scientific reports

Check for updates

OPEN

Comparing storm resolving models and climates via unsupervised machine learning

Griffin Mooers[1]✉, Mike Pritchard[1,2], Tom Beucler[3], Prakhar Srivastava[4], Harshini Mangipudi[4], Liran Peng[1], Pierre Gentine[5] & Stephan Mandt[4]

Global storm-resolving models (GSRMs) have gained widespread interest because of the unprecedented detail with which they resolve the global climate. However, it remains difficult to quantify objective differences in how GSRMs resolve complex atmospheric formations. This lack of comprehensive tools for comparing model similarities is a problem in many disparate fields that involve simulation tools for complex data. To address this challenge we develop methods to estimate distributional distances based on both nonlinear dimensionality reduction and vector quantization. Our approach automatically learns physically meaningful notions of similarity from low-dimensional latent data representations that the different models produce. This enables an intercomparison of nine GSRMs based on their high-dimensional simulation data (2D vertical velocity snapshots) and reveals that only six are similar in their representation of atmospheric dynamics. Furthermore, we uncover signatures of the convective response to global warming in a fully unsupervised way. Our study provides a path toward evaluating future high-resolution simulation data more objectively.

The Earth's atmosphere is a complex system, with many different factors influencing its dynamics on scales ranging from microns to thousands of kilometers. Thanks to modern high-resolution global Earth system models, much of this complexity can now be captured with unprecedented accuracy, down to the "storm-resolving" scale of several kilometers[1–4]. By explicitly resolving fundamental nonlinear and high-resolution processes like deep convection (precipitating clouds) formation, these models can address longstanding issues with cloud and precipitation patterns in conventional climate simulations[5–10]. However, despite these advances, there remain substantial differences in how these models are designed, which contribute to uncertainty in their weather and climate predictions[4]. While attempts have been made to validate and compare ensembles of these models, this has traditionally been done using coarsened statistics, such as annual averages, guided by physically informed approaches. A community goal is to directly compare models at the scale of storm formation, which could improve understanding of the consequences of different design decisions and help narrow the uncertainty of cloud-climate feedback[4,11–13].

One of the biggest challenges with understanding those simulations' output is the massive amount of high-resolution data produced. This can quickly become overwhelming, as seen in the first inter-comparison study of Global Storm-Resolving Models (GSRMs), the DYAMOND project[4]. For just 40 days of hourly simulation output, nearly two petabytes per GSRM were generated. This means that storing the data is a significant hurdle and analyzing it is even more challenging. To get around these barriers to understanding those simulations' results simpler dimensionality reduction methods such as clustering and projections are traditionally used. However, these methods may not fully capture the non-linear relationships embedded in small-scale physical processes, which are what make these simulations so valuable[14–16].

To gain more insight and confidence in these climate predictions, we need objective ways to quantify changes in convective organization, identify models that are outliers, and more comprehensively analyze modern GSRMs[4,17]. As intercomparisons of multiple GSRMs across multiple climates were not available at the time of this work, this paper proposes a novel kind of comparison: we compare models based on their high-resolution simulation data of the *present* climate. In machine learning terminology, we quantify differences between GSRMs

[1]Department of Earth System Science, University of California at Irvine, Irvine, CA 92697, USA. [2]NVIDIA, Santa Clara, CA 95050, USA. [3]Institute of Earth Surface Dynamics, University of Lausanne, 1015 Lausanne, Switzerland. [4]Department of Computer Science, University of California at Irvine, Irvine, CA 92617, USA. [5]Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA. ✉email: gmooers96@gmail.com

based on the notion of *distribution shifts* across different simulated data sets. This approach enables a fully data-driven approach towards model inter- and intra-comparisons.

Our contributions are threefold. (1) We introduce novel methods and metrics utilizing unsupervised machine learning techniques, specifically variational autoencoders (VAEs) and vector quantization, to systematically analyze and compare high-resolution climate models. This approach compliments traditional physically informed analysis allowing for a detailed inter-comparison of nine diverse GSRMs informed by the small-scale convective organization unique to these detailed simulations. (2) Our analysis uncovers inconsistencies in the representations of tropical convection among GSRMs, highlighting the need for further investigation into parameterization choices. (3) Our study provides insights into the impact of climate change on high-resolution simulations. In a fully data-driven fashion, we identify distinct signatures of global warming, including the expansion and intensification of arid, dry zones over the continents and the concentration of deep convection over warm waters.

Data: storm-resolving models and preprocessing

This paper examines high-resolution atmospheric model data (5 kilometers or less horizontally) provided by the DYAMOND project[4]. To simplify modeling, we focus our new unsupervised method on the vertical velocity variable, giving us information about updraft and gravity wave dynamics across different scales and phenomena. Specifically, we consider eight different DYAMOND GSRMs: the Icosahedral Nonhydrostatic Weather and Climate Model (ICON), the Integrated Forecasting System (IFS), the Nonhydrostatic ICosahedral Atmospheric Model (NICAM), the Unified Model (UM), the System for High-resolution modeling for Earth-to-Local Domains (SHIELD), the Global Environmental Multiscale Model (GEM), the System for Atmospheric Modeling (SAM), and the Action de Recherche Petite Echelle Grande Echelle (ARPEGE). In addition, we include SPCAM, a Multi-Model Framework (MMF) that embeds many miniature 2D GSRMs in a host global climate model[18,19].

