PSD_k} \stackrel{\mathrm{def}}{=} \frac{2\Delta \mathrm{t}}{\mathrm{N_t}} \left| \mathcal{F}\left(\mathrm{y}\right)_{\mathrm{k}} \right|^2, \tag{2.3}$$

defined as the square complex modulus of the Fourier transform:

$$\mathcal{F}\left(\mathrm{y}\right)_{\mathrm{k}} \stackrel{\mathrm{def}}{=} \sum_{\mathrm{j}=0}^{\mathrm{N_t}-1} \mathrm{y_j} \mathrm{e}^{\frac{-2\pi \mathrm{ijk}}{\mathrm{N_t}}}, \tag{2.4}$$

where y is a time series of values of convective heating or moistening tendency at a given location, Δt is the sampling time interval, and i is the imaginary unit so that $i^2 = -1$ [35]. The PSDs of heating and moistening tendencies are analyzed both regionally and globally as follows: we mass weight each vertical level and then a PSD value is calculated for each frequency bin at each latitude, longitude, and pressure grid cell. We focus on timescales up to a month to examine variations in convection ranging from sub-diurnal to synoptic timescales. Next, the spectral coefficients at each latitude, longtiude, and pressure level are combined into a single, averaged PSD for the globe. We repeat this same analysis twice more, once with a land mask and once with an ocean mask. We also perform corresponding spectral analysis in the spatial domain i.e. calculating the PSD as a function of zonal and meridional wavenumbers, separately for every vertical level and model time-step over the same month of July. For the zonal spectrum, we restrict our average to just tropical locations from 20S to 20N and weight by the cosine of latitude to make an approximate Cartesian plane assumption. This enables us to sidestep the unequal grid spacing that would be a problem in this analysis

if we included the mid-latitudes in our spatial average.

Finally, to hone in on a regime in the lower tropical atmosphere that our R^2 analysis suggests is especially difficult for our DNN to emulate, we will analyze the temporal autocorrelations of the sub-grid scale tendencies. Our goal is to understand the regions where the DNN emulation of superparameterized convection is detectably worse than the global average performance. As a proxy for the "stochasticity" of atmospheric motions, we calculate the autocorrelation function (ACF) - a measure of the self similarity between a given signal and a delayed version of itself - using the previously calculated PSD:

$$\mathrm{ACF_j} \stackrel{\mathrm{def}}{=} \frac{1}{2\Delta \mathrm{t}} \sum_{\mathrm{k}=0}^{\mathrm{N_t}-1} \left(\mathrm{PSD_k} \times \ \mathrm{e}^{\frac{2\pi \mathrm{ijk}}{\mathrm{N_t}}} \right) \tag{2.5}$$

Fast signals that decorrelate quickly are more likely to be of stochastic nature. We thus compare the time to e-folding decay in ACF with R^2 skill score in the planetary boundary layer to test for correlations between DNN skill and the timescale of dominant atmospheric signals (diurnal cycle, Rossby waves) visible in vertically resolved heating and moistening tendencies (Equation 2.3). We also use the inverse of the e-folding decay timescale, which we will refer to as the "autocorrelation frequency", to examine the patterns between R^2 coefficient of determination globally and the stochasticity of the dominant convective signals. This comparison offers a possible explanation for much of the variations in the performance of our DNN throughout the planetary boundary layer.

To better quantify the differences between true and predicted PSDs, we rely on the Log-Spectral Distance (D) [162]:

$$\mathrm{D} \stackrel{\mathrm{def}}{=} \sqrt{\frac{1}{\mathrm{N_t}} \sum_{\mathrm{k=1}}^{\mathrm{N_t}-1} \left[\ln \left(\frac{\mathrm{PSD_{true_k}}}{\mathrm{PSD_{pred_k}}} \right) \right]^2} \tag{2.6}$$

where the summation is done in frequency space. We examine the neural network performance not just for the convective heating and moistening tendencies but also for precipitation, for which we both calculate the PDF and global error between DNN predicted precipitation and SPCAM5 target data. Furthermore, we determine the diurnal timing of maximum precipitation globally, but we seek to filter out noise in the mid-latitudes. To that end, when determining the hour of maximum precipitation rate, we look only at locations that pass the following threshold:

$$\max\left(\mathrm{Precip}\right) - \min\left(\mathrm{Precip}\right) > \frac{2}{\sqrt{N_{\mathrm{day}}/4 - 1}} \max\left[\mathrm{std}\left(\mathrm{Precip}\right)\right] \tag{2.7}$$

where *Precip* refers to the model output precipitation rate in mm/day, N_{day} refers to the number of days examined and the max refers to the local maximum of the precipitation in the temporal dimension at a given latitude, longitude grid cell. Our assumption is that the effective degrees of freedom in the diurnal composite is 1/4 of the apparent degrees of freedom. More empirically, if we relax the threshold anymore we detect unrealistic signals in the marine zones of the mid-latitudes.

2.3.4 Formal Hyperparameter Tuning

In several previous studies, small volumes of training data (as low as three months) and manual hyperparameter tuning were sufficient to achieve acceptable Machine Learning (ML) emulator performance [140, 50]. Here, we make the hypothesis that with real-geography

boundary conditions, neural networks benefit from considerably more training data and formal hyperparameter tuning [123]. To fully exploit our 10-year simulation, we split it into a training data set spanning the first eight years, a validation data set spanning the ninth year, and a test data set spanning the tenth year.

As a first step we subsampled by a factor of ten after sensitivity tests (not shown) indicated little difference in the fit skill from manual tuning attempts, likely due to redundant information from temporally autocorrelated state data. This subsampling in our preprocessing reduced the training data volume to a size that could be managed on a single GPU. Our initial architecture was inspired by previous literature, i.e. composed of five fully connected layers with 256 nodes each. However, this manual configuration yielded poor performance (Figure 2.1), and other manual attempts to explore alternative choices of hyperparameters and learning rate variations were likewise unsuccessful (not shown). A higher fit skill is desirable before undertaking the difficult task of coupling the neural network to the host climate model for online analysis.

We attained much better results after adopting a formal hyperparameter tuning. Automated neural network architecture searches have just begun to prove their value in climate modeling - both for optimizing offline fits [16] and even prognostic online coupled performance [123]. Using similar approaches, we implemented a resource-intensive automated DNN training process, conducting a formal search over the following hyperparameters: batch normalization, dropout, LeakyReLU coefficient [99], learning rate, learning rate decay, number of layers, nodes per layer, and the optimizer [79]. All parameters and their corresponding ranges for the search are shown in Table 2.3.

This hyperparameter search took place in two stages, using "Sherpa" [62], a Python library for hyperparameter tuning. First, we fit a large suite (over 200) candidate DNN models using a random search algorithm. The random search has the advantage of making no assumptions about the network architecture or the task of interest. In this stage, all hyperparameters,

Name	Range	Parameter Type	Best Model
Batch Normalization	yes, no	Choice	yes
Dropout	[0., 0.25]	Continuous	0.01
LeakyReLU	[0., 0.4]	Continuous	0.15
Learning Rate	[0.00001, 0.01]	Continuous (log)	0.000227
Learning Rate Decay	[0.5, 1.]	Continuous	0.91
Number of Layers	[3, 12]	Discrete	7
Number of Nodes	128, 256, 512	Choice	512
Optimizer	Adam, SGD, RMSProp	Choice	Adam

Table 2.3: Hyperparameter Space. The resulting best model configuration is shown in the right-most column.

except the learning rate and learning rate decay, are modified. Excluding learning rate parameters in the first stage is strategic to ensure that any increases in performance are due to more skilled architectures.

Following the initial search, we conducted a second search on the best performing model uncovered during the first stage. This secondary investigation, which tested another fifty models, focused exclusively on the learning rate and learning rate decay settings. This procedure allowed us to train the network with the best possible learning schedule so as to maximize the network's performance while fixing the best-performing architecture uncovered in the first stage.

In total, we tested more than 250 network architectures. We noticed a dramatic improvement in performance from the hyperparameter search quantified by the difference between the initial model's MSE and the MSE of the stage 2 model. We also observed the benefit of tuning the learning rate and the learning rate decay in stage two. The validation loss of the stage 2 model descends smoothly and consistently compared to the more archaic original model or stage 1 model. The final result of the hyper-parameter search is shown in Table 2.3. We discuss below in the results section the extent to which hyperparameter tuning improves the benchmarks discussed above. To help quantify the improvements from the formal tuning we compare this "Best" DNN that was the result of the formal hyperparameter search against the

"Manual" DNN that was designed similarly to neural networks used in previous aqua-planet studies [140]. Additionally, we run tests on a baseline "Linear" model that is identical in all ways to the "Manual" DNN except for the fact that all activations are replaced with the identity function prior to training.

While it would be interesting to know whether skillful models could have been obtained with less data volume, this is impossible to precisely quantify without performing Sherpa hyperparameter tuning on a smaller dataset – something we opted not to do due to the heavy GPU requirements of applying Sherpa. We strategically only utilized it on our richest training dataset to conserve resources. For context on the resource requirement, each candidate neural network architecture required roughly twenty-four hours to train (about one hour per epoch). Eight models could be run in parallel on a single GPU and thus, using four GPUs, we could train about thirty two models per day. In total, with four GPUs (12 GB memory each), it took eight days to train all 250 models.

2.4 Results

Here we use the diagnostics outlined in Section 2.3.3 to benchmark the performance of our DNNs. We quantify the overall performance of our DNNs in emulating atmospheric sub-grid heating and moistening tendencies in Sections 2.4.1, 2.4.2, and analyze the emulated hydrological cycle in Section 2.4.3. Note that since we used the first eight years to train the network and the ninth year to optimize the hyperparameters, we benchmark the performance of our DNN on the remaining tenth year that we held out for testing.

Label	Training Data	Region	Variable	timestep	25th	50th	75th
a	aqua-planet	Ocean	Heating	15 min.	0.05	0.27	0.55
b	real-geog. (Linear)	Ocean	Heating	15 min.	-0.30	0.00	0.21
c	real-geog. (Linear)	Land	Heating	15 min.	-0.93	-0.06	0.25
d	real-geog. (Manual)	Ocean	Heating	15 min.	-0.26	0.00	0.31
e	real-geog. (Manual)	Land	Heating	15 min.	-0.93	-0.06	0.35
f	real-geog. (Best)	Ocean	Heating	15 min.	0.28	0.54	0.76
g	real-geog. (Best)	Land	Heating	15 min.	0.41	0.65	0.82

Table 2.4: Statistical breakdown of skill score. We show quartiles of the skill distribution in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw convective heating tendency data at the 15-minute sampling scale. We compare a neural network trained on aqua-planet data (a) with three different neural networks trained on more realistic SPCAM5 data. These models include a baseline "Linear" model (b-c), a manually tuned neural network (d-e), and a neural network formally tuned by Sherpa (f-g).

2.4.1 Spatial Structures

Differences between the aqua-planet and full complexity real-geography test beds become clear when we analyze the performance of neural networks without any spatial averaging in statistics. While even a manually tuned feed-forward neural network can fit much of the variations in convective tendencies in an aqua-planet ($R^2 > 0.5$ at the 75th percentile– Table 2.4a), an identical manually tuned neural network architecture performs far worse in emulating convection with land masses (Table 2.4a vs. b-e). However, the effects of hyperparameter tuning are dramatic, boosting the 50th percentile R^2 to over 0.5 and 75th percentile R^2 to over 0.75 for convective heating tendencies (Table 2.4 d,e vs. f,g). The fact that one quarter of the domain of convective heating tendency is emulated with $R^2 > 0.75$, even prior to any averaging in space or time, suggests our final DNN setup (full training data volume and hyperparameter tuning) generates a good fit. In the Appendix, we include corresponding statistics for convective moistening as well as for both heating and moistening tendencies on the diurnal time scale – all of which show similar relationships between the four models. The differences in Table 2.4 between land and ocean (f vs. g) indicate spatial variations in the skill, suggesting certain regions are preferentially fit by our DNNs.

A first look at spatial structures in the skill affirms a generally close fit with familiar structures relative to aqua-planet expectations but also some interesting differences. Figure 2.1 presents the skill of zonally averaged DNN predictions. Here we again compare our prototype manually tuned DNN's skill on aqua-planet target data (Figure 2.1 a,b) representative of what was used in [140] against our "Linear" baseline model (Figure 2.1 c,d), our "Manual" DNN (Figure 2.1 e,f), and our optimized "Best" DNN (Figure 2.1 g,h) all trained in real-geography. Achieving realistic performance on zonal means is an easier objective due to the averaging between land and marine atmosphere and the smoothing of the sharpest temporal variations in convection. However, this zonal mean perspective still provides a useful composite view of the emulation of the atmosphere with land masses. Here the R^2 for zonal mean net diabatic heating and moistening is greater than 0.7 throughout the free troposphere, agnostic to latitude (Figure 2.1g and h). This widespread skill in the upper troposphere is further amplified by cores of R^2 greater than 0.9 around mid-latitude storm tracks and locations of deep tropical convection above the southern and northern bounds of the ITCZ. Our "Best" DNN can skillfully emulate heating and moistening tendencies of convection across latitude and vertical level close to SPCAM5 target data (Figure 2.1). It is also reassuring that the best skill (R^2 over 0.9 for zonal mean predictions) occurs in important regions of the troposphere (ITCZ, mid-latitude storm tracks) where mean diabatic heating couples to the general circulation (Figures 2.1).

When interpreting the different DNNs in Figure 2.1, it is important to consider the two combined factors that can influence skill: the dataset and the quality of the DNN fit. To separate their influence, Figure 2.1a,b vs. e,f shows the effect of switching training data at fixed architecture. We note a skill increase in the continental boundary layer, consistent with the existence of new deterministic signals in real-geography settings likely associated with the strong, predictable diurnal cycle over land. However, the overall atmosphere is now more complex with both land and sea regimes, as well as interactions between the two. We suspect this causes the decrease in upper-tropospheric emulation skill, particularly

outside the Hadley Circulation and for the moistening tendency. Despite a superior fit in the continental boundary layer, the inability of the "Manual" neural network to capture deep convection would be a significant hurdle to online coupling. Fortunately, Figure 2.1g,h shows dramatic skill improvement when migrating from a manual tuning environment to a formal hyperparameter search, underscoring the crucial role that Sherpa can play in identifying the optimal or "Best" DNN. Our preliminary analysis suggests that even under increasingly complex conditions, the "parameterizability" of convection can be cast locally in space and time for a fit by a feed-forward DNN.

These results have much in common with the findings from aquaplanet trained DNNs in [140] and our own aqua-planet benchmark (Figure 2.1a and b) is further evidence of these similarities. However, unlike the aqua-planet, there is a new region of high skill in the real-geography emulator with R^2 greater than 0.9 in the planetary boundary layer for heating tendency emulation. This signal in convection appears deterministic enough that even our "manual" neural network can emulate it with $R^2 > 0.7$ (Figure 2.1e). This looks to be a continental signal, evidenced by both higher skill in the northern hemisphere (Figure 2.1h) and comparatively lower near-surface zonal mean skill at the latitudes of the Southern Ocean. Though less skillful overall, there is a similar pattern in the convective moistening tendency in the boundary layer, where the highest DNN performance at the surface (R^2 over 0.7) is confined to the continent heavy northern hemisphere (Figure 2.1h).

We confirm the existence of some distinct land-sea spatial structures in emulation skill by examining maps of predictions prior to any spatial averaging. At the lowest model level, our "Best" DNN predictions achieve R^2 greater than 0.7 (greater than 0.9 in continental interiors; Figure 2.2a). However, this spatial pattern is inverted when examining skill on a model layer in the midtropopshere, near 500 hPa. At this altitude, our "Best" DNN now makes the most accurate predictions over the extratropical marine atmosphere but struggles over continents and deep convecting regions of the tropics (Figure 2.2b). We speculate that

a strong, deterministically predictable component of diurnal variability in surface heating and moistening associated with large surface flux diurnal variations over land could allow the low-level heating skill to be enhanced there. Meanwhile, in the upper troposphere we see the expected skill deficits in regions of tropical and continental convection (on this 15-min timescale). Diurnal signals will be examined in greater detail in Section 2.4.3.

We now focus on the spatial structures where even the formally optimized "Best" neural network still struggles. For this, we return to assessing zonal mean predictions, since these are less exposed to details of stochasticity but are particularly important to emulate accurately when using neural networks prognostically. The greatest emulation challenge for our "Best" DNN is fitting mean temporal variance throughout the lower troposphere (excluding the continental boundary layer) where R^2 falls below 0.3 (Figure 2.3a). This is especially challenging in the case of convective moistening (Figure 2.3b). Our results here are consistent with previous aqua-planet simulations [140, 50], and the study of [59]. Boundary layer moistening is an especially challenging target for machine learning emulation, particularly when focusing on the 15-minute sampling interval. Further evidence of this challenge can be observed by the temporal standard deviations of the heating and moistening tendencies, where much of the spatial field is emulated well, but our neural network nevertheless under-predicts values of moistening tendency in the lower troposphere (Figure 2.12).

We conclude with an animation demonstrating unfiltered, non-composited views of the convective tendency emulation of over a two week period in July, the link to which can be found in the Supplemental Materials (Movie 1.1). The animation shows the evolution of total diabatic heating and moistening on a model level near 600 hPa (the lower-to-mid troposphere). In the three lower panels, the diurnal cycle of peak nocturnal radiative cooling can be seen propagating from east to west tracking the earth's rotation. It is punctuated by local features of positive diabatic heating from latent heating within slow moving weather systems, as well as the stationary lagged diurnal convective response to the passage of the

sun over Central Eurasia and America. No geographic distortions of synoptic disturbances are detectable – even heating tendencies from tropical convective clusters and the Atlantic Convergence Zone and Pacific ITCZ are all closely emulated from this perspective. On the three upper panels, the associated moisture perspective provides an especially clear view of the lack of lower amplitude motion captured in emulated convection. The main distortion compared to truth is the lack of stochasticity, which manifests as geographic static in the benchmark test data (center panels) but is absent in the DNN emulation (right panels).

2.4.2 Temporal Variability

Why do we see such considerable variations in the skill of our DNN as a function of geographic location and altitude? One hypothesis is that the DNN fits "mode-specific" fluctuations. A first test of this is re-examining spatial skill structures after averaging predictions from their native timescale of 15 minutes to the daily mean timescale instead. Figure 2.3 shows the corresponding skill for daily-mean, zonal-mean predictions. From this view, the vast majority of the atmosphere can be emulated in terms of both heating and moistening tendencies with R^2 greater than 0.9. Meanwhile, compared to the faster timescale, the skill deficit in the lower troposphere for the convective moistening tendency is not nearly as dramatic. The fact that structures in spatial skill appear sensitive to temporal averaging is consistent with the hypothesis that the DNN performance might be "mode-specific".

To better answer understand whether our neural network only fits a dominant mode or two of convection at the expense of lower amplitude variations, we now turn to spectral analysis (Figure 2.4) on the SPCAM5 target data and DNN predictions. Switching to frequency space is a clean way to determine if specific modes of variation such as the diurnal heating cycle and synoptic storm propagation are driving preferential modes of DNN fit. The PSD is calculated separately at each unique latitude, longitude, and vertical level from the CAM5

data, SPCAM5 data, and the corresponding DNN output for both our Best neural network and "Linear" baseline model, and weighted by layer mass. These location-specific PSD are then averaged together horizontally and vertically to arrive at a globally representative power spectra.

In contrast to our hypothesis, the spectral analysis does not reveal any major mode-specificity in the DNN skill on the hourly-to-weekly timescale. All of the most important spectral features in the target data exhibit comparable power in the DNN predictions, including the main signals from disturbances slower than one day, but also the discrete variance from diurnal, semi-diurnal, and other harmonics of the daily cycle of convection. While there is an expected under prediction of total variance for sub-diurnal modes, the DNN skill is not obviously preferential to any modes at the diurnal timescale or longer.

Our DNN emulation performance can be further analyzed by taking the PSD from a spatial, rather than temporal, domain. Here the DNN also shows skill at capturing variations in convection at large scales but does not emulate all the details at the small spatial scales (Figure 2.5). We note that this is very much in line with the findings of [175] in which their Random Forests achieved poorer fits at smaller spatial scales compared to wider ones when performance was tested on different course-graining length scales.

While even our simple baseline "Linear" model (orange line in Figure 2.5) can capture the variance in convective tendencies on a global scale, we see evidence of the benefits of automated, formal hyper-parameter at the model grid cell scale. While our "Best" neural network still underestimates smaller signals in convection, it is much closer to the SPCAM5 test data with respect to both convective heating and moistening.

We can quantify these differing degrees of skill captured in the spatial spectra by calculating the total log spectral difference (LSD). Quantitatively, the LSD between the SPCAM5 target data and the Sherpa "Best" neural network predictions is 1.19 for mass-weighted, averaged

tropical zonal heating tendency spectra and 1.55 for mass-weighted, averaged tropical zonal moistening tendency spectra from Equation 2.6. This is a far smaller deviation than when the baseline "Linear" model is compared to the target SPCAM5 data and the difference between the tropical averaged, mass weighted zonal spectra are 2.71 and 3.61 for heating and moistening respectively. Reassuringly, our "Best" neural network also has a lower LSD than the difference between CAM5 and SPCAM5 data (1.20 and 2.30 for heating and moistening zonal spectra). From the temporal domain, the "Best" neural network has a far smaller LSD for both the heating and moistening spectra than the "Linear" baseline. However, the LSD between the CAM5 and SPCAM5 is actually the smallest (in the temporal domain), though from the figure it is clear that this is due to behavior at the shortest time scales that produce exponentially less variance, but which are up-weighted by this metric (not shown).

Even with the sophisticated hyperpameter tuning, the very weak signals of fast variability on timescales less than 2-6 hours is where our DNN is still challenged most (Figure 2.4). Evidently, there is something native to high (spatial or temporal) frequency heating and moistening convective tendencies that challenges our DNN. This suggests an alternate hypothesis that geographic structures in our DNNs skill might be an artifact of variance sorting (i.e. the fact that the fastest signals are also the weakest potentially downweights their contribution to the loss function) or of stochasticity (since fast variations can often be stochastic in origin).

To further test these hypotheses, we construct a proxy of stochasticity and compare its geographic structure to the geographic structure of DNN R^2 skill. The proxy is the e-folding time at which signals of atmospheric variance decouple into noise based on the autocorrelation, i.e. the decorrelation timescale. Quantitatively, the spatial structures of R^2 heating skill and the de-correlation timescale are similar with a pattern correlation coefficient of 0.50. The fact that the regions of least DNN skill are also the regions of fastest signal decorrelation (purple contours in Figure 2.6) supports the view that imperfect emulation of fast, stochastic signals is mainly responsible for sculpting the spatial structures in the DNN's skill score.

To illustrate this further within the challenging tropical regime, Figure 2.7 examines temporal autocorrelations in different DNN skill regimes as follows: First, all tropical grid cells having the poorest skill (the bottom 10 percent) are identified, and the temporal autocorrelation of the benchmark time series data is calculated from time lags 0 to +0.4 days. This is done separately for each tropical grid cell and then composited into a single autocorrelation plot (red line). Repeating the procedure for those horizontal tropical grid cells where the DNN fit has the highest skill (top 10 percent, the blue line) reveals the characteristic difference in de-correlation timescale in the high-skill vs. low-skill spatial regions. Repeating this procedure globally (Figure 2.7b) confirms the same timescale-selectively of skill exists across multiple geographic regimes, even though this was not obvious in Figure 2.6. The relationship is robust and clear – locations where signals tend to decorrelate faster are locations where DNN skill is lower. This is consistent when examining both the planetary boundary layer of the oceans and the continents each in isolation as well as globally (not shown).

2.4.3 Hyperparameter Optimization vs. Physical Constraints for Emulating the Diurnal Cycle

So far we have shown comprehensive skill for our optimized DNN except for some fast-varying convective signals that de-correlate quickly. On the one hand, it is not obviously a problem to have skill deficits for the stochastic component of superparameterized convection, since it does not appear to be critical to most emergent behaviors of SPCAM [72] – though some have suggested that a close representation of the stochastic component is still necessary to properly model large scale weather phenomena [118]. On the other hand, some fast-varying signals are also critical to regional climate simulation and should be deterministically predictable. The diurnal cycle should provide us with a perfect test-bed. To what extent does our DNN emulate the details of the diurnal cycle of convection?

A first look at the composite height vs. time-of-day structure of convective moistening (Figure 2.8) is reassuring – the DNN captures the coherent temporal transition of shallow to deep convection in the afternoon over land (moistening above drying growing after sunrise into the mid-troposphere; Figure 2.8a vs. b). Our DNN also shows a good fit over the ocean, where it captures the opposite phase of peak moistening-above-drying which happens during the night between 8pm and 6am (Figure 2.8c vs. d). Thus our first impression is that the DNN correctly recreates the land-sea contrast of diurnal convection present in the SPCAM5 target data (Figure 2.8).

We now hone in on the full geographic structure, focusing on the diurnal cycle of *precipitation*, which reveals some interesting surprises. The benchmark SPCAM5 target data (Figure 2.9b) resembles observations. Over land regions our test data shows a strong, predictable diurnal precipitation cycle over tropical rainforests and continents in the northern hemisphere (experiencing boreal summer), with lagged afternoon maximum precipitation (Local Solar Time between 13:00 and 18:00). In contrast, a weak diurnal cycle of precipitation occurs over the oceans that peaks at the end of night into early morning, and is especially detectable in subtropical stratocumulus regions. We observe the familiar benefits of superparameterization relative to conventional parameterization (Figure 2.9b vs. a) including a reduced detectability of the diurnal mode except where it is supposed to be strong such as over tropical rainforests or where it is especially consistent such as over marine subtropical drizzle regimes. Here we see an interesting result in the optimized DNN's precipitation predictions (Figure 2.9c): Although the stratocumulus marine drizzle cycle appears to be well emulated, consistent with the diurnal moistening composites seen in Figure 2.8d, over land there is incorrect timing and spatial extent of maximum precipitation (Figure 2.9c). This is paradoxically at odds with Figure 2.8a and b which indicated excellent emulation of the diurnal cycle of convective moistening over land regions. DNN detection of a cycle of precipitation over desert regions is physically unrealistic and the timing of the onset of deep convection and heavy precipitation on land is several hours premature, much like CAM5 data (Figure 2.9c, b, and a).

Why is precipitation emulated less skillfully? Our working hypothesis is that in hindsight our DNN architecture did not respect an important physical constraint that distinguishes this variable. Unlike moistening and heating tendencies, precipitation should be positive definite. To test this hypothesis we introduce four additional neural networks in Figure 2.9 (d,e,f,g). Each new neural network has a different positive constraint (nonlinear activation function) on the last precipitation node to ensure rainfall predictions remain positive definite in line with physical reality. Additionally, in these new constrained DNNs we alter the training data used. We require less data overall (just three months) but empirically find that we should no longer selectively sample as we did previously (e.g. take a day every 10 days). We note that restricting the training dataset to less than a full year does not seem to cause problems with out of sample test data; it emulates the diurnal cycle of precipitation just was well over boreal winter even when the training data is restricted to boreal summer (not shown). We compare our previous DNNs (Best, Manual, Linear baseline model) that ignored positive definite nature of precipitation but included differing hyperparameter tunings (Figure 2.9c,h,i). against these new positively constrained, but not formally tuned networks.

Our results show the different positive constraints induce improvements over different regions of the globe. However, there are substantial variations between different constraint choices with little systematic effects other a realistic enforcement of the timing of the onset of maximum precipitation over land and (except ReLU) a tendency towards poor emulation in dry regions of Africa and the Middle East where these three constrained (by Softmax, Exponential, and Sigmoid functions) neural networks invent a diurnal cycle of precipitation. Unsurprisingly, the DNNs with neither positive constraints nor formal hyperparameter tuning (Figure 2.9h and i) perform the worst. The timing of maximum daily precipitation is premature over land: noon-centric rather than peaking in the mid-afternoon (Figure 2.9h and i). Also, these neural networks fail to detect a diurnal cycle of precipitation over much of the globe (Figure 2.9b vs. h,i). Adding the positive-definite constraint alone produces dramatic improvements over land (Figure 2.9d-g vs. h,i)– but there is the aforementioned variation in the magnitude

of improvement between choices of activation function as a constraint. The DNN with a ReLU activation on precipitation seems to emulate the diurnal cycle of precipitation the best (Figure 2.9d). The hour of maximum precipitation is correct and the neural network emulates a diurnal cycle only where it should be strong, over tropical rain forests, the Southeast United States, and mid-continental summertime Eurasia (Figure 2.9d vs. e,f,g). Unlike the other three constrained neural networks (Figure 2.9e,f,g) and our Sherpa neural network (Figure 2.9d), it does not fabricate diurnal precipitation over the deserts of north Africa and the Middle East – but nevertheless, its emulation of precipitation over continental Africa is still imperfect. But taken on balance, a constrained (by ReLU) neural network appears to solve most of the problems that our original DNNs suffered over land. However, the positive constraint alone is unable to emulate the more subtle marine stratocumulus diurnal cycle well in both the Atlantic and Pacific oceans where the Sherpa DNN emulates the correct time and spatial coverage of this lower amplitude cycle of precipitation (Figure 2.9c vs. d).

We have highlighted the power of automated hyperparameter tuning on convective tendencies in Figure 2.1, but discovered that even our "Best" Sherpa DNN did not take into account physical laws governing its simultaneous prediction of precipitation, instead corrupting it. Difficulty emulating details of precipitation cycles are certainly not unique to this Chapter but do point to larger growing pains in the machine learning and climate science communities. Similar problems with capturing the physics behind precipitation through neural networks have been discussed in [173, 23] where neural networks created non-trivial negative precipitation as well. As in [90], we show that augmenting our DNN with a positive constraint could reduce the errors in land precipitation emulation (Figure 2.9b vs. d). Without formal hyperparameter tuning, these constrained DNNs emulated the land-sea contrast in the timing of peak precipitation: nocturnal over oceans, late-afternoon over the hottest and moistest continental regions. In hindsight, it would be logical to complement the benefits of hyperparameter tuning with such constraints – an important topic for future work. It is also possible that skill in the precipitation field would benefit from enforcing consistency between

it and the column moistening that is better emulated, as in [16] or [163].

To further assess these trade-offs we now look beyond just the diurnal cycle to examine the full PDF of precipitation across our sensitivity tests. The formally tuned "Best" DNN does outperform all other neural networks in capturing the global precipitation amount distribution (Figure 2.10 blue vs. green). Consistent with the diurnal cycle analysis this "Best" DNN performs especially well over the ocean (Figure 2.10d blue vs. green). However, issues over land are even more striking from the viewpoint of the full PDF where the DNNs have radically different values for the amount mode – i.e. rainfall rate delivering maximal precipitation. Whereas the diurnal cycle analysis had suggested a positive definite constraint alone brought continental precipitation into focus, we can see from the amount distribution that beyond diurnal timing its statistics are incorrect (dashed line vs. solid green). In fact, our constrained neural network has a more accurate PDF over the ocean despite its established struggle fitting the nocturnal cycle of precipitation over the marine stratocumulus regions of the globe (Figure 2.10d and Figure 2.9). Meanwhile, the formally tuned DNN has a pronounced problem of producing too much drizzle over land which is also a problem seen in precipitation from standard parameterization.

Taken as a whole our precipitation results suggest that this is an area where further refinement of even our "Best" DNN is needed. Over the oceans, the DNN captures much of the PDF of precipitation (Figure 2.10), including moderate to heavy regimes at the tail that challenge many climate models, as well as the diurnal cycle of precipitation over the oceans (Figure 2.9b vs. c). But there is substantial corruption of the emulated signal over continental locations, particularly with regards to the timing of onset of heaviest precipitation and the intensity of rain delivering most surface accumulation. For an even more information-rich view, we have attached as Supplemental Information an animation showing two weeks of July precipitation from CAM5 data, target SPCAM5 data, and DNN emulation (Movie 1.2) as well as a version of Figure 2.10 with all four constrained DNNs (Figure 2.14).

Comparing the fit sensitivity of adding a constraint vs. leveraging hyperparameter optimization methods [157], both methods provide unique, disparate performance enhancements with the Constrained DNN performing better over land and the "Best" Sherpa DNN doing better overall (Figures 2.9 and 2.10). But synthesizing both tools may ultimately be necessary since neither a physically constrained neural network architecture nor an automated hyperparameter tuned network on its own could capture the full complexity and timing of the diurnal cycle of precipitation over both land and ocean. We recommend an integration of both these tools for future attempts at this work. Meanwhile, it is worth recalling that these corruptions are less obvious in the diurnal cycle of heating and moistening which is better emulated, perhaps because it dominates the loss function, or perhaps because – unlike for precipitation – there is no internal inconsistency with the values of these variables and the assumed DNN architecture. But since precipitation is a critical input to land surface models, resolving the issues revealed in this section will be an important next step towards realizing successful prognostic behavior. Other issues at the frontier of coupling emulators to land surface models are discussed next.

2.4.4 Towards Interactive Land Coupling

Taken together, most of the above results look promising enough that it is natural to wonder if prognostic tests using an emulator like this might produce stable simulations as was shown for an aqua-planet by [140], but in a real-geography setting. This would be exciting to test but our view is that as yet it is premature to try. For instance, beyond its corruptions of continental precipitation, the DNN we have described does not predict everything that would be needed to drive an interactive land surface model in practice. It is even unknown whether the imperfections in the near-surface state of the DNN's predictions would even be compatible with land surface modeling.

As a first credibility test on the latter front we thus report some results from "offline" standalone land surface model simulations driven with actual (vs. emulated) surface state data. These simulations predate the real-geography fits here but use the downwelling surface solar radiation, precipitation, surface pressure, near-surface humidity, and temperature, as well as wind speed from a previous neural network powered aqua-planet GCM to drive several land model integrations. These simulations are easier to perform than fully interactive land-atmosphere coupled simulations and provide a quick test of the null hypothesis that corruptions of the surface state by the DNN might be incompatible with land modeling in general. The idealized offline land model test-bed assumes Amazon-like properties and 5-year simulations are repeated for 112 separate grid cells driven by atmospheric inputs spanning the tropical band [169].

The neural network used here is philosophically similar in design to the architecture we use in Figure 2.1 (a and b). Among these similarities are the training data, which is also of T42 spectral truncation and a native 30 minute model time step. However, this neural network is trained on a full year of simulation data, rather than three months. It is also a larger neural network with 8 hidden layers of 512 nodes each and an input vector of 124 (the dynamic tendencies of temperature and humidity over the entire column are included for an extra length of 60 in the input). Likewise, the output vector includes the longwave and shortwave tendencies over the column for a total vector size of 120. The complete details can be found in [169].

The results in Figure 2.11 reaffirm the potential for prognostic tests. Figure 2.11 shows the carbon cycle flux responses from the resulting ensemble of Community Land Model (CLM) simulations, each with Amazon-like conditions, driven by high frequency forcing data taken from different tropical grid cells of actual vs. emulated SPCAM aqua-planet data. Relative to CLM's conventional coupled behavior (orange lines) these integrations drift to an unusual attractor, which can be understood by the unusual aqua-planet surface state (e.g. high wind

speeds from a frictionless surface). Despite this idealization, the key point is that the CLM drifts to the same new attractor regardless of whether the emulated surface inputs or the actual surface inputs are used to drive it, including details of multiple nonlinear cycles that we have traced to threshold physics associated with wildfire and carbon cycle feedbacks interior to CLM's biogeochemistry modules. These similar trajectories, despite the nonlinearities inherent to CLM physics, are strong evidence against the null hypothesis. This supports the idea of trying DNN convection emulators like this in fully interactive real-geography simulations if they can be adapted to produce all necessary output fields, including separately tracking snow vs. liquid precipitation as well as separating diffuse vs. direct downwelling solar radiation fluxes.

2.5 Conclusion

We find that a feed-forward deep neural network can skillfully emulate the deterministic part of sub-grid scale diabatic heating and moistening tendencies from global superparameterization with the inclusion of land. For the zonal mean, neural network emulated convective tendencies capture over 70% of the actual variance at the 15-minute sampling scale and over 90% of the actual variance at the daily-mean sampling scale throughout most of the mid-to upper troposphere. On regional scales, heating skill is best at low altitudes over land, and at mid-levels over extratropical oceans – both regions where we expect convection to be deterministically set by the large-scale thermodynamic state. On diurnal timescales, convective responses to solar heating are emulated correctly, including land-sea contrast and vertical structure. Full temporal and spatial spectral analyses reveal no obvious "mode-specifity" to what is vs. isn't emulated other than imperfections in the goodness of fit on small spatial (less than 10^3 km in both the zonal and meridional directions) and temporal (less than 3 hours) scales (Figures 2.4 and 2.5). A Pearson Correlation Coefficient of 0.50 between DNN

skill and autocorrelation statistics suggests these errors are highest in stochastic regions where the deterministic component of diabatic tendencies is weaker such as the tropical, marine boundary layer, and the mid-to-upper troposphere over convective land regions. But on longer timescales, particularly where there are distinct, deterministic patterns of atmospheric variation like the diurnal heating of the continents or baroclinic Rossby wave disturbances along mid latitude storm tracks, our DNN effectively emulates superparameterized diabatic processes. We find the highest R^2 coefficient of determination values (typically greater than 0.9) for daily and zonal-mean predictions especially compelling (Figure 2.3c and d). Despite issues in precipitation emulation, our DNN captures much of the marine PDF of precipitation, though it has an unexpected drizzle bias over land that can be partially reduced via a positive constraint on precipitation (Figure 2.9). Despite these imperfections, precipitation statistics produced by the DNN are superior to conventional parameterization.

The accuracy achieved by our neural network suggests that feed-forward DNNs may still be the best way to create next generation, hybrid climate emulators. Skip-connections in conjunction with convolutions would seem to have possible advantages in allowing multi-scale structures to be simultaneously prioritized in the loss function [59]. But our DNN achieves similar (Figures 2.12) to superior (Figure 2.13b vs. e), skill compared to the more sophisticated Convolutional (in the vertical direction) Neural Networks and Resnets trained recently on similar data in [59]. This would suggest that model architecture choices like skip connections and 1D convolutional layers are not critical to achieving a good fit for a neural network in the emulation of convection.

More broadly, these results also speak to an ongoing question of what sets the "parameterizability" of deep convection, which can be inferred from the success of machine learning methods trained on superparameterized simulations (recognizing that despite their constraints SP includes nontraditional degrees of freedom like convective memory and organization). Our findings suggest that convective memory may not be essential [59, 34], at least for feed-

forward DNNs. That is to say, our feed-forward DNNs did not require memory from previous timesteps in converging on skillful fits to convective tendencies – predictions independent of space and time may be the better way forward to achieve successful moist convection emulation. We find that our DNN, with no memory used in training, preferentially fits the atmospheric modes of variation where convective memory would be most helpful (diurnal cycle, onset of afternoon deep convection/heavy precipitation, synoptic storms). Our issues in DNN emulation are greatest over regions where the controlling signals happen at the shortest temporal (or spatial) scales – especially in the tropical, marine boundary layer. These are exactly the places convective memory would be least helpful.

Looking ahead, we believe a feed-forward deep neural network, powered by automated hyperparameter tuning as well as physical constraints, may be the most realistic way for ML to emulate superparameterized moist convection in a realistic atmosphere with real-geography boundary conditions. This is also a more direct way to achieve two way coupling with a host climate model since feed-forward DNNs can be rapidly deployed today as prognostic Fortran hybrid models thanks to new automated software [123]. More broadly, for general deep learning applications, we believe our experience sheds light on the importance of incorporating physical science knowledge while exploiting machine learning methods when designing appropriate neural networks to tackle problems such as moist convection emulation. We have found like many others [157] that while each of these design choices show notable improvements to emulation performance on their own, both are likely needed in conjuncture to utilize DNNs to their fullest potential.

Though not our primary focus, our findings also point to some of the challenges ahead in neural network emulation of the stochastic component of convection. To replace CRMs in a convection simulation, deep neural networks will likely eventually need to fit not just deterministic but also stochastic parameterizations, which are crucial to error and bias reduction [81, 127]. Even in superparameterized climate simulations such stochastic effects,

while not critical to mean climate, have been linked to some important regional precipitation extremes [72]. Our results indicate there are high variance modes of moist convection that will be difficult for any feed-forward DNN to represent perfectly without a faster time step interval in training data or stochastic, generative modeling.

