Environmental Data Science (2024), 3: e3, 1–10
doi:10.1017/eds.2024.1

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Understanding precipitation changes through unsupervised machine learning

Griffin Mooers[1], Tom Beucler[2,3], Mike Pritchard[1,4] and Stephan Mandt[5]

[1]Earth System Science, University of California, Irvine, CA, USA
[2]Faculty of Geosciences and Environment, University of Lausanne, Lausanne, Switzerland
[3]Expertise Center for Climate Extremes, University of Lausanne, Lausanne, Switzerland
[4]NVIDIA Research, Santa Clara, CA, USA
[5]Department of Computer Science, University of California, Irvine, CA, USA
Corresponding author: Griffin Mooers; Email: gmooers96@gmail.com

Received: 02 March 2023; **Revised:** 24 November 2023; **Accepted:** 09 January 2024

Keywords: atmospheric dynamics; climate change; heavy precipitation; unsupervised learning; variational autoencoders

Abstract

Despite the importance of quantifying how the spatial patterns of heavy 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, spatial machine-learning framework to quantify how storm dynamics affect changes in heavy precipitation. We find that changes in heavy precipitation (above the 80th percentile) are predominantly explained by changes in the frequency of these events, rather than by changes in how these storm regimes produce precipitation. Our study 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.

Impact Statement

Heavy precipitation causes severe flooding and water storage issues across the globe. A better understanding of how the heavy precipitation will change with warming is a priority. However, the details of the heavy precipitation are controlled by storm-scale atmospheric dynamics which are normally neglected or averaged out. We present a data-driven approach using machine learning to leverage the spatial details of dynamics controlling heavy precipitation. This allows us to determine the degree to which changes in convection type are responsible for shifts in heavy precipitation as the climate warms. With most machine learning focusing on prediction problems, our study highlights the impact of unsupervised machine learning on knowledge discovery in climate science.

1. Introduction: understanding the changing spatial patterns of heavy precipitation

According to the latest Intergovernmental Panel on Climate Change report (Edelman et al., 2014), "there is high confidence that heavier 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

This research article was awarded an Open Materials badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

associated increases in flood-related risk (Hettiarachchi et al., 2018) and challenges in water management (Adger et al., 2007, 2009). To first order, bounds of heavy precipitation 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 (O'Gorman and Schneider, 2009). This is referred to as the "thermodynamic contribution" to heavy precipitation changes (Emori and Brown, 2005) and gives a solid theoretical foundation for *spatially averaged* changes in heavy precipitation.

Yet climate change adaptation requires knowledge of how heavy precipitation will change at the *local* scale, that is, understanding the *changing spatial patterns* of heavy precipitation under warming. Focusing on the tropics, where most of the vulnerable world population lives (Edelman et al., 2014), 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 heavy precipitation changes (Emori and Brown, 2005).

A comprehensive understanding of this "dynamic contribution" remains elusive. Approximate scalings can be derived based on quasi-geostrophic dynamics (O'Gorman, 2015; Li and O'Gorman, 2020) and convective storm dynamics (Muller et al., 2011; Abbott et al., 2020). As well, some of the effects from dynamics can be approximated using observational and reanalysis data (Tan et al., 2015). However, actionable findings require Earth-like simulations of the present and future climates (e.g., Pendergrass and Hartmann, 2014), 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; Khairoutdinov and Kogan, 1999; Khairoutdinov and Randall, 2003) simulations we use in this study (Section 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 1 for examples).

2. Theory: heavy precipitation decomposition

In this section, we outline our strategy to facilitate the analysis of heavy 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. Our analysis transforms these convection types into "cluster probabilities." For the total N data points being assigned to a set of K clusters, we first compute a vector of counts N_k, indicating the number of data points assigned to each cluster index by k: $\sum_{k=1}^{N} N_k \equiv N$. We now convert these numbers to a normalized probability vector as $\pi \equiv (\pi_1, \ldots, \pi_k) \equiv (N_1/N, \ldots, N_k/N)$.

To calculate these π_i's, we use variational autoencoders in conjunction with k-means clustering, also called vector quantization (Lloyd, 1982; Gray, 1984). Details on the exact coarse-graining procedure will be presented in Section 3.2; for now, we take the N different convection types and their probabilities as given.