We extract two-dimensional image-like snapshots of the original 3D vertical velocity data (pressure/altitude vs. longitude), which are taken every three hours. We use 285,000 randomly selected samples from each model (160,000 for training, 125,000 for testing), spanning the 15 S–15 N latitude belt and representing diverse tropical convective regimes. The GSRMs' varying horizontal and vertical resolutions and other sub-grid parameterization choices are detailed in Tables 1 and 2 of[4]. Figure 2 and Movie S1 provide example data. These selected datasets provide us with a comprehensive testbed of vertical velocity imagery.

Besides comparing different GSRMs on the *present* climate, we also consider data produced by a single model, but for different simulated climates. Here, we use SPCAM to simulate global warming by increasing sea surface temperatures by four Kelvin. We treat this as a proxy for climate change, where we consider spatial and intensity shifts between convective updrafts in two simulated climates. The use of the SPCAM model is a pragmatic choice which facilitates exploration of climate change emulation, due to its computational efficiency compared to GSRMs[20] that allows sampling of multiple climates, and the known characteristics of its climate change behavior[10]. The use of the SPCAM model is essential for climate change emulation as at this point no climate change simulations exist from DYAMOND[4].

Unsupervised model intercomparison

Our approach is based on variational autoencoders (VAEs)[21], a deep learning approach to dimensionality reduction and density estimation. (For more details, see "Methods".) VAEs are probabilistic autoencoders that use neural networks to *embed* data in a low-dimensional "latent" bottleneck representation termed the "latent space". From there, the VAE attempts to reconstruct the original data with minimal information loss. At the same time, VAEs impose a regularization on the latent space that encourages the latent representation to have a simple structure so that the latent representation can be used to discover patterns in high-dimensional data. The tradeoff between both tasks is a manifestation of the rate-distortion tradeoff from information theory[22] and forms the basis for deciding on an architecture.

In order to facilitate the discovery of hidden structure in the latent space, we additionally cluster the embedded data using k-means clustering. In machine learning terminology, such an approach is also called vector quantization (see "Methods" for details), in particular if the number of clusters is large. We find that VAEs are essential to our dimensionality reduction task. Directly attempting the clustering in the raw data space does not result in stable and reproducible clusters. Likewise, a simpler dimensionality reduction technique such as PCA also fails to create robust results (Fig. 8). Furthermore, we find that the VAE-based clusters are interpretable and correspond to different convective and geographical phenomena, which will be discussed next. Finally, we show that working with a large number of clusters gives rise to natural similarity metrics across GSRMs (Fig. 3). See the Supplementary Information for more details.

Latent space inquiry uncovers differences among storm-resolving models

As follows, we will provide evidence that the learned low-dimensional representations are semantically meaningful and can be well-described using only three learned latent clusters that correspond to distinct convective organizations.

Cluster characterization

As a first qualitative analysis, we can learn a shared clustering across the dimensionality-reduced data of all nine GSRMs (Fig. 3). Since the latent space is 1000-dimensional, we plot the dominant two (or three) principal components for visualization purposes. Each data point is colorized according to its cluster *assignment*, i.e., its nearest cluster, where each cluster has a unique color. We find that the VAE organizes convection in the way an atmospheric scientist might[23,24]. By analyzing each cluster in the latent space's vertical velocity kinetic energy $\sqrt{\overline{w'w'}}$ profiles (which can be thought of as a measure of the variance in vertical velocity at each vertical level of

the atmosphere), we find a clear distinction between top-heavy (*deep*) and bottom-heavy (*shallow*) convection types. Furthermore, plotting the proportion of each of the three clusters for every spatial coordinate separately reveals a distinction of one cluster dominating over land, and two over oceans. We thus find that the three dominant clusters represent *marine shallow convection* (blue), *deep convection* (red), and *continental shallow convection* (green) (Fig. 4).

Qualitative model intercomparison

Inspecting the dimensionality-reduced data along with the learned latent clustering and spatial visualization (Fig. 4) gives unique qualitative insights into commonalities and differences across GSRMs. While most GSRMs share similar distributions in the latent space, Fig. 3 reveals that the SPCAM and SAM models show systematic differences compared to the other ones (Fig. 3 g, j vs. all). SAM reveals a differently-shaped *deep convection* cluster (Fig. 3j, red regime). SPCAM shows an unusual *deep convection* cluster adjacent to the *marine shallow* (blue) mode. A closer inspection of the $\sqrt{\overline{w'w'}}$ profile shows a unique regime of continental convection with a short horizontal scale of variability for SPCAM, particularly near the surface of the Earth (Fig. S9b, red line vs. all). For SAM, the $\sqrt{\overline{w'w'}}$ profile of *deep* convection is much more intense than that of other GSRMs, especially in the upper atmosphere (Fig. S9b; blue line). These differences in intensity statistics and vertical structure help explain the unusually wide extent of the *deep convection* cluster on the latent space projection (Fig. 3j, red cluster vs. all).