Meanwhile, a next step for the specific case of emulating superparameterization should be an online test of the performance of our trained DNN in prognostic mode to determine if the neural network is skilled enough to produce physically plausible outputs from the coupled run. In this limit, the secondary effects of stochasticity noted by [72] argue deterministic DNNs are appropriate. Although coupling emulated atmospheres to prognostic land models is mostly an unexplored frontier, we are optimistic based on our first pilot tests that it is readily approachable; imperfections of the fit do not break standalone land model simulations. But carrying this forward into fully prognostic coupled tests will require significant work, such as expanding this prototype DNN's output vector to include additional variables needed to allow fully interactive land model coupling, and associated tuning. Even if this next step proves successful, feed-forward DNNs should not be thought of as a panacea for all flaws in climate models – they cannot in their present application resolve biases induced by imperfect microphysics parameterization and the resulting errors in associated turbulence and cloud-radiative effects produced by superparameterized models. However, neural networks do still have broad use for sidestepping the computational bottlenecks that currently limit the global modeling community's ability to approach eddy-resolving scales. We remain excited about that potential, especially given our findings here that such approaches can be made remarkably skillful beyond aqua-planets, at least in tests of offline hold-out test skill.

Though much work remains to improve the synthesis between physical climate models and neural networks for successful online testing, we have shown that the complexity of earth's atmosphere should not in and of itself be an insurmountable barrier to this endeavor. But we now leave further improvements for future works. Chapter 2 has focused on the "engineering

potential" of neural networks to work in parallel with climate models, but we will now turn in Chapters 3 and 4 to their potential for objective analysis of high-resolution weather and climate data.

2.6 Appendix A: Performance Comparison with Existing Literature

Figure 2.12 shows the extent to which the variations of the atmosphere, particularly those driven by deep convection and latent heating can be captured by a feed-forward DNN with minimal under-prediction. The spatio-temporal patterns are replicated over the annual data.

Overall, when looking at the annual mean, our DNN performs well globally. These are some imperfections with emulation of intense tropical precipitation, but the heating and moistening tendency predictions are very close to the target data. In particular, Figure 2.13 shows modes of variation in the planetary boundary layer can be captured by our DNN. The DNN seems to fit at least as well as the Resnet throughout the latitude-pressure cross section, and perhaps marginally better the boundary layer when moisture variations are examined (Figure 2.13 b vs. e).

2.7 Appendix B: Supporting Tables, Figures, and Movies

Appendix B provides an expanded version of Table 2.4 for additional context on neural network performance and comparison between the baseline neural network, the manually tuned neural network, and the formally tuned (Sherpa) neural network. We also include an expanded version of Figure 2.10 to contrast the performances between constrained neural networks looking at the totality of the precipitation PDF instead of just the hour of maximum

Label	Training Data	Region	Variable	timestep	25th	50th	75th
a	real-geog. (Linear)	Land	Heating	15 min.	-0.96	-0.06	0.25
b	real-geog. (Linear)	Land	Heating	Daily	-3.09	-1.22	-0.09
c	real-geog. (Manual)	Land	Heating	15 min.	-0.93	-0.06	0.35
d	real-geog. (Manual)	Land	Heating	Daily	-2.87	-1.00	0.00
e	real-geog. (Best)	Land	Heating	15 min.	0.41	0.64	0.82
f	real-geog. (Best)	Land	Heating	Daily	0.42	0.71	0.85

Table 2.5: Statistical breakdown of skill score showing percentiles summarizing skill variability in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw data at the 15-minute sampling scale or the daily mean sampling scale. This table highlights the three neural networks trained on the SPCAM5 real geography data: A "Linear" baseline model (a-b), a manually tuned neural network (c-d), And our formally tuned Sherpa neural network (e-f). The table depicts convective heating K/s over continental locations.

precipitation view offered in Figure 2.9. We have also embedded public links for several animations showing neural network emulation of convective tendencies and precipitation in the tropics.

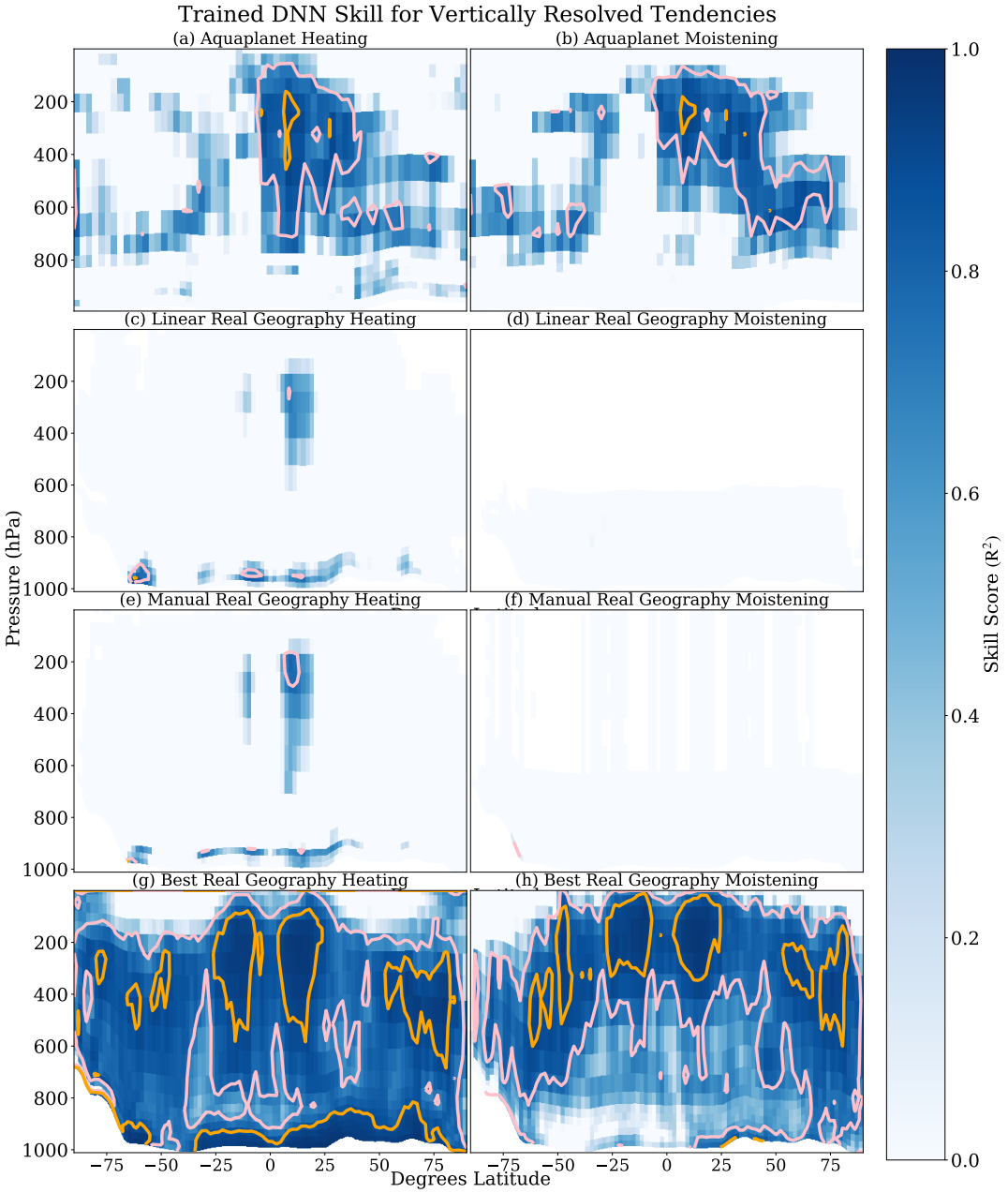

Figure 2.1: The R^2 coefficient of determination for zonally averaged DNN predictions. We contrast the performance of a manually tuned deep neural network emulating aqua-planet target data (a and b) with three comparable neural networks trained on full complexity real-geography data. These include our baseline linear model (c and d), a manually tuned neural network (e and f) and our semi-automated, formally tuned Sherpa neural network (g and h). Skill is shown separately for heating tendency in (K/s) (a, c, e, g) and moistening tendency in $(kg/kg/s)$ (b, d, f, h). Areas where R^2 is greater than 0.7 agreement between are contoured in pink and areas greater than 0.9 in orange.

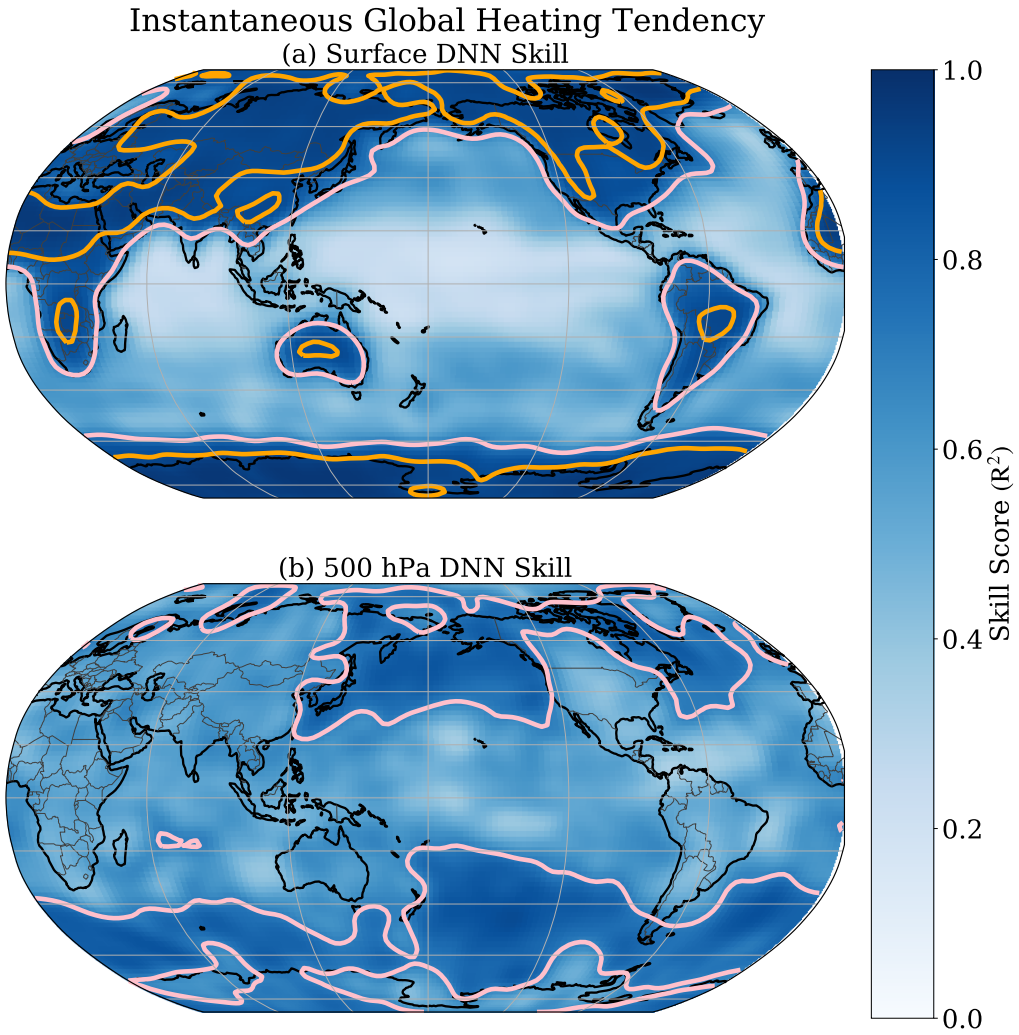

Figure 2.2: The neural network skill in emulating sub-grid heating at (a) the lowest model level and (b) the model level closest to 500 hPa, both at the native 15 minute timestep interval. The neural network fits locations over continents and the mid-latitudes best down at the surface, while locations of mid latitude storm tracks are best fit by our neural network in the mid-to-upper troposphere above 500 hPa. The tropics, in particular tropical locations over oceans, create the greatest challenge for the neural network emulation of sub-grid heating tendencies. Areas where the coefficient of determination R^2 is greater than 0.7 are contoured in pink and areas greater than 0.9 are in orange. To facilitate reading, the map was smoothed using a 2D Gaussian averaging kernel with a standard deviation of 2 grid cells in both latitude and longitude (y and x). Each Gaussian filter was additionally truncated at 4 standard deviations. For ease of visualization and cleaner comparison with previous work, we show the plot of $\max(0,R^2)$.

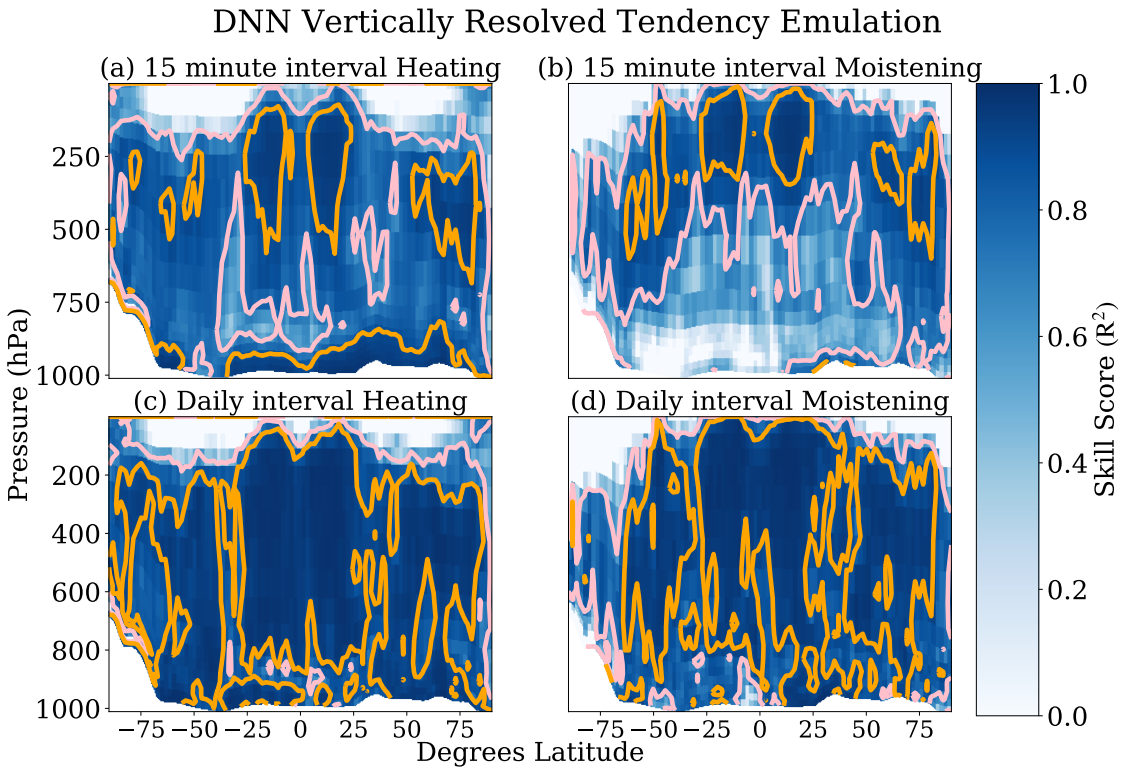

Figure 2.3: Neural network performance at time step interval (a and b – also seen in Figure 2.1 g and h) is contrasted with performance at the diurnal scale (c and d). Representation of heating tendency in (K/s) (a and c) and moistening tendency in $(kg/kg/s)$ (b and d) are both examined. Zonal averages are again taken upstream of R^2 calculation. In both vertically resolved heating and moistening, there is an across the board gain in skill at longer timescales. Areas where R^2 is greater than 0.7 are contoured in pink and areas greater than 0.9 in orange. For ease of visualization and cleaner comparison with previous work we show the plot of $\max(0,R^2)$.

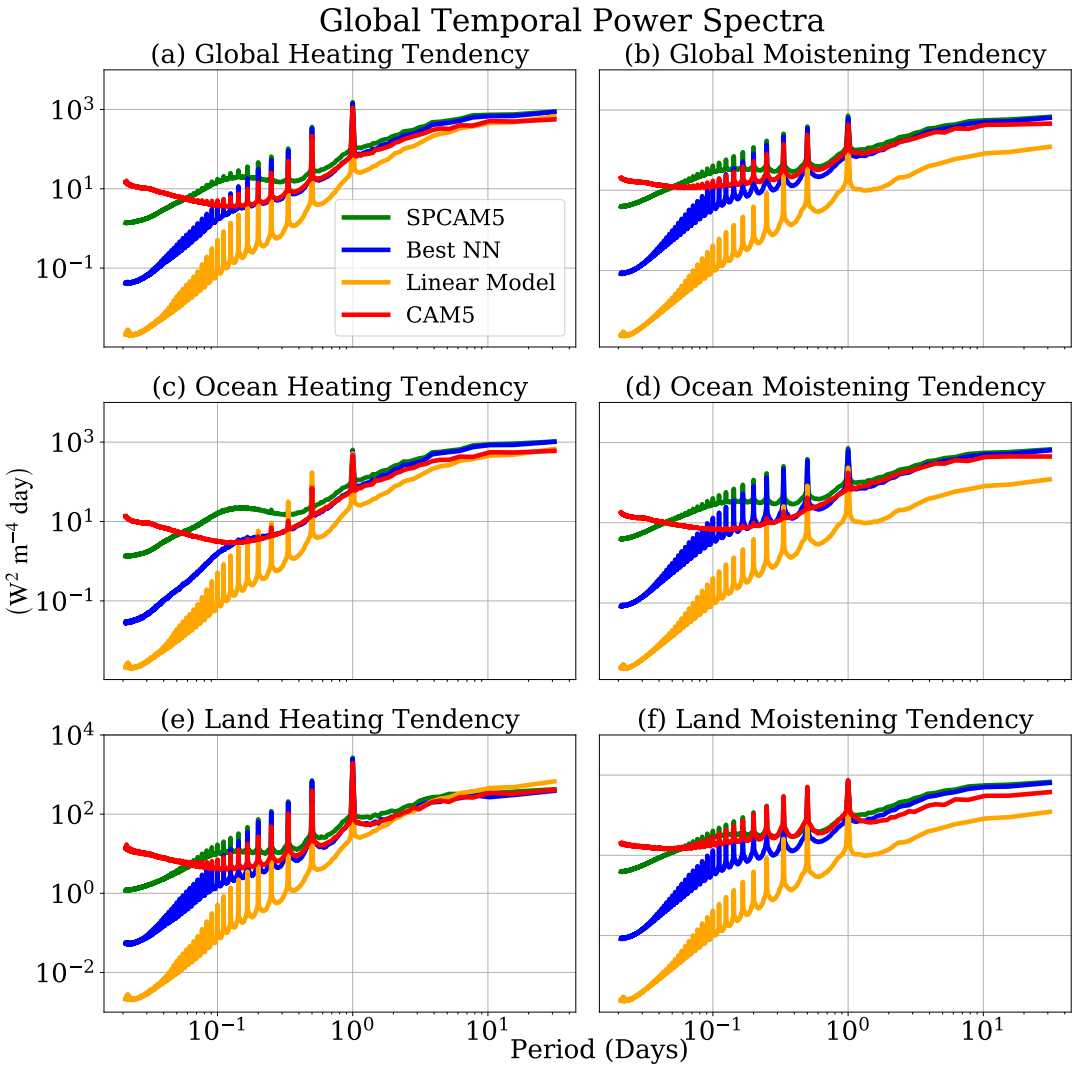

Figure 2.4: The temporal power spectrum for vertically resolved heating tendency in $(W^2/m^4 day)$ (a) and vertically resolved moistening tendency in $(W^2/m^4 day)$ (b) are calculated at each latitude, longitude, and elevation across the globe. These spectra are then averaged together to see how much variance the linear baseline model captures globally compared to our formally tuned Sherpa neural network. Results from SPCAM5 test data and CAM5 data are also plotted for perspective. Further tests are done exclusively over marine locations (c and d) and over continental ones (e and f). The peaks correspond to the solar radiation driving the diurnal cycle, though this is stronger on land (e and f) than in marine locations (c and d). Multi-taper spectra were also calculated for both tendencies but showed no qualitative difference with the results above calculated through the numpy fft package.

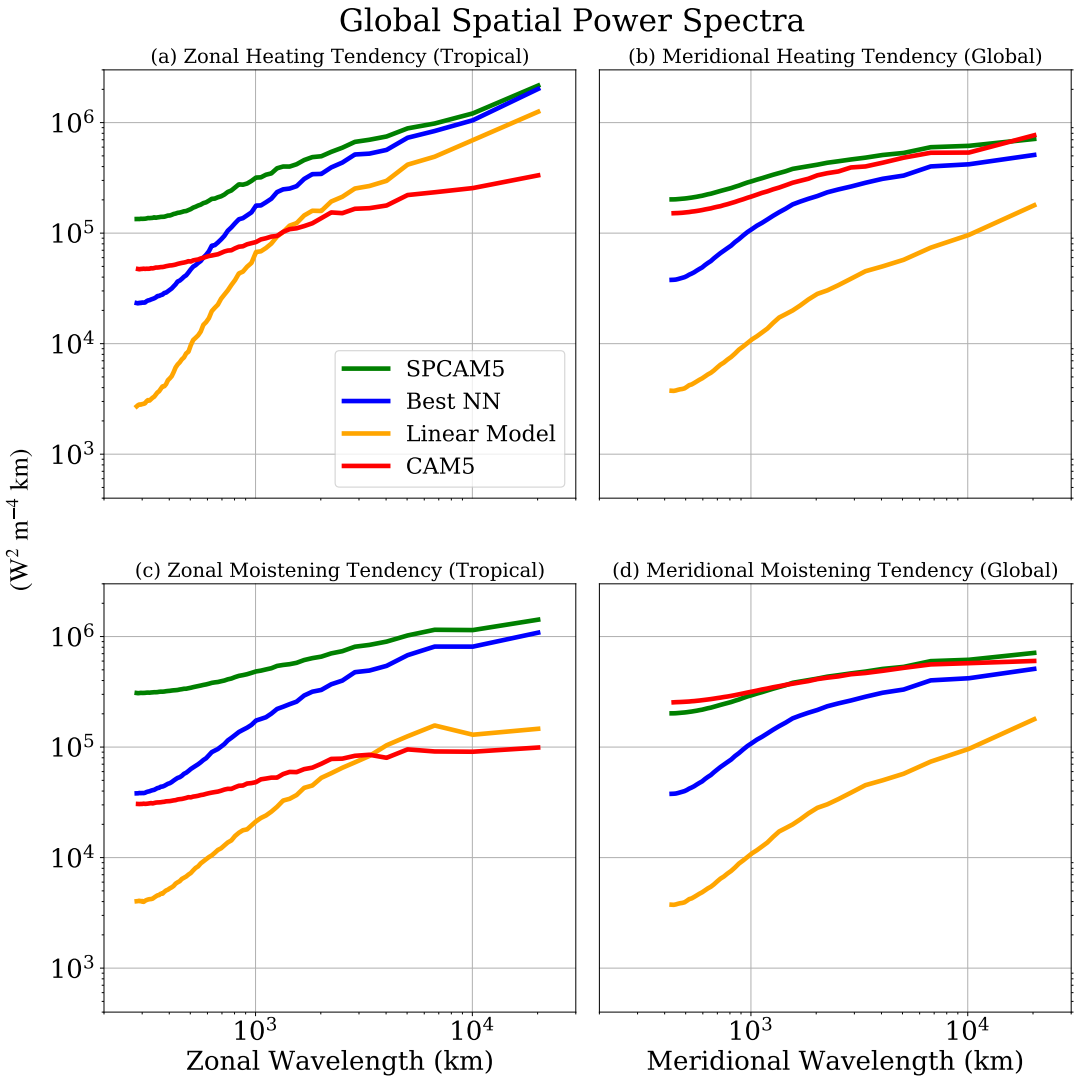

Figure 2.5: The spatial power spectrum for vertically resolved heating tendency in (W^2/m^4km) (a and b) and vertically resolved moistening tendency in (W^2/m^4km) (c and d) are calculated at each vertical level and time step across the simulation data. These spectra are then averaged together to see how much variance the linear baseline model captures globally compared to our formally tuned Sherpa neural network. Results from SPCAM5 test data and CAM5 data are also plotted for perspective. We take a 1D fft in both the x (zonal) (a and c) and y (meridional) (b and d) directions. However, we restrict our zonal cross-section to just a tropical belt (20N-20S) so we can assume a cartesian plane and neglect variable grid spacing. These results tie in with Figure 2.4 in that capturing the variations in convective tendencies at small scales proves more difficult for our neural networks than at large scales.

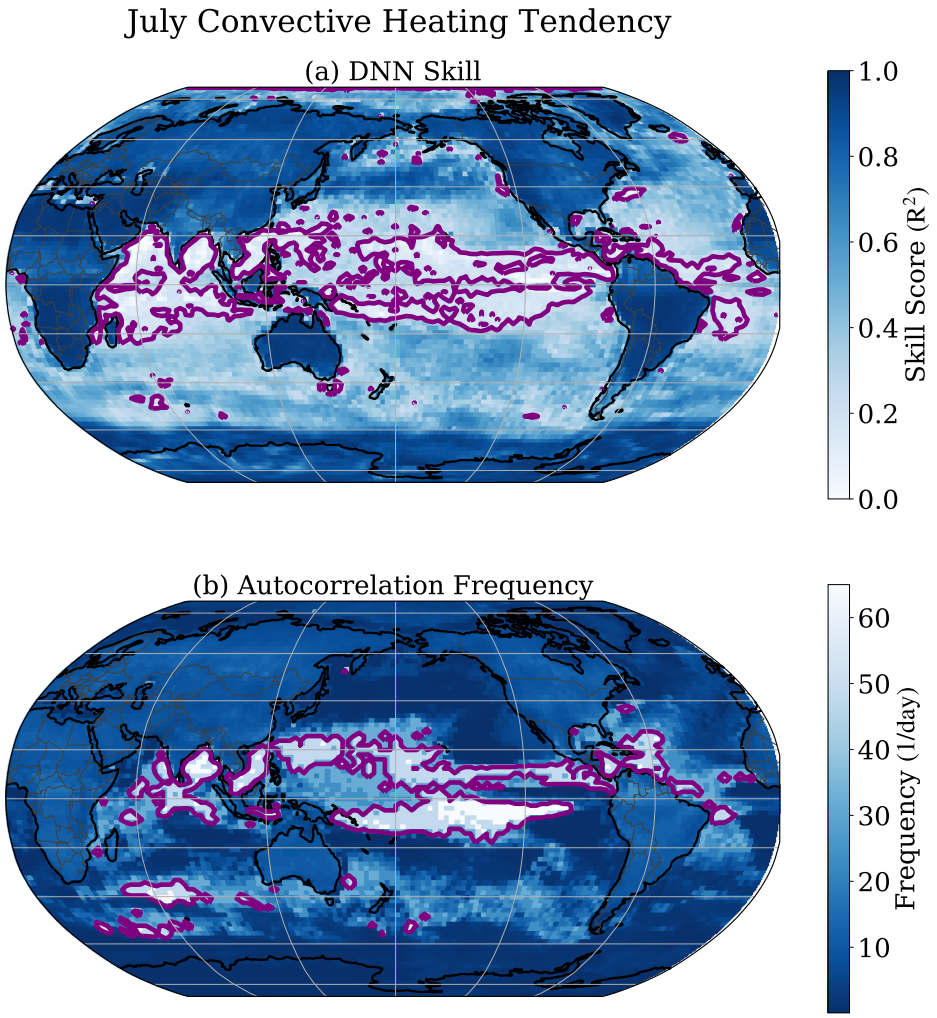

Figure 2.6: A comparison between the neural network R^2 skill in emulating the vertically resolved heating tendency in (K/s) (a) and the autocorrelation frequency of the SPCAM5 heating tendencies (b). Both cross sections are taken at the lowest pressure level in the model. Qualitatively the patterns closely match. The areas of lowest skill score (bottom tenth percentile) and highest autocorrelation frequency (90th percentile) are both contoured in purple. For ease of visualization and cleaner comparison with previous work we show the plot of $\max(0,R^2)$ in panel a.

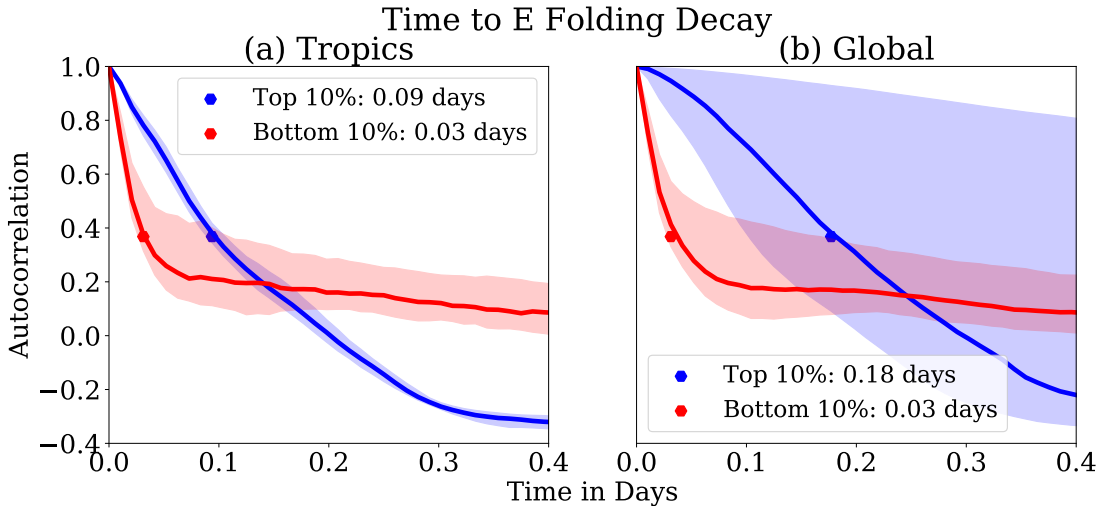

Figure 2.7: The solid lines represent the median autocorrelation as a function of time at every surface location where the R^2 skill score of heating tendency in (K/s) is in the top 10 percent (blue) and the bottom 10 percent (red). We restrict our comparison to surface locations in the tropics (15°S to 15°N) (a) and then examine the entire surface of the earth (b). The corresponding inter-quartile regions are shaded in as a marker for statistical significance. The dots show the time to e-folding decay. The test data spans the month of July.

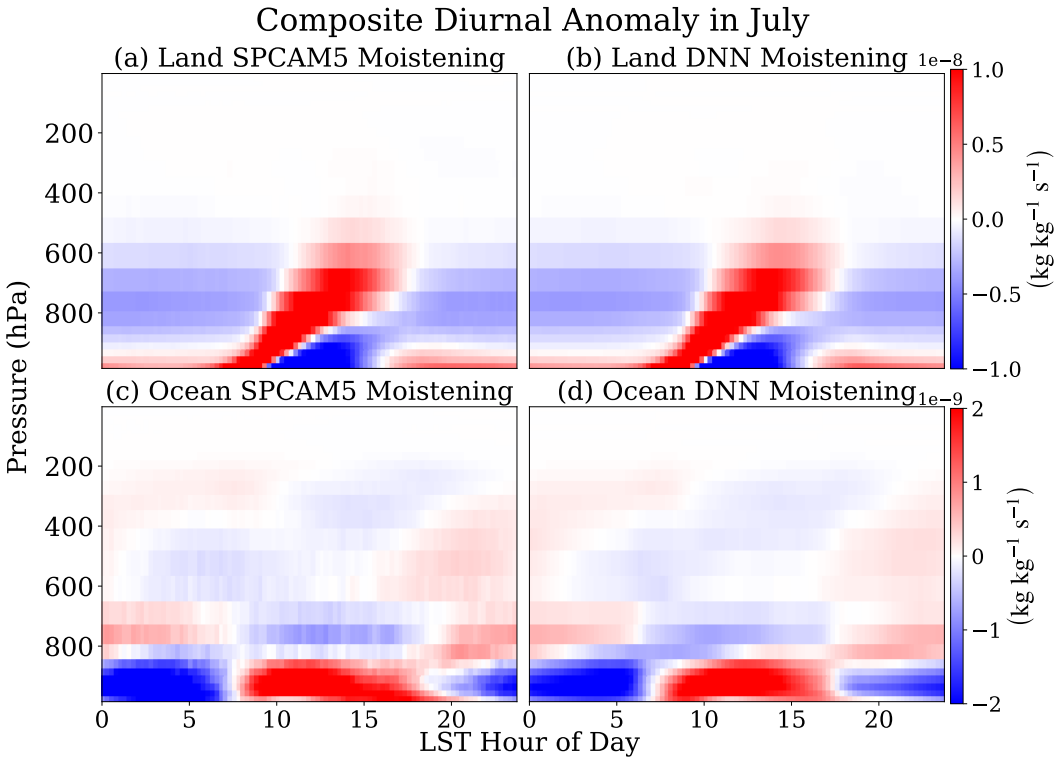

Figure 2.8: A comparison between the moistening tendency of SPCAM5 target data (a and c) and DNN predictions (b and d) in $(kg/kg/s)$ over continental (a and b) and marine (c and d) locations respectively. The composite is taken over the month of July and we choose to show the anomaly of the diurnal cycle in which the mean is subtracted out.

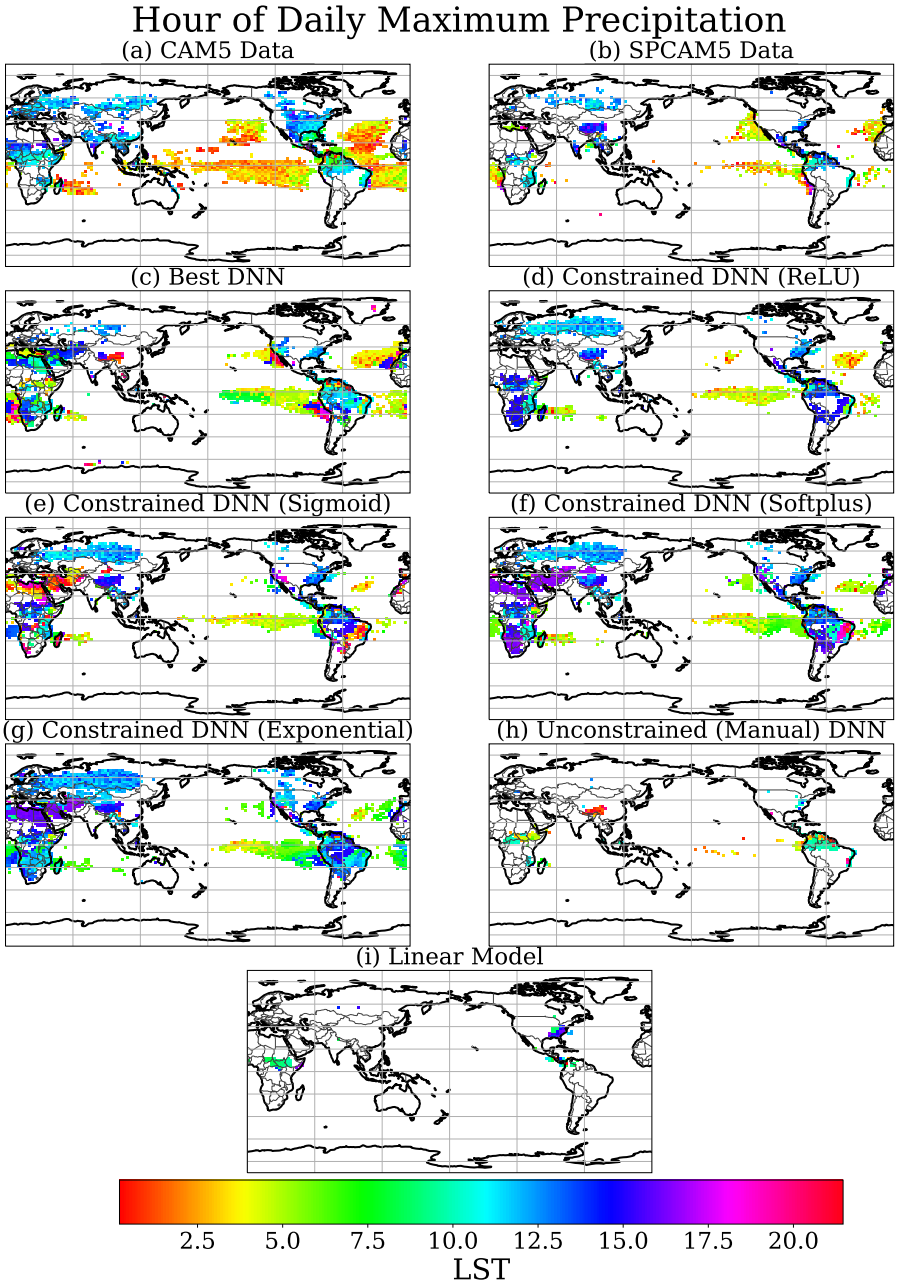

Figure 2.9: A comparison between CAM5 data (a), SPCAM5 test data (b), and our overall best neural network with automated hyperparameter tuning (c), neural networks with different positive constraints on the precipitation output (d, e, f, g), an archaic version of our DNN without automated hyperparameter tuning or physical constraints (Manual) (h), and our linear baseline model (i). The figures show the hour of maximum precipitation in (mm/day) during the boreal summer (months of June, July, and August). The time of maximum precipitation is colored in only over areas with a significant diurnal amplitude in precipitation rate as defined in Equation 2.7.

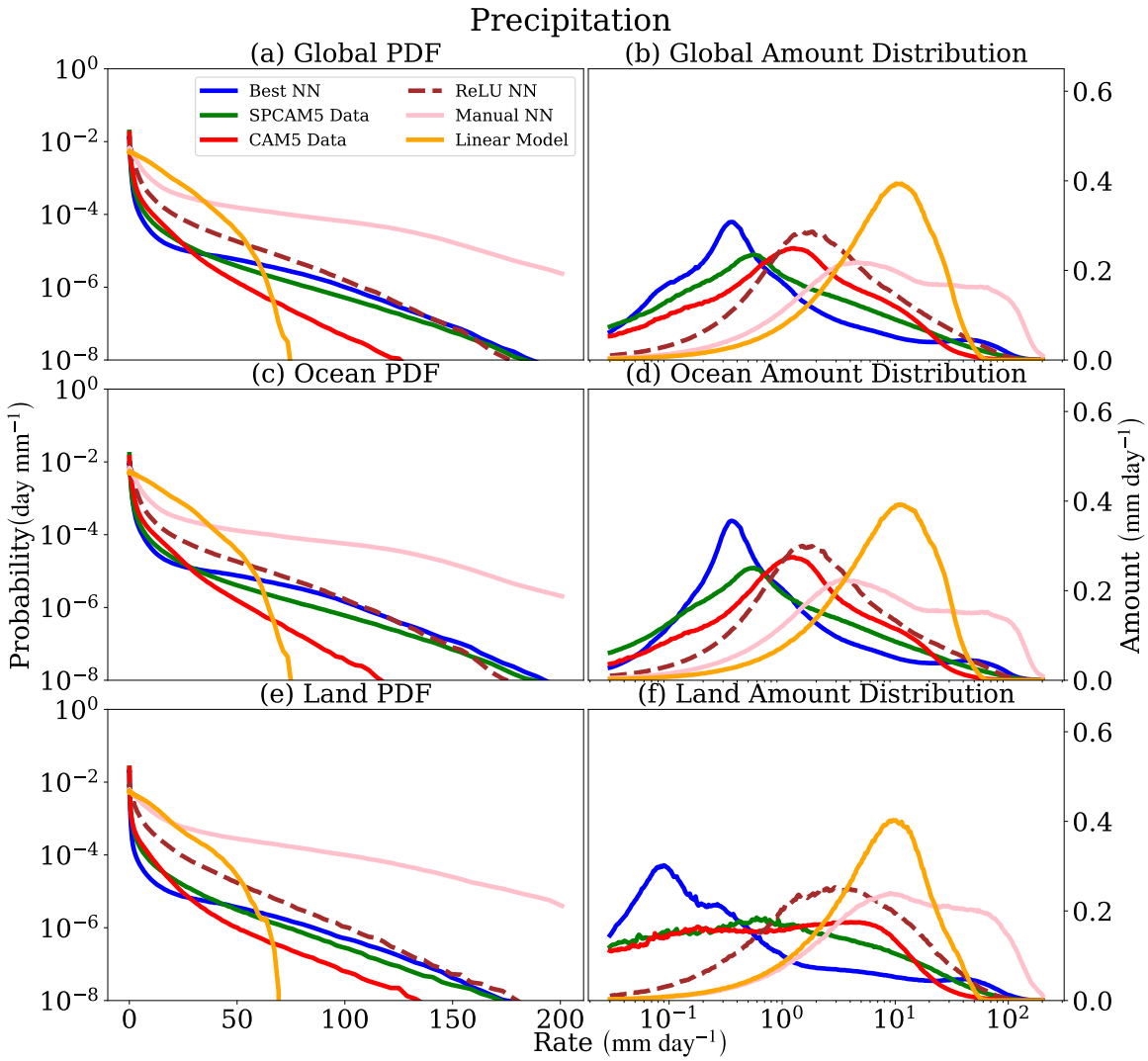

Figure 2.10: The Probability Density Function across the range of simulated precipitation rates (a) and the corresponding amount distribution (b) of precipitation in which the probability density function is multiplied by the bin-averaged values of precipitation. We design the histograms based on the methods outlined in [164], which have been widely adopted in literature including in formative works such as [132]. We implement logarithmically distributed rain-rate bins. In our case, each bin width grows by 3 percent to ensure the entirety of the precipitation PDF is reflected. For more detail, we include an archaic version of our neural network without an automated hyperparameter tuning or physical constraints (Manual), our best constrained neural network (dashed line), and our overall best (Sherpa) DNN discussed previously in the methods section. Comparisons are also made exclusively over marine areas (c and e) and continental ones (d and f).