This unsupervised quantization of regimes of convection through ML allows us to quantitatively understand changes in heavy precipitation (P_{heavy}) from both changes in convection regime characteristics and probability. Here we define P_{heavy} 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 ΔP_{heavy} as its absolute change from the "Control" to the "Warmed" climate, and show below that relative changes in heavy precipitation can be decomposed using changes in π (equation (7)).

We derive a decomposition of heavy precipitation changes for global warming by making a series of simple physical assumptions about precipitation. While these approximations facilitate our results' interpretability and lend physical meaning to each term, they need not hold to derive equation (5),

[1] Five kilometers or less horizontal grid spacing.

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

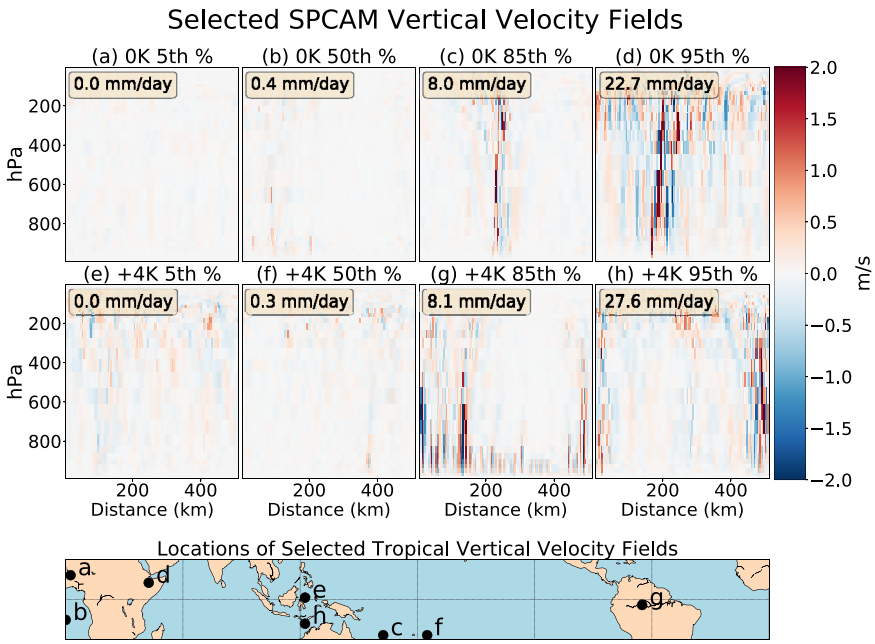

Figure 1. *Selected vertical velocity fields from our "Control" (0K, a–d) and "Warmed" (+4K, e–h) SPCAM simulations. By sampling the precipitation distribution, we show instances of vertical velocity fields associated with no precipitation (a,e), drizzle (b,f), heavy rainfall (c,g), and intense storms (d,h).*

the final form of the decomposition as it could also be derived by decomposing the heavy precipitation field into its spatial average and an anomaly, before further decomposing the anomaly using the objectively identified dynamical regimes. To first order, precipitation (P) scales like condensation rate (C), which depends on the full vertical velocity (w) and atmospheric water vapor (here quantified using specific humidity q) fields:

$$P \approx C(w, q). \tag{1}$$

Note that equation (1) neglects the dependence on microphysical processes (see e.g., Muller and Takayabu, 2020) to focus on the thermodynamical and dynamical components of precipitation. When focusing on heavier 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 (1) in the case of high quantiles of P:

$$P_{\text{heavy}} \approx P_{\text{heavy}}(w, q_{\text{sat}}). \tag{2}$$

We now make the assumption that the thermodynamic dependence on q_{sat} can be factored out of the right-hand side of equation (2) and denote the dynamical pre-factor as $\mathscr{D}(w)$:

$$P_{\text{heavy}} \approx q_{\text{sat}} \times \mathscr{D}(w). \tag{3}$$

The previous assumption has been historically justified by assuming a moist adiabatic temperature profile and scaling the vertical velocity by using a single value for the entire vertical profile for intense events (O'Gorman and Schneider, 2009; Muller et al., 2011). However, it is more accurate to assume vertical velocity profiles collapse when changing the vertical coordinate from pressure to the normalized integral of the moisture lapse rate (Abbott et al., 2020).