A further inspection of the GSRMs' relative cluster proportions (Fig. S10) confirms this perspective. SPCAM and SAM differ significantly from the other models (Fig. S10, second and third rows vs. bottom six). These two divergent GSRMs contain high proportions of stronger convection types, consistent with our previous analysis (Fig. 3 and Figs. S5–7, S9). For ICON, we find similarly pronounced differences in cluster proportions, showing a higher proportion of strong convection types (*continental shallow* and *deep*). While these were primarily qualitative findings, we will quantify distributional differences across GSRMs next.

Dynamic consistency between high-resolution climate models

In our analysis, we delve into a comprehensive inter-comparison of various GSRMs on a *distributional level*, aiming to uncover both commonalities and disparities across their entire simulated datasets. The idea behind the following approach is to consider model dissimilarities or distances as *distribution shifts*. In the machine learning literature[25], such shifts occur in various contexts (e.g., changing lighting conditions in videos, medical data from different hospitals, etc.) and are usually associated with a degradation of the trained classifier. In contrast, we consider an *unsupervised* version of distribution shift assessment and use it to assess similarities between simulation data sets.

ELBO scores

To initiate this comparison, we turn our attention to the VAE's training objective, the Evidence Lower Bound (ELBO) (Eq. 3). As detailed in "Methods", this metric serves as a reflection of the model's likelihood estimate for each observation, indicating the probability of a particular sample's occurrence. Examining the probability density function (PDF) of ELBO scores offers a distinct and unique fingerprint for each GSRM. The ELBO also aids in measuring disparities between different data distributions, making it a pivotal tool in our analysis. Utilizing a common encoder model, we visualize the PDF of each GSRM test dataset, providing valuable insights into the intricacies of their respective data distributions.

Figure 5a shows nine resulting PDFs, where the red lines corresponding to ICON, SPCAM, and SAM have different distributions than the (blue lines denoting the) other six GSRMs. Specifically, the ELBO PDFs of ICON, SPCAM, and SAM are more right-skewed and less symmetric, confirming our earlier findings of a "majority" group involving most GSRMs, and a "minority"/"outlier" group involving ICON, SPCAM, and SAM.

Assessing GSRM distances using vector quantization

In order to further quantify the distribution shifts between different GSRMs, we revisit our non-linear dimensionality reduction and clustering technique from before. But crucially, for a more quantitative comparison, we partition the latent space into a large number of regions, essentially through k-means clustering with a large ($K = 50$) number of clusters. As before, we then attribute each data by their nearest cluster centroid. This technique is called *vector quantization* and is commonly used in the context of data compression[27,28]. This discrete representation has the advantage of making certain computations tractable. In particular, it allows computing statistical distance measures between (discrete) data distributions, such as the symmetrized Kullback–Leibler (KL) divergence. See "Methods" for technical details. Using this approach, we present a matrix of pairwise similarities among the nine GSRMs (Fig. 5b–g).

Figure 5g shows the results of the analysis, where a dark red indicates a high distance between models. We make two observations: firstly, three GSRMs (SAM, SPCAM, and ICON) exhibit a significant dissimilarity with respect to each other and with the rest of the models. Secondly, a group of "similar" models (GEM, UM, NICAM, IFS, SHIELD, ARPEGE) shows a relatively high degree of mutual similarity. It is worth noting that Fig. 5b shows similar results; here we use a lower but physically interpretable cluster count (K = 3).

Our results obtained from vector quantization align well with our earlier investigations in "Latent space inquiry uncovers differences among storm-resolving models". In both approaches ("Latent space inquiry uncovers differences among storm-resolving models", "Dynamic consistency between high-resolution climate models"), we found a split between six similar GSRMs and three divergent GSRMs. Specifically, our analysis revealed that ICON had a lower proportion of shallow convection compared to other GSRMs, SAM contained unusually intense "Deep Convection", and SPCAM exhibited small scale turbulence with distinct profiles of $\sqrt{\overline{w'w'}}$ showing unusual updraft intensity near the earth's surface not seen in other GSRMs.

Though we have put much of the focus on using our framework to identify unique GSRMs and hone in on the causes behind these inter-GSRM differences, the apparent similarity among the GEM, UM, NICAM, IFS, SHIELD, and ARPEGE models is another key finding of our approach. This conformity mirrors what we found by inspecting the latent representations (Figs. 3, S5–7), the vertical structure of the leading three convection regimes (Fig. S9), and the proportion of each type of convection in the simulation (Fig. S10). It would be worth elucidating the degree to which the similarity between these GSRMs is a reflection of DYAMOND GSRMs better representing observational reality than coarser GCMs or an artifact of the inter-dependence of climate-models occluding the interpretation of a multi-model ensemble[29], but this question is outside the scope of our present work. Instead, we will move on from inter-GSRM comparisons in the same climate state to a comparison of different climate states.