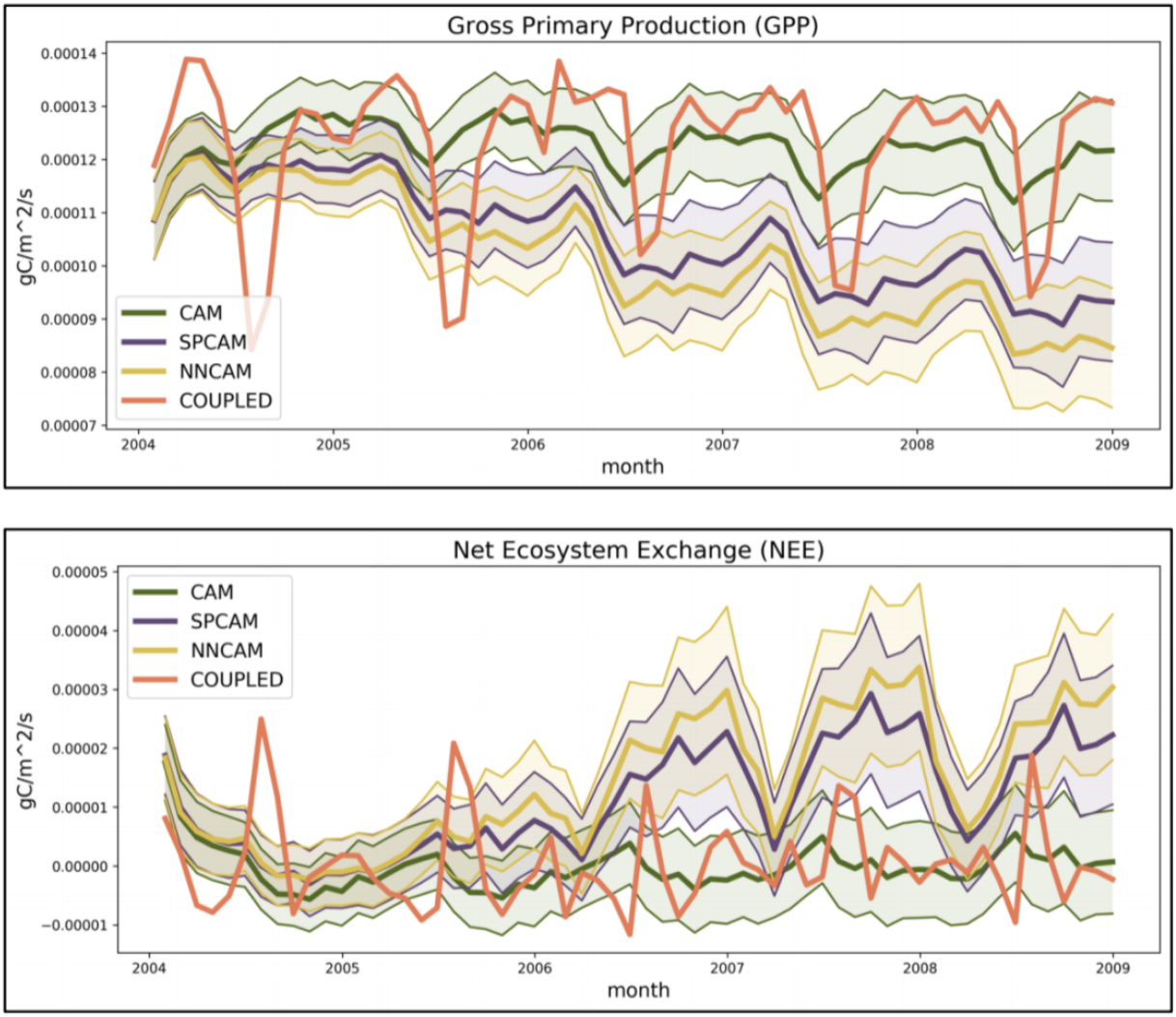

Figure 2.11: The Gross Primary Production (GPP) and Net Ecosystem Exchange (NEE) monthly based on CAM data (also in aqua-planet mode) are contrasted against SPCAM (aquaplanet) and a neural network (aquaplanet trained), the results of which are derived from one way land coupling. The solid lines correspond to mean values while the shading encompasses the extent of the monthly mean standard error at each time step.

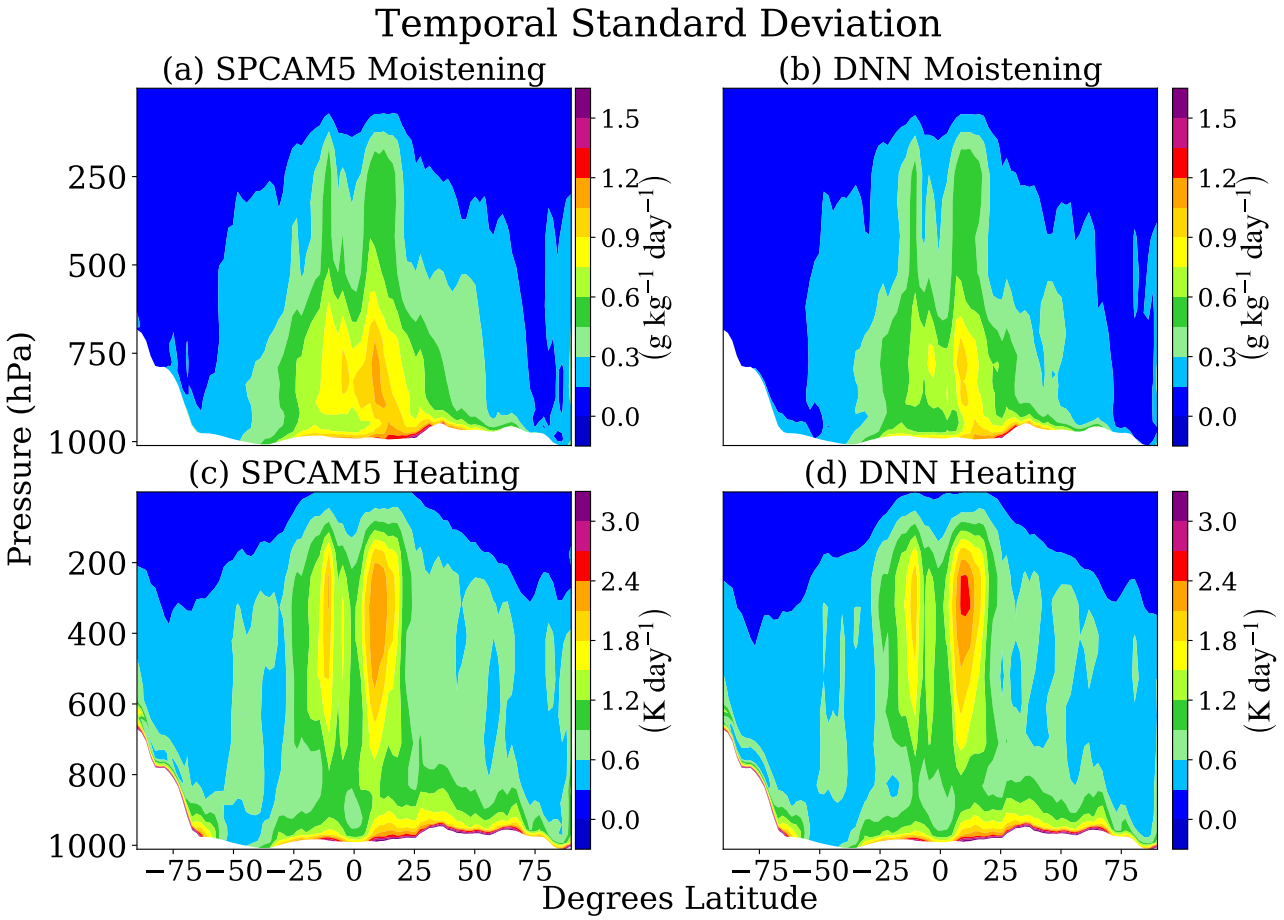

Figure 2.12: The temporal standard deviation of annual heating and moistening tendencies. Units converted to (K/day) and $(g/kg/day)$ respectively to contrast with the performance of a Resnet [59].

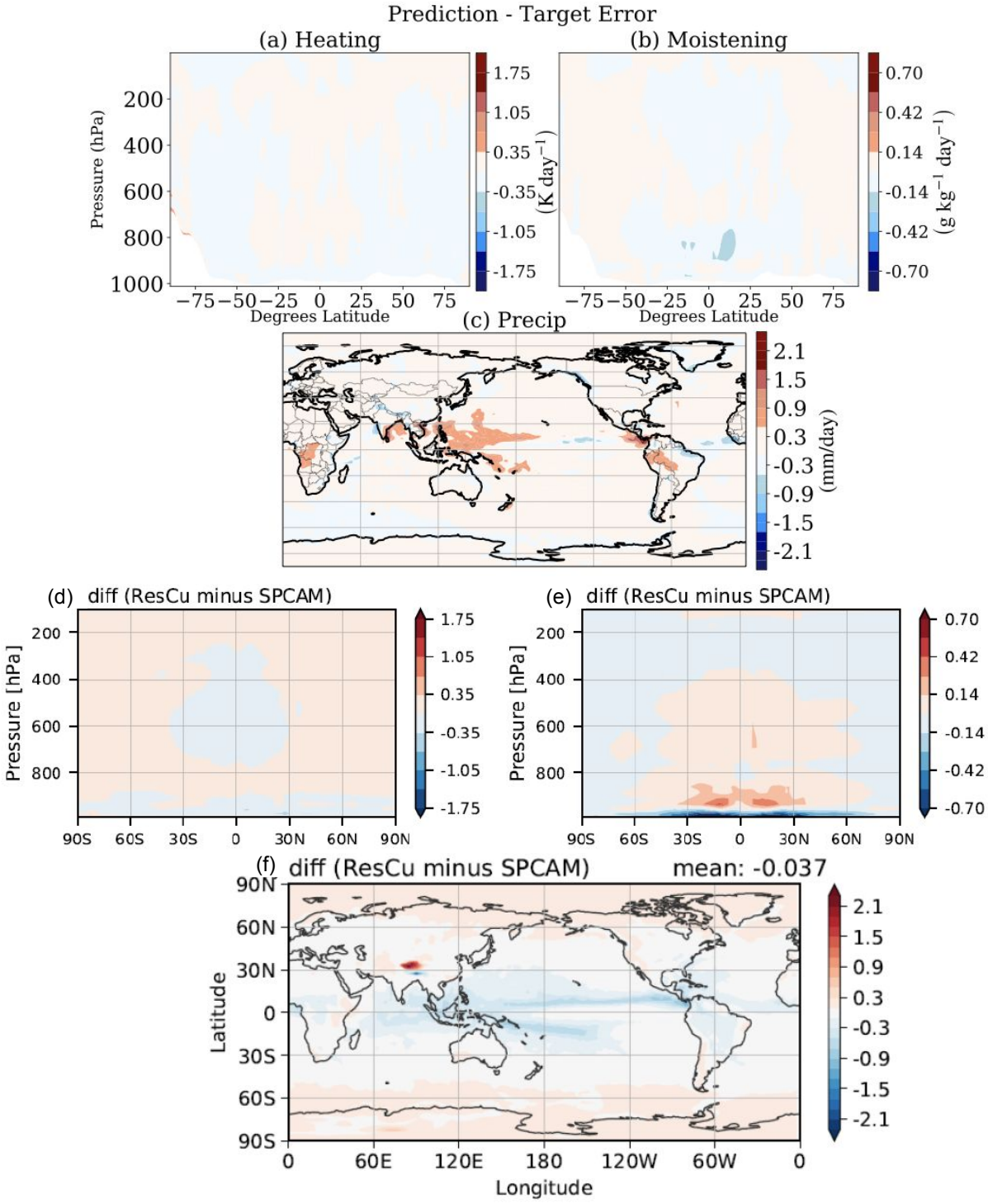

Figure 2.13: The difference between annual target SPCAM5 data and the DNN predictions for heating tendency (K/day), moistening tendency ($g/kg/day$) and precipitation (mm/day). The 3 panels on the bottom have been taken from [59] to provide direct comparisons between the performance of our DNN and the [59] Resnet on full complexity, real-geography simulation data.

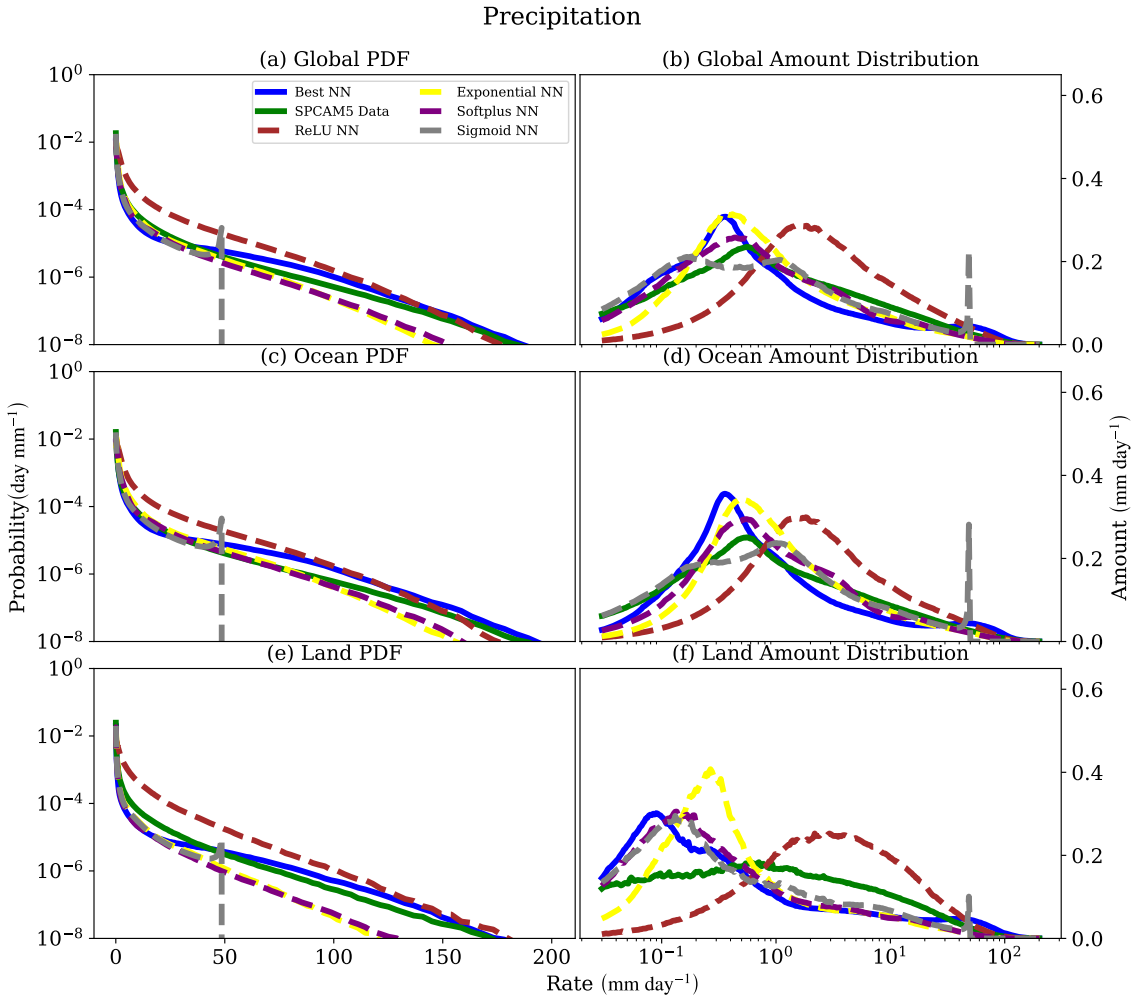

Figure 2.14: An extension of Figure 2.10, but this time contrasting four constrained neural networks (dashed lines) against the SPCAM5 target data (green line) and the Sherpa NN (blue line). The Probability Density Function across the range of simulated precipitation rates (a, c, e) and the amount distribution (b, d, f) of precipitation in which the probability density function is multiplied by the bin-averaged values of precipitation.

Label	Training Data	Region	Variable	timestep	25th	50th	75th
a	real-geog. (Linear)	Land	Moistening	15 min.	-0.07	-0.02	0.00
b	real-geog. (Linear)	Land	Moistening	Daily	-0.27	-0.09	-0.02
c	real-geog. (Manual)	Land	Moistening	15 min.	-0.07	-0.02	0.00
d	real-geog. (Manual)	Land	Moistening	Daily	-0.27	-0.09	-0.02
e	real-geog. (Best)	Land	Moistening	15 min.	-5.5	0.10	0.55
f	real-geog. (Best)	Land	Moistening	Daily	-12.9	0.14	0.76

Table 2.6: Statistical breakdown of skill score showing percentiles summarizing skill variability in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw data at the 15-minute sampling scale or the daily mean sampling scale. This table highlights the three neural networks trained on the SPCAM5 real geography data: A "Linear" baseline model (a-b), a manually tuned neural network (c-d), And our formally tuned Sherpa neural network (e-f). The table depicts convective moistening kg/kg/s over continental locations.

Label	Training Data	Region	Variable	timestep	25th	50th	75th
a	aqua-planet	Ocean	Heating	15 min.	0.05	0.27	0.55
b	aqua-planet	Ocean	Heating	Daily	-0.41	0.24	0.59
c	real-geog. (Linear)	Ocean	Heating	15 min.	-0.30	-0.01	0.21
d	real-geog. (Linear)	Ocean	Heating	Daily	-2.26	-0.58	0.03
e	real-geog. (Manual)	Ocean	Heating	15 min.	-0.26	0.00	0.31
f	real-geog. (Manual)	Ocean	Heating	Daily	-1.82	-0.33	0.32
g	real-geog. (Best)	Ocean	Heating	15 min.	0.28	0.54	0.76
h	real-geog. (Best)	Ocean	Heating	Daily	0.30	0.66	0.85

Table 2.7: Statistical breakdown of skill score showing percentiles summarizing skill variability in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw data at the 15-minute sampling scale or the daily mean sampling scale. This cumulative table compares results from a neural network trained on SPCAM3 aqua-planet data (a-b). It also highlights three neural networks trained on the SPCAM5 real geography data: A "Linear" baseline model (c-d), a manually tuned neural network (e-f), and our formally tuned Sherpa neural network (g - h). The table depicts convective heating K/s over marine locations.

Label	Training Data	Region	Variable	timestep	25th	50th	75th
a	aqua-planet	Ocean	Moistening	15 min.	-4e6	-0.40	0.23
b	aqua-planet	Ocean	Moistening	Daily	-4e5	-2.19	0.40
c	real-geog. (Linear)	Ocean	Moistening	15 min.	-0.06	-0.01	0.48
d	real-geog. (Linear)	Ocean	Moistening	Daily	-0.22	-0.06	0.00
e	real-geog. (Manual)	Ocean	Moistening	15 min.	-0.05	0.00	0.04
f	real-geog. (Manual)	Ocean	Moistening	Daily	-0.22	-0.06	0.00
g	real-geog. (Best)	Ocean	Moistening	15 min.	-3.45	0.16	0.48
h	real-geog. (Best)	Ocean	Moistening	Daily	-11.0	0.49	0.49

Table 2.8: Statistical breakdown of skill score showing percentiles summarizing skill variability in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw data at the 15-minute sampling scale or the daily mean sampling scale. This cumulative table compares results from a neural network trained on SPCAM3 aqua-planet data (a-b). It also highlights three neural networks trained on the SPCAM5 real geography data: A "linear" baseline model (c-d), a manually tuned neural network (e-f), and our formally tuned Sherpa neural network (g - h).The table depicts convective moistening kg/kg/s over marine locations.

Movie 1.1

An animation of convective heating and moistening tendencies between CAM5 data, SPCAM5 data, and the Sherpa tuned "Best" neural network. This video contains every other 15 minute timestep for 14 days in July from 35S-35N. The data have all been converted to W/m^2. The complete animation can be accessed at `https://drive.google.com/file/d/17MdO7Lb7DusakuT_2WcqW7iakVYUr7hh/view`

Movie 1.2

An animation of precipitation from CAM5 data, SPCAM5 data, and the Sherpa tuned "Best" neural network. This video contains every other 15 minute timestep for 14 days in July from 35S-35N. The data have all been converted to mm/day and the data is shown on a log scale. The complete animation can be accessed at `https://drive.google.com/file/d/1jLccgEBkeIK-ciCvoKKtt0JOdlNPcvRl/view`

Chapter 3

Generative Modeling of Atmospheric Convection

3.1 Abstract

While storm-resolving models can explicitly simulate the details of small-scale storm formation and morphology, these details are often ignored by climate models for lack of computational resources. In Chapter 3, we explore the potential of generative modeling to cheaply recreate small-scale storms by designing and implementing a Variational Autoencoder (VAE) that performs structural replication, dimensionality reduction, and clustering of high-resolution vertical velocity fields. Trained on $\sim 6 \cdot 10^6$ samples spanning the globe, the VAE successfully reconstructs the spatial structure of convection, performs unsupervised clustering of convective organization regimes, and identifies anomalous storm activity, confirming the potential of generative modeling to power stochastic parameterizations of convection in climate models.

3.2 Introduction

With modern, storm-resolving models there are challenges beyond our inability to run these models for the ∼100-year timescales needed [69, 74, 108]. We also require the ability to analyze the simulation output when we do run them on short-time scales. The detail revealed by these simulations has historically been too minute to explicitly resolve in GCMs [139, 71, 146]. Likewise, because key physics driving convection and cloud formation occurs on the scale of meters to a few kilometers, there is also no (both spatial and temporally) satisfactory observational record to rely on. This increases the value of these "storm-resolving" simulation outputs as they contain a treasure trove of information about the physics and behavior of storm systems not available elsewhere. But we have yet to fully leverage these model outputs because they are overwhelming to analyze at a native scale. This gridlock leaves significant gaps in knowledge about many of the details of cloud-climate feedbacks as well as the relationship between storm organization and its thermodynamic environment [139, 108]. However, deep learning, and in particular generative models, may provide a path to a better understanding of these phenomena and their role driving the weather and climate of our world.

The application of machine learning in the physical sciences has increased exponentially in recent years but with important avenues still largely unexplored. Similar to Chapter 2, other deep neural networks have been re-purposed to emulate the large-scale consequences of storm-level heating and moistening over the atmospheric column to replicate mean climate and expected precipitation patterns and extremes in the wider field of climate modeling [140, 50, 108, 114, 121]. However, much of the previous Chapter and this body of work in general have been confined to simple, feed-forward neural networks that ignore the interesting stochastic details of eddy and storm organization. The recent application of Generative Adversarial Networks (GANs, [53]) to the Lorenz '96 Model suggests a potential, under-explored role for generative models in atmospheric sciences – particularly towards stochastic

parameterizations [49, 36]. There have also been initial successes using various types of GAN architectures to generate plausible Rayleigh-Bernard convection. In particular, adding informed physical constraints to GAN loss functions seem to improve the generation of these non-linear fluid flow systems [170, 166, 153, 172]. While promising, such techniques have thus far been restricted to idealized turbulent flows of reduced dimension and complexity; there is ample room to explore generative modeling methods for representing convective details amidst settings of realistic geographic complexity. Meanwhile, generative modeling besides GANs have not been as thoroughly considered for turbulent flow emulation and could potentially power climate models down the line.

VAEs may prove more appropriate than GANs for these climate applications given their design containing both a generative and representational model, their often superior log-likelihoods and reconstruction simulations, and practical advantages including stabler training results, easier performance benchmarking, and more interpretable latent manifold representations [167, 109, 65]. Modified VAEs can reconstruct plausible two-dimensional laminar flow with computational efficiency beyond what is common when numerically solving linear differential equations [46]. There has been preliminary work using VAEs for the clustering of atmospheric dynamics – a gain again relying on simplified Lorenz '96 model data as well as potential vorticity fields and geopotential heights [155, 156]. This application of representation learning across a variety of simplified simulations suggests VAEs offer great potential as both an engineering tool to help escape computational limits on the generative side and may provide the ability to learn and extract hidden organizational details in atmospheric dynamics on the representation side. However, to the best of our knowledge, this is the first work to use a VAE for representational learning on the details of convective organization and associated gravity wave radiation[1] as revealed by spatial snapshots of vertical velocity – an inherently chaotic and bimodal variable [51] – across a dataset large enough to nonetheless encompass

[1]Here we are referring to internal gravity waves, which are horizontally-propagating disturbances in the atmosphere generated by density perturbations, e.g. from deep convection, frontogenesis, or topography.

Layer	Filters	Kernel	Stride	Activation
2D Conv	64	3x3	2	relu
2D Conv	128	3x3	2	relu
2D Conv	512	3x3	2	relu
2D Conv (μ)	64	3x3	2	relu
2D Conv (σ)	64	3x3	2	relu

Table 3.1: Our Encoder architecture. Conv refers to a convolutional hidden layer. The first hidden Conv layer receives an input vector of 32x128 (30x128 expanded by padding) representing a vertical velocity snapshot.

the spatiotemporal diversity of turbulence regimes in the atmosphere. As far as we know, this is also the first study to constrain a VAE's output statistics by adding a covariance constraint term to its loss function to improve representation and capture variance details at small spatial scales in the turbulent atmospheric boundary layer, which can be considered one of the most difficult locations for climate models. Our results demonstrate the power of VAEs to accurately reconstruct high-resolution climate data, as well as their ability to leverage dimensionality reduction for high level feature learning and anomaly detection. This casts VAEs as promising tools for both dynamical analysis and stochastic parameterization of fine-scale atmospheric processes from storm-resolving data.

3.3 Methods

In this Section, we discuss the architecture of the three machine-learning models used here, the design of our covariance constrained VAE loss function, and the generation and preprocessing of the atmospheric simulation data.

Layer	Filters	Kernel	Stride	Activation
2D Conv-T	1024	3x3	2	relu
2D Conv-T	256	3x3	2	relu
2D Conv-T	64	3x3	2	relu
2D Conv (μ)	1	3x3	2	sigmoid
2D Conv (σ)	1	3x3	2	linear

Table 3.2: Our Decoder architecture. Conv-T refers to a transposed convolutional hidden layer.

3.3.1 Architecture

Our VAE takes vertical velocity fields formatted as (30×128) 2D images. We adopt a fully convolutional design[2] to preserve local information, which is essential in atmospheric convection modeling (Tables 3.1 and 3.2). We obtain meaningful reconstruction performance by ensuring that the information bottleneck in the VAE is not too severe, i.e. that the latent space is still wide enough to preserve enough fine features of the vertical velocity fields (in our case of dimension 1024), and by implementing annealing techniques outlined in [64, 6]. Here, we analyze two successful VAEs: One with a traditional negative ELBO in the loss, and one with an additional covariance constraint in the loss. As a baseline, we also implemented a regular autoencoder of the same design as above, with two key differences: All activations were replaced with the identity function and our covariance constrained loss was replaced with the mean-squared error. We refer to this model as the "linear" model and use it to better quantify the added value of VAEs for modeling atmospheric convection.

3.3.2 VAE Loss Implementation

The total loss is the sum of two terms: the negative of the Evidence Lower Bound (ELBO), commonly used as the total VAE loss, and a covariance constraint loss term [46, 153, 1] on

[2]Earlier experiments used architecture similar to models used for CIFAR-10 data [89] with fully connected dense layers separating the encoder and the decoder from the latent space but led to discouraging reconstructions plagued by posterior collapse and an inability to represent the spatial patterns of convection.

the covariance matrix that we weigh by $\lambda \in \mathbb{R}^+$:

$$\text{Loss} \stackrel{\text{def}}{=} -\text{ELBO} + \lambda \times \text{CC}, \tag{3.1}$$

where CC is a "covariance constraining" term using the Frobenius norm $||\cdot||$ to measure the distance between the covariance, Σ, of the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ and the covariance, Σ, of the true data distribution $p(\mathbf{x})$. θ refers to model parameters and $\mathbf{x}$ refers to observed vertical velocity fields:

$$\text{CC} = ||\Sigma(p_\theta(\mathbf{x}|\mathbf{z})) - \Sigma(p(\mathbf{x}))||. \tag{3.2}$$

Unconstrained VAEs ($\lambda = 0$), henceforth referred to as "VAE" for short, maximize the ELBO, defined as the sum of the log-likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, and the Kullback-Leibler (KL) Divergence between $p(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\begin{aligned} \text{ELBO}(\mathbf{x}; \theta, \phi, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})), \end{aligned} \tag{3.3}$$

where ϕ are our variational parameters which are learned jointly with the model parameters, θ. $p(\mathbf{z})$ refers to the prior and $q_\phi(\mathbf{z}|\mathbf{x})$ refers to the estimated posterior. We denote hidden variables as $\mathbf{z}$. Minimizing the KL loss term regularizes the variational parameters in the model and makes the VAE posterior more similar to the VAE prior. Maximizing the log-likelihood enables the VAE to produce realistic vertical velocity fields where the output will be more closely aligned with the latent variable of the model. Following [80], we assume that

the prior over the parameters and the hidden variables are both centered isotropic Gaussian and calculate ELBO using equation (24) of [80].

To control the rate-distortion trade-off [6], we implement linear annealing to the KL loss term following [21], where the KL term is multiplied by an annealing factor linearly scaled from 0 to 1 over the course of training. In our VAE, linear annealing results in significantly lower KL losses and more interpretable latent spaces.

Finally, to generate vertical velocity fields with realistic spatial variability, we additionally implement covariance-constrained VAEs. Following Equation 3.2, the covariance constraint is defined as the Frobenius norm of the covariance matrix error, which we estimate over each batch during optimization. We choose a pre-factor $\lambda = 10^6$ so that the magnitude of the covariance constraint matches that of the reconstruction loss, resulting in a covariance-constrained VAE "CC-VAE" that generates more faithful covariance matrices.

3.3.3 Data & Preprocessing

Storm-Resolving Data

To train and test our VAE, we rely on snapshots of vertical motions with explicitly-resolved moist convection and gravity wave radiation obtained from ∼15k instances of a Storm-Resolving Model (CRM) [75, 77] embedded within a host Global Climate Model (GCM). The CRMs operate at a 20s native timestep data and we extract state snapshots from it every 15 minutes, the frequency with which its horizontal average state is permitted to interact with its host GCM. We perform a 100-day multi-scale climate simulation to generate data showing details of atmospheric convection within a tropical belt from 20N to 20S latitudes. Specifically, at each $1.9^\circ \times 2.5^\circ$ horizontal grid cell of the Super-Parameterized Community Atmosphere Model (SPCAM5), we embed a 128-column System for Atmospheric Modeling

(SAM) micro model with kilometer-scale horizontal resolution; both the host and embedded models use 30 vertical levels. This entire dataset comes to a size of 1.3 Tb. For our purposes, there is 30 level by 128 CRM-column "snapshot" or "image" of a convective-scale vertical velocity field at each latitude-longitude grid cell that we feed into the encoder of our neural network. We train our VAEs on sub-samples of this data staged on UC Irvine's GreenPlanet Super-computing node and our machine learning simulations are powered by two NVIDIA Tesla V100 and one NVIDIA Tesla T4 GPUs.

Preprocessing

To reduce data volume for efficient training and to ensure our VAE is exposed to a plethora of convective motion, we selectively sample from the initial 1.3Tb SAM dataset. We restrict our initial data volume to the 144 latitude/longitude coordinates with a detectable diurnal cycle of precipitation where the amplitude of daily precipitation is greater than two times its standard deviation within the larger-scale host model. This precipitation filtering ensures samples of strong convection get placed into the training dataset, as a persistent diurnal cycle of precipitation often indicates deep convection and the presence of mesoscale convective systems [33]. Within these selected grid cells, the vertical velocity values range from 37.3m s^{-1} to -17.4m s^{-1} and are then scaled from 0-1 by subtracting the minimum and dividing by the range.

We shuffle data in the spatial and temporal dimensions prior to training. We use An 80%/20% training/test split for all models. To ensure a balanced dataset of different convective types, we apply K-means clustering with two centroids to group data with active and inactive vertical velocity fields. We then sample equally from both clusters without replacement to design a balanced dataset for the VAE. This new 4.3Gb dataset has a 111206/27802 training/test split. Since the horizontal domain is doubly-periodic, two vertical velocity updrafts of equal magnitudes and size located at different horizontal locations are physically identical. To

prevent the VAE from treating them as different at the expense of reconstruction magnitude and variance, we preprocess all samples so that the center of the vertical velocity field is the location of the strongest convection present in the sample. We define the "strongest convection" as the largest absolute value of spatially-averaged vertical velocity, from 400hPa to 600hPa in the vertical and using a moving average of 10km horizontally.

3.3.4 Quantifying Reconstruction Performance

We quantify the reconstructions of our final VAE and CC VAE as well as our linear baseline using the following metrics:

Hellinger Distance

We calculate the Hellinger distance H between the discrete distributions to gauge similarity [113]:

$$H(p,q) = \sqrt{\sum_{i=1}^{k} \frac{(\sqrt{p_i} - \sqrt{q_i})^2}{2}} \tag{3.4}$$

where p is the distribution of the original vertical velocity fields and q is the distribution of the corresponding reconstruction.

Mean Squared Error (MSE)

To provide an overall skill of the reconstruction, the MSE is calculated between each original sample and its corresponding reconstruction.

Model	MSE	Hellinger Distance	Frobenius Norm
Linear	4.2e-6	2.0e-3	8.0e-3
VAE	1.1e-5	3.1e-4	3.2e-4
CC VAE	4.5e-6	2.0e-3	8.0e-6

Table 3.3: **Quantitative Reconstruction Metrics.** We compute the MSE and Hellinger Distance between true and predicted reconstructions. This shows the baseline is equally good at predicting the mean reconstruction. We also compute the Frobenius Norm of the error in the covariance matrices of the true data and the reconstructions. Both VAEs capture more of the covariance structure of the data than the linear baseline.

Spectral Analysis

To better understand the skill of the VAE reconstruction from a spatial perspective, we perform one-dimensional spectral analysis on each sample and reconstruction at all 30 levels in the vertical dimension. We examine four vertical levels commonly used in meteorology: 850hPa (top of the boundary layer), 700hPa (lower troposphere), 500hPa (mid-troposphere), and 250hPa (upper-troposphere) to see how our VAEs capture the spatially-resolved vertical velocity variance throughout the atmosphere. We calculate the power spectral density Φ_{k} using:

$$\Phi_k \stackrel{\text{def}}{=} \frac{\Delta n}{N} \left| \sum_{j=0}^{N-1} y_j e^{\frac{-ijk}{NT}} \right|^2 \tag{3.5}$$

where N is the length of the x dimension, $\mathrm{y_j}$ is the sample or reconstruction, T is 1/length, i is the imaginary unit, and k is the vertical level of interest in hPa (850, 700, 500, or 250) [35].

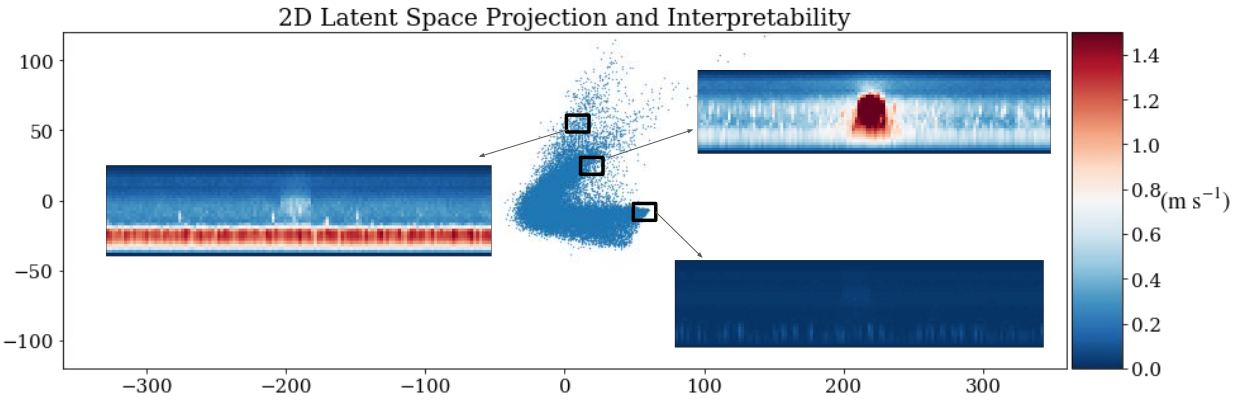

Figure 3.1: **Visualization of the latent space** originally in dimension 1024, but reduced to dimension 2 by Principle Component Analysis (PCA) [130]. The standard deviations of different types of convection the VAE learns to cluster are embedded near corresponding clusters. This suggests the VAE learns an interpretable clustering of the data, with means and variances both contributing to the results.

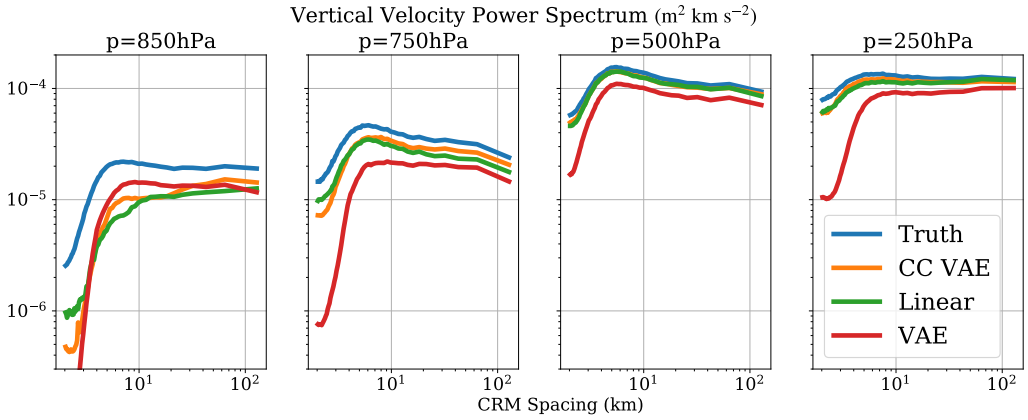

Figure 3.2: **Spectral Analysis** at 4 different levels of the atmosphere comparing the test data to our best VAE and CC VAE as well as a linear model. At small spatial scales, we see the importance of the Covariance Constraint to capture the variance native to convection (orange vs. red).

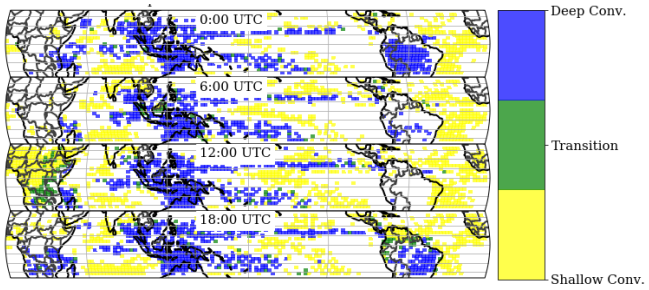

Figure 3.3: **Convection Type Predictions** The diurnal composite from a ten day average at four unique times of day are shown above. The VAE predicts the type of convection occurring in tropical locations over the course of a typical Boreal Winter Diurnal Cycle. Blue coloring refers to a VAE prediction of deep convection, yellow to a VAE prediction of shallow convection, and green to a convective type transitioning between shallow or deep convection. Areas where the VAE detects little convection are blanked out. Semantic similarities of the VAE latent space are reflected in the global geospatial weather patterns.