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

We can now linearly decompose the dynamical pre-factor $\mathscr{D}(w)$ into the N regimes identified by our unsupervised learning framework:

$$\mathscr{D}(w,P) \approx \mathscr{D}_0 + \sum_{i=1}^{N} \mathscr{D}_i \cdot \pi_i, \tag{4}$$

where π_i is the normalized probability of each dynamical regime determined by the vertical velocity field information. Combining equations (3) and (4), and taking a logarithmic derivative with respect to climate change allows us to decompose relative changes in upper precipitation quantiles as follows:

$$\frac{\Delta P_{\text{heavy}}}{P_{\text{heavy}}} \approx \frac{\Delta q_{\text{sat}}}{q_{\text{sat}}} + \frac{\Delta\left(\mathscr{D}_0 + \sum_{i=1}^{N} \pi_i \mathscr{D}_i\right)}{\mathscr{D}_0 + \sum_{i=1}^{N} \pi_i \mathscr{D}_i}, \tag{5}$$

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

$$q_{\text{sat}} = q_{\text{sat}}(T_s, p_s) \Rightarrow \frac{\Delta q_{\text{sat}}}{q_{\text{sat}}} \approx 7\%, \tag{6}$$

where T_s is near-surface temperature and p_s near-surface pressure. Expanding equation (5) and substituting $\mathscr{D}(w)$ using equation (3) yields the desired decomposition of heavy precipitation changes in climate:

$$\frac{\Delta P_{\text{heavy}}}{P_{\text{heavy}}} = \overbrace{\underbrace{\frac{\Delta q_{\text{sat}}}{q_{\text{sat}}}}_{\text{From theory}}}^{\text{Thermodynamic}} + \overbrace{\underbrace{\frac{q_{\text{sat}}}{P_{\text{heavy}}}}_{\text{From current climate}} \left(\Delta \mathscr{D}_0 + \underbrace{\sum_{i=1}^{N} \Delta \pi_i \mathscr{D}_i}_{\text{Regime prob.shifts}} + \underbrace{\sum_{i=1}^{N} \pi_i \Delta \mathscr{D}_i}_{\text{Intra-regime changes}} \right)}^{\text{Dynamic}}. \tag{7}$$

Equation (7) shows that relative changes in P_{heavy} are the sum of a well-understood, spatially uniform "thermodynamic" increase in saturation specific humidity (q_{sat}—see equation (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 probability or "cluster sizes"—covered in more detail in Section 3), and N changes in regime characteristics $\Delta\mathscr{D}_i$ (changes in the "dynamic contribution" pre-factors, in precipitation units).

Our simulation data already contain P_{heavy} 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 $\mathscr{D}_i$ and their changes with warming. Using equation (4), we linearly regress $\frac{P_{\text{heavy}}}{q_{\text{sat}}}$ on the regime probabilities π_i in both the reference and warm climates to derive the pre-factors $\mathscr{D}_i$, 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 heavy precipitation. Understanding these changes is critical to trust local climate change predictions.

3. Machine learning approach

We will now discuss the data, models, and statistical techniques used in this paper. Additional details can be found in the Supplementary Material.

3.1. Data: high-resolution, earth-like simulations of global surface warming

The multi-scale modeling framework (Randall et al., 2003) used to generate our training and test data is composed of small, locally periodic 2D subdomains of explicit high-resolution physics that are embedded

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

within each grid column of a coarser resolution ($1.9^\circ \times 2.5^\circ$ degree) host planetary circulation model (Khairoutdinov and Kogan, 1999). 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 extract snapshots every four hours. 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 1.

3.2. VAE training

Our ML methodology identifies dynamical regimes on the basis of two million two-dimensional vertical velocity fields. We first nonlinearly reduce the dimensionality of the data using a fully convolutional VAE design, whose architecture is depicted in Figure 2. To train the VAE, we perform beta-annealing (Bowman et al., 2016; Higgins et al., 2017), expanding the Evidence Lower Bound (ELBO) traditionally used to train the VAE by including a β parameter and linearly anneal β from 0 to 1 over 1600 training epochs. The number of layers and channels in the encoder and decoder are depicted in Figure 2 (4 layers in each, stride of 2). 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.

We refer the reader to SI B for the details of the VAE's benchmarking and performance evaluation and proceed to use that VAE to investigate how convective regimes respond to climate change.

3.3. Quantization procedure

The goal of the following analysis is to analyze upper quantile precipitation patterns arising from climate shifts, leveraging vertical velocity fields. The sheer amount of data requires additional statistical 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 (Nasrabadi and King, 1988). 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 (Yang et al., 2022). 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 Mooers et al. (2022). We cluster convection into three distinct regimes: (1) Marine Shallow Convection, (2) Continental Shallow Cumulus Convection, and (3) Deep Convection. We have found that higher counts of N created repetitive regimes of Deep Convection while lower counts do not properly separate out these three distinct species of convection (Mooers et al., 2022).