VAEs extract planetary patterns of convective responses to global warming

The assessment of the distribution shift is a powerful tool for comparing different climate models, but also for investigating the impact of global warming on atmospheric convection. In this section, we apply our approach to the SPCAM model, which provides simulation data for two different global temperature levels: present-day conditions and a scenario with +4 K of sea surface temperature warming. Besides predicting changes to the vertical velocity profiles, we can also identify geographic regions that are most affected by climate change.

In order to investigate the geographic effects of global warming on convection and specific regions where convection undergoes the most significant changes, we build on the methods described in "Latent space inquiry uncovers differences among storm-resolving models" by first learning global convection clusters and initializing three cluster centers ($K = 3$) for physical interpretability. We then stratify the SPCAM data by their latitude/longitude gridcell and calculate location-specific *cluster proportions* based on the fixed cluster centers. These proportions (π_1, π_2, π_3) with $\pi_1 + \pi_2 + \pi_3 = 1$ indicate the fraction of the data being assigned to each cluster $K \in \{1, 2, 3\}$; see "Methods" for details. We can now visualize the geographic distribution of these cluster proportions and identify the dominant convection types in each region (see Fig. S11).

When we examine the latent space of SPCAM, we again three distinct regimes of convection. The first mode corresponds to deep convection over the Pacific Warm pool, almost identical to the other GSRMs. A second mode of shallow convection dominates over areas where air is descending, both over continents and the oceans. In contrast to the other GSRMs, which treat continental connection as a single regime, we have identified a third unique mode that we call "Green Cumulus," which is exclusively found over specific sub-regions of semi-arid tropical land areas (see Fig. S11a).

Changing probabilities of convective modes in response to global warming

We again use technical notation to measure the shift in convection patterns between the control and warmed climates. We first encode our dataset into a latent space and cluster the encoded data using K-means. The fraction of data assigned to each cluster represents the prevalence of each convective regime in the dataset. We can use these "cluster assignment" vectors to identify the spatial pattern of each type of convection across the tropics. By comparing these normalized probabilities between the control and warmed climates, we can objectively quantify the change in the atmosphere's structure with warming, which we refer to as a *distribution shift*. Specifically, let $(\pi_1^{0K}, \pi_2^{0K}, \pi_3^{0K})$ denote the cluster proprotions at present temperatures, and $(\pi_1^{+4K}, \pi_2^{+4K}, \pi_3^{+4K})$ the corresponding quantities in a climate globally warmed by four Kelvin. Then, the probability shifts $\Delta\pi_k = (\pi_k^{+4K} - \pi_k^{0K})$ for $K \in \{1, 2, 3\}$ reveal the effects of climate change on convection patterns.

The most prominent signal of climate change that our analysis captures are the shifts in deep and shallow convection across different geographic regions. Figure 6a shows that shallow convection is increasing over areas of subsidence, while Fig. 6b shows a corresponding decrease in deep convection over these less active oceanic regions. Simultaneously, Fig. 6b depicts an expected increase in the proportion and intensity of deep convection over warm ocean waters and particularly the Pacific Warm pool[30], with shallow convection becoming less prevalent in these unstable areas. Finally, as shown in Fig. 6c, the rare "Green Cumulus" mode becomes more common over semi-arid land masses, consistent with the overall intensification and expansion of arid zones (dry get drier mechanism)[31,32].

We find evidence of the vertical shift in the structure of each convective regime as temperatures warm, as shown in Fig. 6d. The upper-tropospheric maximum in $\sqrt{\overline{w'w'}}$ shifts upwards with warming. This finding is consistent with the expected tropopause vertical expansion induced by climate change[33,34]. Additionally, a reduction in mid-tropospheric $\sqrt{\overline{w'w'}}$ can be explained by the decrease in vertical transport of mass in the atmosphere due to the enhanced saturation vapor pressure in a warmer world[35,36]. The decrease in lower-tropospheric $\sqrt{\overline{w'w'}}$, indicated by the blue lines, corresponds to a decrease in marine shallow convection intensity, which we believe is evidence of marine boundary layer shoaling[37]. Finally, beyond the median $\sqrt{\overline{w'w'}}$ statistics, we see an increase in the upper percentiles of deep convection (Fig. S12b), revealing an intensification of already powerful storms over warm waters, consistent with observational trends[30].

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 rare "Green Cumulus" convection

Finally, we hone in on the unique ways in which "Green Cumulus" Convection changes with a warming climate as inferred from our unsupervised framework. Within SPCAM, this sub-group of continental convection corresponds to a rare form of convection that was first identified by[38]. We choose to formally adopt the unique label of "Green Cumulus" here due to the near total overlap between the geographic domain of this subsection

of continental convection in SPCAM and the regions of the highest proportion of "Green Cumulus" convection identified in satellite imagery (Figure 6a in[38]). Both our results and[38] identify this convection primarily over semi-arid continents (Fig. S11a). Despite its existing identification in literature, is not traditionally included in the analysis of tropical convection[23,24,39]. 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[40]. However, the clustering of the latent space of SPCAM immediately separates "Green Cumulus" out into its own unique mode distinct from the rest of the continental convection.