3.4 Results

Our VAE trained on storm-resolving climate data produces accurate vertical velocity field reconstructions. When we provide the high resolution training dataset and appropriate convolutional architecture, our VAE learns remarkably accurate representations of any type of convection found within the test dataset. Our VAE captures the magnitude, proper height, and structure across deep convective regimes, shallow convective regimes, and non-convecting regimes (Figure 3.4). When the "Covariance Constraining" term is added to create a physically informed loss, the CC VAE performance improves enough to match a linear baseline (Table 3.3). But unlike many other image recognition tasks generative models perform, reconstructing the mean of the convection is necessary but not sufficient – we must capture the variance and correlation in the vertical velocity fields. The CC VAE reconstructs variance better than a traditional convolutional VAE and at least on par with the linear baseline (Table 3.3, Figure 3.2). Our CC VAE is the most versatile of our models with an accurate reconstruction performance overall at different levels of the atmospheric column and different convective spatial scales based on the power spectra of the three models (Figure 3.2). This precision across both small and large spatial scales reveals our CC VAEs ability to

emulate both the overall large pattern of convective plumes and the details within convective composition. Our CC VAEs results replicate disparate structures of convection in both areas of high stochasticity near the atmospheric boundary layer, characteristic of shallow convection, as well as in the upper troposphere, where deep convective regimes dominate. At this stage, CC VAEs match the performance of our linear baseline but do not exceed it.

However, unlike the linear baseline, our VAE and CC VAE discover the details of the convective organization by representation learning via dimensionality reduction and feature extraction. A 2D, deterministic PCA projection of our CC VAE latent space clusters and separates different convective types (Figure 3.1). In particular, the distinction between deep and shallow convective regimes and non-convective regimes is encouraging (Figure 3.1, please visit this link for a complete animation of the 2D Projection of the latent space). The physical knowledge represented in our CC VAEs latent space stands alone from other forms of dimensionality reduction (PCA and t-SNE on the preprocessed data) where there is no evidence of distinction based on convective type. Furthermore, CC VAE predictions of convective type based solely on latent space location map back to a physically sensible pattern over the tropics with deep convection concentrated on land over the Amazon and African Rainforests as well as over the Pacific Warmpool (Figure 3.3). These predictions from latent space location not only map convection type in a spatially coherent pattern, but also capture the change in convection type with the diurnal cycle over moist, tropical continents (Figure 3.3, please visit this link for a complete animation of the tropical diurnal cycle). When we exclusively restrict the test dataset to an Amazon Diurnal Composite, the known coherent transitions from shallow to deep convection that occur over tropical rain-forest in response to solar heating of the diurnal cycle correspond to monotonic trajectories in the latent space projection, verified using both t-SNE and PCA (Figure 3.5). Further tests are required on more complex convective transitions to understand the extent of the physical meaning of the CC VAE latent space, but these initial positive results suggest great potential for physically constrained VAEs as a tool in atmospheric dynamics to uncover information

about convective transitions, storm morphology, and propagation.

We also evaluate ELBO (Equation 3.3) for each sample of our test data to find unusual storm development and activity in the dense CRM data.

ELBO allows us to determine the degree to which a vertical velocity field, drawn from our model's latent variables is an aberration in the data. Our VAEs inherent ability to detect anomalies in the vertical velocity data proves to be an elegant way to identify deep convection in a more thorough manner than traditional vertical velocity thresholding. An example of one such anomaly we identify is Figure 3.6 – in this case an instance of two moderate storms developing in one CRM array. These phenomena would be less straightforward to locate through conventional methods, particularly given the size and density of data involved. Our VAEs attribute of anomaly detection learns characteristics of the data instead of naively thresholding based on priors experiences that may not reflect the composition of the dataset. This feature provides the potential to help identify interesting and unexpected weather phenomena from noise – artifacts that might otherwise never be studied in overwhelmingly large and rich datasets.

3.5 Conclusion

We develop a VAE to reconstruct immaculate convection images from a high-resolution, storm-resolving dataset. Our VAE, particularly once a statistically constrained loss function is added, captures the variance and magnitude of distinct convective regimes. The latent space of the VAE proves to be a potent tool for making physically sensible predictions of convection type that accurately reflect the tropical atmosphere and capture the effects of solar heating through the diurnal cycle. The unique VAE loss function allows us to use ELBO to find anomalous storm development in a dense, high resolution dataset that traditional

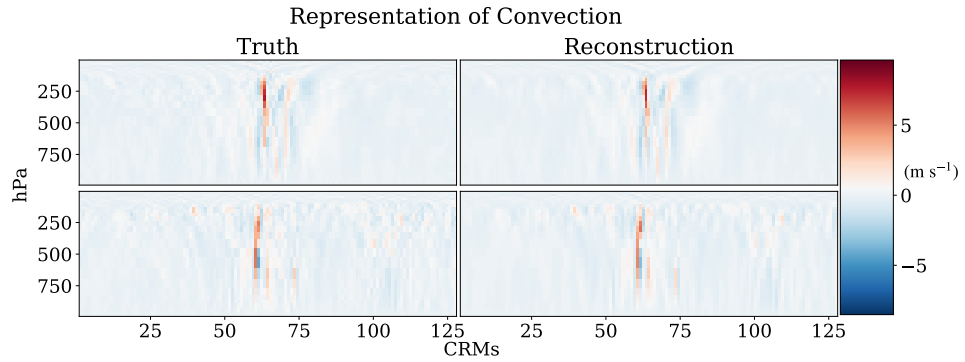

Figure 3.4: **Reconstructions** The trained VAE reconstructions closely resemble those from the test dataset and accurately predict the location, magnitude, and spatial structure of convective plumes.

methods might miss. But there is much work to be done before a VAE could be implemented to power stochastic parameterizations for a climate model, likely requiring to condition the VAE on large-scale thermodynamics via expansion of the input vector. If successful, the ability to quickly and efficiently generate synthetic, detailed vertical velocity fields to help run climate models would be a valuable resource for the atmospheric sciences and meteorology communities. But improvements in the generative capabilities would likely come at the expense of the representation learning and the VAEs diagnosis of the physics of convection. We believe these preliminary physical intuitions achieved via latent space analysis represent a promising avenue for the broader application of generative modeling for advancing the field of atmospheric dynamics [6, 64] and warrant further investigation to understand their full potential. In the following Chapters (4 and 5) we will show how the VAE encoder and latent space can create a framework to both better understand model design choices and consequences as well as identify mechanisms driving changes in precipitation.

3.6 Appendix A: Additional Figures

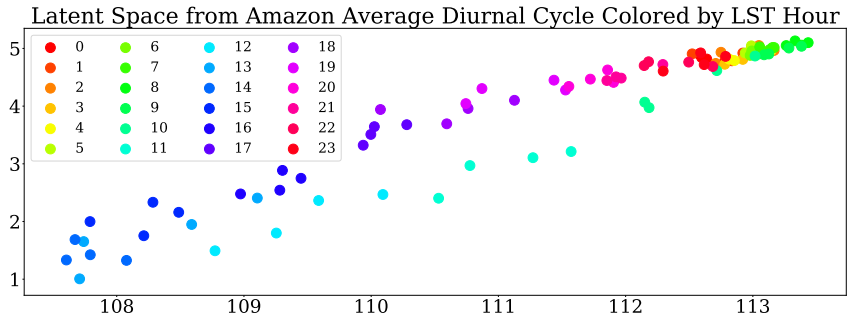

Figure 3.5: **2D PCA Temporal Projection** All spatial locations comprising the Amazon Rainforest are averaged together from November to February to get a single composite diurnal cycle that is fed through our trained VAE. The colors above correlate to the time of day (Local Solar Time). The results show a clear separation in representation within the latent space of the timing of deepest convection and maximum precipitation (mid-afternoon) from when shallow convection and calmer conditions dominate (early morning).

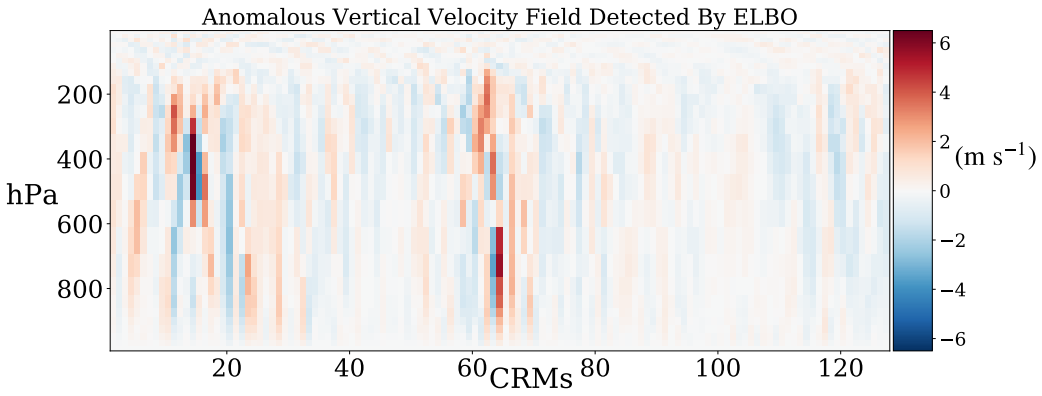

Figure 3.6: **Anomaly Detection** We use the ELBO in the VAE Loss function to identify the most anomalous vertical velocity fields. We show the 9th most anomalous field because it exhibits multiple deep convective plumes.

Chapter 4

Comparing Storm Resolving Models and Climates

4.1 Abstract

Storm-resolving models (SRMs) have gained widespread interest for the unprecedented detail with which they resolve the global climate. However, while many different SRMs have been created, it remains difficult to quantify objective differences in how these SRMs resolve complex atmospheric formations. Understanding the design choices that lead to these differences is also an opaque process. The lack of appropriate tools for quantifying model similarities and differences is not unique to climate science but encountered in many disparate fields that involve complex data simulation tools. This Chapter develops a new unsupervised machine learning workflow to analyze and intercompare different SRMs based on their high-dimensional simulation data. Instead of drawing on domain knowledge, the approach automatically learns appropriate notions of similarity from low-dimensional latent data representations that the different models produce. To quantify such inter-SRM "distribution shifts", we use variational

autoencoders in conjunction with vector quantization. Our analysis involving nine different global SRMs reveals that only six of them are aligned in their representation of atmospheric dynamics. Our analysis furthermore reveals regional and planetary signatures of the convective response to global warming in a fully unsupervised, data-driven way. In particular, this approach can help elucidate the effects of climate change on rare convection types, such as "Green Cumuli". This Chapter provides a path toward evaluating future high-resolution global climate simulation data more objectively and with less human intervention than has historically been needed.

4.2 Introduction

Modern storm-resolving models better represent the totality of the atmosphere on scales ranging from microns (cloud microphysics) to kilometers (convective storms that mediate floods) to many thousands of kilometers (organized storm systems), there is still a spread in the representation of atmospheric dynamics [25, 18, 146, 152]. This does not mean these "storm-resolving models" are not still of great use – Features like deep convective updraft formation can be resolved explicitly and we can improve the emulation of clouds and precipitation patterns in conventional climate simulations by reducing excessive drizzle and correctly representing the onset of deep convection in the afternoon [139, 32, 37, 96, 95, 84]. But substantial differences in SRM design choices, including treatment of sub-km scale shallow convection, initialization of soil moisture and the land surface, and inconsistent vertical coordinate systems, all contribute to uncertainty in SRM weather and climate predictions [152]. Comprehensively separating groups of self-similar from dissimilar models remains challenging. Attempts have been made before to validate and intercompare ensembles of global SRMs, but have been limited to traditionally coarsened statistics guided by physically informed approaches [73, 24, 103, 152]. A longstanding community goal is to directly inter-

compare models at the native scale of storm formation, which would improve understanding of the consequences of different model design decisions and help narrow the uncertainty of cloud-climate feedback.

A key part of the problem is the sheer amount and complexity of the data created by the simulation output, which quickly becomes overwhelming. The extent of this challenge can be observed in the first inter-comparison study of global SRMs, DYAMOND (the DYnamics of the Atmosphere general circulation Modeled on Non-Hydrostatic Domains). Preserving hourly simulation output for just 40 days necessitated nearly 2 PBytes per SRM for the DYAMOND Initiative [152]. This makes storage, let alone any detailed analysis, a significant hurdle. Traditional dimensionality reduction methods, including clustering and projections, are used to help analyze SRM simulations to cope with such issues. However, the required assumptions (such as linearity and scale selectivity) may fail to fully capture the non-linear relationships embedded in small-scale coherent structures that make these simulations so valuable to begin with [19, 168, 165].

To gain more insights and confidence in model predictions, we need objective ways to quantify changes in convective organization, highlight nonphysical artifacts in simulations, and more thoroughly analyze these modern SRMs [152, 126]. This Chapter proposes to compare models based on their high-resolution simulation data, i.e., by quantifying *distribution shifts* among the outputs of different SRMs. Machine learning methods, typically grounded in the assumption that training and test data are drawn from the same distribution, can aid in this endeavor. Since test data often deviates from training data [13], distribution shifts are a frequently occurring problem, and methods for detecting such "out-of-distribution" data have been developed [144, 145]. Yet, with few exceptions [138], most of this work has focused on detecting individual data instances and not on comparing data distributions as a whole. This Chapter develops a methodology based on a combination of nonlinear dimensionality reduction and vector quantization [55, 171] as used in data compression to

estimate distributional distances. Chapter 4 will build on the unsupervised learning methods of Chapter 3 for a new way to build inter and intra-model comparisons with only minimal physical domain knowledge.

4.3 Methods

4.3.1 Data and Preprocessing

We begin by discussing the mechanics of the Multi-Model Framework (MMF) in more detail. The MMF responsible for generating our SPCAM data is composed of small, locally periodic 2D subdomains of explicit high resolution physics that are embedded within each grid column of a coarse resolution ($1.9^\circ \times 2.5^\circ$ degree) host planetary model [77]. The simplifications behind the MMF-imposed scale separation provide a useful contrast to the cutting-edge DYAMOND SRMs. The resolution and geography of the data will be the same as Chapter 3, but here we performed six simulations of present-day climate launched from different initial conditions (but identical resolution) using the MMF [139], configured with storm resolving models that are 512 km in physical extent, each with 128 grid columns spaced 4 km apart. This ensures our results will not be overly influenced by the biases of a particular sample.

We also expand beyond the scope of Chapter 3 by exploring state-of-the-art high-resolution[1] atmospheric model data, archived by the DYAMOND Project [152]. We consider eight SRMs from the DYAMOND initiative: ICON, IFS, NICAM, UM, SHIELD, GEM, SAM, and ARPEGE. As a ninth data set, we also extract high-resolution vertical velocity fields produced by SPCAM MMF that embeds many miniature 2D SRMs in a planetary climate model) [75, 77]. We focus exclusively on the vertical velocity state variable and its 3D structure, which contains information about complex updraft and gravity wave dynamics

[1] 5 kilometers or less horizontally

across multiple scales and phenomena. We eliminate one spatial (latitude) axis from the DYAMOND datasets by extracting 2D "image" snapshots (pressure vs. longitude), which are temporally spaced by three hours; this set-up allows us to directly compare the DYAMOND and SPCAM outputs. In each test dataset from each model, 125,000 samples are selected randomly (with respect to space and time) across the 15S-15N latitude belt, comprising diverse tropical convective regimes. The nine SRMs differ in horizontal and vertical resolution and other sub-grid parameterization choices (see [152] Tables 1 and 2). These SRMS form a comprehensive testbed of vertical velocity imagery. For visualization of these convective updrafts (which comprise our training and test data) in nine different SRMs, see Figure 4.1 and Movie 4.1.

To directly compare distributions between our DYAMOND simulations as well as with SPCAM, we preprocess accordingly:

- In all nine simulations, we draw data from boreal winter and focus exclusively on the tropic belt (15°S to 15°N latitude).
- We extract 2D snapshots of vertical velocity in the pressure-longitude plane from the 3D data (all eight DYAMOND datasets) in order to compare to the MMF dataset in its native form.
- Each pixel in a 2D snapshot is scaled from its original velocity value in meters per second (m/s) to values between 0 and 1 based on the highest pixel value and the lowest among all the snapshots.
- For pairwise comparisons, we interpolate the data to a common vertical (pressure) and horizontal grid and use consistent normalizations to generate directly comparable test datasets.
- We create the training dataset by extracting 1.6e5 sample images randomly with respect

to time and geography to ensure we densely sample the rich spatial-temporal diversity of tropical weather, turbulence, and cloud regimes.

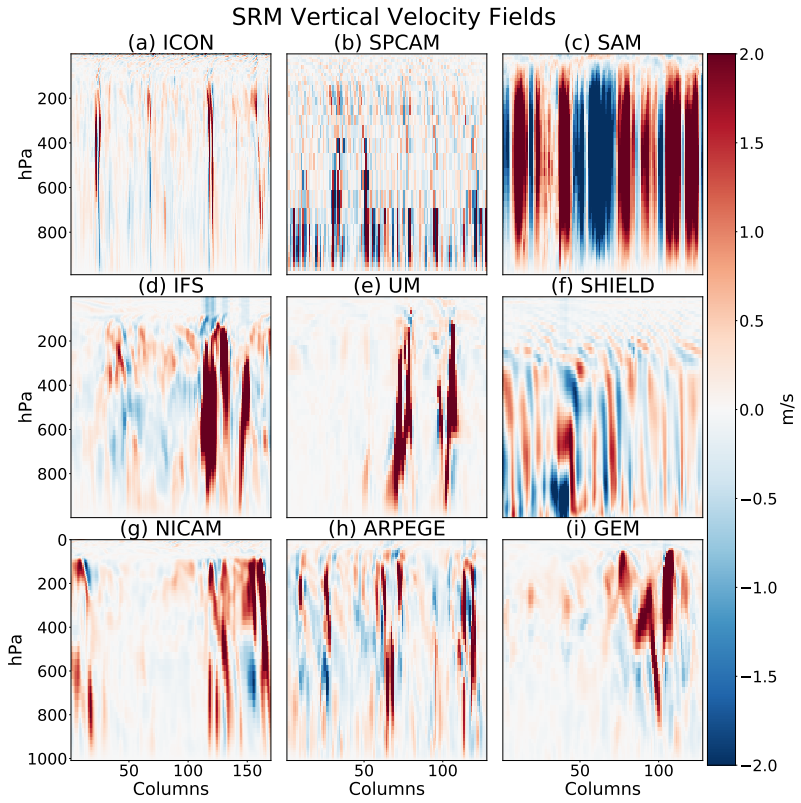

Figure 4.1: A randomly selected vertical velocity field from each of the nine SRMs used in this intercomparison. Atmospheric pressure is denoted on the y-axis and the number of embedded columns in a given snapshot is shown on the x-axis. We see a rich mix of turbulent updrafts (red) of various scales and types. Each model has a different native horizontal spatial resolution. For more examples, see Movie 4.1

Figure 4.1 provides vertical velocity snapshots for various models used in this Chapter. For more examples, see Movie 4.1

Unlike the DYAMOND SRMs, we can use the MMF model to simulate global warming through uniformly increasing sea surface temperatures by four Kelvin. We treat this setup as a proxy for climate change, which we can better understand by examining spatial and

intensity shifts between the turbulent updrafts within the two SPCAM climates.

4.3.2 Variational Autoencoders

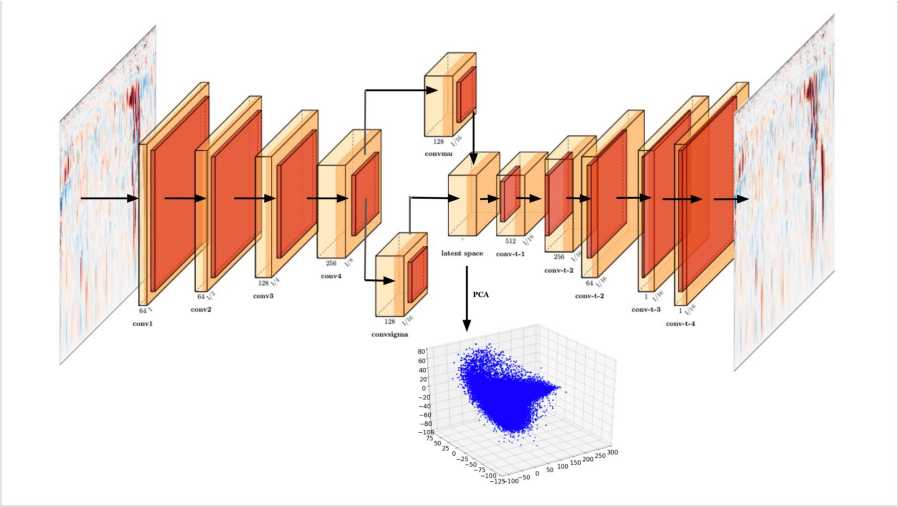

Figure 4.2: Typical VAE architecture used in Chapter 4. Given a vertical velocity field $\mathbf{x}$, the VAE reduces the input dimensionality, resulting in a latent representation $\mathbf{z}$. We use PCA to further reduce and visualize the latent structure in two (Figures 2, 4.12-4.14) or three (Movies 4.2-4.6) components. Based on $\mathbf{z}$, the VAE is trained to reconstruct $\mathbf{x}$ as $\hat{\mathbf{x}}$.

VAEs are among the most popular models for high-dimensional density estimation and non-linear dimensionality reduction [80]. In contrast to regular autoencoders, VAEs are probabilistic models that assign likelihoods to individual data instances and can be used for density estimation and anomaly detection [144, 145]. Readers familiar with the concepts established in Chapter 3 can skip to the next paragraph.

VAEs model data points $\mathbf{x}$ in terms of a *latent variable* $\mathbf{z}$, i.e., a low-dimensional vector representation, through a conditional likelihood $p(\mathbf{x}|\mathbf{z})$ and a prior $p(\mathbf{z})$. Integrating over the latent variables (i.e., summing over all possible configurations) yields the data log-likelihood as $\log p(\mathbf{x}) = \log \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. This integral is intractable, but can be lower-bounded by a quantity termed evidence lower bound (ELBO),

$$\mathcal{L}\left(\theta;\mathbf{x}\right) := \mathbb{E}_{q_\theta\left(\mathbf{z}|\mathbf{x}\right)}\left[\log p_\theta\left(\mathbf{x}|\mathbf{z}\right)\right] - \mathrm{KL}\left[q_\theta\left(\mathbf{z}|\mathbf{x}\right)||p_\theta\left(\mathbf{z}\right)\right]. \tag{4.1}$$

This involves a so-called variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, also called "encoder", while $p_\theta(\mathbf{x}|\mathbf{z})$ is also called "decoder". Both encoder and decoder are parameterized by neural networks, see [80] for details. The first term measures the expected log-likelihood of a data point $\mathbf{x}$ upon first stochastically mapping it to a latent state $\mathbf{z}$ and then decoding back to $\mathbf{x}$. The term therefore measures a reconstruction error and forces the latent variable $\mathbf{z}$ to be informative of $\mathbf{x}$. In contrast, the second term measures the distance between the distribution of the latents $\mathbf{z}$ to the prior $p(\mathbf{z})$. As discussed in [64], this term encourages the autoencoder to "disentangle" the input images for added human interpretability shown below. For typical machine learning parameterizations and training details, we refer to the literature [80].

To ensure the local correlations in the updrafts of our vertical velocity fields are preserved, we rely on a fully convolutional VAE design. For training the VAE, we perform beta-annealing [64, 21]. To this end, we expand the ELBO of Equation 4.1 by including a β parameter and linearly anneal β from 0 to one:

$$\mathrm{ELBO}\left(\mathbf{x};\theta,\phi\right) = \mathbb{E}_q\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\,||\,p(\mathbf{z})\right). \tag{4.2}$$

We anneal over all 1600 training epochs. The number of layers and channels in the encoder and decoder are depicted in Figure 4.2 (4 layers in each, stride of two). We use ReLUs as activation functions in both the encoder and the decoder. We pick a relatively small kernel size of 3 in order to preserve the small-scale updrafts and downdrafts of our vertical velocity fields. The dimension of our latent space is 1000.

4.3.3 Understanding Convection via Vertical Structure

The sheer volume of the vertical velocity fields from DYAMOND makes analyzing differences in physical properties of convection difficult. To analyze their physical coherence, we therefore use vertical velocity moments as summary statistics. More specifically, we create the anomaly profiles of the vertical velocity fields. This approach is grounded on the principle that for convection the vertical (v) dimension will be far more important than the horizontal (h) dimension for preserving the key physical signatures. We calculate the first-moment statistic:

$$\overline{\mathrm{w'w'_i}} \stackrel{\mathrm{def}}{=} \sqrt{(\mathrm{W_i} - \bar{\mathrm{W}}_{\mathrm{i,h}})^2}, \tag{4.3}$$

where $\bar{W}_{i,h}$ is the mean of the vertical velocity field upon averaging-out the horizontal dimension. Equation 4.3 effectively creates a low dimensional portrait of the full snapshot. We can average these statistics across a cluster to approximate the convective structures organized within. More specifically we use this metric to quantify the average physical properties sorted by the VAE latent space in Figures 4.3, 4.5, 4.8, 4.10, 4.15, and 4.18.

4.3.4 The Horizontal Extent of Convection

The following metric measures the spatial extent of our turbulent updrafts across the vertical velocity snapshots. We will rely on a derivation of the Turbulent Length Scale (TLS) [15]. More specifically we calculate the power spectrum of weighted averaged length:

$$\mathrm{TLS_i} \stackrel{\mathrm{def}}{=} \frac{2\pi\sqrt{\mathrm{n}}}{\langle\varphi_\mathrm{i}\rangle}\left\langle\frac{\varphi_\mathrm{i}}{||\mathrm{k}||}\right\rangle, \tag{4.4}$$

where φ represents the power spectra, $||k||$ is the modulus, n is the number of dimensions, and we calculate over the vertical integral $\langle\rangle$. We derive a unique TLS at each vertical level, i, before summing to get a composite of a given vertical velocity field. This serves as a proxy for the width of an updraft or downdraft and allows us to analyze the degree of coherent organization in the horizontal dimension.

4.3.5 K-Means Clustering of Tropical Convection

We apply the K-Means Clustering algorithm to partition the latent space of our VAE and analyze which physical properties are clustered in this reduced order $\mathbf{z}$ space. This approach first randomly assigns centroids, C, to locations in the $\mathbf{z}$ space to maximize the initial distances between the centroids). Latent representations of each sample $\mathbf{z}_i$, in the test dataset of size N, are assigned to their nearest centroid. The second stage of the algorithm moves each centroid to the middle of its assigned cluster. The process repeats until the sum of the square distances (or the Inertia, I) between the latent space data points and the centroids are minimized [98, 100] such that:

$$\bar{\mathrm{I}} \stackrel{\mathrm{def}}{=} \sum_{\mathrm{i}=0}^{\mathrm{N}} \min_{\mathrm{l}\in\mathrm{C}} ||\mathrm{z_i} - \bar{\mathrm{z}}_\mathrm{l}||^2, \tag{4.5}$$

in which $\bar{z}_l$ is the mean of the given samples belonging to a cluster l for the total number of cluster centers C. We always calculate ten different K-means initializations and then select the initialization with the lowest inertia. This process allows us to derive the three data-driven convection regimes within a SRM highlighted in Figure 4.7h.

To confirm the robustness of these clusters we perform a hyper-parameter sweep over the clustering routine type (k++ or true K-Means) and the number of initializations. From one

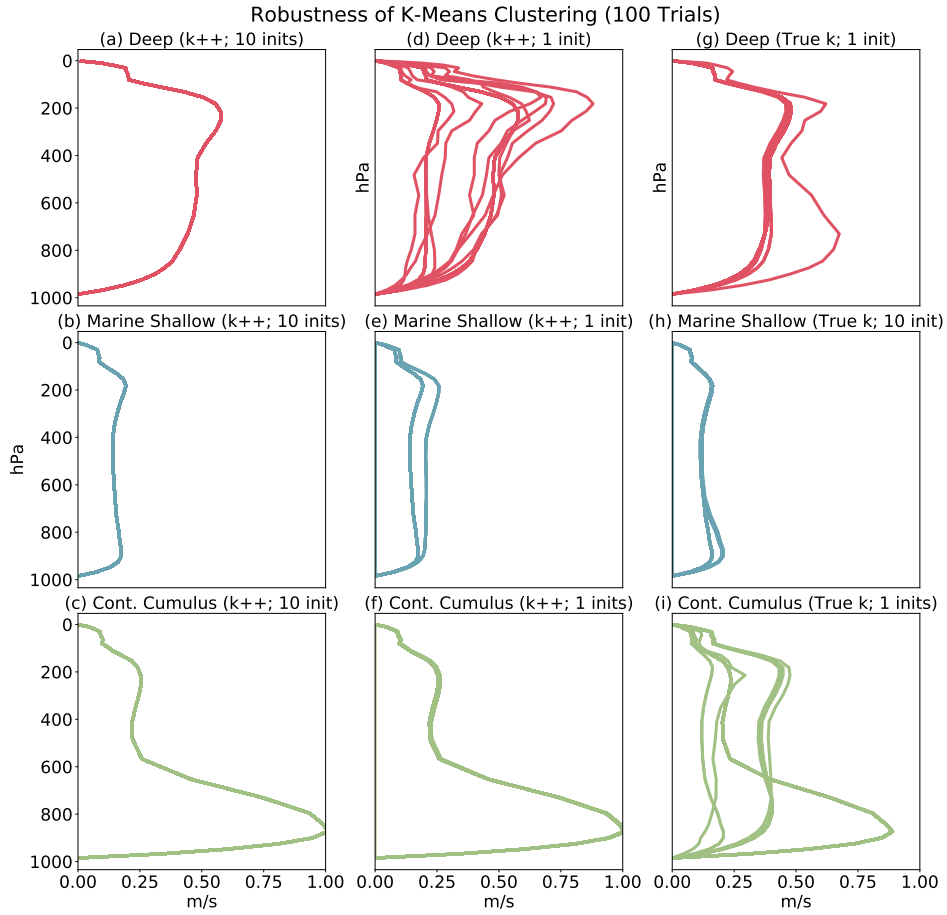

Figure 4.3: Our hyperparameter sweep for the k-means clustering algorithm. In all cases, we set k=3, but sweep over algorithm choice (k++ vs. true k-means) and a number of initializations. Each panel shows a cluster's median vertical structure. Fewer profiles indicate more robust clusters between different trials.

hundred trials we observe a combination of the more modern k++ algorithm [9] and sufficient initializations (ten) yields three reproducible clusters (Figure 4.3)

We qualitatively choose an optimal number of cluster centroids (centers), k by incorporating domain knowledge rather than a traditional approach relying on the rate of decrease in I as k increases or a single quantitative value such as a Silhouette Coefficient [143] or Davies-Bouldin Index [38]. More specifically, we identify the maximum number of "unique clusters". We define a "unique cluster" of convection as a group in the latent space where the typical physical properties (vertical structure, intensity, and geographic domain) of the vertical velocity fields are not similar to the physical properties of another group elsewhere in the latent space.

Empirically this exercise enables us to create three unique regimes of convection (Figure 4.8). When we increase k above three, we get sub-groups of "Deep Convection" without differences in either vertical mode, intensity, or geography. Thus we don't consider $k > 3$ to be physically meaningful for our purposes.

A key benefit of our method is the ability to create directly comparable clusters of convection between different SRMs. Because we seek to contrast clusters between different data, we do not use Agglomerative (hierarchical) Clustering, unlike other recent works that cluster compressed representations of clouds from machine learning models [40, 91]. Using the K-means approach, we can save the cluster centroids at the end of the algorithm. This provides a basis for cluster assignments for latent representations of out-of-sample test datasets when we use a common encoder as in Claim Two of our results section. More specifically, we only use the cluster centroids to get label assignments in other latent representations. We don't move the cluster centroids themselves once they have been optimized on the original test dataset (the second part of the K-means algorithm). Keeping the center of the clusters the same between different types of test data ensures we can objectively contrast cluster differences through the lens of the common latent space. This process allows us to create interpretable regimes of convection across nine different SRMs Figure 4.7 (d-l).

4.3.6 Vector Quantization

We seek to approximate differences between data distributions by directly estimating their Kullback-Leibler (KL) divergence. The KL divergence is always non-negative and only zero if two distributions match, but note it is non-symmetric and therefore not a proper distance. For any two continuous distributions $p^A(\mathbf{x})$ and $p^B(\mathbf{x})$, the KL divergence is defined as $KL(p^A||p^B) = \mathbb{E}_{p^A(\mathbf{x})}[\log p^A(\mathbf{x}) - \log p^B(\mathbf{x})]$. However, if both distributions are only available in the form of samples, the KL divergence is intractable since the probability densities are

unavailable.

In theory, the KL divergence between data distributions can be approximated well by a technique called vector quantization [55]. This amounts to coarse-graining an empirical distribution to a discrete one obtained from clustering and then working in a discrete space where the KL divergence is tractable. In more detail, we can perform a K-means clustering on the union of both data sets. This results in K cluster centers μ_k and N cluster assignments $m_{ik} \in \{0, 1\}$, where $m_{ik} = 1$ if data point $i \in \{1, \cdots, N\}$ is assigned to cluster k, and $m_{ik} = 0$ otherwise. $\pi_k = \frac{1}{N} \sum_{i=1}^{N} m_{ik}$ count the fraction of data points being assigned to each cluster and thus represent "cluster proportions". By increasing the number of clusters (making enough bins), we can more and more confidently quantize continuous distributions into discrete ones. The two data distributions $p^A(\mathbf{x})$ and $p^B(\mathbf{x})$ result in two distinct cluster proportions π^A and π^B for which we can estimate the KL as

$$\mathrm{KL}\left(p^A\left(\mathbf{x}\right)||p^B\left(\mathbf{x}\right)\right) \geq KL\left(\pi^A||\pi^B\right) = \sum_{k=1}^{K} \pi_k^A \log \frac{\pi_k^A}{\pi_k^B}. \tag{4.6}$$

The inequality comes from the fact that any such discrete KL estimate lower-bounds the true KL divergence [44].

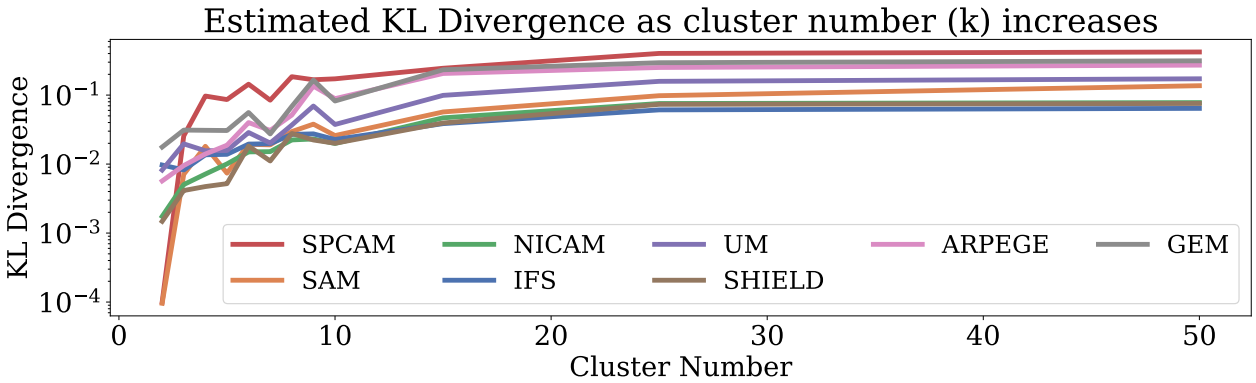

Figure 4.4: Approximating the KL divergence using vector quantization (VQ) based on k-means clustering, using a variable number of clusters. The VQ lower-bounds the KL and becomes asymptotically exact for large k. We considered the distributional divergence between ICON and the eight other SRMs. Empirically, the KL approximation seems to saturate at $k = 50$.

Vector quantization suffers from the curse of dimensionality. To mitigate this issue, we work in the latent space of a VAE and cluster the latent representations of the data instead (i.e., we replace $\mathbf{x}$ by $\mathbf{z}$ in Eq. 4.6). Our VAE's latent space still has sufficiently high dimensionality (typically 1000) to allow for a reliable KL assessment. In the Supplementary Information provided, we investigate the required cluster size to get convergent results and find that $K = 50$ gives reasonable results (Figure 4.4).

4.3.7 Computing Pairwise SRM Dissimilarities.

To quantify similarities and dissimilarities among the data that different SRMs produce, we adopt the Vector Quantization approach for computing KL divergences. Since the KL divergence is not symmetric, we explicitly symmetrize it as $KL(q||p) + KL(p||q)$ (termed *Jeffreys divergence*). Since we adopt vector quantization in the latent space, this amounts to training nine different VAEs, one for each SRM. Briefly, to compare Models A and B, we (i) save the K-means cluster centers from the latent vector of the VAE trained on Model A, (ii) feed both models' outputs into Model A's encoder as test data, (iii) obtain discrete distributions of cluster proportions for Model A and Model B, and (iv) compute symmetrized KL divergences based on the discrete distributions using the right-hand side of Eq. 4.6.

4.3.8 Baselines

Our approach for a data-driven inter-comparison of SRMs involves two key uses of machine learning: (1) The use of a VAE Encoder for non-linear dimensionality reduction and (2) K-Means Clustering of the latent representation of the SRM to approximate the true lower bound of the KL Divergence. We now test baselines to ensure this is the appropriate workflow for achieving our inter-comparison results. We know from testing that the clustering of the latent representation is both robust (Figure 4.3) and yields clusters of convection with

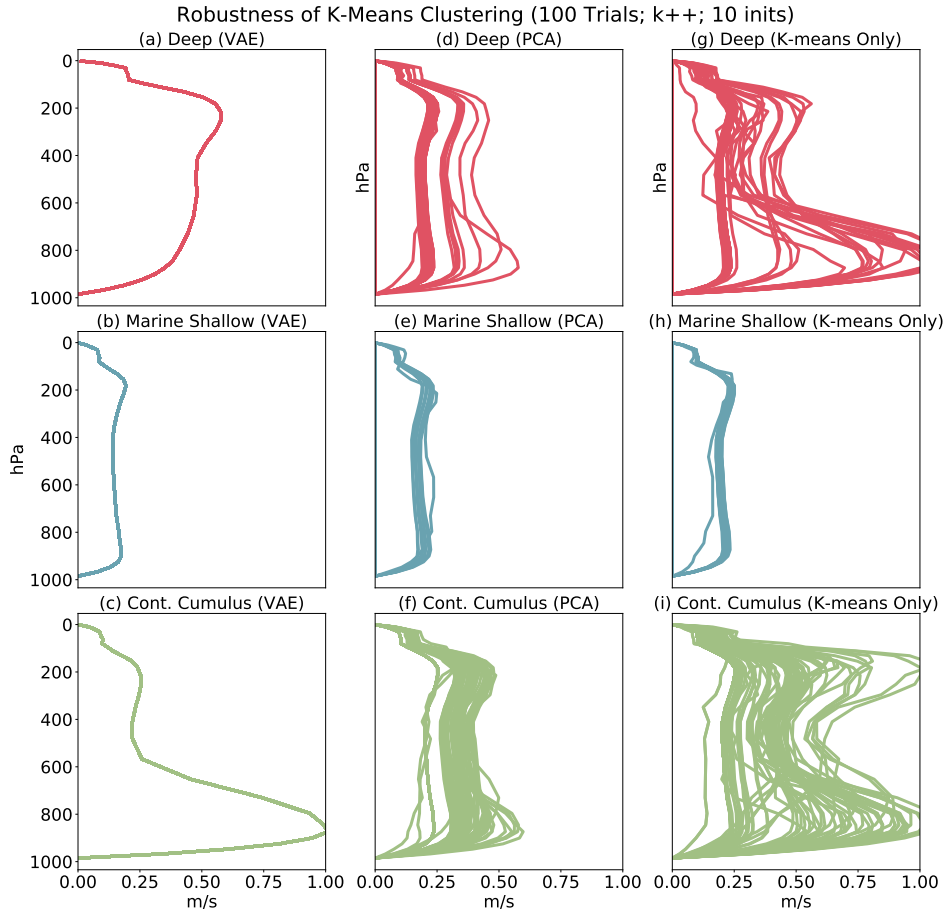

Figure 4.5: K-means clustering performed on the latent representation of convection from a VAE encoder (a, b, c), clustering on convection after dimensionality reduction from PCA (d, e, f), and clustering directly on full resolution vertical velocity fields (g, h, i). In all cases, we set k=3, use the k++ algorithm, and ten initializations. Each panel shows a cluster's median vertical structure. Fewer profiles indicate more robust clusters between different trials.

distinct, recognizable physical properties (Figure 4.8 b-d).

We now stress test the validity of our assumption that non-linear dimensionality reduction is needed prior to clustering the SRM simulation outputs for reproducible, physically consistent results. Instead of clustering the latent representations of outputs, we directly cluster the full vertical velocity fields in the test datasets. What we find is that even with less stochastic hyperparameter choices (ten unique initialization, k++ algorithm), reproducible clusters are no longer possible when 100 trials are performed (Figure 4.5 a,b,c vs. g,h,i). Simultaneously, varying the number of clusters, k, leads to disparate results. (Figure 4.6, bottom two rows).