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

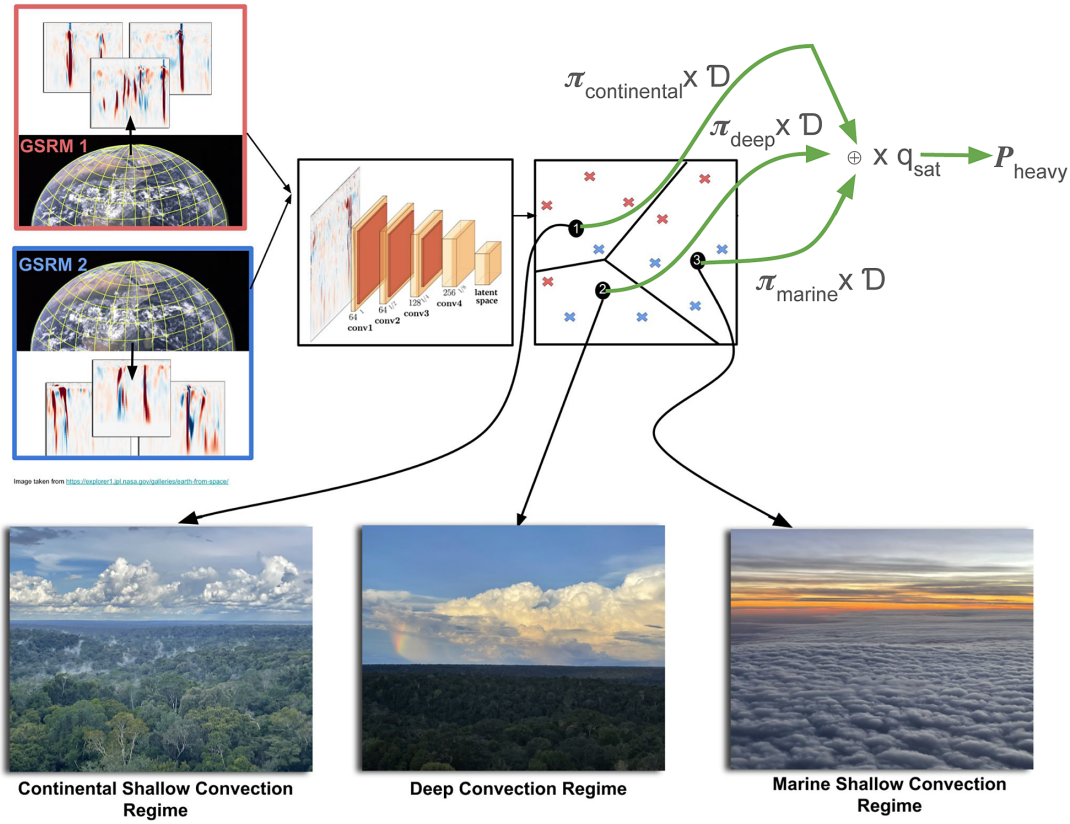

Figure 2. *The VAE-based approach to understand changes in heavy precipitation, P_{heavy}. Given a vertical velocity field w, the VAE nonlinearly reduces the input dimension to yield a latent representation z. We can cluster the simulated data sets (control and warmed) through their latent representations into three recognizable regimes of convection (continental shallow, deep, and marine shallow). The corresponding cluster assignment probabilities, π's allow us to also quantify the dynamical contribution, D to heavy precipitation. Photos taken by Griffin Mooers.*

We are especially interested in *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. This yields the present cluster assignment probabilities 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 probabilities π^{4K}. The difference vector of assignment probabilities, 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 heavy precipitation below.