By geographically conditioning the latent space cluster associated with "Green Cumuli" we can not only confirm the regional patterns of the mode, 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[41]) (Fig. S13). This unique atmospheric state at locations of this convective mode, combined with its very distinct $\sqrt{\overline{w'w'}}$ profile (Green lines in Fig. 6d), suggests it does in fact deserve to be separated out from other types of convection despite its scarcity.

Although other studies have made note of this convective form[42–44], our distribution shift analysis shows that "Green Cumuli" expand as global temperatures rise (Fig. 6c). We observe that both the proportion and geographic localizations of "Green Cumulus" increase in a hotter atmosphere—this is likely aided by expected dry-zone expansions[31,32]. Comparison of these "Green Cumuli" $\sqrt{\overline{w'w'}}$ cluster profiles between the control and warmed climates also shows a substantial increase in the associated boundary layer turbulence (Fig. S12c). 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 rare-event "Green Cumuli" and capturing its climate change signals, synthesizing dynamic analysis and allowing new discovery.

Discussion

We introduced new methods and metrics to compare high-resolution climate models (global storm-resolving models—GSRMs) based on their very large output data by using unsupervised machine learning. Systemically comparing models and providing an understanding of the effect of climate change in such high-fidelity high-resolution simulations has been challenged by their enormous dataset sizes and has limited progress. Our new unsupervised approach relied on a combination of non-linear dimensionality reduction using variational autoencoders (VAEs) and vector quantization for an unsupervised inter-comparison of these storm resolving models. Beyond inter-model comparisons, we also compared global climates at different temperatures and developed new insights into the changes in convection regimes.

Our data-driven method provides a complementary viewpoint to physics-based climate model comparisons, potentially less susceptible to human biases. 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.

Our distributional comparisons identify consistency in only six of the nine considered storm resolving models. 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 GSRM simulation outputs. Note we cannot rule out the possibility that one of the divergent GSRMs may still be reflecting observational reality better than the majority group. We leave this comparison to observations for future work.

Our work suggests the need to further investigate the parameterization choices in these high-resolution simulations. In the DYAMOND initiative, ICON was configured at an unusually high resolution (grid-cell dimension of 2 km) so that typical sub-grid orography and convection parameterizations were deactivated[45]. In the design of both SPCAM and SAM, there are approximations required for the anelastic formulations of buoyancy[46,47]. 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 GSRMs. However, further investigation is needed to confirm the true root causes of the differences between GSRMs we have identified.

When comparing different climates, convolutional variational autoencoders identify two distinct signatures of global warming: (1) an expansion and (at the atmosphere's boundary layer) an intensification of "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.

The present study has focused on vertical velocity fields in high-resolution climate models as one of the most challenging data to analyze. Improved performance could be obtained by jointly modeling multiple "channels" (i.e. variables) of spatially-resolved data such as temperature and humidity. While we have performed preliminary analysis of these results here[48], we leave more detailed conclusions for further studies. Our study could also be extended to alternative data sets, such as the High Resolution Model Inter-comparison Project (HighResMIP)[49,50] and observational satellite data sets. Besides variational autoencoders, future studies could also focus on other methods such as hierarchical variants, normalizing flows, or diffusion probabilistic models. Ultimately, we hope that our work will motivate future data-driven and/or unsupervised investigations in the broader scientific fields where Big Data challenges conventional analysis approaches.

Methods

A broad overview of our approach can be seen in Fig. 1, with more details discussed below.