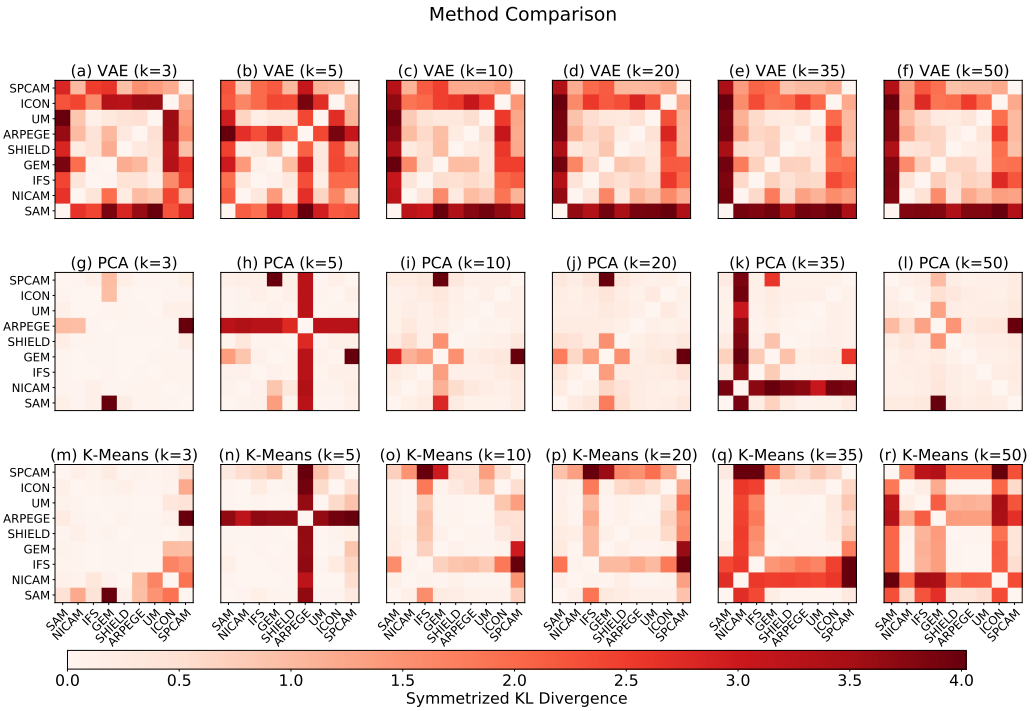

Figure 4.6: Symmetrized KL divergences between DYAMOND models obtained through nonlinear dimensionality reduction and vector quantization (top row), only vector quantization (bottom row), and a combination of Principal Component Analysis and vector quantization (middle row). We test the results for a physically interpretable k (k=3), a converged k (k=50), and intermediate values. We see that only the VAE-based approach (a-g) shows consistency between different k values.

Having determined a form of dimensionality reduction is necessary, we must now ask if it must be a VAE encoder; Could a simpler form of compression work? We will now reduce the SRM test data through Principle Component Analysis (PCA) to the same size as the latent representations of the VAEs (1000) and then follow an identical clustering procedure. There is less variation in the clusters than simply clustering the full fields (Figure 4.5d, e, f vs. g, h, i). However, they are still not reproducible unlike the clusters from the latent representation of the VAE (Figure 4.5a, b, c vs. d, e, f). Another limitation of this approach is that it is sensitive to the number of clusters chosen for analysis – results are not robust (Figure 4.6, middle row).

Only the synthesis of the VAE encoder and the k-means clustering together yield the three

robust, physically consistent clusters that can be used for vector quantization and produce consistent findings even when the hyperparameters of the k-means algorithm are altered (Figure 4.3 and 4.6).

4.4 Results

Our unsupervised learning framework uncovers 4 key findings when analyzing all nine SRM simulations which we discuss in more detail below.

4.4.1 Unsupervised Machine Learning Reveals Physically Interpretable Convection Clusters

We first observe that non-linear dimensionality reduction uncovers patterns in the organization of convection without human supervision. We use Variational Autoencoders: a probabilistic autoencoder architecture that maps data into a lower dimensional latent space using a neural network and allows us to reconstruct the data from the latent space, using another neural network. VAEs regularize this latent space by including a term in their loss function that encourages disentanglement of the latent space (See Section 4.3.2 for more details).

For our purposes, VAEs extract low-dimensional representations of SRMs in their latent spaces allowing us to investigate the details of high-resolution convection. We will first demonstrate that for convection the disentanglement achieved by our VAE separates is based on identifiable physical properties. Among these features that organize the latent space are the intensity and horizontal extent of vertical velocity updrafts as well as differences between maritime vs. continental and deep vs. shallow modes of convection.

Visualizing the VAE's Latent Space. Encoding DYAMOND SRM simulation data with a VAE reveals interpretable structure and enables rich visualizations. While this section focuses on data from the SRM "UM", we found that this analysis generalizes to all nine simulation data sets.

To show that our latent space *disentangles* the data according to meaningful criteria, we label each data point according to various widely accepted metrics for differentiating unique types of convection such as intensity and geography. We then *colorize* these data points but the chosen properties in two-dimensional Principle Component Analysis (PCA) projections of the latent space.

The clusters of convection formed in the latent projection can be seen in Figure 4.7. Figure 4.7a shows the data colorized by overall convective intensity[2]; this quantity is strongly correlated with the y-axis ($R^2 > 0.6$) and explains the main variation in the data. The x-axis shows correlations with a metric of the dominant turbulent horizontal length scale [15] (or put more simply how wide a vertical velocity updraft is). On the x-axis, we also observe geographic disentanglements between continental and maritime convection, despite the fact that information about the geographic location or land-sea contrast was not included in the training data.

Interpretable Clustering. We first confirm that statistically distinct convection regimes exist by applying a K-means clustering algorithm to the latent space. Clustering with $k = 3$ isolates three unique regimes of convection in SRMs (Figure 4.7d-h). Having found that there are three well-defined clusters of convection that appear to generalize across SRMs, we now hone in on just the UM simulation output and look at the differences in properties between clusters of convection that form in our UM latent space projection. Figure 4.8 shows the geography information and intensity statistics of each convection cluster.

[2]summed absolute magnitude of vertical velocity across the input image

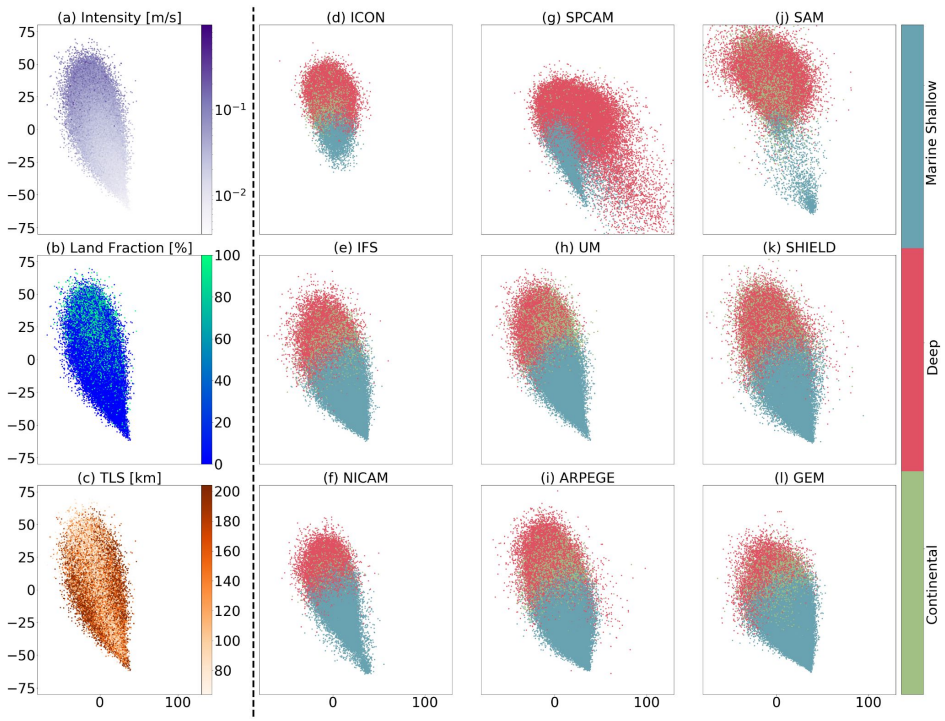

Figure 4.7: Two-dimensional PCA plots of DYAMOND data encoded with a shared VAE (trained on UM data). The left column (panels a-c; see also Figures 4.12-4.14) shows data points colorized by physical convection properties, including convection intensity (a), land fraction (b), and turbulent length scale (c). The VAE visibly disentangles all three properties. The right columns (panels d-i) show data points from different DYAMOND data sets, colorized by convection type (as found by clustering). The top panels (g and j) show clear differences in their latent organization compared to the remaining models; see Section 4.4.3 for a discussion. Movies 4.2-4.6 show additional animations of the latent space.

The statistics of the intensity of convection in each regime are distinct (Figure 4.8a) and each regime is composed of convection from different parts of the planet (Figure 4.8b-d). A first regime, "Continental" Convection, is from tropical convection over land (Figure 4.8b). This cluster is defined by a bottom-heavy vertical velocity variance profile (henceforth abbreviated as $\overline{w'w'}$ based on Equation 4.3) that one expects from shallow morning continental convection over drier surfaces (Figure 4.8a, green line). A second regime, "Deep" Convection, captures intense tropical convection from over the warmest ocean surfaces, such as the Indian Ocean and West Pacific Warm pool (Figure 4.8c). These zones have long been known to promote convection defined by especially top-heavy and intense $\overline{w'w'}$ profiles (Figure 4.8a, red line).

Our final regime, “Marine Shallow” Convection, captures the rest of the tropical ocean where less intense $\overline{w'w'}$ profiles and low clouds are known to occur, especially on the western coasts of subtropical latitudes (Figure 4.8a, blue line and d).

Our VAE works much like an atmospheric scientist, logically grouping convection based on differences in geography and physical properties. Having now opened the "Black Box" of our unsupervised machine learning approach and gained confidence in its findings, we now determine whether these latent space exploration can also reveal not just differences in convection in a single SRM but differences *between* SRMs.

4.4.2 Latent Space Inquiry Uncovers Differences among Storm-Resolving Models

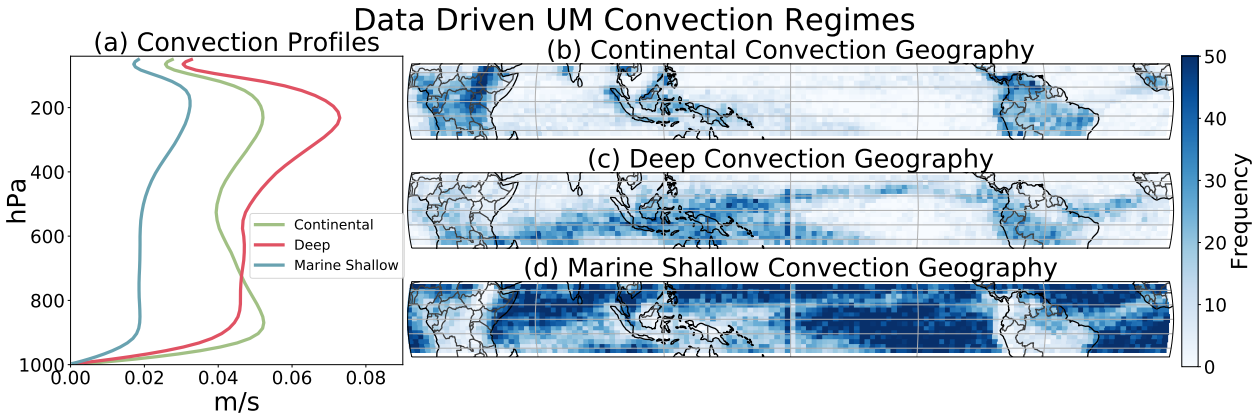

Figure 4.8: Unsupervised clustering results ($k = 3$) obtained on UM data, resulting in three distinct regimes of convection. Panel (a) shows each cluster's median vertical structure, calculated by $\overline{w'w'}$. Panels (b)-(d) show the frequency of occurrence of each convection type at each lat/lon grid-cell of a sample assigned to a particular regime, showing distinct geographical patterns. Additional evidence of this disentanglement can be seen qualitatively in Figure 4.7a,b,c,h.

More formally, our goal is to exploit the learned clustering for SRM inter-comparison. Here we begin with a qualitative intercomparison; an in-depth quantitative extension will follow in Section 4.4.3.

Visual Model Intercomparison. We adopt Section 4.4.1's approach to cluster and visualize the latent space—this time to compare different SRMs in 2D through the lens of a common VAE encoder. If the different SRMs are producing similar simulation outputs, we expect the latent representation of each SRM to show only slight variations between themselves. Results from all nine SRMs are shown in Figure 4.7.

Upon closer inspection, we find that while most SRMs share similar appearances when projected onto a latent space, SPCAM and SAM do not (Figure 4.7 g, j vs. all). More specifically, SAM seems to have a different cluster of "Deep Convection" (Figure 4.7j, red regime), with an intensity not seen in other SRMs (Figure 4.12c dark purple vs. all). SPCAM shows an unusual extension of the "Deep Convection" cluster to locations adjacent to the "Marine Shallow" (blue) mode. These outgrowths hint at something fundamentally different in the convection generated by these two simulations compared to other SRMs in the DYAMOND Project.

As we did in Section 4.4.1 with the UM data, we will now examine the physical properties of the SPCAM and SAM clusters – this will help inform us if these visual differences in latent representations indicate significant differences in how the SRMs represent the atmosphere. For SPCAM, this further latent space analysis (Figure 4.13b vs. all) reveals a unique regime of continental convection with a short horizontal scale (Figures 4.14b vs. all) and an intense $\overline{w'w'}$ profile (Figure 4.15a red curve vs. all), particularly compared to other SRMs near the surface of the earth. For SAM, the $\overline{w'w'}$ profile of "Deep" Convection is much more intense than that of other SRMs, especially in the upper atmosphere (Figure 4.15b; blue line). This distinction helps explain the unusually wide extent of this "Deep Convection" cluster on the latent space projection in (Figure 4.7j, red cluster vs. all).

Investigating the details of latent spaces has allowed us to begin to separate SPCAM and SAM from other SRMs in terms of the way the models represent the dynamics of the tropical atmosphere. However, we can take this analysis further by looking at the other aspects of

the latent space clusters.

Cluster Size Comparison. Another way of comparing latent representations of different SRMs by their convection clusters is to consider each model's relative cluster proportions, i.e., the fraction of the data assigned to each convection regime across the nine simulations. From this perspective, Figure 4.16 shows that SPCAM and SAM are very different from a super-majority of models (Figure 4.16, second and third rows vs. bottom six). These two SRMs have high proportions of stronger convection types, findings that are consistent with our earlier latent space explorations (Figure 4.7 and Figures 4.12- 4.15).

But categorized by the relative cluster proportions, ICON is also unique from other SRMs. Despite having a latent representation visually similar to the other six self-similar DYAMOND SRMs, we find differences upon closer inspection of the frequency of each type of simulated convection in ICON (Figure 4.16; top three columns vs. all). More specifically, ICON's output contains a higher proportion of stronger convection types ("Continental" & "Deep") and a lower proportion of less intense convection ("Marine Shallow").

To gather more evidence of differences between these SRMs and others, we can inspect the distribution of Evidence Lower Bound (ELBO) scores (Equation 4.1), i.e., the approximated Probability Density Function (PDF) of the models. We use a common encoder model (Figure 4.9a) to visualize the PDF of each SRM test dataset. When comparing the shape of the nine PDFs in Figure 4.9a, the red lines corresponding to ICON, SPCAM, and SAM have very different shapes than the blue lines denoting the other six SRMs. More specifically, ICON, SPCAM, and SAM are more right-skewed than left-skewed and less symmetric. These ELBO PDFs (Figure 4.9a), latent space projections (Figures 4.7 and 4.12-4.14), and clustered regimes of convection (Figures 4.15 and 4.16) all reveal differences exist between a subset of SRMs and the majority in the representation of atmospheric dynamics. But to complete our data-driven SRM inter-comparison of the DYAMOND Initiative, we will now formally

quantify this dynamic split between the models with "Distribution Shift" measurements.

4.4.3 Only Six DYAMOND SRMs are Dynamically Consistent

Our analysis so far has mainly focused on *qualitative* aspects, but a formal assessment of model differences requires quantitative tools. We define distances between SRMs based on their high-dimensional output data distributions, using the Kullback-Leibler (KL) divergence. We use nonlinear dimensionality reduction and vector quantization to approximate the latter in a tractable form.

Quantifying Distribution Shifts between SRMs. The most natural way to compare different SRMs is through their high-dimensional simulation data. While these distributions share many similarities, they will show slight but consistent differences, e.g., in the intensity and geography of their predicted atmospheric convection. Borrowing machine learning terminology [138], we will denote such distributional differences as *distribution shifts*. Most machine learning work around distribution shifts previously focused on supervised learning, but we put forward a methodology focusing on the unsupervised setup.

We primarily rely on a technique called vector quantization, partitioning the latent space into K different cells using a clustering technique (see Section 2.3). We coarse-grain the number of embedded data points by their cluster assignment frequencies, resulting in a K-dimensional vector representation of the data set that we use to assess mutual similarities. We use an information-theoretical notion of similarity termed KL divergence (see Section 2.3).

Using the *distribution shift* based approach outlined above we now present a quantitative comparison of the degree of similarity and dissimilarity between each of the nine SRMs. Figure 4.9b shows a matrix of pairwise similarities among SRMs (See Section 2.3 for additional details) as measured by this symmetrized KL divergence approximated from the data

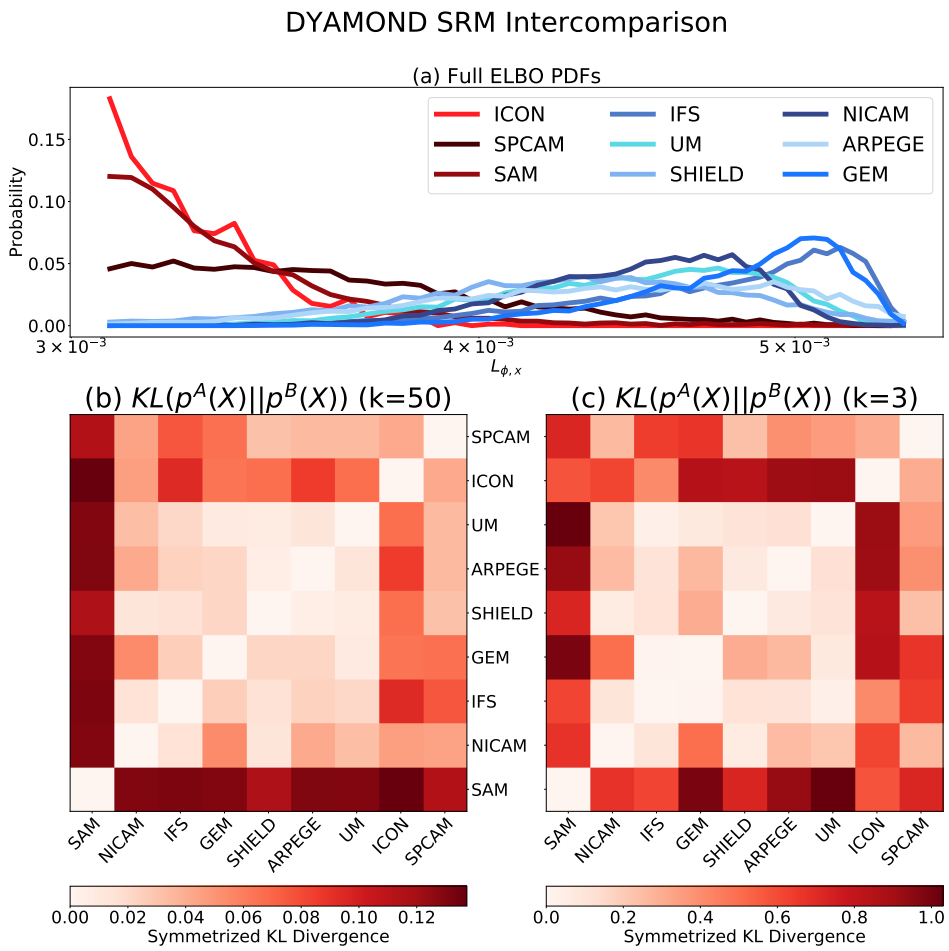

Figure 4.9: Unsupervised storm-resolving model inter-comparison. The top panel (a) shows the ELBO (Eq. 4.1) score distribution of data from different DYAMOND simulations. (The VAE encoder is shared and trained on UM data.) We see that three model types (ICON, SPCAM, and SAM) have qualitatively different ELBO score distributions than the remaining models. Panels (b) and (c) show symmetrized KL divergences between DYAMOND models obtained through nonlinear dimensionality reduction and vector quantization (see main text). Panel (b) shows results obtained from $k = 50$ clusters, while panel (c) shows results obtained from $k = 3$ clusters. Both methods yield similar results. To better highlight the structure, we apply agglomerative clustering to the columns [147] and symmetrize the rows. We find dynamical consistency between six of the nine SRMs we examine (6x6 light red sub-region corresponding to NICAM, IFS, GEM, SHIELD, ARPEGE, UM), which is in agreement with panel (a).

discretized by a high cluster count (k=50).

From this result, we make two observations: (1) three SRMs, namely SAM, SPCAM, and ICON, show a large dissimilarity (Dark red colors corresponding to large KL Divergences in

Figure 4.9b) with respect to each other as well as to the remaining models; and (2) there appears to be a cluster of "self-similar" models (GEM, UM, NICAM, IFS, SHIELD, ARPEGE) that show a comparatively high degree of mutual similarity (Light red colors corresponding to small KL Divergences in Figure 4.9b). Note that similar results are evident from Figure 4.9c, when we use a lower but physically interpretable cluster count (k=3; same as for the analysis in Sections 4.4.1 and 4.4.2).

Our results obtained from vector quantization thus align well with our earlier latent space investigation (Sections 4.4.1 and 4.4.2) with all three approaches (Section 4.4.1, 4.4.2, and 4.4.3) showing a split between six self-similar SRMs and three divergent SRMs. To summarize these differences, we found ICON had a lower proportion of shallow convection than other SRMs, SAM contained unusually intense "Deep Convection", and SPCAM had small scale turbulence and distinct profiles of $\overline{w'w'}$ with unusual intensity near the earth's surface not seen in other SRMs.

Though we have put much of the focus on using our framework to identify unique SRMs and hone in on the causes of inter-SRM differences, the apparent similarity among GEM, UM, NICAM, IFS, SHIELD, and ARPEGE is another key finding of our approach. This conformity mirrors what we found by inspecting the latent representations (Figures 4.7, 4.12-4.14), the vertical structure of the leading three convection regimes (Figure 4.15), and the proportion of each type of convection in the simulation (Figure 4.16). It would be worth elucidating the degree to which the similarity between these SRMs is a reflection of better representing observational reality or model herding, but this question is outside the scope of our present work. Instead, we will move on from inter-SRM comparisons in the same climate state to a comparison of different climate states.

4.4.4 VAEs Extract Planetary Patterns of Convective Responses to Global Warming

VAEs not only enable inter-model comparisons, but they also help us understand distribution shifts in convection caused by global warming. In this section, we focus on a single model (SPCAM) that produced simulation data at two different global temperature levels (present conditions vs. $+4K$ of sea surface temperature warming). Our approach identifies both changes to the $\overline{w'w'}$ of convection and the geographic regions where convection shifts the most with climate change.

To visualize the effects of global warming on convection, we adopt the methods from Sections 4.4.1 and 4.4.2 but this time emphasize geographic aspects. We first learn global convection clusters, where we again initialize three cluster centers ($k = 3$) for physical interpretability. Given these fixed cluster centers, we stratify the SPCAM data by their unique latitude/longitude gridcell and compute location-specific cluster proportions. We can now visualize these cluster proportions geographically and identify which convection type is dominant in a given region (Figure 4.17).

As was the case for the other SRMs, we can classify each of the latent space clusters as a unique convection species. A first again corresponds to "Deep Convection" over the Pacific Warm pool. Meanwhile, a second mode, "Marine Shallow Convection" dominates over subsiding zones (locations of descending air). Different from other SRMs, we find the third mode to be a unique form we will call "Continental Shallow Cumulus" for now. It is found exclusively over certain sub-regions of semi-arid tropical land masses (Figure 4.17a).

Changing Probabilities of Convective Modes in Response to Global Warming. To measure the shift between our control climate and a warmed world, we convert the cluster assignments of SPCAM into normalized probabilities requiring some technical notation (See

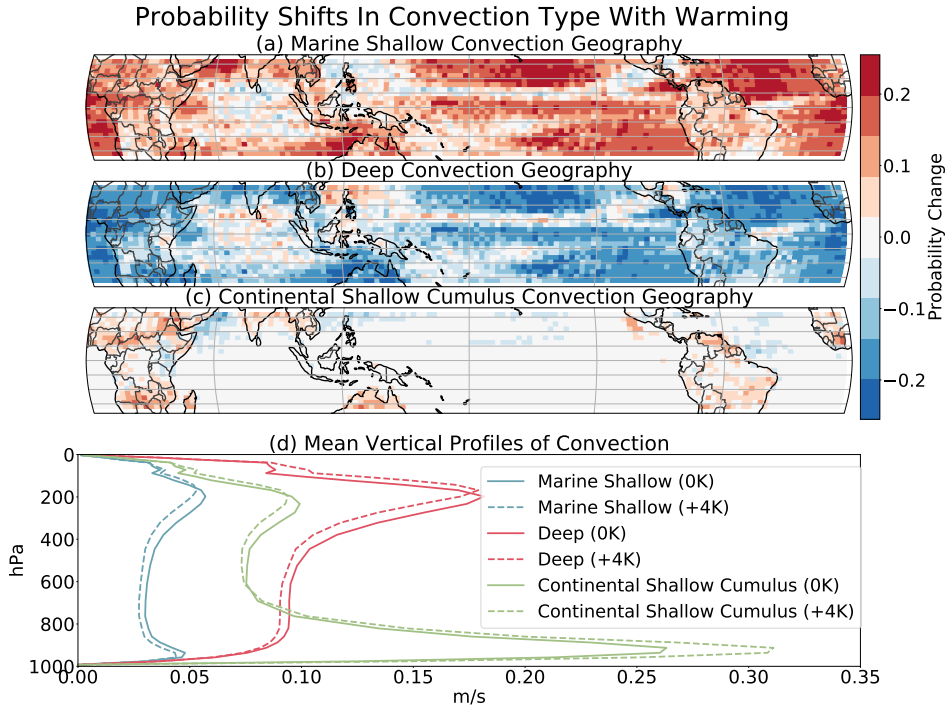

Figure 4.10: Convection type change induced by $+4K$ of simulated global warming (see main text). Panels (a-c) show differences in convection type frequency (see main text), where we stratified and plotted the data by latitude/longitude grid cell. Each panel displays probability shifts in the three convection types found through clustering with $k = 3$, corresponding to "Marine Shallow" Convection (a), "Deep" Convection (b), and "Continental Shallow Cumulus" Convection (c). Panel (d) shows the shift in the mean vertical structure of each convection type with warming (solid vs. dashed lines). This unsupervised approach captures key signals of global warming, including geographic sorting of convection (a, b), expansion of arid zones over the continents (c), and anticipated changes to turbulence in a hotter atmosphere (d).

Methods for a more detailed explanation). After auto-encoding our data into a latent space and clustering the encoded data with K-means, we can represent our dataset by the fraction of data π_k assigned to each cluster k. These "cluster assignment" vectors thus summarize the prevalence of each cluster (convective regime) in each dataset. We can geographically condition these probabilities to elucidate the spatial pattern of each type of convection across the tropics. When analyzing global warming, we can now visualize how the convective regimes change in their overall frequency and their spatial domain. Looking at the differences between these normalized probabilities provides an objective measure of the cumulative change in the structure of the atmosphere with warming through the lens of a "distribution shift". We will show the probability shifts $\pi_k^{+4K} - \pi_k^{0K}$ in convection type k reveal many of the expected effects of climate change.

The dominant signal of climate change captured by these distribution shifts is the increased separation of deep and shallow convection by geographic domain. Figure 4.10a shows shallow convection increasing over regions of subsidence while Figure 4.10b shows a corresponding decrease in "Deep" Convection across these less active oceanic zones. At the same time, Figure 4.10b also reveals the anticipated increase in the frequency and concentration of "Deep Convection" over warm ocean waters and especially the Pacific Warm pool [7] while shallow convection becomes less common in these unstable regions. Finally, Figure 4.10c shows the rare "Continental Shallow Cumulus" mode, which increases in probability over semi-dry land masses, consistent with overall arid zone expansion and intensification [117, 31].

We are also reassured by how the vertical structure of each convective regime shifts as temperatures warm, which is shown in Figure 4.10d. Comparing the dashed and solid lines shows that the upper-tropospheric maximum in $\overline{w'w'}$ shifts upwards with warming. This is a finding consistent with tropopause expansion induced by climate change [128, 177]. Relatedly, a decrease in mid-tropospheric $\overline{w'w'}$ can be explained by the expected reduction in convective mass flux due to more latent heat release from enhanced saturation vapor pressure in a

warmer world [149, 142]. Blue lines show a decrease in lower-tropospheric $\overline{w'w'}$, corresponding to a decrease in "Marine Shallow" convection intensity. We believe this is evidence of marine boundary layer shoaling [94]. Finally, when we look beyond the median $\overline{w'w'}$ statistics to the upper percentiles of "Deep Convection" (Figure 4.18b), we see an increase in $\overline{w'w'}$ magnitude congruent with observational trends that show an intensification of already powerful storms over the warm waters, aided by greater moisture convergence [7].

The expected geographic and structural effects of climate change become apparent by inspecting the latent space's leading three clusters, showing that VAEs can quantify distribution shifts due to global warming in a meaningful and interpretable way.

Global Warming Impacts on Green Cumulus Convection. Finally, we hone in on the unique ways in which "Continental Shallow Cumulus" Convection changes with climate in SPCAM according to our unsupervised framework. Within this model, this "Continental" regime corresponds to a rare form of convection, "Green Cumulus", that was first identified by [42]. We choose to more formally adopt the label of "Green Cumulus" here due to the near total overlap between the geographic domain of SPCAM's "Continental" convection and the regions of most frequent "Green Cumulus" convection identified in satellite imagery (Figure 6a in [42]). Both our results and [42] identify this convection primarily over semi-arid continents (Figure 4.17a). Despite its existing identification in literature, is not traditionally included in the analysis of tropical convection [70, 159, 105]. This is due both to its rarity and the fact that previous efforts to rigidly classify it fail to identify statistically significant differences in physical properties between "Green Cumuli" and other existing convection types [43].

However, by geographically conditioning the latent space cluster associated with "Green Cumuli" we can not only confirm the regional patterns of the species, but we can begin to uncover unique physical properties behind its formation and growth. Looking at the condition of the atmosphere in these geographic regions during the times when "Green Cumuli"

dominate, we identify consistent signatures of very high sensible heat flux, relatively low latent heat flux, and the smallest lower tropospheric stability values (as defined in [22]) (Figure 4.19). This unique atmospheric state at locations of this convective mode, combined with its very distinct $\overline{w'w'}$ profile (Green lines in Figure 4.10d), suggests it does in fact deserve to be separated out from other types of convection despite its scarcity.

Although other studies make note of this convective form [41, 178, 5], our distribution shift analysis provides a view of expanding "Green Cumuli" as global temperatures rise (Figure 4.10c). We observe the frequency and geographic habitat of "Green Cumulus" both increase in a hotter atmosphere – this is likely aided by expected dry-zone expansions [117, 31]. Comparison of these "Green Cumuli" $\overline{w'w'}$ cluster profiles between the control and warmed climates also shows a substantial increase in the associated boundary layer turbulence (Figure 4.18c). This suggests two trends as the climate changes: (1) "Green Cumuli" will become more frequent over larger swaths of semi-arid continents in the future and (2) When "Green Cumuli" occur, they will be even more intense. Unsupervised machine learning models here proved capable of isolating "Green Cumuli" and capturing its climate change signals, synthesizing dynamic analysis and discovery.

4.5 Discussion

We introduced new methods and metrics to compare storm-resolving models (SRMs) based on their high-resolution simulated output data by using unsupervised machine learning. Our approach relied on a combination of non-linear dimensionality reduction using variational autoencoders (VAEs) and vector quantization. Beyond inter-model comparisons, we also compared global climates at different temperatures.

While avoiding human biases and subjective physical thresholds, our data-driven method

provides a complementary viewpoint to physics-based climate model comparisons. For example, we could independently reproduce known types of tropical convection verified through examination of the geographic domain and vertical structure. At the same time, our machine learning methods facilitate an intuitive understanding of simulation differences in SRMs.

Our distributional comparisons identify consistency in only six of the nine considered SRMs. The other three (SAM, SPCAM, ICON) deviate from the larger group in their representations of the intensity, type, and proportions of tropical convection. These divergences temper the confidence with which we can trust SRM simulation outputs. Note we cannot rule out the possibility that one of the non-conformist SRMs may still be reflecting observational reality better than the majority group.

Our findings suggest the need to further investigate the parameterization choices in these SRM simulations. In the DYAMOND Initiative, ICON was configured at an unusually high resolution (grid-cell dimension of 2km) so that typical sub-grid orography and convection parameterizations were deactivated [82]. In the design of both SPCAM and SAM, there are approximations required for the anelastic formulations of buoyancy [120, 10]. When these formulations are ultimately used to calculate vertical velocity, they could be causing the deviations between models in the intensity of updraft speeds. We believe there is a high chance these specific distinctions between parameterizations could be causing the split in the dynamics of the SRMs. However, further investigation is needed to confirm the true root causes of the differences between SRMs we have identified.

When comparing different climates, convolutional VAEs identify two distinct signatures of global warming: (1) An expansion and (at the atmosphere's boundary layer) an intensification of "Continental Shallow Cumulus" Convection and (2) An intensification and concentration of "Deep Convection" over warm waters. We argue that the first signal contributes to distribution shifts in the enigmatic "Green Cumulus" mode of convection.

Chapter 4 has focused on vertical velocity fields as one of the most challenging data to analyze. Improved performance could be obtained by jointly modeling multiple "channels" of spatially-resolved data such as temperature and humidity, which we leave for further studies. This Chapter could also be extended to alternative data sets, such as the High Resolution Model Inter-comparison Project (HighResMIP) [48, 57] or observational satellite data sets. Besides variational autoencoders, future studies could also focus on hierarchical variants, normalizing flows, or diffusion probabilistic models. Ultimately, we hope that this Chapter will motivate future data-driven and/or unsupervised investigations in the broader scientific field anywhere that Big Data challenges conventional analysis approaches. But we now choose to focus our efforts on using the analytical framework developed in Chapters 3 and 4 to better understand the physical mechanisms of our atmosphere.

4.6 Appendix A: Leveraging the VAE Encoder

4.6.1 Latent Space Projections

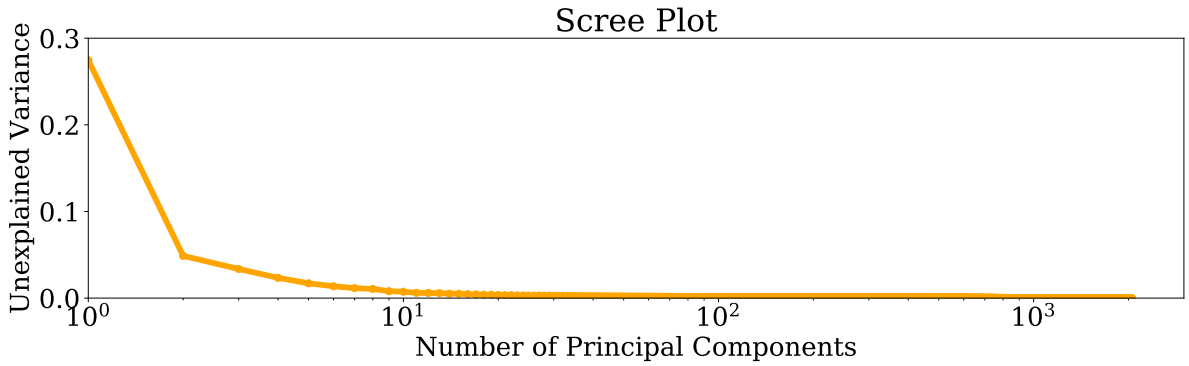

Figure 4.11: The proportion of variance of the full $1e3$ dimensional encoding left unexplained as we project down from the full z vector to visualize the latent representation in 2D or 3D Space. We see the first three principal components are the most important for preserving the information from the latent vector.

For much of our qualitative analysis, we rely on visual inspections of the latent space. Because

of this, we need to verify sufficient information is preserved in these representations. Given the comparatively high dimension of $\mathbf{z}$, visualization is only possible with further compression. We rely on Principle Component Analysis (PCA) to linearly project $\mathbf{z}$ to just two (Figures 2, 4.12- 4.14) or three (Movies 4.2-4.6) components for visualization. We acknowledge there will be a degree of information loss through this process. But we can quantify this compromise by examining a Scree [27] plot of the data. The Scree plot reveals how much of the variance of the full $\mathbf{z}$ vector can be explained by each principal component. Figure 4.11 suggests the first three, and in particular, the first two principal components are orders of magnitude more important than the others and thus we can project the latent representation down to a visible dimension and still conduct meaningful analysis.

4.6.2 A Common Encoder for Analysis

We elaborate on the process by which we use a single VAE to create a qualitative comparison between all nine high resolution SRM data simulations. Since we cannot directly quantify differences between two high dimensional DYAMOND SRM simulations, we can treat the VAE as a density model to approximate the qualitative differences. To that end, let $p_{\theta_A}(\mathbf{x}^A)$ be a generative model (VAE) trained on dataset A with learned dependence on the data from parameters θ_A. We demonstrated above we can leverage the encoder, $q_{\theta_A}(\mathbf{z}^A|\mathbf{x}^A)$, of the model to visualize the encoding of datatype A as $\mathbf{z}_A$ for novel dynamic analysis. But we now use the trained model encoder on another data, B, such that $q_{\theta_A}(\mathbf{z}^B|\mathbf{x}^B)$, so as to get a comparable latent encoding of $\mathbf{x}^B$, $\mathbf{z}_B$. This common density model encoder allows us to elucidate differences in the simulations not visible in the high dimensional dataspaces, $\mathbf{x}^A$ and $\mathbf{x}^B$

We use the same three common physical metrics (Intensity, TLS, land fraction) in all nine simulations for a consistent view of convective organization across the different latent

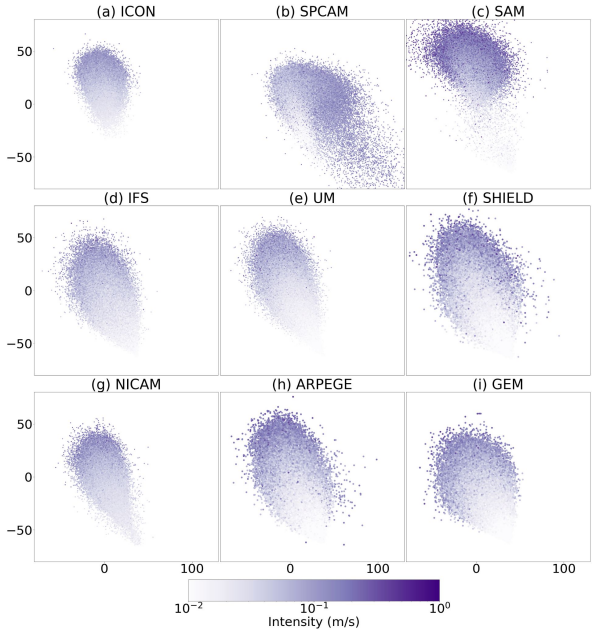

Figure 4.12: Two-dimensional PCA plots of DYAMOND data encoded with a shared VAE (trained on UM data). Data points colorized by the mean of the absolute value of all updrafts in the vertical velocity field. We see a clear separation in the latent space of convection by the intensity of updraft (light purple vs. dark). SAM data (c) shows greater intensity (darker purples) compared to other DYAMOND SRMs. Movie 4.4 shows a 3D visualization.

representations produced by our encoder. The amount of disentanglement varies somewhat between different test datasets and different physical metrics but we see common themes across all latent spaces suggesting generalizability to this approach we can leverage to investigate high dimensional SRMs.