4. Results

4.1. Unsupervised machine learning reveals convective responses to climate change

Figure 3 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 3c). More specifically, deep convection becomes more probable over warm ocean waters and especially the Pacific Warm Pool (Allan et al., 2014) while shallow convection becomes less common in these unstable regions (Figure 3a). This result is consistent

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

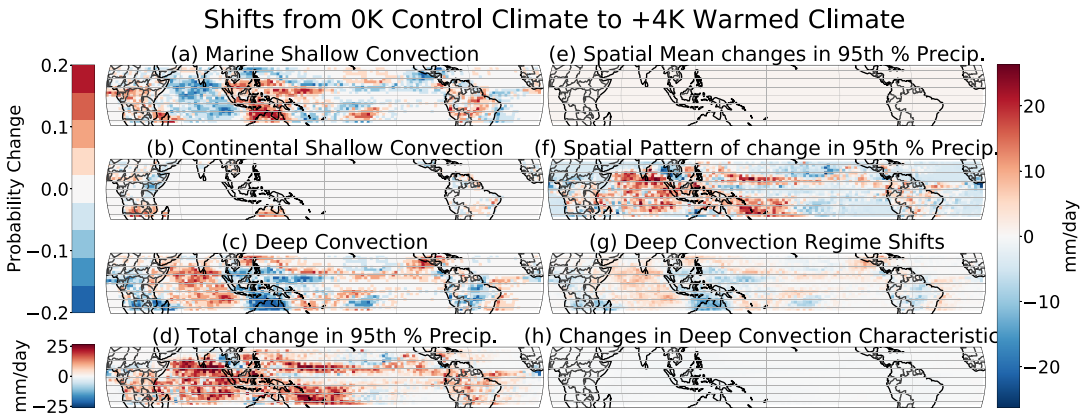

Figure 3. *Changes induced by $+4^{\circ}C$ of simulated global warming. Panels (a–c) display normalized probability shifts ($\Delta\pi$'s) in the three dynamical regimes found through clustering with $N = 3$, corresponding to (a) "marine shallow," (b) "continental shallow cumulus," and (c) "deep" convection. The difference in the 95th percentile in precipitation between the control and warmed simulations is shown in (d). We subtract the spatial-mean change (e, the "thermodynamics" from equation (6)) from the total change (d) to yield the "dynamic" contribution (f). Using equation (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 Supplementary Figure S5. The patterns of storms change (a–c), which changes the patterns of heavy precipitation (f), mostly because deep convective storms shift location (g).*

with observational trends showing an intensification of already powerful storms over the warm tropical waters (Allan et al., 2014). At first glance, the pattern of this unsupervised deep convection shift with warming ($\Delta\pi_1$) looks quite similar to shifts in upper precipitation quantiles (Figure 3c vs. f).

With just the information from the regime probabilities, we can model the spatial patterns of precipitation changes at upper percentiles (Supplementary Figure S4). Our model becomes less accurate at lower precipitation quantiles, partly because we are not using specific humidity information (the approximation of equation (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 heavy precipitation, which should be further tested in diverse climate change modeling frameworks.

4.2. Decomposing the dynamic contribution to heavy precipitation changes

We now isolate the dynamical contributions to dynamical changes in heavy precipitation by decomposing the spatial patterns (3d) into changes in regime probability π_i and changes in regime characteristics $\mathscr{D}_i$. 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 3, so we must now calculate changes in how each regime produces precipitation, which involves the following two steps. First, we empirically estimate $\mathscr{D}_i$ by using the probabilities of deep and shallow convection.[2] Second, we estimate changes in "deep" and "shallow" convection dynamical pre-factors as $\Delta\mathscr{D}_i = \mathscr{D}_i^{4K} - \mathscr{D}_i^{0K}$.

[2] More specifically, we estimate the dynamical pre-factors (D_i) by regressing $P_{\text{heavy}}/q_{\text{sat}}$ 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 (Dror et al., 2022).