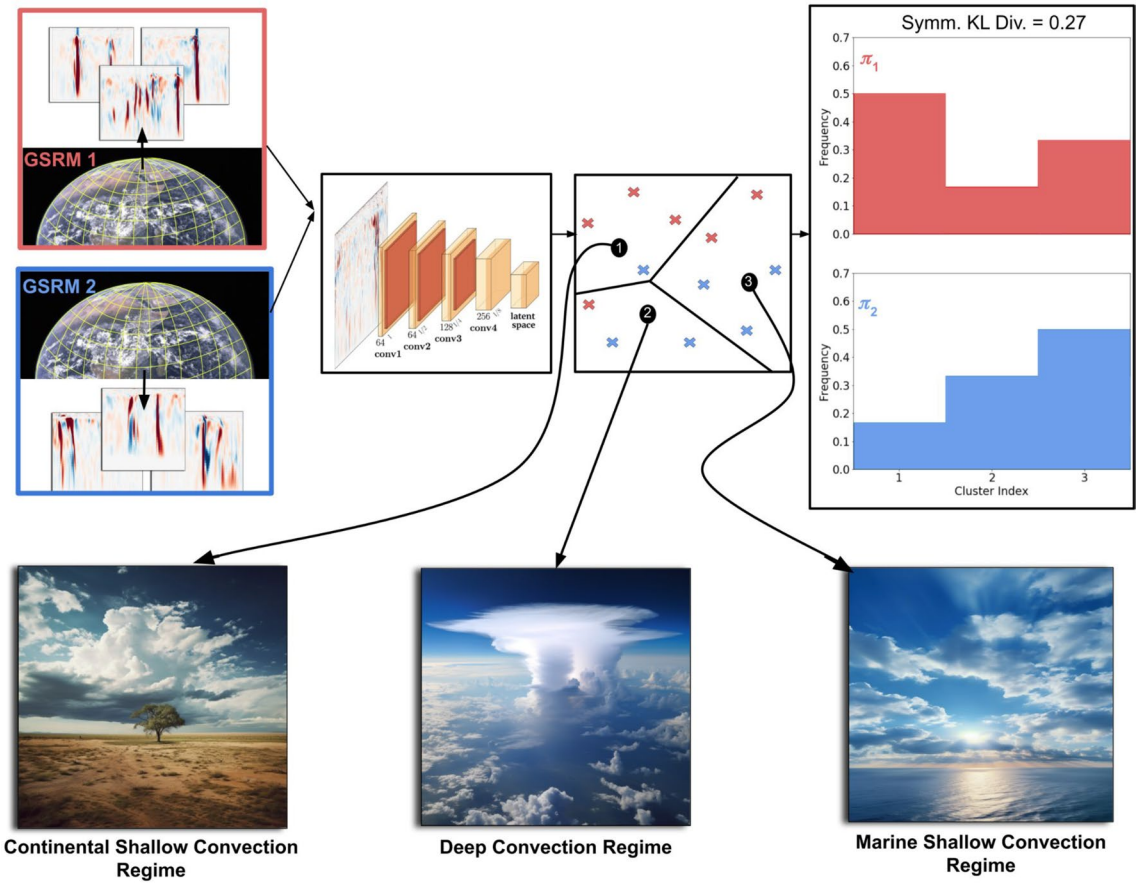

Figure 1. An overview of our machine learning based approach. We extract 2D vertical velocity fields from GSRM's across the tropics. We use a variational autoencoder to reduce these high dimensional vertical velocity fields to low dimensional latent representations for analysis. Clustering of these latent representations reveals three unique regimes of tropical convection. We can compare the "Distance" between these GSRMs by looking at the symmetrized KL divergence between the normalized PDF of convection type probabilities. The image of earth's surface was taken from https://explorer1.jpl.nasa.gov/galleries/earth-from-space/.

Simulation data and preprocessing

We examine the data on vertical velocity generated by high-resolution km-scale global storm resolving models (GSRMs) from the DYAMOND archive, and a multi-scale modeling framework (MMF)[51,52]. GSRMs are numerical simulators that provide uniform high-resolution simulations of the entire atmosphere. On the other hand, MMFs are a specialized type of coarse-resolution global climate model that incorporate small, periodic 2D subdomains of local storm resolving dynamics (LSRMs)[5,53]. In our study, we utilize the Super Parameterized Community Atmosphere Model (SPCAM) v5 as our MMF. It is consistent with the code base of REF[33] but configured at a coarser exterior resolution, consisting of 13,824 local 2D (vertical level—longitude cross sections) GSRMs, with each spanning 512 km and composed of 128 grid columns spaced 4 km apart. Since we are only using the DYAMOND II GSRMs data covering the boreal winter (though future work could include the DYAMOND III GSRM data when it is publically released as the next phase could cover the entire year), we generate six separate realizations of boreal winter for the MMF by introducing perturbed initial conditions to gather more data points. Although there is DYAMOND I data modeling the boreal summer, it is not with the exact same set of models and many models in common between DYAMOND I and II were configured differently making a synthesis of data across DYAMOND data generations challenging[54].

To preprocess the input, we follow these steps: we convert the 4D vertical velocity data from the DYAMOND GSRMs into 2D input samples of horizontal width and vertical level. To do this, we extract the 2D instantaneous subsets that are aligned in the pressure-longitude plane. This allows for a direct comparison with the MMF, which uses 2D LRSMs aligned in the same way. We restrict our data sampling to the tropical latitudes between 15° S and 15° N during boreal winter. This results in a dataset of 160,000 training sample images that is large enough to capture the diverse spatial-temporal patterns of tropical weather, turbulence, and cloud regimes.

We normalize the input values by scaling each pixel's original velocity value in meters per second (m/s) to a normalized range between 0 and 1. We do this consistently across all samples using the range measured across the entire dataset. To ensure uniform structure across all samples, we interpolate the input images onto a standardized vertical (pressure) and horizontal grid. This is necessary to account for differences in the GSRMs' respective grid structures when performing pairwise comparisons.