4.6.3 Additional details on Convection Types assigned by our Common VAE Encoder

Our procedure to uncover the physical characteristics of the convection in each cluster of the latent space is covered here in greater depth. While the latent space visuals (Figures, 4.12-4.14) are useful for highlighting where disagreements exist among SRMs, we also need to elucidate the specific nature and cause of these differences. Unfortunately, this is challenging for the human eye at the native resolution of the vertical velocity fields, particularly when

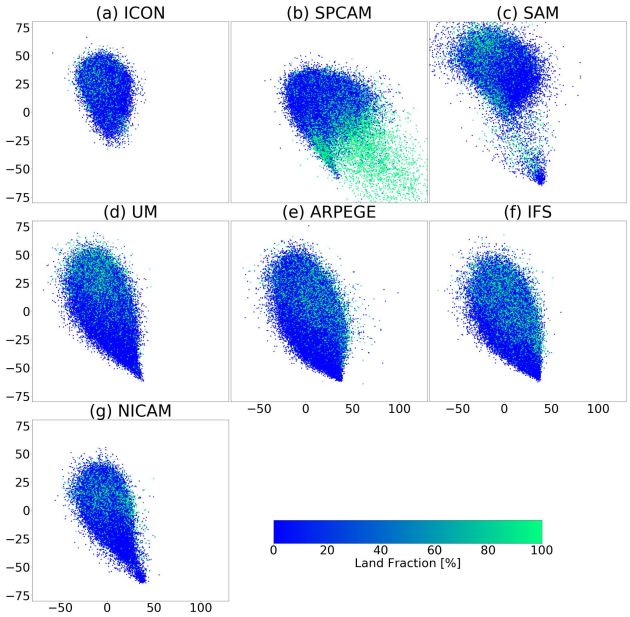

Figure 4.13: Two-dimensional PCA plots of DYAMOND data encoded with a shared VAE (trained on UM data). Data points are colorized by the surface type (continent or ocean) of each vertical velocity field. We see disentanglement in the latent space between convection occurring over land and convection occurring over the ocean (green vs. blue). In SPCAM (b) we see a unique regime of continental convection. GEM and SHIELD were left off due to missing land masks in the data. See Movie 4.5 for a full animation of the latent space.

the data volume (test datasets each in the hundreds of thousands) is high.

But we can still summarize this information in an interpretable way. First, we can average out the horizontal dimension of all the vertical velocity fields in the test data, reducing the 2D fields down to just first moment statistics ($\overline{w'w'}$ from Equation 4.3). Then at each cluster, we can average these $\overline{w'w'}$ profiles together to get a representation of the typical vertical structure of each regime of convection (Figure 4.15).

We can use this approximation for both intra and inter SRM comparisons based on the latent space clustered convection types. More specifically, we can look at the degree to which different regimes of convection within a single SRM have meaningfully different vertical profiles (Figure 4.15, a vs. b vs. c – different subplots but same color profiles). Additionally, we can determine how similar the same species of convection is across different SRM simulation

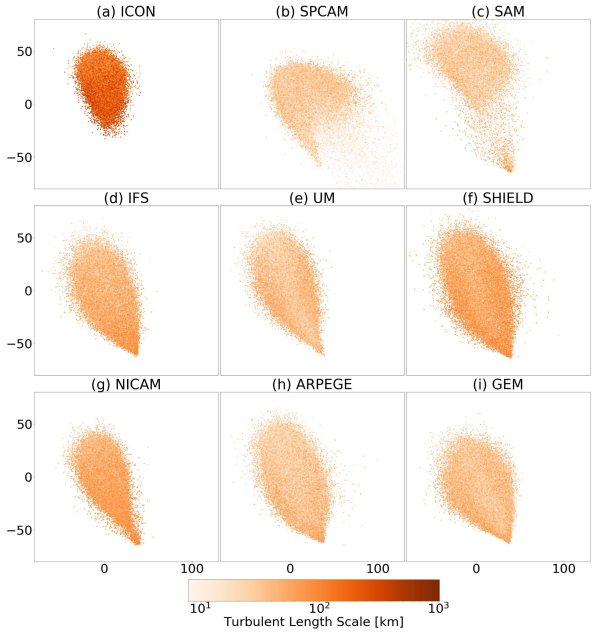

Figure 4.14: Two-dimensional PCA plots of DYAMOND data encoded with a shared VAE (trained on UM data). Data points are colorized by the Turbulent Length Scale of each vertical velocity field (See Equation 4.4). The latent space separates out vertical velocity fields by the horizontal extent of convective updrafts (light orange vs. dark). This perspective reveals the unique land regime of convection in SPCAM (Figure 4.13b) to be defined by small-scale horizontal organization. See Movie 4.6 for a full animation of the latent space.

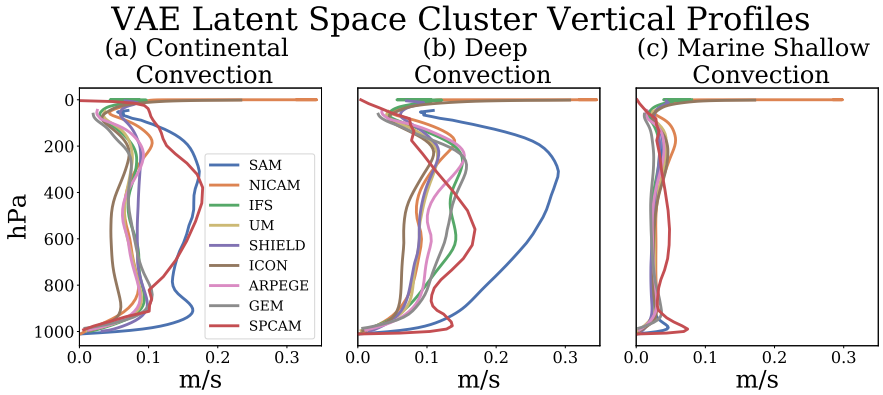

Figure 4.15: The mean $\overline{w'w'}$ (Equation 4.3) profile of each cluster of convection across all nine SRM simulation outputs. The centroids used to organize the other eight SRM simulations are fixed by initial clustering on the UM latent space. Overall, we see common types of convection identified across SRMs (similar vertical velocity fields clustered in the same parts of the latent space regardless of input data type). SAM (blue curves) and SPCAM (red curves) stand out as unique from the typical vertical structure of a SRM convection regime.

outputs (Figure 4.15, differences in vertical profiles in the same subplot). We admit there is information loss from neglecting the horizontal dimension as well as from the data extremes

because we choose to visualize the mean. But this framework remains useful for a composite view of how interpretable these unsupervised regimes of convection are and the degree of generalizability of these clusters across different SRMs.

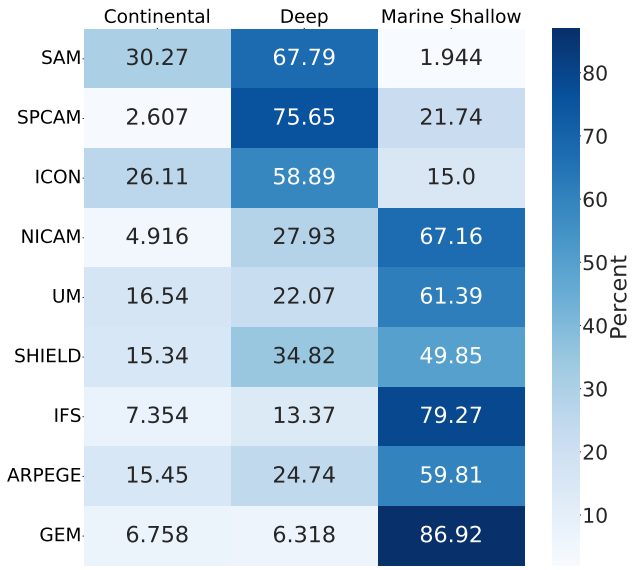

Figure 4.16: The proportion of vertical velocity fields assigned to each of the three regimes of convection across the nine simulations. As in Figure 4.15, the centers initialized in $\mathbf{z}_{UM}$ are used to assign labels to data in across all nine simulations. We see a split across the DYAMOND simulations (top three rows vs. all). SAM, SPCAM, and ICON all assign much high proportions of convection in their test datasets to the more intense regimes compared to other DYAMOND SRMs.

Another issue caused by the relatively large test dataset sizes is density differences in the latent spaces become harder to view in just two dimensions. Therefore, we look at cluster probabilities to better understand the nature of SRM simulation outputs. Two SRM simulations could appear to have the same three types of convection based on the results of Figures 4.12-4.15. But this does not necessarily guarantee these regimes of convection will be present at the same frequency in both simulations. We include Figure 4.16, which quantifies the proportion of convection assigned to each cluster, to help us better compare these large test datasets. ICON (Figure 4.16, row 3), has vertical profiles and a latent representation similar to most other SRMs (Figures 4.12-4.15). But Figure 4.16 shows that within ICON the frequency of occurrence of different species of convection is different than from other SRMs.

4.7 Appendix B: Additional Analysis of Convection in SPCAM

4.7.1 On the Nature of Convection Types in SPCAM

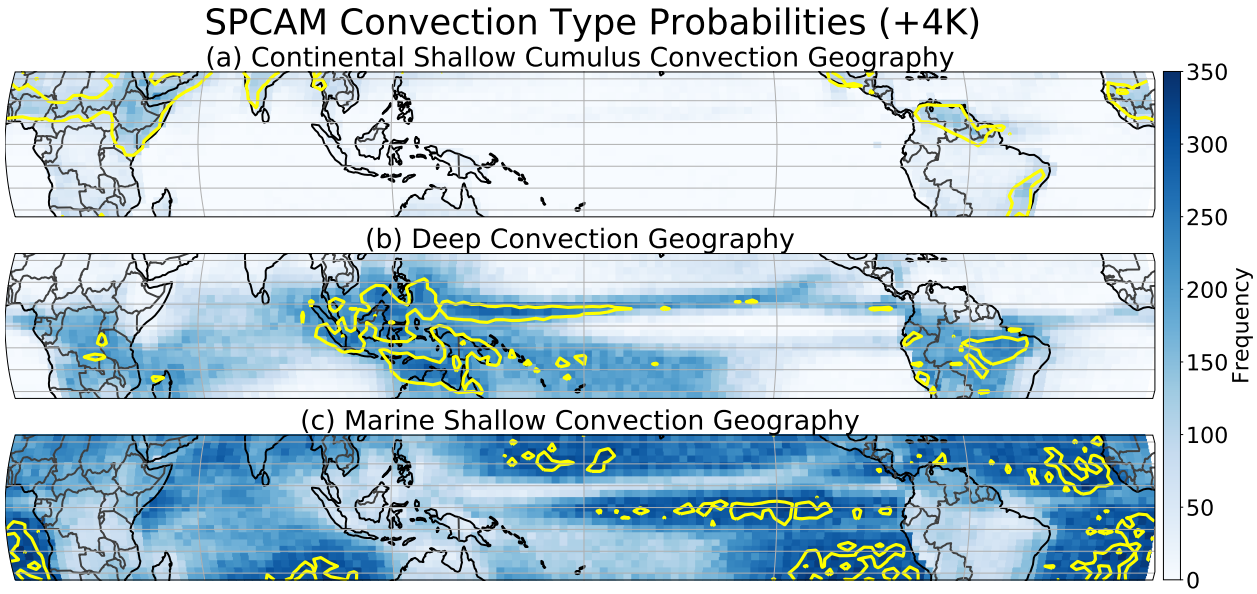

Figure 4.17: The geographic domain of each of the three regimes of convection organized by the VAE latent space in SPCAM. More specifically, we total the number of instances of a regime of convection identified at each lat/lon grid cell. Results are shown for SPCAM +4K data (Not shown for the 0K control climate but findings are similar). Yellow contour lines encompass the 92.5 percentile for each regime. Though not the convective species typically identified by physically informed approaches, these convection types found by the VAE all have distinct physical properties and geographic extents which would justify their separation from a domain perspective.

We examine the frequency of each type of convection at every latitude-longitude grid-cell (Figure 4.17 for our +4K SPCAM simulation; not shown for control climate). This analysis yields three physically distinct and interpretable geographic patterns of convection.

These three convective species organized by the latent space of our VAE provide for a clean comparison with previous literature on tropical meteorology which also typically identifies three distinct convection types [70, 159, 105]. Both approaches isolate a cluster of "Deep

Convection" (Figure 4.17b). However, historically the remainder of tropical convection, visualized from the two baroclinic modes of vertical velocity or other summary statistics (cloud top height, precipitation, or maximum updraft intensity), is classified as cumulus congestus (or stratiform) and shallow cumulus [78, 101, 135]. Our VAE latent space unites these two groups into one regime ("Marine Shallow Convection") while at the same time isolating an unusual "Continental Shallow Cumulus" mode of convection (Figure 4.17a and c).

In summary, this suggests that the organization of tropical convection by unsupervised methods will yield different patterns than those found in traditional physically informed approaches. But both methods deliver results that are logical from a domain perspective demonstrating how unsupervised machine learning models can complement and even augment traditional analysis.

4.7.2 Expanded analysis of Convection Cluster shifts with warming

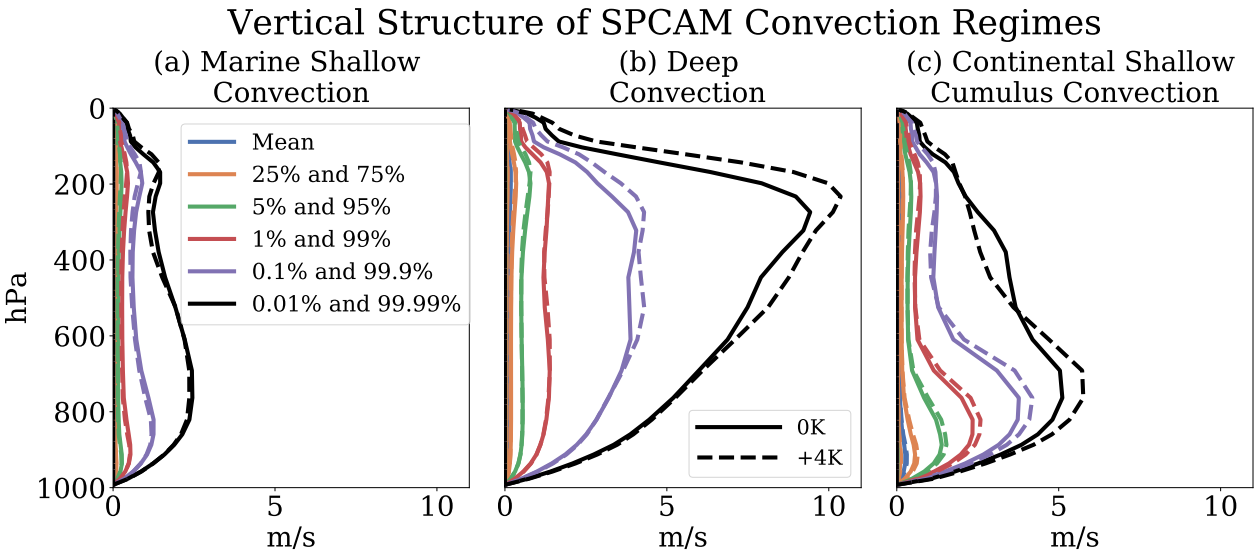

Figure 4.18: A comprehensive view of the vertical structure of each type of convection in SPCAM and how it changes as temperatures rise (solid vs. dashed lines). But instead of only restricting ourselves to a view of the mean, we look at percentiles across the test data in each convection cluster. The VAE anticipates both an increase in the most intense deep convection with warming (b) and a strengthening of turbulent updrafts in the boundary layer (c).

The effects of climate change are often most visible at the extremes so we must look beyond the means of the SRM simulation data to the tail of the PDFs. We can see the value of this expanded analysis when examining the vertical structure of convection. If we looked just at the means (Figure 4.10d) we would think that convective intensity systematically decreased with global warming. But by also looking at the profiles of the extreme vertical velocity fields, we can see that the most powerful convective structures in the Deep Convection regime (above the 99th percentile) actually intensify with climate change (Figure 4.18b). Furthermore, this analysis highlights the otherwise hidden signal of intensification of boundary layer turbulence of arid continental zones (Figure 4.18c).

4.7.3 Additional details on SPCAM's Green Cumulus Convection

We expand on efforts to diagnose the physical properties that compose "Green Cumulus" convection. It has been difficult historically to make the case that "Green Cumulus" deserves its own convective classification. This type of convection is both rare in occurrence and its physical attributes including potential temperature, specific humidity, relative humidity, and large scale omega only slightly differ from other established types of convection [43]. But through both analysis of its vertical structure, as defined by small-scale vertical velocity (Figure 4.18c), and by examining its associated surface fluxes (Figure 4.19), we can better quantify this mode of convection. It is defined by intense updrafts in the boundary layer which are far larger than any other mode of convection. However, in the upper troposphere, we find very weak updrafts during periods of "Green Cumulus" (Figure 4.18c). Conditions for the growth of "Green Cumuli" are most favorable when Lower Troposphere Stability is small, Sensible Heat Flux is high, and Latent Heat Flux is relatively low (Figure 4.19). This suggests "Green Cumulus" is the dominant regime over land regions where conditions are semi-dry but not extremely dry like deserts.

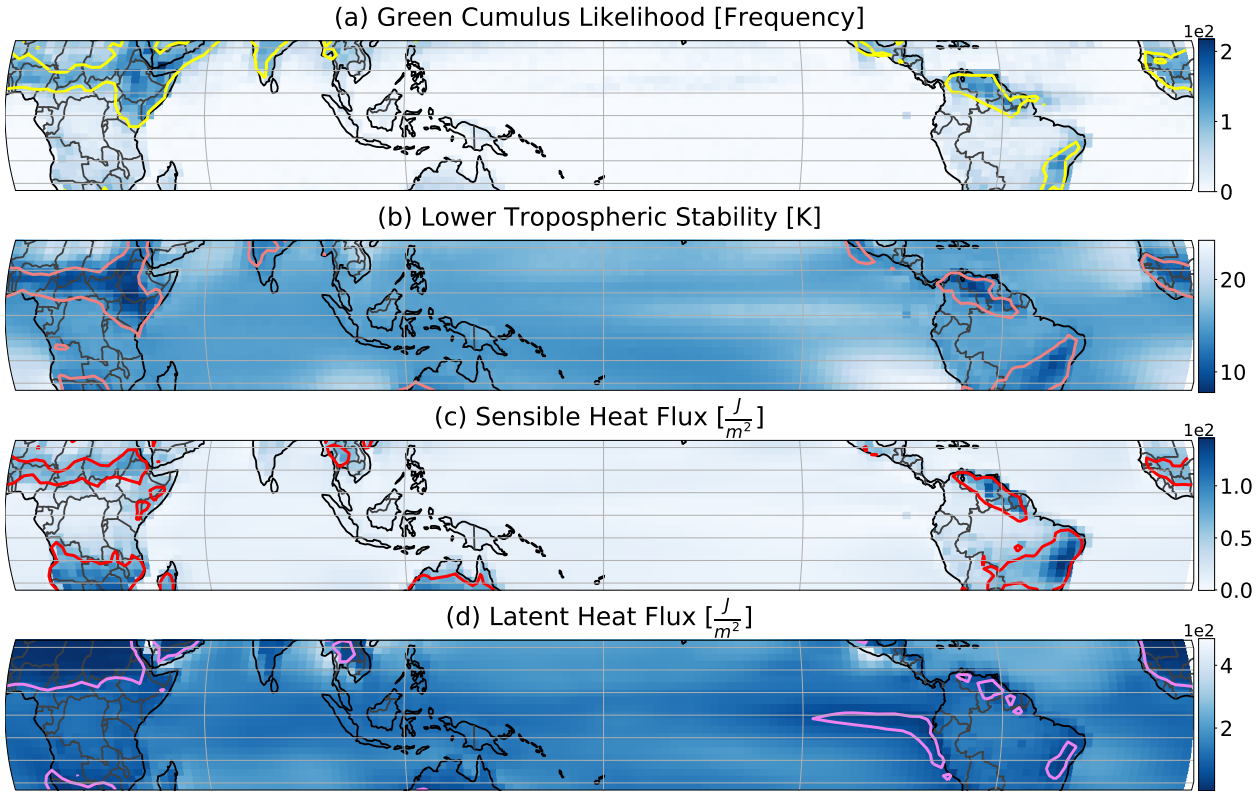

Figure 4.19: We identify the atmospheric conditions that enable the growth and development of "Continental Shallow Cumulus" (or "Green Cumulus"). The regions where "Green Cumulus" convection occurs most frequently (a) are contoured against the patterns of various physical measures of atmospheric conditions (b,c,d). We find "Green Cumulus" can be classified by small Lower Tropospheric Stability (b), large Sensible Heat Flux (c), and low Latent Heat Flux (d). Contours cover the 92.5 percentile (a,c) and the 7.5 percentile (b,d).

We believe our efforts to better understand "Green Cumulus" can yield benefits for atmospheric modeling and physical process understanding. While our VAE identifying Green Cumulus is not a novel discovery, it is valuable to investigate because it is an understudied form of convection [42, 178, 5] compared to the marine variant which is much easier to simulate given its lack of a diurnal cycle and weaker surface fluxes allowing for an assumption of quasi-equilibrium when modeled [180]. While some campaigns like AmazonGO and ARM have provided observational data and a basis for some simulation of this "Green Cumulus" [179, 61], these studies have been restricted geographically (to just the Southern Great Plains and the Amazon Basin). Analysis using satellite data offers a spatially richer view but lower temporal

resolution [42, 41]. The study of this convective regime is further limited because it is missing in the GOES ABI cloud mask [154]. Our VAE extracts this unique "Green Cumulus" mode regardless of its geographic domain in large SPCAM simulations with high temporal frequency (15 minute time-step). This can improve our understanding of "Green Cumuli" behavior with respect to both short temporal transitions and full seasonal timescales, which are currently lacking to due sampling limitations and inconsistencies [61, 181]. This physical understanding is crucial because this mode of convection has a typical domain size on the order of just one kilometer [178, 93] necessitating its parameterization in models. Improvements in physical understanding of this "Green Cumulus" through more rigorous spatial and temporal analysis could help build superior parameterizations, potentially resolving downstream problems stemming from unconstrained shallow cloud representation, including premature shallow-to-deep convection transition and associated temporal precipitation inaccuracies in current climate simulations [76, 11, 174].

4.8 Appendix C: Movies

- 250 examples of vertical velocity snapshots used as training data from each of the nine SRMs we examine in the scope of this Chapter. We observe a variety of convection formations and species. The movie can be viewed at the link here.

- Three-dimensional PCA animation of UM Data encoded by a VAE. Data points are colorized by physical convection properties, including convection intensity (a), the land fraction (b), turbulent length scale (c), and convection type (as found by clustering) (d). We see evidence of disentanglement in all four metrics. For a different visual perspective, we increase transparency to 99.9 % (a,b,d) or 99 % (c) to better show the latent representation of the full test dataset (size 125,000). The movie can be viewed at the link here.

- Three-dimensional PCA animation of DYAMOND data encoded with a shared VAE (trained on UM data). Latent data is colorized by convection type (as found by clustering). The top panels (b and c) show clear differences in their latent organization compared to the remaining models. The movie can be viewed at the link here

- Three-dimensional PCA animation of DYAMOND data encoded with a shared VAE (trained on UM data). Latent data is colorized by the mean of the absolute intensity of the vertical velocity field. The latent representation of SAM (c) shows much greater intensities than other SRMs. The movie can be viewed at the link here.

- Three-dimensional PCA animation of DYAMOND data encoded with a shared VAE (trained on UM data). Latent data is colorized by the surface type (land or ocean) of the vertical velocity field. In the latent representation of SPCAM (b) we see a unique regime of continental convection (green). GEM and SHIELD left off due to missing land masks in the data. The movie can be viewed at the link here.

- Three-dimensional PCA animation of DYAMOND data encoded with a shared VAE (trained on UM data). Latent data is colorized by the Turbulent Length Scale of each vertical velocity field (See Equation 4.4). The latent space separates vertical velocity fields by the horizontal extent of convective updrafts (light orange vs. dark). This perspective reveals the unique land regime of convection in SPCAM (Movie 4.5) to be defined by small-scale horizontal organization. The movie can be viewed at the link here.

Convection Classification Disagreements		
VAE Classification	$\overline{w'w'}$ Classification	Test Dataset %
Deep	MS	9.74
Deep	CSC	0.04
CSC	Deep	0.10
CSC	MS	0.74
MS	Deep	0.68
MS	CSC	0.04

Table 4.1: Of our 1e6 test dataset, we examine all vertical velocity fields where the results of K-Means Clustering algorithm applied to VAE latent space and the $\overline{w'w'}$ fields yields a different classification. While the disagreements are normally small, the exception is 10% of our data that clustering on the latent space classifies as deep, and the $\overline{w'w'}$ suggest Marine Shallow (MS). CSC abbreviates Continental Shallow Cumulus.

4.9 Appendix D: VAE Based Convection Clusters vs. Statistical Moment Based Clusters

To test the hypothesis that significant differences exist in the clusters derived from the latent representation $\mathbf{z}$ compared to other baselines we look at both the cluster sizes (Table 4.1) from each approach (VAE vs. $\overline{w'w'}$). We see immediately that the VAE includes significantly more samples in its "Deep Convection" regime compared to the baseline $\overline{w'w'}$ approach (Table 4.1, column 3). So which approach is more physically consistent?

By examining the physical properties of the vertical velocity fields where the approaches disagree, we can determine which approach is classifying convection correctly. Immediately it is clear the median $\overline{w'w'}$ of these convective samples is intense much like the other samples classified as "Deep Convection" (Figure 4.20a). Looking not just at the intensity, but stability and moisture, we see these vertical velocity fields are most often characterized by high Q values and moderate LTS, classic signs of conditions favorable for "Deep Convection" and storm formation (Figure 4.20b) [22]. Likewise, the precipitation associated with these samples that the VAE clustered in with "Deep Convection" but the baseline approach would not be

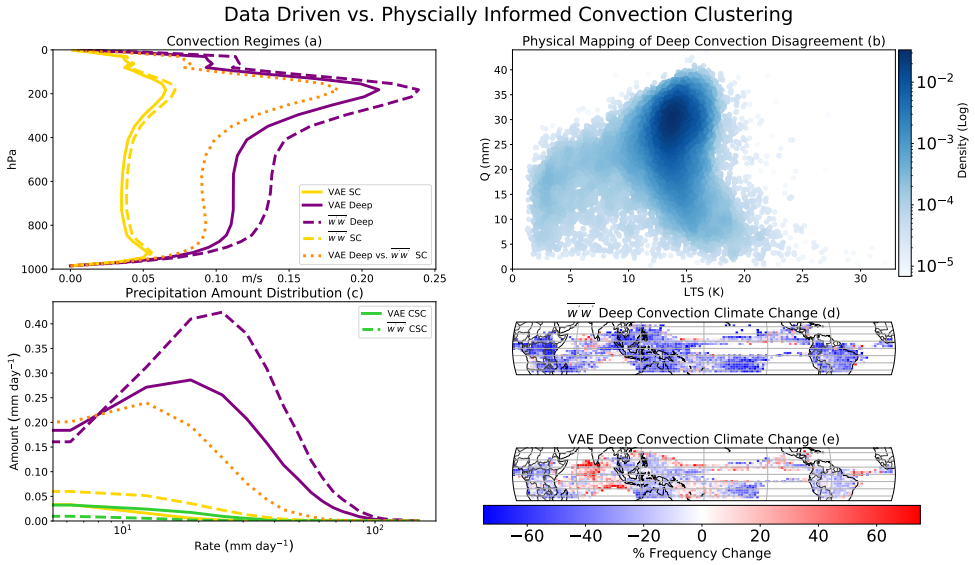

Figure 4.20: A comparison of three regimes of convection in SPCAM identified by clustering the latent representation of the VAE Encoder $\mathbf{z}$ compared against clustering the first moment statistic (the $\overline{w'w'}$ profiles) of the same vertical velocity fields. Three similar groups are identified, but there is disagreement over roughly 10 % of the test data that the VAE approach classifies as Deep Convection but the first moment statistic would be grouped in with the "Marine Shallow" convection. The median vertical profile of these is shown above by the orange dotted line (a). These same vertical velocity fields where the approaches disagree are mapped individually onto LTS-Q Space (Lower Tropospheric Stability in Kelvin and Moisture in mm) (b). These samples are then colored by their density. We also look at the amount distribution of precipitation in each of the convection regimes (c). The geographic shifts with climate change in the regime of Deep Convection identified by each method are shown in (d) and (e). While the vertical profiles look similar (a; dashed vs. solid purple and yellow lines), the geographic regime shifts with climate change diverge with only the VAE convection clusters capturing the expected signals (d vs. e).

associated with stronger storms and convection based on the median amount distribution (Figure 4.20c).

The physical properties of these eristic vertical velocity fields strongly suggest the VAE latent space based organization of convection is more physically consistent, particularly for "Deep Convection". But why is this the case; what information in the vertical velocity fields does the VAE leverage that is not available when we cluster the $\overline{w'w'}$ of the full fields?

4.10 Appendix E: VAEs use the full Vertical Velocity Field

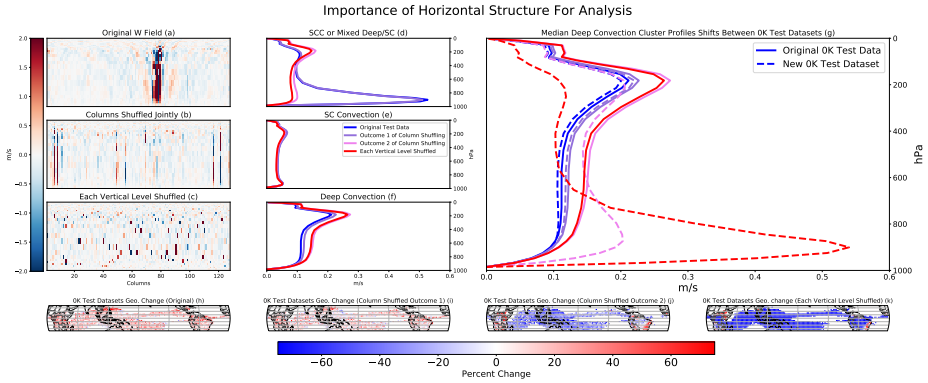

Figure 4.21: We test the importance of the horizontal structure in the vertical velocity fields to the organization of the latent space of the VAE. The vertical velocity fields typically included in the test dataset (a) have their columns shuffled (b) and the horizontal dimension of each vertical level shuffled (c). We cluster latent representations of a,b,c and examine the physical properties of the clusters (d, e, f). We also see how much these cluster centers shift if used to initialize clusters on different, randomly selected test data (g). The change in geographic frequency of (original, column shuffled, and vertical level shuffled) Deep Convection regime are shown as well (h,i,j,k). The results show the original test data leads to the most physically interpretable and robust regimes of convection.

The key difference between the two approaches (clustering $\mathbf{z}$ vs. clustering $\overline{w'w'}$) is that the VAE encoder preserves information about the *horizontal* structure as well as the vertical structure. We will now examine this theory in detail with "scrambled" test datasets. Highlighting the location of the most intense updrafts and downdrafts vertically is thought to be essential to identifying the convection phenomena while we believe the location of the updraft horizontally should be much less important. However, the less comprehensive extraction of Deep Convection by moment statistics compared to the clustering of our latent space suggests that the treasure trove of detail our VAE leverages in the horizontal dimension is important for thorough analysis and organization of Deep Convection and storms in a high resolution simulation.

More concretely, a way we can test the importance of coherent horizontal structure to our VAE is to intentionally corrupt the test dataset in the horizontal dimension. We can "scramble" our test dataset in two ways. First, we shuffle the order of the 128 columns in each image, disrupting the horizontal structure but leaving the vertical information untouched (Figure 4.21a vs. b). We also more radically alter the structure of the field (but critically not the $\overline{w'w'}$ moments) by shuffling each of the 30 vertical levels in the horizontal direction, disrupting both horizontal organization and column coherence (Figure 4.21a vs. c). We can now use feed these "scrambled" test datasets to our trained VAE and examine the latent space. On these three latent spaces (the original and the ones altered by the two scrambling approaches) we perform one hundred trials of k-means clustering to objectively analyze these new low-dimensional representations of the data and then pose the question: Is the VAE still able to separate out distinct regimes of physically interpretable convection with the horizontal information intentionally perturbed?

We immediately see differences in the regimes of convection the VAE identifies in these corrupted vertical velocity fields (Figure 4.21d, e, f; blue lines vs. red and purple). When both the information in each column and the columns themselves are scrambled (red lines), the VAE no longer identifies a mode of "shallow" Continental Cumulus. Instead, it splits the Deep Convection into 2 regimes, a first of only the most intense storms around the tropical rainforests and Pacific Warmpool (Figure 4.21f, red line), and another regime mixing both some of the Stratocumulus and the rest of the Deep Convection into one bucket (Figure 4.21d). In these results, just like in the three regimes identified originally, the vertical structure, geographic distribution, and cluster counts are robust under the 100 repetitions of the K-Means algorithm.

But when we perform the same clustering routine on the latent representation of the vertical velocity fields where only the column order (Figure 4.21b) was shuffled we lose this robustness. In two-thirds of the trials, the clusters formed on the latent space of this partially shuffled

data closely match those found in the original data (in both count, vertical structure, and geographic distribution). But in the other third, the clusters that form match those from the latent space whose data was scrambled both vertically and horizontally (Figure 4.21d, e, f, purple vs. blue and red lines).

We introduce one last test dataset (in original form, horizontally shuffled, and both horizontally and vertically shuffled) to determine the stability of the three sets of clusters discussed above. Since both this new test dataset and the original are drawn randomly from the same simulations (both spatially and temporally), they should have similar compositions. Therefore, if the latent space regimes we cluster are robust, we should expect only small shifts in the cluster centers and physical properties of the clusters between test datasets. Anything more than small changes associated with natural sampling variation would suggest unstable clusters of little physical meaning.

Figure 4.21g summarizes the median vertical structure of the Deep Convection clusters on the original test dataset (solid lines) and the new clusters with the original centers for initialization on the alternative test dataset (dashed lines). Reassuringly, in the original, unshuffled test dataset, the median vertical profile of each cluster shifts very little between test datasets (Figure 4.21g, blue lines; and h). But when the vertical levels are scrambled (and often when just the column order is scrambled) there are large shifts in the median vertical profile indicating instability in clusters. More specifically, for both types of scrambled test data, clusters that once corresponded to Deep Convection shift significantly, incorporating large amounts of Shallow Cumulus and simultaneously losing lots of Deep Convection samples. The effect of this is that the physical properties of each cluster become more muddled and it is more difficult to justify calling the clusters of scrambled data distinct species of convection.

The divergence in results between a test dataset composed of unblemished vertical velocity fields (Figure 4.21a), and one distorted at each vertical level (Figure 4.21c) confirms what we suspected when we first uncovered that the latent space organizes deep and shallow convection

separately – the VAE is highly sensitive to the vertical structure. Yet the inability of the latent space to consistently organize the same stable regimes of convection when input data column order (horizontal structure) is disrupted suggests something equally as profound: the VAE learns to identify its physically distinct convection regimes and anomalies in part from the coherent horizontal structure of the vertical velocity fields it receives through the encoder.

With this information, The disparate results in clustered convection regimes between the VAE and the $\overline{w'w'}$ profiles finally become clear. The $\overline{w'w'}$ method sacrifices all the horizontal detail that helps inform the VAE latent space organization; it is this critical added information that allows the VAE to extract more physically interpretable convection regimes, particularly for Deep Convection. Relying solely on dimensionality reduction approaches such as moment statistics to approximate high resolution dynamic and thermodynamic processes, comes at a cost as the horizontal structure is lost. Deep generative models like VAEs, which have access to the totality of the images offer a more holistic and pragmatic analysis of high-resolution climate simulation data than the moment statistics or other simple physical approaches. We believe this is a strong case for the widespread deployment of VAEs for more in-depth analysis of high resolution climate simulations and observational datasets.

Chapter 5

Understanding Extreme Precipitation Changes

5.1 Abstract

Despite the importance of quantifying how the spatial patterns of extreme precipitation will change with warming, we lack tools to objectively analyze the storm-scale outputs of modern climate models. To address this gap, we develop an unsupervised machine learning framework to quantify how storm dynamics affect changes in precipitation extremes, without sacrificing spatial information. For the upper precipitation quantiles (above the 80th percentile), we find that the spatial patterns of extreme precipitation changes are dominated by spatial shifts in storm dynamical regimes rather than changes in how these storm regimes produce precipitation. This Chapter shows how unsupervised machine learning, paired with domain knowledge, may allow us to better understand the physics of the atmosphere and anticipate the changes associated with a warming world.

5.2 Introduction

In Chapters 3 and 4 we investigated the idea that VAE-powered analysis of vertical velocity fields yields more comprehensive results than the estimation of the same fields through traditional moments. Our hope is that this unsupervised learning approach we developed has the potential to deepen our physical understanding of changes in convective organization and extreme precipitation with climate change. Our knowledge of these phenomena is limited because the dynamic information controlling powerful storms and convection is normally averaged prior to analysis. In particular, when studying the dynamic control on precipitation, second order effects of vertical velocity are neglected in the process of analyzing changes to the vertical structure and magnitude of atmospheric convergence [119, 47, 28, 2]. This approximation creates large spread in the projected changes in the dynamic tendency of precipitation (especially compared to the thermodynamic tendency). This creates significant model disagreement about the magnitude of the expected increase of extreme precipitation in the tropics [122, 136]. Given the anticipated impacts on vulnerable populations, it is critical we better understand the mechanisms by which global warming will impact extreme precipitation. We hope the objective VAE analysis, unlike statistical approximations, which is provided the full information from the vertical velocity fields, can more comprehensively track changes of Deep Convection, and provide a superior lens through which to visualize changes in extreme precipitation in tropical regions from dynamic effects (Figures 4.10e,f). We will test this theory now in Chapter 5.

5.2.1 Background

According to the latest Intergovernmental Panel on Climate Change report [45], "there is high confidence that extreme precipitation events across the globe will increase in both intensity and frequency with global warming". As the severity of storms and tropical cyclones

magnifies, there will be associated increases in flood-related risk [63] and challenges in water management [3, 4]. We know that these changes can be highly variable from region to region [39]. To first order, heavy precipitation extremes are limited by the water vapor holding capacity of the atmosphere, which increases by about 7% per 1K (Kelvin) of warming following an approximate Clausius-Clapeyron scaling [125]. This is referred to as the "thermodynamic contribution" to extreme precipitation changes [47] and gives a solid theoretical foundation for *spatially-averaged* changes in precipitation extremes.

Yet climate change adaptation requires knowledge of how precipitation extremes will change at the *local* scale, i.e., understanding the *changing spatial patterns* of precipitation extremes under warming. Focusing on the tropics, where most of the vulnerable world population lives [45], these changing spatial patterns are primarily dictated by atmospheric vertical velocity ("dynamical") changes because horizontal spatial gradients in temperatures are weak. This is referred to as the "dynamic contribution" to extreme precipitation changes [47].

A comprehensive understanding of this "dynamic contribution" remains elusive. Approximate scalings can be derived based on quasi-geostrophic dynamics [97, 124] and convective storm dynamics [112, 2]. But actionable findings require Earth-like simulations of the present and future climates (e.g., [133]), which can resolve regional circulation changes and their effects on storms in their full complexity. These simulations are computationally demanding and output large amounts of multi-scale, three-dimensional data that challenge traditional data analysis tools. For example, the state-of-the-art storm-resolving[1], SPCAM (Super Parameterized Community Atmospheric Model, [75, 77]) simulations we will use in this Chapter (Section 5.3.1) output 3.4 Terabytes of data over 90 days, with 76,944,384 samples of precipitation and the corresponding storm-scale vertical velocity fields (see Figure 5.1 for examples).

[1] 5 kilometers or less horizontal grid spacing

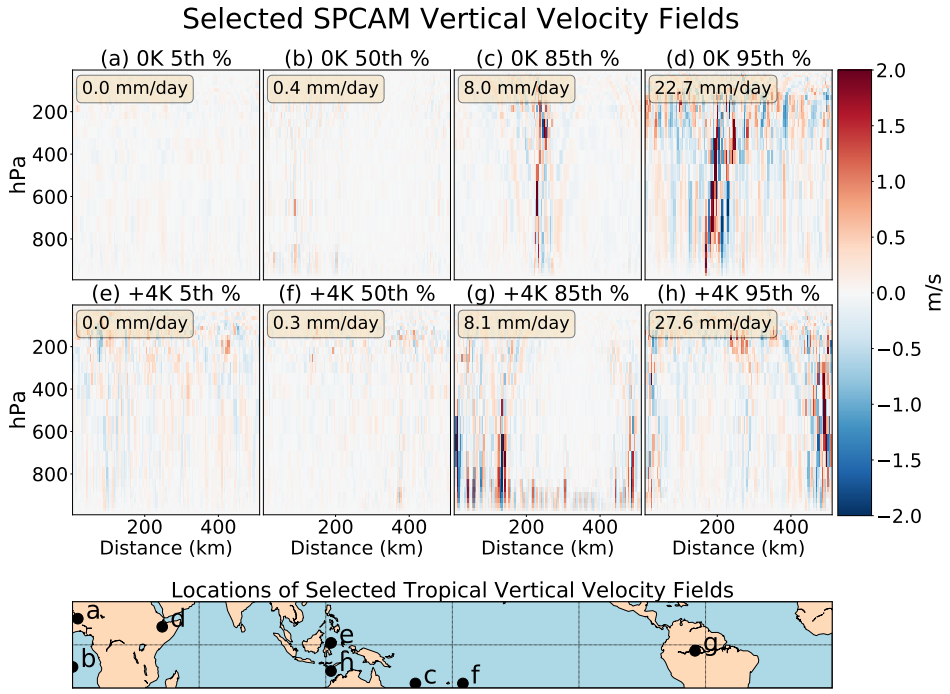

Figure 5.1: Selected vertical velocity fields from our "Control" (0K, a-d) and "Warmed" (+4K, e-h) SPCAM simulations. By sampling the precipitation distribution, we show instances of vertical velocity fields associated with no precipitation (a, e), drizzle (b, f), heavy rainfall (c, g), and intense storms (d, h).