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

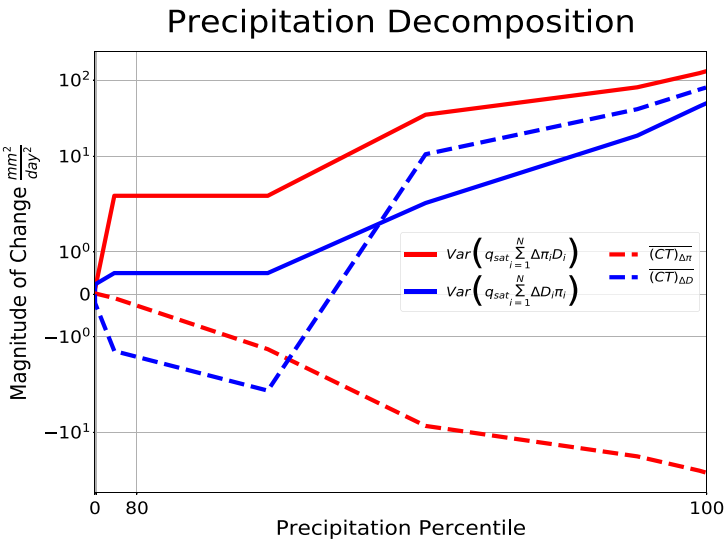

Figure 4. *Based on equation (8), we decompose the total magnitude of the change in heavy precipitation. In our decomposition, we compare the mean of the spatial anomaly of convective probability shifts ($\Delta\pi$) to the changes in the dynamical prefactors (ΔD). We find that the convective regime shifts are of greater importance to explain the changes in heavy precipitation (80th–99.99th percentiles).*

We now have the requisite information to understand the drivers of heavy precipitation changes themselves. We ask: *Did the patterns of heavy 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 heavy precipitation ΔP_{heavy} can be explained by changes in convection probability ($\Delta\pi$), and how much of it can be explained by changes in the dynamical prefactors ($\Delta\mathscr{D}$). This comparison relies on the following decomposition of heavy precipitation variance $\text{var}\left(\Delta P_{\text{heavy}}\right)$, derived in Sec D of the SI:

$$\text{var}\left(\Delta P_{\text{heavy}}\right) = \underbrace{\text{var}\left[\left(q_{\text{sat}}\sum_{i=1}^{N}\Delta\pi_i\mathscr{D}_i\right)\right] + \overline{(\text{CT})}_{\Delta\pi}}_{\text{Shift in regime location}} + \underbrace{\text{var}\left[\left(q_{\text{sat}}\sum_{i=1}^{N}\pi_i\Delta\mathscr{D}_i\right)\right] + \overline{(\text{CT})}_{\Delta\mathscr{D}}}_{\text{Intra-regime changes}} + \mathscr{R}, \quad (8)$$

where CT are cross-terms and R groups all terms of the decomposition that are not needed to compare differences in precipitation from regime shifts versus intra-regime changes. To understand what is most crucial to precipitation changes, we depict the spatial mean of equation (8)'s terms in Figure 4.

For high precipitation quantiles where our model works best (especially above the 80th percentile), changes in the upper quantiles are dominated by regime probability shifts rather than by intra-regime changes (Figure 4): The "$\Delta\mathscr{D}$ term" is smaller than the "$\Delta\pi$ term" (red line vs. blue line).[3] This result aligns well with our qualitative analysis in Figure 3 showing the similarity between the spatial patterns of deep convection regime changes and heavy precipitation changes. The spatial patterns of heavy 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.

We are also confident in the robustness of these results based on the degree of similarity in the findings of the importance of deep convection shifts in both our work and Tan et al. (2015). The fact that in satellite and reanalysis data probability shifts in convection type are also the drivers of changes in extreme

[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 Supplementary Figure S3.

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press

precipitation is a reassuring measure of support for our findings as well as for the performance of the GSRM representation of the global high-resolution vertical velocity structure more broadly. At the same time, it suggests that machine learning approaches like the one we deploy with minimal domain knowledge are just as faithful to the physics of the atmosphere as more traditional convection analysis.

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 heavy precipitation. This highlights the importance of leveraging the full spatial extent of this information (traditionally averaged out) to derive accurate regional and local changes. Although our findings were achieved from unsupervised learning, we are reassured that similar findings can be derived in observational satellite data from regime-based analysis of deep convection as well (Tan et al., 2015). 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 intense 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) (Eyring et al., 2016; Haarsma et al., 2016) and determine its ability to explain spatial variations in heavy precipitation with climate change.

Acknowledgments. The authors thank Paul O'Gorman, Harshini Mangipudi, Pierre Gentine, Veronika Eyring, Prakhar Srivastava, and Liran Peng for useful conversations that helped advance this work.

Author contribution. Conceptualization: G.M., T.B., S.M., M.P.; Data curation and visualizations: G.M., T.B.; Methodology: G.M., T.B.; Writing manuscript: G.M., T.B., S.M., M.P.

Competing interest. The authors have no conflicts of interest to declare.

Data availability statement. The codebase for running the "SPCAM5" simulations is the same employed by Parishani et al. (2017), 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. Training and postprocessing scripts, as well as saved model weights and Python environments, are available on GitHub at 10.5281/zenodo.7693052.

Ethics statement. The research meets all ethical guidelines.