5.2.2 Theory

In this section, we outline our strategy to facilitate the analysis of extreme precipitation changes including spatial details elucidating storm formation.

The crux of this analysis rests on the assumption that we can meaningfully cluster different vertical velocity fields into N different convection types. These convection types may have different frequencies or probabilities, indicated by π_i (so that, $\sum_{k=1}^{N} \pi_i = 1$). To calculate these π_i's, we use variational autoencoders in conjunction with k-means clustering, also called vector quantization [55, 98]. Details on the exact coarse graining procedure will be presented in Section 5.3.2; for now, we take the N different convection types and their frequencies as given.

This unsupervised quantization of regimes of convection through ML allows us to quantitatively

understand changes in precipitation extremes (P_{extreme}) from both changes in convection regime characteristics and probability. Here we define P_{extreme} as a fixed *high* quantile of precipitation (e.g., 80th-99.99th percentiles) at a given spatial location. To model the effects of climate change, we define $\Delta P_{\text{extreme}}$ as its absolute change from the "Control" to the "Warmed" climate, and show below that relative changes in precipitation extremes can be decomposed using changes in π (Equation 5.7).

We derive a decomposition of extreme precipitation changes for global warming by making a series of simple physical assumptions about precipitation. Note that while these assumptions help give physical meaning to each term, this decomposition could also be derived by decomposing the extreme precipitation field into its spatial-average and an anomaly, before further decomposing the anomaly using the objectively-identified dynamical regimes. This means that the assumptions made in this section only need to *approximately* hold to physically interpret the results of our decomposition. To first order, precipitation (P) scales like condensation rate, which depends on the full vertical velocity (w) and atmospheric water vapor (here quantified using specific humidity q) fields:

$$P \approx P\left(w, q\right). \tag{5.1}$$

Note that Equation 5.1 neglects the dependence on microphysical processes (see e.g., [111]) to focus on the thermodynamical and dynamical components of precipitation. When focusing on extreme precipitation, we de facto sample atmospheric columns that are so humid that the specific humidity q equals its saturation value q_{sat}. This allows us to further simplify Equation 5.1 in the case of high quantiles of P:

$$P_{\text{extreme}} \approx P_{\text{extreme}}\left(w, q_{\text{sat}}\right). \tag{5.2}$$

We now make the assumption that the thermodynamic dependence on q_{sat} can be factored out of the right-hand side of Equation 5.2 and denote the dynamical pre-factor as $\mathcal{D}\left(w\right)$:

$$P_{\text{extreme}} \approx q_{\text{sat}} \times \mathcal{D}\left(w\right). \tag{5.3}$$

The previous assumption can be justified quickly by assuming a moist adiabatic temperature profile and a vertically-uniform vertical velocity profile for extreme events [125, 112]. It can also be justified more accurately by noting that such vertical velocity profiles collapse when changing the vertical coordinate from pressure to the normalized integral of the moisture lapse rate [2]. We can now linearly decompose the dynamical pre-factor $\mathcal{D}\left(w\right)$ into the N regimes identified by our unsupervised learning framework:

$$\mathcal{D}\left(w\right) \approx \mathcal{D}_0 + \sum_{i=1}^{N} \pi_i \mathcal{D}_i, \tag{5.4}$$

where π_i is the frequency/probability of each dynamical regime. Combining Equation's 5.3, 5.4, and taking a logarithmic derivative with respect to climate change allows us to decompose relative changes in extreme precipitation as follows:

$$\frac{\Delta P_{\text{extreme}}}{P_{\text{extreme}}} \approx \frac{\Delta q_{\text{sat}}}{q_{\text{sat}}} + \frac{\Delta\left(\mathcal{D}_0 + \sum_{i=1}^{N} \pi_i \mathcal{D}_i\right)}{\mathcal{D}_0 + \sum_{i=1}^{N} \pi_i \mathcal{D}_i}, \tag{5.5}$$

where Δ denotes absolute changes from the reference to the warm climate. Lastly, we approximate the thermodynamic contribution to precipitation extremes as the relative changes in near-surface saturation specific humidity, which can be further approximated as spatially uniform:

$$q_{\text{sat}} = q_{\text{sat}}\left(T_s, p_s\right) \Rightarrow \frac{\Delta q_{\text{sat}}}{q_{\text{sat}}} \approx 7\%, \tag{5.6}$$

where T_s is near-surface temperature and p_s near-surface pressure. Expanding Equation 5.5 and substituting $\mathcal{D}\left(w\right)$ using Equation 5.3 yields the desired decomposition of precipitation extremes changes with climate:

$$\frac{\Delta P_{\text{extreme}}}{P_{\text{extreme}}} = \overbrace{\underbrace{\frac{\Delta q_{\text{sat}}}{q_{\text{sat}}}}_{\text{From theory}}}^{\text{Thermodynamic}} + \overbrace{\underbrace{\frac{q_{\text{sat}}}{P_{\text{extreme}}}}_{\text{From current climate}} \left(\Delta\mathcal{D}_0 + \underbrace{\sum_{i=1}^{N} \Delta\pi_i \mathcal{D}_i}_{\text{Regime prob. shifts}} + \underbrace{\sum_{i=1}^{N} \pi_i \Delta\mathcal{D}_i}_{\text{Intra-regime changes}} \right)}^{\text{Dynamic}} \tag{5.7}$$

Equation 5.7 shows that relative changes in P_{extreme} are the sum of a well-understood, spatially-uniform "thermodynamic" increase in saturation specific humidity (q_{sat} – see Equation 5.6), and a spatially-varying term. This spatially-varying term is the sum of N regime probability shifts $\Delta\pi_i$ (changes in our unsupervised ML-derived convection cluster assignment frequencies or "cluster sizes" – covered in more detail in Section 5.3), and N changes in regime characteristics $\Delta\mathcal{D}_i$ (changes in the "dynamic contribution" pre-factors, in precipitation units).

Our simulation data already contain P_{extreme} and q_{sat}, and we can derive π_i from our unsupervised learning framework, giving us all the information we need to calculate the elusive pre-factors $\mathcal{D}_i$ and their changes with warming. Using Equation 5.4, we linearly regress

$\frac{P_{\text{extreme}}}{q_{\text{sat}}}$ on the regime frequencies π_i in both the reference and warm climates to derive the pre-factors $\mathcal{D}_i$, which are the weights of the multiple linear regression. This is a step toward understanding how the spatial patterns of storm-scale dynamical changes, which are notably hard to analyze, can affect the spatial patterns of extreme precipitation. Understanding these changes is critical to trust local climate change predictions.

5.3 Methods

We will now discuss the data, models, and statistical techniques used in this Chapter. Additional details can be found in the Supplemental Information.

5.3.1 Data: High-resolution, Earth-like Simulations of Global Surface Warming

We acquire SPCAM data following the same procedure as Chapter 4, however, we also analyze other output variables necessary to build the relationship between Deep Convection and Extreme Precipitation including specific humidity, temperature, and pressure information. Readers familiar with the data can skip the remainder of this section.

The multi-scale modeling framework [139] used to generate our training and test data is composed of small, locally periodic 2D subdomains of explicit high-resolution physics that are embedded within each grid column of a coarser resolution ($1.9^\circ \times 2.5^\circ$ degree) host planetary circulation model [77]. In total, we performed six simulations of present-day climate launched from different initial conditions (but consistent resolution), configured with storm resolving models that are 512 km in physical extent, each with 128 grid columns with a horizontal resolution of 4 km. We approximate the atmosphere with a simple bulk

one-moment microphysical scheme and thirty vertical levels. We then perform six additional simulations but increase the sea surface temperatures by 4K. We compare the "Control" simulations against those with uniform increases in sea surface temperatures ("Warmed"). For our purposes, this creates a testbed for climate change, but we acknowledge that surface warming is only an approximation for the thermodynamic consequences of CO_2 concentration increase.

To investigate the "dynamic mode" of precipitation, we choose vertical velocity to represent the state of the atmosphere. These vertical velocity fields contain information about complex updraft and gravity wave dynamics across multiple scales. We considered the entire 15S-15N latitude band containing diverse tropical convective regimes. Examples of these vertical velocity snapshots, selected by precipitation percentile, can be seen in Figure 5.1.

5.3.2 VAE Training

Our ML methodology objectively defines dynamical regimes from two million two-dimensional vertical velocity fields, for which we proceed with the creation of a latent manifold to ensure the local correlations in the updrafts of our vertical velocity fields are preserved. For this, we rely on a fully convolutional VAE design, whose architecture is depicted in Figure 4.2 and was discussed in more detail in Chapter 4. But to summarize:

We train the VAE, we perform beta-annealing [64, 21], expanding the Evidence Lower Bound (ELBO) traditionally used to train the VAE by including a β parameter and linearly anneal β from 0 to one over 1600 training epochs. The number of layers and channels in the encoder and decoder can be found in Figure 4.2 (4 layers in each, stride of two). After manual hyperparameter tuning, we choose ReLUs as activation functions in both the encoder and the decoder. We pick a relatively small kernel size of 3 to preserve the small-scale updrafts and downdrafts of our vertical velocity fields. The dimension of our latent space is 1000.

5.3.3 Quantization Procedure

Our aim is to objectively interpret extreme precipitation from climate shifts based on detailed vertical velocity fields. But on the full vertical velocity snapshots, this problem is intractable so we must rely on additional statistical techniques for analysis.

Although the use of a VAE encoder makes our high-resolution simulation data more manageable, we require additional work to derive the formal convective probability information, π. The main idea is to convert a high-dimensional, continuous probability distribution over velocity fields into a fixed-size, discrete probability distribution over quantization points [115]. Then, we use the coarse-grained, discrete distribution to compute various quantities of interest.

We use a convolutional VAE to nonlinearly embed our 2D input data (vertical velocity fields) into a lower-dimensional latent space. To quantize the emergent latent space, we employ k-means clustering (More details in SI-A): we encode our training data into the latent space and cluster them into N clusters. We also define a vector of *cluster assignment probabilities* π_i for $i = 1, \cdots, N$ as the percentage of training data assigned to each cluster i. This dimensionality reduction and clustering can be thought of as a lossy compression of the data [171]. As we will see, the discrete structure helps us compute various quantities of interest. While the quantization approximation can, in principle, be made arbitrarily precise using a large N, we use $N = 3$ in practice for interpretability based on the findings of Chapter 4. We cluster convection into three distinct regimes we are familiar with from Chapter 4: (1) Marine Shallow Convection, (2) Continental Shallow Cumulus Convection, and (3) Deep Convection.

But we go beyond our previous work by honing in on the *changes* of the cluster assignment probabilities under global warming. In order to compare the present and future data distributions, we train the VAE and learn the cluster centers based on present climate data.

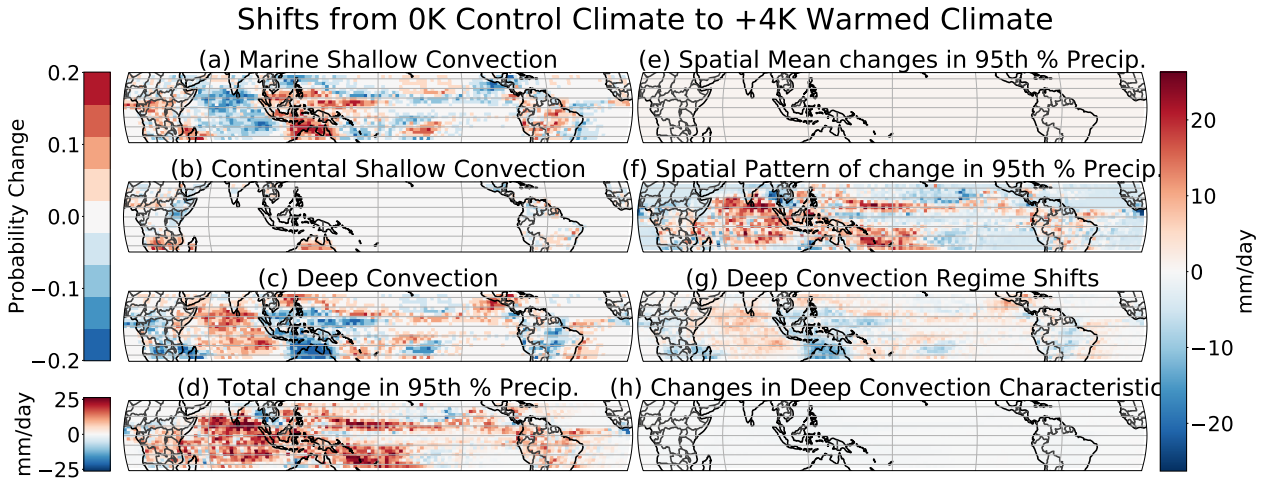

Figure 5.2: Changes induced by +4°C of simulated global warming: **The patterns of storms change** (a-c), **which changes the patterns of extreme precipitation** (f), **mostly because deep convective storms shift location** (g). Panels (a-c) display probability shifts in the three dynamical regimes found through clustering with $N = 3$, corresponding to (a) "marine shallow", (b) "continental shallow cumulus", and (c) "deep" convection. We subtract the spatial-mean change (e, the "thermodynamics") from the total change (d) to yield the "dynamic" contribution (f). Using Equation 5.7, we decompose the changing spatial patterns (f) into five terms, including (g) probability changes in deep convection, (h) changes in deep convective precipitation, and three additional terms depicted in Figure 5.7

This yields the present cluster assignment frequencies π^{0K}. In a second step, we encode all future climate data into the latent space and assign each datum to the nearest (control) cluster center, yielding the future cluster assignment frequencies π^{4K}. The difference vector of assignment frequencies, before and after global warming, is given by $\Delta\pi = \pi^{4K} - \pi^{0K}$. This information can then be used as a proxy for dynamical regime shifts with warming. We visualize these shifts and interpret their implications for extreme precipitation below.

5.4 Results

5.4.1 Unsupervised Machine Learning Reveals Convective Responses to Climate Change

Figure 5.2 shows the probability shifts in convection type ($\Delta\pi_i$) from the "Control" to the "Warmed" climate. The dominant climate change signal captured by our unsupervised framework is the increased geographic concentration of deep convection (Figure 5.2c). More specifically, deep convection becomes more frequent over warm ocean waters and especially the Pacific Warm Pool [7] while shallow convection becomes less common in these unstable regions (Figure 5.2a). This result is consistent with observational trends showing an intensification of already powerful storms over the warm tropical waters [7]. At first glance, the pattern of this unsupervised deep convection shift with warming ($\Delta\pi_1$) looks quite similar to shifts in extreme precipitation (Figure 5.2c vs. f).

With just the information of the regime probabilities, we can model the spatial patterns of precipitation changes at upper percentiles (Figure 5.6). Our model becomes less accurate at lower precipitation quantiles, partly because we are not using specific humidity information (the approximation of Equation 5.2 is only valid for high precipitation quantiles). This degree of accuracy at the upper percentiles suggests that changes in the location of convective dynamical regimes can explain a large fraction of changes in extreme precipitation, which should be further tested in diverse climate change modeling frameworks.

5.4.2 Decomposing the Dynamic Contribution to Extreme Precipitation Changes

We now isolate the dynamical contributions to dynamical changes in extreme precipitation by decomposing the spatial patterns (5.2d) into changes in regime probability π_i and changes in regime characteristics $\mathcal{D}_i$. Unlike traditional approaches that spatially average information, we use our fully-convolutional encoder and latent space clustering to leverage storm-scale variability.

We calculated changes in regime probability (how regimes move in space) in section 5.3, so we must now calculate changes in how each regime produces precipitation, which involves the following two steps. First, we empirically estimate $\mathcal{D}_i$ by using the probabilities of deep and shallow convection[2]. Second, we estimate changes in "deep" and "shallow" convection dynamical pre-factors as $\Delta\mathcal{D}_i = \mathcal{D}_i^{4K} - \mathcal{D}_i^{0K}$.

We now have the requisite information to understand the drivers of extreme precipitation changes themselves. We ask: *Did the patterns of extreme precipitation simply follow the changing patterns of the convective regime, or are there more complex changes in how deep convection produces rain?* We address this question by comparing how much of the spatial variance in extreme precipitation ΔP_e can be explained by changes in convection probability ($\Delta\pi$), and how much of it can be explained by changes in the dynamical prefactors ($\Delta\mathcal{D}$). This comparison relies on the following decomposition of extreme precipitation variance $\text{var}\left(\Delta P_{\text{extreme}}\right)$, derived in Sec D of the SI:

[2]More specifically, we estimate the dynamical pre-factors ($\mathcal{D}_i$) by regressing $P_{\text{extreme}}/q_{\text{sat}}$ on π_1 and π_2, neglecting the "Continental Shallow Cumulus" regime as it concentrates over arid continental zones with high lower tropospheric stability and low latent heat fluxes, making conditions unfavorable for precipitation [43].

$$\text{var}\left(\Delta P_{\text{extreme}}\right) = \underbrace{\text{var}\left[\left(q_{\text{sat}}\sum_{i=1}^{N}\Delta\pi_i\mathcal{D}_i\right)\right] + \overline{(\text{CT})_{\Delta\pi}}}_{\text{Shift in regime location}} + \underbrace{\text{var}\left[\left(q_{\text{sat}}\sum_{i=1}^{N}\pi_i\Delta\mathcal{D}_i\right)\right] + \overline{(\text{CT})_{\Delta\mathcal{D}}}}_{\text{Intra-regime changes}} + \mathcal{R} \tag{5.8}$$

where CT are cross-terms and $\mathcal{R}$ groups all terms of the decomposition that are not needed to compare differences in precipitation from regime shifts vs. intra-regime changes. To understand what is most crucial to precipitation changes, we depict the spatial mean of Equation 5.8's terms in Figure 5.3.

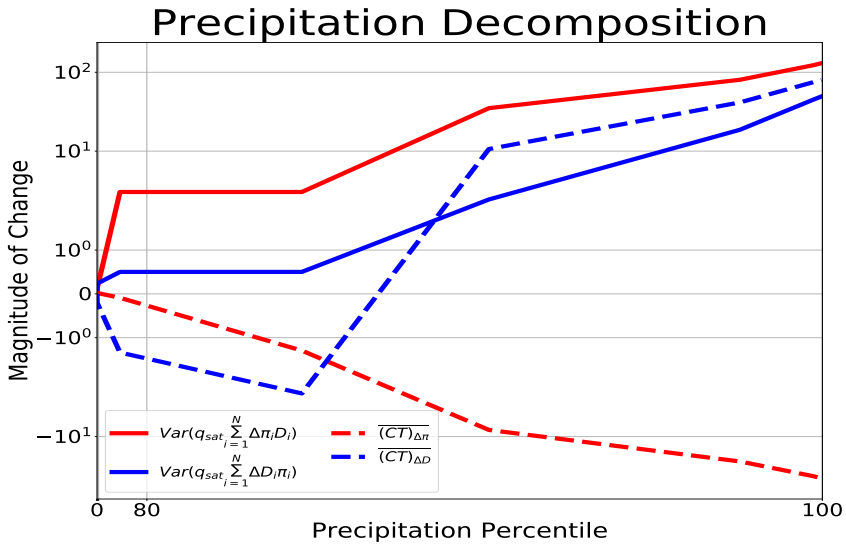

Figure 5.3: Derived from Equation 5.8 we compare the mean of the spatial anomaly of convective probability shifts ($\Delta\pi$) to the changes in the dynamical prefactors (ΔD). We find that the convective regime shifts are of greater importance to explain the changes in extreme precipitation (80th-99.99th percentiles)

For high precipitation quantiles where our model works best (especially above the 80th percentile), extreme changes are dominated by regime probability shifts rather than by intra-regime changes (Figure 5.3): The "$\Delta\mathcal{D}$ term" is smaller than the "$\Delta\pi$ term" (red line

vs. blue line) [3]. This result aligns well with our qualitative analysis in Figure 5.2 showing the similarity between the spatial patterns of deep convection regime changes and extreme precipitation changes. The spatial patterns of extreme precipitation changes are dominated by the changing patterns of storm characteristics identified by our unsupervised framework while changes in how each regime produces precipitation are less important.

5.5 Conclusion

Based on our findings, proper prediction of spatial shifts in deep convection with global warming should allow us to anticipate regional and local changes in precipitation extremes. This highlights the importance of leveraging the full spatial extent of this information (traditionally averaged out) to derive accurate regional and local changes. The necessity of understanding this rich spatial information indicates a role for ML methods like variational encoders and clustering routines for the analysis of storm-scale climate information to deepen our understanding of extreme events. Our next step to build credibility in this unsupervised model is to deploy the workflow on more diverse climate change data like the High-Resolution Model Inter-comparison Project (HighResMIP) [48, 57] and determine its ability to explain spatial variations in extreme precipitation with climate change.

5.6 Appendix A: VAE Benchmarking and Performance Evaluation

We train our VAE on 160,000 unique vertical velocity fields and use an additional 125,000 samples to validate and optimize the model hyperparameters. Finally, we leverage 1,000,000

[3]Note that at precipitation quantiles larger than 0.99, we lack samples for the analysis to work properly as evidenced by the pixelation of the changing patterns in Figure 5.5

Mean Squared Error m^2/s^2			
Model	Training Set	Validation Set	Test Set
VAE	$3.79 * 10^{-4}$	$1.11 * 10^{-3}$	$3.33 * 10^{-3}$
Linear Baseline	$3.10 * 10^{-3}$	$4.70 * 10^{-3}$	$5.10 * 10^{-2}$

Table 5.1: The MSE of both of our models ("linear baseline" and VAE) calculated across training/validation/test data. For both training and test data, we see low reconstruction errors, suggesting satisfactory skill and generalization ability. Overall, the VAE outperforms the "linear" baseline

Structural Similarity Index Metric			
Model	Training Set	Validation Set	Test Set
VAE	0.998	0.995	0.987
Linear Baseline	0.990	0.986	0.981

Table 5.2: The mean SSIM [161] of both of our models across training/validation/test data. The models both generalize well to our test data. Again, the VAE outperforms the "linear baseline"

vertical velocity fields in the test dataset for robust analysis. The high count in the test dataset is necessary both due to the high spatio-temporal correlations common in meteorological data but also because of the geographic conditioning in our analysis – we need enough samples at each lat/lon grid cell, not just globally. To determine whether our data are nonlinear enough to warrant the use of a VAE we also train a baseline model of the same architecture but with all activation functions replaced by "linear". The fact that the VAE reconstructs the vertical velocity snapshots with both lower error and a higher degree of structural similarity suggests significant non-linearity is involved in compressing and rebuilding the 2D fields (Tab 5.1 and Tab 5.2). This problem is therefore well suited for the non-linear dimensionality reduction of the VAE encoder and less so for linear models.

Tab 5.1 and Tab 5.2 show 160,000 is enough training samples to create reconstructions of high-resolution vertical velocity fields with both a low MSE and a high degree of overall structural similarity. Though there is a small amount of overfitting, we see that performance remains strong for a test dataset containing multiple species of convection from all parts of the tropics ranging from deserts to rainforests; oceans to continents. Furthermore, what we

are most concerned with is not the reconstruction quality itself, but the interpretability of the latent space for clustering.

5.7 Appendix B: Full Decomposition of the Spatial Variance of Extreme Precipitation

We derive the decomposition of the spatial variance of extreme precipitation in four steps. First, we multiply Equation 5.7 by P_{extreme}:

$$\Delta P_{\text{extreme}} = \underbrace{P_{\text{extreme}} \frac{\Delta q_{\text{sat}}}{q_{\text{sat}}}}_{\mathcal{Q}} + q_{\text{sat}} \left(\Delta \mathcal{D}_0 + \sum_{i=1}^{N} \Delta \pi_i \mathcal{D}_i + \sum_{i=1}^{N} \pi_i \Delta \mathcal{D}_i \right). \tag{5.9}$$

For convenience, we use $\mathcal{Q}$ to denote the first term of Equation 5.9. We then take its spatial anomaly by applying the spatial anomaly operator (X'):

$$\Delta P'_{\text{extreme}} = \mathcal{Q}' + \Delta \mathcal{D}_0 q'_{\text{sat}} + \left(q_{\text{sat}} \sum_{i=1}^{N} \Delta \pi_i \mathcal{D}_i \right)' + \left(q_{\text{sat}} \sum_{i=1}^{N} \pi_i \Delta \mathcal{D}_i \right)', \tag{5.10}$$

where we have used the fact that $\Delta \mathcal{D}_0$ is uniform in space ($\Delta \mathcal{D}_0 = \overline{\Delta \mathcal{D}_0}$ and $\Delta \mathcal{D}'_0 = 0$). Squaring Eq 5.10 yields:

$$\left(\Delta P'_{\text{extreme}}\right)^2 = \left(\mathcal{Q}'\right)^2 + \left(\Delta \mathcal{D}_0\right)^2 \left(q'_{\text{sat}}\right)^2 + \left[\left(q_{\text{sat}} \sum_{i=1}^{N} \Delta \pi_i \mathcal{D}_i \right)' \right]^2 + \left[\left(q_{\text{sat}} \sum_{i=1}^{N} \pi_i \Delta \mathcal{D}_i \right)' \right]^2 + \text{CT}, \tag{5.11}$$

where the cross-terms CT can be decomposed into cross-terms involving spatial shifts in regime probability:

$$(\mathrm{CT})_{\Delta\pi} \stackrel{\mathrm{def}}{=} 2\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\Delta\pi_i\mathcal{D}_i\right)'\left[\mathcal{Q}'+\Delta\mathcal{D}_0 q'_{\mathrm{sat}}\right], \tag{5.12}$$

cross-terms involving changes in how each regime produces precipitation:

$$(\mathrm{CT})_{\Delta\mathcal{D}} \stackrel{\mathrm{def}}{=} 2\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\pi_i\Delta\mathcal{D}_i\right)'\left[\mathcal{Q}'+\Delta\mathcal{D}_0 q'_{\mathrm{sat}}\right], \tag{5.13}$$

and additional cross-terms:

$$(\mathrm{CT})_{\mathrm{other}} \stackrel{\mathrm{def}}{=} 2\Delta\mathcal{D}_0 q'_{\mathrm{sat}}\mathcal{Q}' + 2\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\Delta\pi_i\mathcal{D}_i\right)'\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\pi_i\Delta\mathcal{D}_i\right)' \tag{5.14}$$

Taking the spatial mean $\left(\overline{X}\right)$ of Eq 5.11 and noting that the spatial variance is defined as the spatial mean of the squared spatial anomaly, we derive the following decomposition:

$$\begin{aligned}\mathrm{var}\left(\Delta P_{\mathrm{extreme}}\right) = {} & \mathrm{var}\left(\mathcal{Q}\right) + \left(\Delta\mathcal{D}_0\right)^2\mathrm{var}\left(q_{\mathrm{sat}}\right) + \mathrm{var}\left[\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\Delta\pi_i\mathcal{D}_i\right)\right] + \mathrm{var}\left[\left(q_{\mathrm{sat}}\sum_{i=1}^{N}\pi_i\Delta\mathcal{D}_i\right)\right] \\ & + \overline{(\mathrm{CT})_{\Delta\pi}} + \overline{(\mathrm{CT})_{\Delta\mathcal{D}}} + \overline{(\mathrm{CT})_{\mathrm{other}}} + \overline{\text{Numerical Residual}},\end{aligned} \tag{5.15}$$

where we have introduced the decomposition's numerical residual, which helps us assess which terms are significant. Grouping the terms irrelevant to the comparison between regime spatial shifts and intra-regime changes into a single term, $\mathcal{R}$, mathematically defined as:

$$\mathcal{R} \stackrel{\mathrm{def}}{=} \mathrm{var}\left(\mathcal{Q}\right) + \left(\Delta\mathcal{D}_0\right)^2\mathrm{var}\left(q_{\mathrm{sat}}\right) + \overline{(\mathrm{CT})_{\mathrm{other}}} + \overline{\text{Numerical Residual}}, \tag{5.16}$$

we recover Equation 5.8 from the manuscript's main text. For additional context on the significance of our decomposition, we plot all the terms in Figure 5.4. We see that as in Figure 5.5, our decomposition is most valid for high precipitation percentiles (percentiles where the residual (blue line) is of lesson magnitude than other quantities).

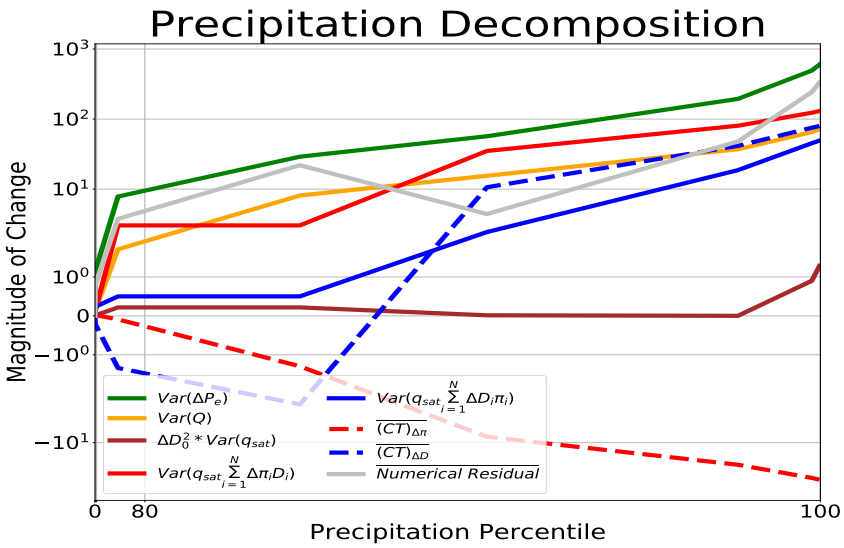

Figure 5.4: Derived from Equation 5.15, we plot each term from the full decomposition for the variance in the change in extreme precipitation, $Var(\Delta P_e^2)$. We focus primarily on precipitation percentiles 80-99, where our model is valid (the numerical residual, grey, is smaller than the key terms) and we have sufficient data (Figure 5.5). Across these extreme precipitation percentiles, we find that the change in probability of convection type ($\Delta\pi$ – red) is of greater importance than changes in the Dynamical Prefactors (ΔD – blue). For additional context compared to Figure 5.3, we include all terms from Equation 5.15

5.8 Appendix C: Supplemental Figures

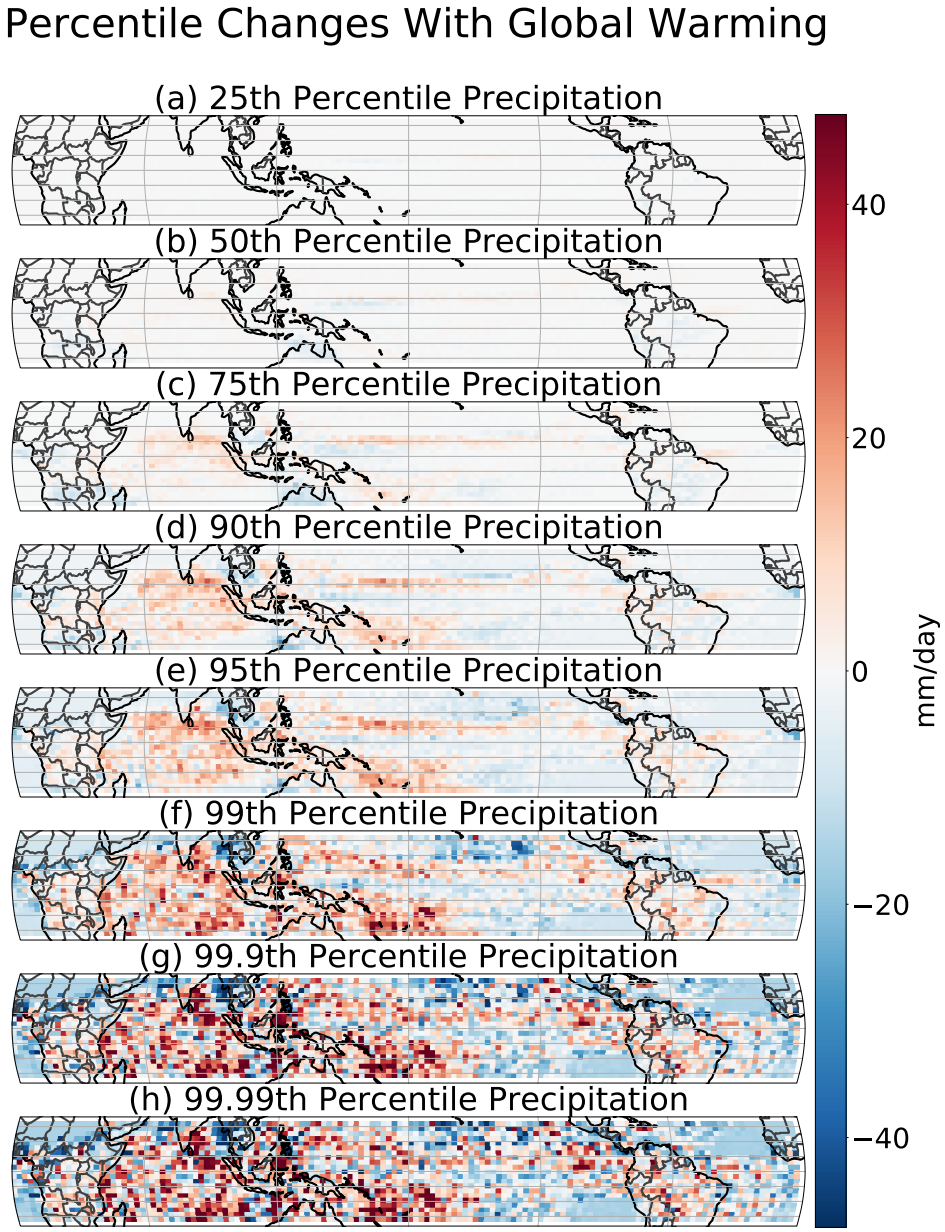

Figure 5.5: The shifts in different percentiles of precipitation with global warming, where we again stratified and plotted the data by latitude/longitude grid cell. As in Figure 5.2d we again remove the mean to highlight the dynamical pattern and see at what threshold the alignment with the VAE identified Deep Convection shifts (Figure 5.2c) is greatest. The top percentiles including (f-h) are pixelated because of a lack of samples that are out on the tail of the PDF.

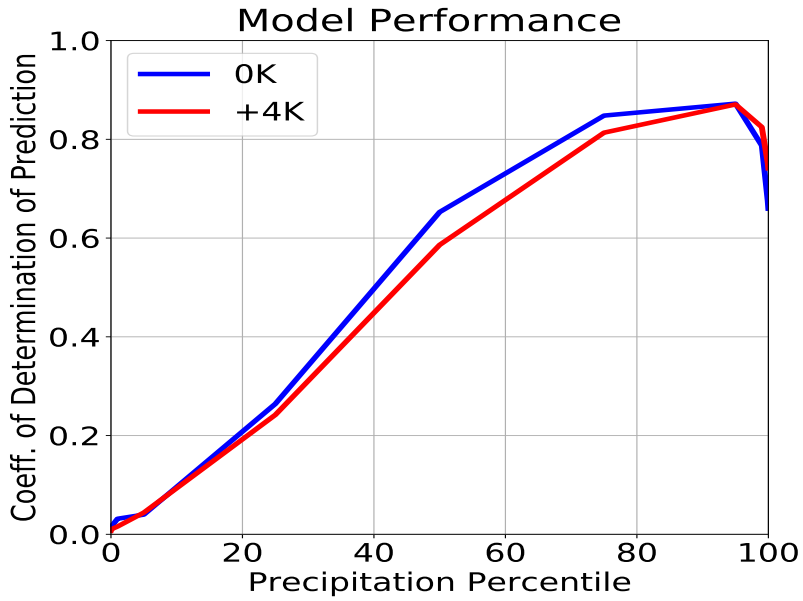

Figure 5.6: The simple results of the simple regression model we use to predict extreme precipitation patterns ($\frac{P_{extrmeme}}{qsat}$) using just the dynamic contributions, $\pi_{Deep\ Convection}$ and $\pi_{Shallow\ Convection}$ identified by our unsupervised ML framework. We see our model works very well for high precipitation percentiles where the dynamic contributions are greatest and less well for lower percentiles where thermodynamics are also important.

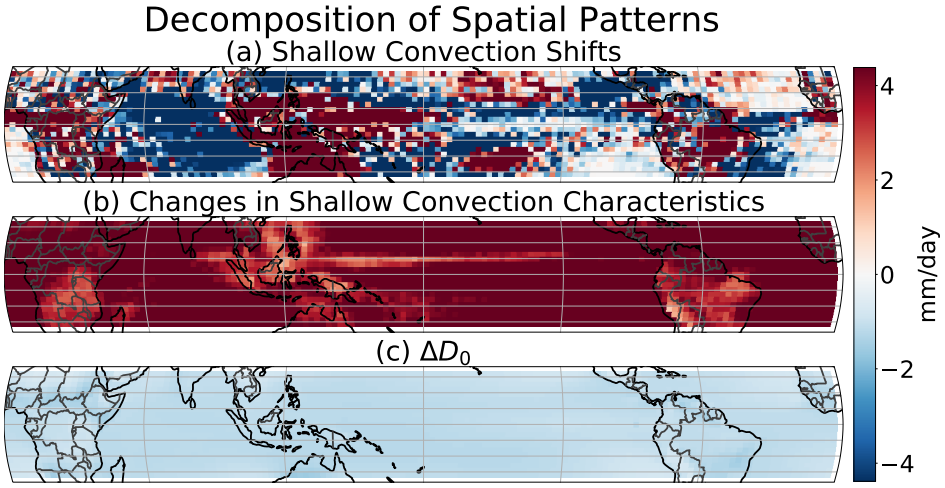

Figure 5.7: From Equation 5.7, we can decompose the changing spatial patterns (Figure 5.2f) into five terms, including probability changes in shallow convection (a), changes in deep convective precipitation (b), and the intercept of Dynamical Prefactor (c).

Chapter 6

Conclusion

6.1 Significance Statement

Global warming poses an imminent risk to people across the globe, but especially in the tropics due to increased precipitation extremes, flooding, droughts, and temperature variations. It is a critical priority to better anticipate and prepare for these changes. However, the models we use to try and represent these changes have large errors and uncertainties even as the field has made significant strides forward. This deadlock is due to two critical problems:

1. The key processes controlling the atmosphere of the Earth occur over scales of meters to kilometers, while typical modern climate models have a resolution of $100 - 200\text{km}^2$ horizontally - meaning important sub-grid processes are parameterized [69, 74, 108]. This is necessary because of sharp limits on the rate of increase in available computing resources [56]. According to Moore's Law, our computing power should double roughly every two years as we increase the number of transistors in computer processors. While any improvement is helpful for running high-resolution climate models, at this pace it will take many decades more to run climate models at "storm-resolving" scales in

hundred-year simulations [146]. But we need to anticipate the climate changes now, not far in the future after the changes are already taking place.

2. A distinct, but related, problem is the output of the world's first generation of climate models that sidestep parameterization by explicitly resolving convection on planetary scales. Such "storm-resolving" models are difficult to analyze since even just a couple of weeks of model output from a global SRM quickly grows to Petabytes of data [152]. By itself, this output is not only challenging to store and access but also simply overwhelming for scientific analysis. Classical dimensionality reduction approaches require averaging out the small-scale structures in the simulation that are most valuable. This limits our understanding of deep convection, convective organization, storm growth, and extreme precipitation.