Funding statement. The authors thank the Department of Energy under grant DE-SC0022331, the National Science Foundation (NSF) Machine Learning and Physical Sciences (MAPS) program and NSF grant 1633631, Office of Advanced Cyberinfrastructure grant OAC-1835863, Division of Atmospheric and Geospace Sciences grant AGS-1912134, Division of Information and Intelligent Systems grants IIS-2047418, IIS-2003237, and 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 cofunding 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. The authors 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 Nos. 675191 and 823988. This work was also supported by gifts from Disney and Qualcomm.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/eds.2024.1.

References

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

Adger W, **Agrawala S**, **Mirza M**, **Conde C**, **O'Brien K**, **Pulhin J**, **Pulwarty R**, **Smit B and Takahashi K** (2007) Assessment of adaptation practices, options, constraints and capacity. Climate change 2007: Impacts, adaptation and vulnerability. In

Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, pp. 717–743.

Adger WN, **Dessai S**, **Goulden M**, **Hulme M**, **Lorenzoni I**, **Nelson DR**, **Naess LO**, **Wolf J and Wreford A** (2009) Are there social limits to adaptation to climate change? *Climatic Change 93*(3), 335–354.

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

Bowman SR, **Vilnis L**, **Vinyals O**, **Dai AM**, **Józefowicz R and Bengio S** (2016) *Generating sentences from a continuous space. CoNLL.*

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

Edelman A, **Gedling A**, **Konovalov E**, **McComiskie R**, **Penny A**, **Roberts N**, **Templeman S**, **Trewin D and Ziembicki M** (2014) *State of the tropics – 2014 report.*

Emori S and Brown S (2005) Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. *Geophysical Research Letters 32*(17), L17706.

Eyring V, **Bony S**, **Meehl GA**, **Senior CA**, **Stevens B**, **Stouffer RJ and Taylor KE** (2016) Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development 9*(5), 1937–1958.

Gray R (1984) Vector quantization. *IEEE ASSP Magazine 1*(2), 4–29.

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

Hettiarachchi S, **Wasko C and Sharma A** (2018) 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.

Higgins I, **Matthey L**, **Pal A**, **Burgess C**, **Glorot X**, **Botvinick MM**, **Mohamed S and Beta-Vae Lerchner A** (2017) Learning basic visual concepts with a constrained variational framework. *ICLR*.

Khairoutdinov MF and Kogan YL (1999) 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.

Khairoutdinov M and Randall D (2003) Cloud resolving modeling of the arm summer 1997 iop: Model formulation, results, uncertainties, and sensitivities. *Journal of The Atmospheric Sciences 60*, 607–625.

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

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

Mooers G, **Pritchard M**, **Beucler T**, **Srivastava P**, **Mangipudi H**, **Peng L**, **Gentine P and Mandt S** (2022) Comparing storm resolving models and climates via unsupervised machine learning. *arXiv preprint arXiv:2208.11843*.

Muller CJ, **O'Gorman PA and Back LE** (2011) Intensification of precipitation extremes with warming in a cloud-resolving model. *Journal of Climate 24*(11), 2784–2800.

Muller C and Takayabu Y (2020) 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.

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

O'Gorman PA (2015) Precipitation extremes under climate change. *Current Climate Change Reports 1*(2), 49–59.

O'Gorman PA and Schneider T (2009) Scaling of precipitation extremes over a wide range of climates simulated with an idealized gcm. *Journal of Climate 22*(21), 5676–5685.

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

Pendergrass AG and Hartmann DL (2014) Changes in the distribution of rain frequency and intensity in response to global warming. *Journal of Climate 27*(22), 8372–8383.

Randall D, **Khairoutdinov M**, **Arakawa A and Grabowski W** (2003) Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society 84*(11), 1547–1564.

Tan J, **Jakob C**, **Rossow WB and Tselioudis G** (2015) Increases in tropical rainfall driven by changes in frequency of organized deep convection. *Nature 519*(7544), 451–454.

Yang Y, **Mandt S and Theis L** (2022) An introduction to neural data compression. *arXiv preprint arXiv:2202.06533*.

Cite this article: Mooers G, Beucler T, Pritchard M and Mandt S (2024). Understanding precipitation changes through unsupervised machine learning. *Environmental Data Science*, 3: e3. doi:10.1017/eds.2024.1

https://doi.org/10.1017/eds.2024.1 Published online by Cambridge University Press