Motivated by these limits, the body of work represented by this thesis has intentionally explored machine learning methods and the broader idea of incorporating the field of atmospheric sciences into the "Deep Learning Revolution" [30] of the past decades, which has been transforming the physical sciences. These neural networks have the potential to skillfully emulate the effects of explicit high-resolution physics by training on large data archives. Once trained, they can be coupled at a fraction of the computational costs, providing a potential path to sidestep Moore's Law. Meanwhile, analogous tools can facilitate novel dynamical discoveries including complementary ways to analyze the mechanisms of climate change without ignoring the native structures of turbulent complexity within high resolution data archives. We hope the frameworks we have developed over these previous Chapters will encourage others in atmospheric sciences to also explore the potential of neural networks for engineering, analysis, and discovery toward the goal of a better understanding of the atmosphere.

6.2 Summary of Results

Overall, results aimed to show the potential for machine learning to aid in running next-generation multi-scale climate models, in analyzing the details of the atmosphere, and for a better characterization of how extreme precipitation processes will change with climate. We also perform dimensionality reduction tasks to enable an objective analysis of massive climate simulations. We summarize the most important results from each of these tasks below.

Simple neural networks emulate realistic convection In Chapter 2, we deployed feed-forward neural networks to replace sub-grid convective parameterizations in SPCAM5. We find that when we test our "hybrid" climate model its representation of convection across the globe is similar to the original physical model. At the time this was one of the first attempts to achieve such skill in a testbed more difficult than statistically steady aquaplanet simulations [50, 140]. We found that success in the more operationally relevant configuration required both significantly more training data for the neural network and a semi-automated hyperparameter tuning sweep. In contrast to contemporary work [59], we found it possible to develop a skillful machine learning framework locally in both space and time, consistent with the paradigm of diagnostic single column parameterization. This is important for the ability to run hybrid machine learning models in a prognostic or "online" mode as the host physical climate model receives the sub-grid convective parameterizations locally as well. It is also a finding that at the very least raises questions about to what extent it is important to account for "convective memory" for modeling storm systems in climate prediction, a topic that remains in debate.

Deep Generative Models help organize the details of tropical convection. In Chapter 3 we deploy the first-ever VAE to learn the details of images of high-resolution convection in the form of 2D vertical velocity fields from SPCAM5. Despite the detail of these snapshots, we find that our VAE can learn not only to reconstruct the details of this

convection with high accuracy but also to construct physically coherent, low dimensional representations of the convection that are human-interpretable. The latent space sorts out species of convection by differences in intensity and vertical structure. At the same time, through "density estimation", the VAE can identify powerful storms and unusual structures of convection out of large test datasets. In Chapter 4 when we extrapolate the VAE approach to an archive of uniformly resolved global storm resolving model simulations and look at geographic characteristics, we find three recognizable regimes of tropical convection similar to what could be identified through classical physical methods. By sorting these physical properties in a data-driven approach, the VAW allows us to analyze the true contents of large climate simulations.

There are significant differences in today's SRM representation of tropical convection In Chapter 4 we leverage the findings of Chapter 3 to measure the "Distribution Shifts" between different SRMs as a proxy for degree of similarity. We combine analysis through a VAE Encoder with a Vector Quantization technique to measure the differences. While this is often done through a "supervised" machine learning setup, the framework we developed allows for an unsupervised, objective analysis. We find that while many SRMs are quite similar in their representation of atmospheric dynamics, some models are notably different. Only six of the SRMs we examine are consistent while three others show differences in the type, proportion, and intensity of convection. While convergence between the majority of ensemble members could be viewed as reassuring, these findings highlight the need to better understand the parameterization choices in these SRMs and why they can manifest in occasionally distinct patterns and preferred turbulent structures impacting the representation of tropical dynamics.

Machine learning methods can capture expected signals of global warming In Chapters 4 and 5 we extend our "Distribution Shift" analysis from comparing different climate models to comparing different climate states themselves. When we measure the shifts

between a simulation of our current climate and a simulation after 4K of warming, we find our unsupervised machine learning approach identifies anticipated changes in the tropical climate. In a purely data-driven way, without any human bias, we find an intensification of powerful storms in the tropics, a concentration of deep convection over the warmest waters, an expansion of dry zones over continents, and a rise in the tropopause. This builds credibility in the use of such methods as a basis for deeper scientific investigation.

Future changes in extreme precipitation are controlled by shifts in dynamical regime probability In Chapter 5 we extend our unsupervised machine learning methods for dynamical analysis, leveraging the VAE encoder's ability to extract information content from the fine-scale spatial information of simulations with explicit convection in the tropics for both control and warmed climates. By leveraging the full detail of these simulations, we are able to isolate the contributions to dynamical changes in extreme precipitation. We separate out the contributions from changes in convection regime probability and regime characteristics. When we decompose these changes, we find that extreme precipitation follows the patterns of the convective regimes. This is a finding that suggests the differences in precipitation with global warming are caused by these convection probability shifts rather than how the deep convection itself produces rain, or the morphology of convective extremes.

6.3 The role of machine learning in climate and atmospheric sciences

Instances of applying machine learning to cloud-related processes and the modeling of the interactions between clouds and climate have increased at a rapid pace over the past several years. Simultaneously, the field of machine learning is also evolving at a rapid pace meaning the tools being deployed to analyze climate data and run climate models are constantly

changing and updating. This stochasticity makes predicting the future of this sub-field extremely challenging. Nevertheless, we will conclude with some remarks based on our experience deploying different machine-learning tools for both climate modeling and analysis of climate data simulations.

6.3.1 Next Generation Hybrid Climate Modeling

Many groups have shown proof of concept for a coupled neural network-physical climate model system [50, 175, 176]. We have discovered, whether through brute force of hyper-parameter sweeps and training data or more sophisticated neural network architecture [59], that skillful "offline" performance is possible for neural network emulation of convective parameterizations. It is worth noting however that even with these approaches neural networks still struggle to capture climate extremes, generalize to out-of-sample data, and accurately capture non-gaussian meteorological variables [17]. But even when offline performance is strong, these hybrid models quickly run into stability issues in "online" prognostic mode [140, 175]. If this obstacle could be addressed, then the promise of climate models that can both explicitly resolve deep, moist convection and be run on 100-year timescales could be feasible in the medium term, possibly within a decade. However, this is no small feat and will likely require a combination of more sophisticated, generative machine learning models to better capture probabilistic variables and physically-informed additions to neural networks. In particular, we believe putting key variables, particularly moisture, in a physically interpretable and generalizable form for the neural network is key to improving simulation stability [17]. However, putting too many physical constraints on energy or momentum has the potential to inhibit the neural network from learning the non-linear relationships that control the atmosphere. A recent promising outgrowth may be the use of neural networks to replace convective parameterizations in weather models, where physical conservation is less important and the timescales needed to successfully run hybrid models for use are much shorter.

6.3.2 Understanding the Details of Global Storm-Resolving Models

The idea of using machine-learning as an engineering tool to replace parameterizations in a weather or climate model is now nearly ubiquitous across the field [50, 175, 176, 59, 17, 163]. However, the deployment of neural networks for analysis is still an under-explored area in our opinion. We are pleased with the results of our initial foray into this space with Deep Generative Models (Chapters 3-5), but we believe we have only scratched the surface of the possibilities. In particular, we believe shifting from single-variate to multi-variate analysis will be critical for better understanding the non-linear relationships in these high-resolution outputs [102]. But there is potential beyond the clustering and non-linear dimensionality reduction work we have performed in Chapters 3 and 4. We believe the utility of the latent space-based analysis can extend further. A better understanding of the details of convection could be found through approaches such as latent space interpolation to better understand convective changes on minute timescales and using the latent space itself for equation-discovery [14]. It is also possible that with the right architecture, deep, generative models like VAEs could also power "Reduced Order Modeling" frameworks and generate synthetic convection data. Additionally, the anomaly-detection talents of these models hold great promise for isolating severe weather, or any other phenomena of interest in massive simulation outputs or observational datasets. A challenge will be advancing this work in parallel with the fields by adopting more modern machine learning approaches and higher resolution datasets [52, 48].

6.3.3 Can Machine Learning Replace Domain Knowledge?

Beyond the use of machine learning to replace components of climate models or as a tool for data analysis, there have been efforts to entirely replace physical climate and weather models with neural networks [141, 92]. This philosophical approach is very attractive when

trying to beat a single metric or score. However, beyond the use of a topline number to quantify performance this approach appears to have great difficulty generalizing [83] beyond the immediate training task given to the neural network(s). This hints at the difficulty of introducing "black boxes" in modeling the physical atmosphere. The neural networks can learn relationships between the training data and the targets, but even if they allow the neural network to perform well on hold-out validation data, the findings could nevertheless be based on spurious and unphysical relationships. We require much more use of explainable-AI [26, 158, 160] to better understand how neural networks are actually operating, and what they are learning from the data provided. But also, we believe there is still, and will be for the foreseeable future, a use and a need for incorporating physical knowledge into machine learning approaches for the best chance of improved knowledge of our atmosphere.

In summary, it is an exciting and important time for interdisciplinary science at the interface of computer and climate sciences.

Bibliography

[1] Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.

[2] T. H. Abbott, T. W. Cronin, and T. Beucler. Convective dynamics and the response of precipitation extremes to warming in radiative–convective equilibrium. *Journal of the Atmospheric Sciences*, 77(5):1637 – 1660, 2020.

[3] W. Adger, S. Agrawala, M. Mirza, C. Conde, K. O'Brien, J. Pulhin, R. Pulwarty, B. Smit, and K. Takahashi. Assessment of adaptation practices, options, constraints and capacity. climate change 2007: impacts, adaptation and vulnerability. *Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*, pages 717–743, 01 2007.

[4] W. N. Adger, S. Dessai, M. Goulden, M. Hulme, I. Lorenzoni, D. R. Nelson, L. O. Naess, J. Wolf, and A. Wreford. Are there social limits to adaptation to climate change? *Climatic Change*, 93(3):335–354, 2009.

[5] M. Ahlgrimm and R. Forbes. The impact of low clouds on surface shortwave radiation in the ecmwf model. *Monthly Weather Review*, 140(11):3783 – 3794, 2012.

[6] A. A. Alemi, B. Poole, I. S. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken elbo. In *ICML*, 2018.

[7] R. P. Allan, C. Liu, M. Zahn, D. A. Lavers, E. Koukouvagias, and A. Bodas-Salcedo. Physically consistent responses of the global atmospheric hydrological cycle in models and observations. *Surveys in Geophysics*, 35(3):533–552, 2014.

[8] N. P. Arnold, M. Branson, M. A. Burt, D. S. Abbot, Z. Kuang, D. A. Randall, and E. Tziperman. Effects of explicit atmospheric convection at high co₂. *Proceedings of the National Academy of Sciences*, 111(30):10943–10948, 2014.

[9] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[10] R. Atlas and C. Bretherton. Aircraft observations of gravity wave activity and turbulence in the tropical tropopause layer: prevalence, influence on cirrus and comparison with

global-storm resolving models. *Atmospheric Chemistry and Physics Discussions*, 2022:1–30, 2022.

[11] P. Bechtold, J.-P. Chaboureau, A. Beljaars, A. K. Betts, M. Köhler, M. Miller, and J.-L. Redelsperger. The simulation of the diurnal cycle of convective precipitation over land in a global model. *Quarterly Journal of the Royal Meteorological Society*, 130(604):3119–3137, 2004.

[12] J. Benedict and D. Randall. Structure of the madden-julian oscillation in the super-parameterized cam. *Journal of The Atmospheric Sciences - J ATMOS SCI*, 66, 11 2009.

[13] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms, 2019.

[14] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer, 2018.

[15] T. Beucler and T. Cronin. A budget for the size of convective self-aggregation. *Quarterly Journal of the Royal Meteorological Society*, 145(720):947–966, 2019.

[16] T. Beucler, M. Pritchard, S. Rasp, P. Gentine, J. Ott, and P. Baldi. Enforcing analytic constraints in neural-networks emulating physical systems. *arXiv preprint arXiv:1909.00912*, 2019.

[17] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.*, 126:098302, Mar 2021.

[18] P. N. Blossey, C. S. Bretherton, A. Cheng, S. Endo, T. Heus, A. P. Lock, and J. J. van der Dussen. Cgils phase 2 les intercomparison of response of subtropical marine low cloud regimes to co2 quadrupling and a cmip3 composite forcing change. *Journal of Advances in Modeling Earth Systems*, 8(4):1714–1726, 2016.

[19] M. B. Blumenthal. Predictability of a coupled ocean–atmosphere model. *Journal of Climate*, 4(8):766 – 784, 01 Aug. 1991.

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

[21] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

[22] N. D. Brenowitz, T. Beucler, M. Pritchard, and C. S. Bretherton. Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, 77(12):4357 – 4375, 2020.

[23] N. D. Brenowitz and C. S. Bretherton. Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45(12):6289–6298, 2018.

[24] C. S. Bretherton and M. F. Khairoutdinov. Convective self-aggregation feedbacks in near-global cloud-resolving simulations of an aquaplanet. *Journal of Advances in Modeling Earth Systems*, 7(4):1765–1787, 2015.

[25] F. Brient and S. Bony. Interpretation of the positive low-cloud feedback predicted by a climate model under global warming. *Climate Dynamics*, 40, 05 2012.

[26] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung. Toward trustworthy ai development: Mechanisms for supporting verifiable claims, 2020.

[27] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. PMID: 26828106.

[28] G. Chen, Y. Ming, N. D. Singer, and J. Lu. Testing the clausius-clapeyron constraint on the aerosol-induced changes in mean and extreme precipitation. *Geophysical Research Letters*, 38(4), 2011.

[29] A. Cheng and K.-M. Xu. Improved low-cloud simulation from a multiscale modeling framework with a third-order turbulence closure in its cloud-resolving model component. *Journal of Geophysical Research*, 116, 07 2011.

[30] F. Chollet. *Deep Learning with Python*. Manning Publications Co., USA, 1st edition, 2017.

[31] C. Chou and J. D. Neelin. Mechanisms of global warming impacts on regional tropical precipitation. *Journal of Climate*, 17(13):2688 – 2701, 01 Jul. 2004.

[32] H. M. Christensen, I. M. Moroz, and T. N. Palmer. Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, 44(7):2195–2214, 2015.

[33] A. Clark, W. Gallus, and T.-C. Chen. Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Monthly Weather Review - MON WEATHER REV*, 135, 10 2007.

[34] M. Colin, S. Sherwood, O. Geoffroy, S. Bony, and D. Fuchs. Identifying the sources of convective memory in cloud-resolving simulations. *Journal of the Atmospheric Sciences*, 76, 12 2018.

[35] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965. URL: `http://cr.yp.to/bib/entries.html#1965/cooley`.

[36] D. Crommelin and W. Edeling. Resampling with neural networks for stochastic parameterization in multiscale systems, 04 2020.

[37] C. L. Daleu, R. S. Plant, S. J. Woolnough, S. Sessions, M. J. Herman, A. Sobel, S. Wang, D. Kim, A. Cheng, G. Bellon, P. Peyrille, F. Ferry, P. Siebesma, and L. van Ulft. Intercomparison of methods of coupling between convection and large-scale circulation: 1. comparison over uniform surface conditions. *Journal of Advances in Modeling Earth Systems*, 7(4):1576–1601, 2015.

[38] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[39] A. T. DeGaetano, G. Mooers, and T. Favata. Temporal changes in the areal coverage of daily extreme precipitation in the northeastern united states using high-resolution gridded data. *Journal of Applied Meteorology and Climatology*, 59(3):551 – 565, 2020.

[40] L. Denby. Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophysical Research Letters*, 47(1):e2019GL085190, 2020. e2019GL085190 10.1029/2019GL085190.

[41] T. Dror, M. D. Chekroun, O. Altaratz, and I. Koren. Deciphering organization of goes-16 green cumulus through the empirical orthogonal function (eof) lens. *Atmospheric Chemistry and Physics*, 21(16):12261–12272, 2021.

[42] T. Dror, I. Koren, O. Altaratz, and R. H. Heiblum. On the abundance and common properties of continental, organized shallow (green) clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):4570–4578, 2021.

[43] T. Dror, V. Silverman, O. Altaratz, M. D. Chekroun, and I. Koren. Uncovering the large-scale meteorology that drives continental, shallow, green cumulus through supervised classification. *Geophysical Research Letters*, 49(8):e2021GL096684, 2022. e2021GL096684 2021GL096684.

[44] J. Duchi. Lecture notes for statistics 311/electrical engineering 377. *Stanford*, 2:23, 2016.

[45] A. Edelman, A. Gedling, E. Konovalov, R. McComiskie, A. Penny, N. Roberts, S. Templeman, D. Trewin, and M. Ziembicki. *State of the Tropics - 2014 Report*. 06 2014.

[46] S. Eismann, S. Bartzsch, and S. Ermon. Shape optimization in laminar flow with a label-guided variational autoencoder. 12 2017.

[47] S. Emori and S. J. Brown. Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. *Geophysical Research Letters*, 32(17), 2005.

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

[49] D. J. Gagne II, H. M. Christensen, A. C. Subramanian, and A. H. Monahan. Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, 12(3):e2019MS001896, 2020. e2019MS001896 10.1029/2019MS001896.

[50] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis. Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45(11):5742–5751, 2018.

[51] V. P. Ghate, B. A. Albrecht, and P. Kollias. Vertical velocity structure of nonprecipitating continental boundary layer stratocumulus clouds. *Journal of Geophysical Research: Atmospheres*, 115(D13), 2010.

[52] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[54] W. W. Grabowski. Coupling cloud processes with the large-scale dynamics using the cloud-resolving convection parameterization (crcp). *J. Atmos. Sci.*, 58:978–997, 05 2001.

[55] R. Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

[56] J. L. Gustafson. *Moore's Law*, pages 1177–1184. Springer US, Boston, MA, 2011.

[57] R. J. Haarsma, M. J. Roberts, P. L. Vidale, C. A. Senior, A. Bellucci, Q. Bao, P. Chang, S. Corti, N. S. Fučkar, V. Guemas, J. von Hardenberg, W. Hazeleger, C. Kodama, T. Koenigk, L. R. Leung, J. Lu, J.-J. Luo, J. Mao, M. S. Mizielinski, R. Mizuta, P. Nobre, M. Satoh, E. Scoccimarro, T. Semmler, J. Small, and J.-S. von Storch. High resolution model intercomparison project (highresmip v1.0) for cmip6. *Geoscientific Model Development*, 9(11):4185–4208, 2016.

[58] P. T. Haertel and G. K. Kiladis. Dynamics of 2-day equatorial waves., 2004.

[59] Y. Han, G. J. Zhang, X. Huang, and Y. Wang. A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002076, 2020. e2020MS002076 2020MS002076.

[60] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[61] A. Henkes, G. Fisch, L. A. T. Machado, and J.-P. Chaboureau. Morning boundary layer conditions for shallow to deep convective cloud evolution during the dry season in the central amazon. *Atmospheric Chemistry and Physics*, 21(17):13207–13225, sep 2021.

[62] L. Hertel, J. Collado, P. Sadowski, J. Ott, and P. Baldi. Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, 2020. In press. Also: arXiv:2005.04048. Software available at: https://github.com/sherpa-ai/sherpa.

[63] S. Hettiarachchi, C. Wasko, and A. Sharma. Increase in flood risk resulting from climate change in a developed urban watershed - the role of storm temporal patterns. *Hydrology and Earth System Sciences*, 22(3):2041–2056, 2018.

[64] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[65] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan. Introvae: Introspective variational autoencoders for photographic image synthesis, 07 2018.

[66] M. Janiskov'a, J.-f. Mahfouf, J.-j. Morcrette, and F. Chevallier. Linearized radiation and cloud schemes in the ecmwf model: Development and evaluation. *Quarterly Journal of the Royal Meteorological Society*, 128(583):1505–1527, 2002.

[67] M. Janssens, J. Vilà-Guerau de Arellano, M. Scheffer, C. Antonissen, A. P. Siebesma, and F. Glassmeier. Cloud patterns in the trades have four interpretable dimensions. *Geophysical Research Letters*, 48(5):e2020GL091001, 2021. e2020GL091001 2020GL091001.

[68] F. Jansson, G. van¬†den¬†Oord, I. Pelupessy, J. H. Gr–nqvist, A. P. Siebesma, and D. Crommelin. Regional superparameterization in a global circulation model using large eddy simulations. *Journal of Advances in Modeling Earth Systems*, 11(9):2958–2979, 2019.

[69] E. J. Jensen, G. Diskin, R. P. Lawson, S. Lance, T. P. Bui, D. Hlavka, M. McGill, L. Pfister, O. B. Toon, and R. Gao. Ice nucleation and dehydration in the tropical tropopause layer. *Proceedings of the National Academy of Sciences*, 110(6):2041–2046, 2013.

[70] R. H. Johnson, T. M. Rickenbach, S. A. Rutledge, P. E. Ciesielski, and W. H. Schubert. Trimodal characteristics of tropical convection. *Journal of Climate*, 12(8):2397 – 2418, 1999.

[71] T. R. Jones, D. A. Randall, and M. D. Branson. Multiple-instance superparameterization: 1. concept, and predictability of precipitation. *Journal of Advances in Modeling Earth Systems*, 11(11):3497–3520, 2019.

[72] T. R. Jones, D. A. Randall, and M. D. Branson. Multiple-instance superparameterization: 2. the effects of stochastic convection on the simulated climate. *Journal of Advances in Modeling Earth Systems*, 11(11):3521–3544, 2019.

[73] F. Judt. Insights into atmospheric predictability through global convection-permitting model simulations. *Journal of the Atmospheric Sciences*, 75(5):1477 – 1497, 2018.

[74] H. Kalesse and P. Kollias. Climatology of high cloud dynamics using profiling arm doppler radar observations. *Journal of Climate*, 26(17):6340–6359, 2013.

[75] M. Khairoutdinov and D. Randall. Cloud resolving modeling of the arm summer 1997 iop: Model formulation, results, uncertainties, and sensitivities. *Journal of The Atmospheric Sciences - J ATMOS SCI*, 60:607–625, 02 2003.

[76] M. Khairoutdinov and D. Randall. High-resolution simulation of shallow-to-deep convection transition over land. *Journal of the Atmospheric Sciences*, 63(12):3421 – 3436, 2006.

[77] M. F. Khairoutdinov and Y. L. Kogan. A large eddy simulation model with explicit microphysics: Validation against aircraft observations of a stratocumulus-topped boundary layer. *Journal of the Atmospheric Sciences*, 56(13):2115 – 2131, 1999.

[78] B. Khouider and A. J. Majda. A simple multicloud parameterization for convectively coupled tropical waves. part i: Linear analysis. *Journal of the Atmospheric Sciences*, 63(4):1308 – 1323, 2006.

[79] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[80] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[81] O. Kisi, V. M. Krasnopolsky, M. S. Fox-Rabinovitz, and A. A. Belochitski. 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*, 2013:485913, 2013.

[82] D. Klocke, M. Brueck, C. Hohenegger, and B. Stevens. Rediscovery of the doldrums in storm-resolving simulations over the tropical atlantic. *Nature Geoscience*, 10(12):891–896, 2017.

[83] D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

[84] G. J. Kooperman, M. S. Pritchard, M. A. Burt, M. D. Branson, and D. A. Randall. Impacts of cloud superparameterization on projected daily rainfall intensity climate changes in multiple versions of the community earth system model. *Journal of Advances in Modeling Earth Systems*, 8(4):1727–1750, 2016.

[85] V. Krasnopolsky, M. Fox-Rabinovitz, Y. Hou, S. Lord, and A. Belochitski. Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: Climate simulations and seasonal predictions. *Monthly Weather Review - MON WEATHER REV*, 138:1822–1842, 05 2010.

[86] V. M. Krasnopolsky and M. S. Fox-Rabinovitz. A new synergetic paradigm in environmental numerical modeling: Hybrid models combining deterministic and machine learning components. *Ecological Modelling*, 191(1):5–18, 2006. Selected Papers from the Fourth International Workshop on Environmental Applications of Machine Learning, September 27 - October 1, 2004, Bled, Slovenia.

[87] V. M. Krasnopolsky, M. S. Fox-Rabinovitz, and D. V. Chalikov. New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, 133(5):1370 – 1383, 2005.

[88] V. M. Krasnopolsky, M. S. Fox-Rabinovitz, H. L. Tolman, and A. A. Belochitski. Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors. *Neural Networks*, 21(2):535 – 543, 2008. Advances in Neural Networks Research: IJCNN '07.

[89] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research).

[90] A. Kumar, T. Islam, Y. Sekimoto, C. Mattmann, and B. Wilson. Convcast: An embedded convolutional lstm based architecture for precipitation nowcasting using satellite data. *PLOS ONE*, 15(3):1–18, 03 2020.

[91] T. Kurihana, E. Moyer, R. Willett, D. Gilton, and I. Foster. Data-driven cloud clustering via a rotationally invariant autoencoder, 2021.

[92] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, A. Pritzel, S. Ravuri, T. Ewalds, F. Alet, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, J. Stott, O. Vinyals, S. Mohamed, and P. Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2022.

[93] K. Lamer and P. Kollias. Observations of fair-weather cumuli over land: Dynamical factors controlling cloud size and cover. *Geophysical Research Letters*, 42(20):8693–8701, 2015.

[94] A. Lauer, K. Hamilton, Y. Wang, V. T. J. Phillips, and R. Bennartz. The impact of global warming on marine boundary layer clouds over the eastern pacific - a regional model study. *Journal of Climate*, 23(21):5844 – 5863, 2010.

[95] G. Li and S.-P. Xie. Origins of tropical-wide sst biases in cmip multi-model ensembles. *Geophysical Research Letters*, 39(22), 2012.

[96] Z. Li, F. Niu, J. Fan, Y. Liu, D. Rosenfeld, and Y. Ding. Long-term impacts of aerosols on the vertical development of clouds and precipitation. *Nature Geoscience*, 4(12):888–894, 2011.

[97] Z. Li and P. A. O'Gorman. Response of vertical velocities in extratropical precipitation extremes to climate change. *Journal of Climate*, 33(16):7125–7139, 2020.

[98] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[99] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[100] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[101] A. J. Majda, B. Khouider, G. N. Kiladis, K. H. Straub, and M. G. Shefter. A model for convectively coupled tropical waves: Nonlinearity, rotation, and comparison with observations. *Journal of the Atmospheric Sciences*, 61(17):2188 – 2205, 2004.

[102] H. Mangipudi, G. Mooers, M. Pritchard, T. Beucler, and S. Mandt. Analyzing high-resolution clouds and convection using multi-channel vaes, 2021.

[103] B. MAPES, S. TULICH, T. NASUNO, and M. SATOH. Predictability aspects of global aqua-planet simulations with explicit convection. *Journal of the Meteorological Society of Japan. Ser. II*, 86A:175–185, 2008.

[104] B. E. Mapes. Gregarious tropical convection. *J. Atmos. Sci.*, 50:2026–2037, 1993.

[105] B. E. Mapes. Convective inhibition, subgrid-scale triggering energy, and stratiform instability in a toy tropical wave model. *Journal of the Atmospheric Sciences*, 57(10):1515 – 1535, 2000.

[106] H. Masunaga and C. D. Kummerow. Observations of tropical precipitating clouds ranging from shallow to deep convective systems. *Geophysical Research Letters*, 33(16), 2006.

[107] T. Matsuno. Quasi-geostrophic motions in the equatorial area. *J. Meteor. Soc. Japan*, 44:25–43, 1966.

[108] B. Medeiros, B. Stevens, and S. Bony. Using aquaplanets to understand the robust responses of comprehensive climate models to forcing. *Climate Dynamics*, 44(7):1957–1977, 2015.

[109] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2391–2400. JMLR.org, 2017.

[110] H. Morrison, M. van Lier-Walqui, A. M. Fridlind, W. W. Grabowski, J. Y. Harrington, C. Hoose, A. Korolev, M. R. Kumjian, J. A. Milbrandt, H. Pawlowska, D. J. Posselt, O. P. Prat, K. J. Reimel, S.-I. Shima, B. van Diedenhoven, and L. Xue. Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8):e2019MS001689, 2020. e2019MS001689 2019MS001689.

[111] C. Muller and Y. Takayabu. Response of precipitation extremes to warming: what have we learned from theory and idealized cloud-resolving simulations, and what remains to be learned? *Environmental Research Letters*, 15(3):035001, 2020.

[112] C. J. Muller, P. A. O'Gorman, and L. E. Back. Intensification of precipitation extremes with warming in a cloud-resolving model. *Journal of Climate*, 24(11):2784–2800, 2011.

[113] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[114] S. K. Müller, E. Manzini, M. Giorgetta, K. Sato, and T. Nasuno. Convectively generated gravity waves in high resolution models of tropical dynamics. *Journal of Advances in Modeling Earth Systems*, 10(10):2564–2588, 2018.

[115] N. Nasrabadi and R. King. Image coding using vector quantization: a review. *IEEE Transactions on Communications*, 36(8):957–971, 1988.

[116] R. B. Neale, A. Gettelman, S. Park, A. J. Conley, D. Kinnison, D. Marsh, A. K. Smith, F. Vitt, H. Morrison, P. Cameron-smith, W. D. Collins, M. J. Iacono, R. C. Easter, X. Liu, M. A. Taylor, C. chieh Chen, P. H. Lauritzen, D. L. Williamson, R. Garcia, J. francois Lamarque, M. Mills, S. Tilmes, S. J. Ghan, P. J. Rasch, and M. Meteorology. Description of the ncar community atmosphere model (cam 5.0), tech. note ncar/tn-486+str, natl. cent. for atmos. In *6of7 ZHAO ET AL.: AEROSOL FIE SIMULATED BY CAMS L08806*, pages 2009–038451, 2010.

[117] J. D. Neelin, C. Chou, and H. Su. Tropical drought regions in global warming and el nino teleconnections. *Geophysical Research Letters*, 30(24), 2003.

[118] J. D. Neelin, O. Peters, J. W.-B. Lin, K. Hales, and C. E. Holloway. Rethinking convective quasi-equilibrium: observational constraints for stochastic convective schemes in climate models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1875):2579–2602, 2008.

[119] J. Norris, G. Chen, and J. D. Neelin. Thermodynamic versus dynamic controls on extreme precipitation in a warming climate from the community earth system model large ensemble. *Journal of Climate*, 32(4):1025 – 1045, 2019.

[120] J. M. Nugent, S. M. Turbeville, C. S. Bretherton, P. N. Blossey, and T. P. Ackerman. Tropical cirrus in global storm-resolving models: 1. role of deep convection. *Earth and Space Science*, 9(2):e2021EA001965, 2022. e2021EA001965 2021EA001965.

[121] P. A. O'Gorman and J. G. Dwyer. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 2018.

[122] P. A. O'Gorman and T. Schneider. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proceedings of the National Academy of Sciences*, 106(35):14773–14777, 2009.

[123] J. Ott, M. Pritchard, N. Best, E. Linstead, M. Curcic, and P. Baldi. A fortran-keras deep learning bridge for scientific computing. *Scientific Programming*, 2020. Article ID 8888811. https://doi.org/10.1155/2020/8888811. Also: arXiv:2005.04048.

[124] P. A. O'Gorman. Precipitation extremes under climate change. *Current climate change reports*, 1(2):49–59, 2015.

[125] P. A. O'gorman and T. Schneider. Scaling of precipitation extremes over a wide range of climates simulated with an idealized gcm. *Journal of Climate*, 22(21):5676–5685, 2009.

[126] T. N. Palmer. A personal perspective on modelling the climate system. *Proc Math Phys Eng Sci*, 472(2188):20150772, Apr 2016.

[127] T. N. Palmer. Stochastic weather and climate models. *Nature Reviews Physics*, 1(7):463–471, 2019.

[128] H. Parishani, M. S. Pritchard, C. S. Bretherton, C. R. Terai, M. C. Wyant, M. Khairoutdinov, and B. Singh. Insensitivity of the cloud response to surface warming under radical changes to boundary layer turbulence and cloud microphysics: Results from the ultra-parameterized cam. *Journal of Advances in Modeling Earth Systems*, 10(12):3139–3158, 2018.

[129] H. Parishani, M. S. Pritchard, C. S. Bretherton, M. C. Wyant, and M. Khairoutdinov. Toward low-cloud-permitting cloud superparameterization with explicit boundary layer turbulence. *Journal of Advances in Modeling Earth Systems*, 9(3):1542–1571, 2017.

[130] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space, Nov. 1901.

[131] M. Peel, B. Finlayson, and T. Mcmahon. Updated world map of the koppen-geiger climate classification. *Hydrology and Earth System Sciences Discussions*, 4, 10 2007.

[132] A. Pendergrass and D. Hartmann. Two modes of change of the distribution of rain*. *Journal of Climate*, 27:8357–8371, 11 2014.

[133] A. G. Pendergrass and D. L. Hartmann. Changes in the distribution of rain frequency and intensity in response to global warming. *Journal of Climate*, 27(22):8372–8383, 2014.

[134] K. Peters, T. Crueger, C. Jakob, and B. Möbis. Improved mjo-simulation in echam6.3 by coupling a stochastic multicloud model to the convection scheme. *Journal of Advances in Modeling Earth Systems*, 9(1):193–219, 2017.

[135] M. E. Peters and C. S. Bretherton. Structure of tropical variability from a vertical mode perspective. *Theoretical and Computational Fluid Dynamics*, 20(5):501–524, 2006.

[136] S. Pfahl, P. A. O'Gorman, and E. M. Fischer. Understanding the regional pattern of projected future changes in extreme precipitation. *Nature Climate Change*, 7(6):423–427, 2017.

[137] M. S. Pritchard, C. S. Bretherton, and C. A. DeMott. Restricting 32,Äì128 km horizontal scales hardly affects the mjo in the superparameterized community atmosphere model v.3.0 but the number of cloud-resolving grid columns constrains vertical mixing. *Journal of Advances in Modeling Earth Systems*, 6(3):723–739, 2014.

[138] S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[139] D. Randall, M. Khairoutdinov, A. Arakawa, and W. Grabowski. Breaking the Cloud Parameterization Deadlock. *Bulletin of the American Meteorological Society*, 84(11):1547–1564, 11 2003.

[140] S. Rasp, M. S. Pritchard, and P. Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018.

[141] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[142] D. M. Romps. An analytical model for tropical relative humidity. *Journal of Climate*, 27(19):7432 – 7449, 2014.

[143] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[144] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[145] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges, 2021.

[146] T. Schneider, J. Teixeira, C. S. Bretherton, F. Brient, K. G. Pressel, C. Schär, and A. P. Siebesma. Climate goals and computing the future of clouds. *Nature Climate Change*, 7(1):3–5, 2017.

[147] M. Schonlau. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *The Stata Journal*, 2(4):391–402, 2002.

[148] C. Schumacher, R. A. Houze, and I. Kraucunas. The tropical dynamical response to latent heating estimates derived from the trmm precipitation radar. *Journal of the Atmospheric Sciences*, 61(12):1341 – 1358, 2004.

[149] S. C. Sherwood, W. Ingram, Y. Tsushima, M. Satoh, M. Roberts, P. L. Vidale, and P. A. O'Gorman. Relative humidity changes in a warmer climate. *Journal of Geophysical Research: Atmospheres*, 115(D9), 2010.

[150] A. Siebesma, P. Soares, and J. Teixeira. A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of The Atmospheric Sciences - J ATMOS SCI*, 64, 04 2007.

[151] M. Sonnewald, C. Wunsch, and P. Heimbach. Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5):784–794, 2019.

[152] B. Stevens, M. Satoh, L. Auger, J. Biercamp, C. S. Bretherton, X. Chen, P. Düben, F. Judt, M. Khairoutdinov, D. Klocke, C. Kodama, L. Kornblueh, S.-J. Lin, P. Neumann, W. M. Putman, N. Röber, R. Shibuya, B. Vanniere, P. L. Vidale, N. Wedi, and L. Zhou. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1):61, 2019.

[153] P. Stinis, T. Hagge, A. Tartakovsky, and E. Yeung. Enforcing constraints for interpolation and extrapolation in generative adversarial networks. *Journal of Computational Physics*, 03 2018.

[154] Y. Tian, Y. Zhang, S. A. Klein, and C. Schumacher. Interpreting the diurnal cycle of clouds and precipitation in the arm goamazon observations: Shallow to deep convection transition. *Journal of Geophysical Research: Atmospheres*, 126(5):e2020JD033766, 2021. e2020JD033766 2020JD033766.

[155] X.-A. Tibau Alberdi, C. Requena-Mesa, C. Reimers, J. Denzler, V. Eyring, M. Reichstein, and J. Runge. Supernovae : Vae based kernel pca for analysis of spatio-temporal earth data. 01 2018.

[156] N. Tilinina, M. Krinitskiy, Y. Zyulyaeva, and S. Gulev. Clustering of the Polar Vortex states using deep convolutional neural networks. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, page 17539, Apr. 2019.

[157] B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002, 2020. e2019MS002002 10.1029/2019MS002002.

[158] B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002, 2020. e2019MS002002 10.1029/2019MS002002.

[159] S. N. Tulich, D. A. Randall, and B. E. Mapes. Vertical-mode and cloud decomposition of large-scale convectively coupled gravity waves in a two-dimensional cloud-resolving model. *Journal of the Atmospheric Sciences*, 64(4):1210 – 1229, 2007.

[160] P. Wang, J. Yuval, and P. A. O'Gorman. Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS002984, 2022. e2022MS002984 2022MS002984.

[161] Z. Wang and A. Bovik. Bovik, a.c.: Mean squared error: love it or leave it? - a new look at signal fidelity measures. ieee sig. process. mag. 26, 98-117. *Signal Processing Magazine, IEEE*, 26:98 – 117, 02 2009.

[162] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[163] O. Watt-Meyer, N. D. Brenowitz, S. K. Clark, B. Henn, A. Kwa, J. J. McGibbon, W. A. Perkins, and C. S. Bretherton. Correcting weather and climate models by machine learning nudged historical simulations. *Earth and Space Science Open Archive*, page 13, 2021.

[164] I. G. Watterson and M. R. Dix. Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution. *Journal of Geophysical Research: Atmospheres*, 108(D13), 2003.

[165] D. S. Wilks. *Statistical methods in the atmospheric sciences*. Elsevier, 2006.

[166] J.-L. Wu, K. Kashinath, A. Albert, D. Chirila, Prabhat, and H. Xiao. Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems. *Journal of Computational Physics*, 406:109209, 2020.

[167] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. On the quantitative analysis of decoder-based generative models. 11 2016.

[168] Y. Xue, M. Cane, S. Zebiak, and M. Blumenthal. On the prediction of enso: a study with a low-order markov model. *Tellus A: Dynamic Meteorology and Oceanography*, 46(4):512–528, 1994.

[169] G. Yacalis. *Artificial Neural Network Impact on Cloud Parameterization and Land-Atmosphere Interactions*. PhD thesis, 2018. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2020-11-24.

[170] L. Yang, D. Zhang, and G. E. Karniadakis. Physics-informed generative adversarial networks for stochastic differential equations. *SIAM J. Scientific Computing*, 42:A292–A317, 2020.

[171] Y. Yang, S. Mandt, and L. Theis. An introduction to neural data compression, 2022.

[172] Z. Yang and H. Xiao. Enforcing deterministic constraints on generative adversarial networks for emulating physical systems, 11 2019.

[173] M.-H. Yen, D.-W. Liu, Y.-C. Hsin, C.-E. Lin, and C.-C. Chen. Application of the deep learning for the prediction of rainfall in southern taiwan. *Scientific Reports*, 9(1):12774, 2019.

[174] J. Yin and A. Porporato. Diurnal cloud cycle biases in climate models. *Nature Communications*, 8(1):2269, 2017.

[175] J. Yuval and P. A. O'Gorman. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, 11(1):3295, 2020.

[176] J. Yuval and P. A. O'Gorman. Use of machine learning to improve simulations of climate, 2020.

[177] M. D. Zelinka, S. A. Klein, and D. L. Hartmann. Computing and partitioning cloud feedbacks using cloud property histograms. part ii: Attribution to changes in cloud amount, altitude, and optical depth. *Journal of Climate*, 25(11):3736 – 3754, 2012.

[178] Y. Zhang and S. A. Klein. Factors controlling the vertical extent of fair-weather shallow cumulus clouds over land: Investigation of diurnal-cycle observations collected at the arm southern great plains site. *Journal of the Atmospheric Sciences*, 70(4):1297 – 1315, 2013.

[179] Y. Zhang, S. A. Klein, J. Fan, A. S. Chandra, P. Kollias, S. Xie, and S. Tang. Large-eddy simulation of shallow cumulus over land: A composite case based on arm long-term observations at its southern great plains site. *Journal of the Atmospheric Sciences*, 74(10):3229 – 3251, 2017.

[180] P. Zhu and B. Albrecht. Large eddy simulations of continental shallow cumulus convection. *Journal of Geophysical Research: Atmospheres*, 108(D15), 2003.

[181] Y. Zhuang, R. Fu, J. A. Marengo, and H. Wang. Seasonal variation of shallow-to-deep convection transition and its link to the environmental conditions over the central amazon. *Journal of Geophysical Research: Atmospheres*, 122(5):2649–2666, 2017.