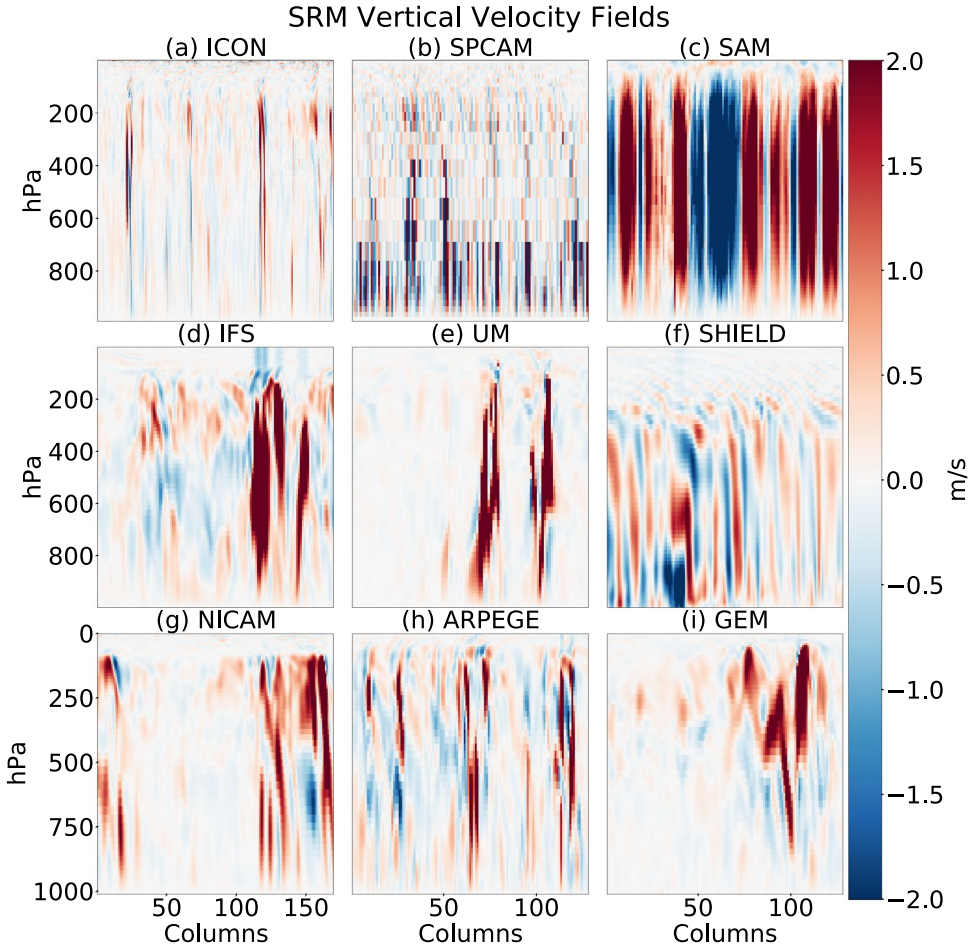

Figure 2. A selected vertical velocity field from each of the nine GSRMs used in this intercomparison. Atmospheric pressure is denoted on the y axis and the number of embedded columns in a given snapshot is shown on the x axis. We see a rich mix of turbulent updrafts (red) of various scales and types. For more examples, see Movie S1.

Figure 2 provides vertical velocity snapshots for various models used in this paper. For more examples, see Movie 1.

Understanding convection via vertical structure

To analyze the dominant vertical structure of convection, we calculate the horizontal variance of vertical velocity within each image. For this, we compute the horizontal mean $\overline{w_i}$ separately at each vertical level (Note this is done over a 2D field at each grid cell, not globally, so $\overline{w_i}$ is not equal to 0), and then subtract it to create the layerwise anomaly $w' = w - \overline{w}$ at a given vertical level. Then the final measure of the variance we are interested in is calculated by

$$\sqrt{\overline{\mathrm{w}'\mathrm{w}'}} \overset{\mathrm{def}}{=} \sqrt{(\mathrm{w} - \overline{\mathrm{w}})^2}, \tag{1}$$

The resulting 1D second-moment vector is widely analyzed in the study of atmospheric turbulence as it helps characterize the altitudes of most vigorous convection[55]. We average it across a cluster to estimate the convective structures present and use it as one metric to discriminate the average physical properties sorted by the VAE latent space in Figs. 4, 6, S3, S8, S9, S12.

The horizontal extent of convection

To distinguish narrow from wide convective structures, it is necessary to separate convective updrafts based on their width. To elucidate these differences, we measure the Turbulent Length Scale (TLS)[56], which is a way to derive the horizontal breadth of the updrafts. We calculate the TLS at each vertical level and then combine the TLS across all layers to get a composite value for the vertical velocity field. We then calculate the power spectrum of the weighted average length of all samples, using φ to represent the power spectra, $||k||$ as the complex modulus, n as the number of dimensions, and $\langle\rangle$ as the vertical integral:

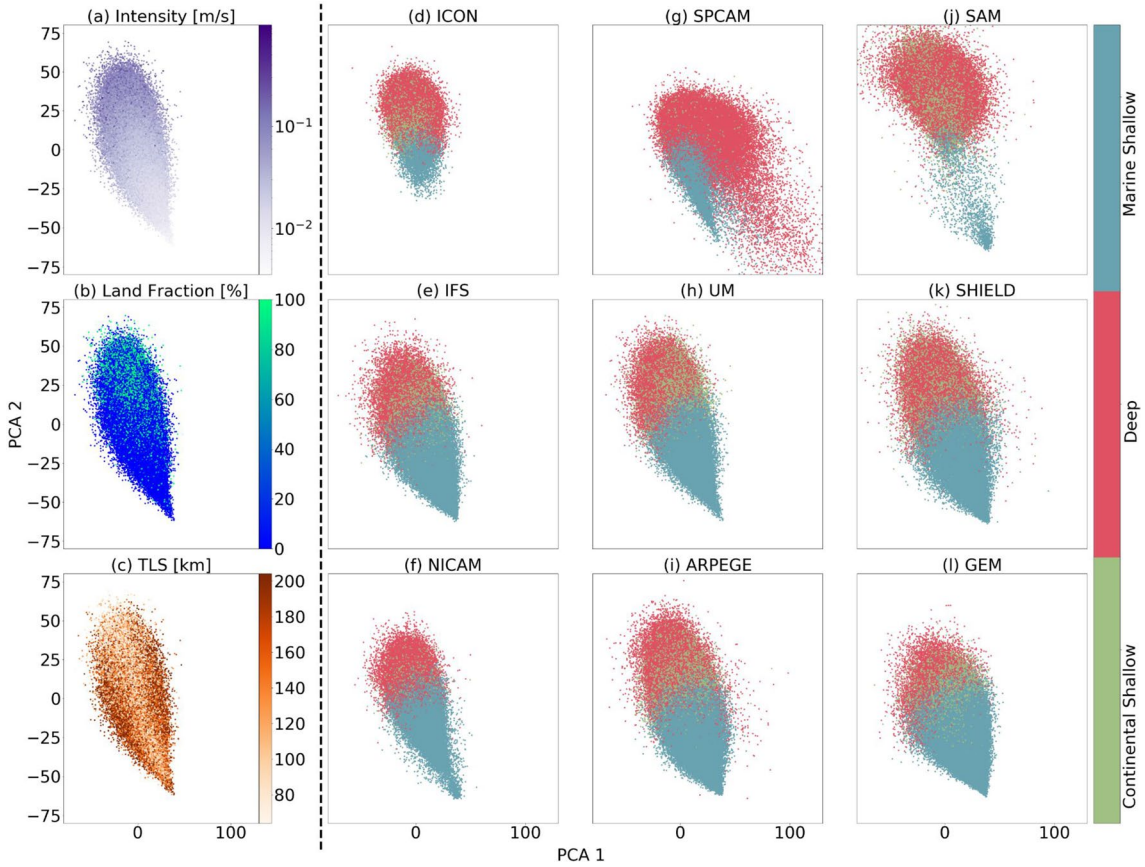

Figure 3. Two-dimensional principal component analysis (PCA) projection plots of DYAMOND data encoded with a shared VAE (trained on UM data). The left column (panels **a**–**c**; see also S5–S7) shows data points colorized by physical convection properties, including convection intensity (**a**), land 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**,**j**) show clear differences in their latent organization compared to the remaining models; see "Dynamic consistency between high-resolution climate models" for a discussion. Movies S2–S6 show additional animations of the latent space.

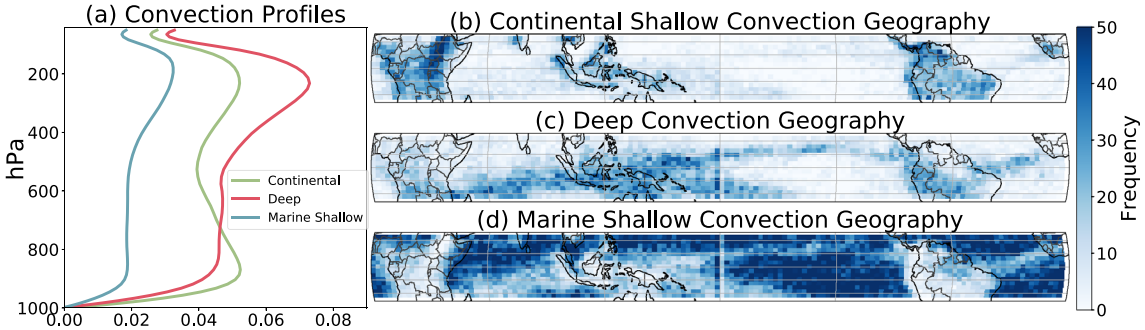

Figure 4. The results from the VAE trained on DYAMOND UM data. Unsupervised clustering ($k = 3$) obtained from UM test data reveals three distinct regimes of convection. Panel (**a**) shows each cluster's median vertical structure, calculated by $\sqrt{\overline{w'w'}}$. Panels (**b**–**d**) show the proportion 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 Fig. 3a,b,c,h.

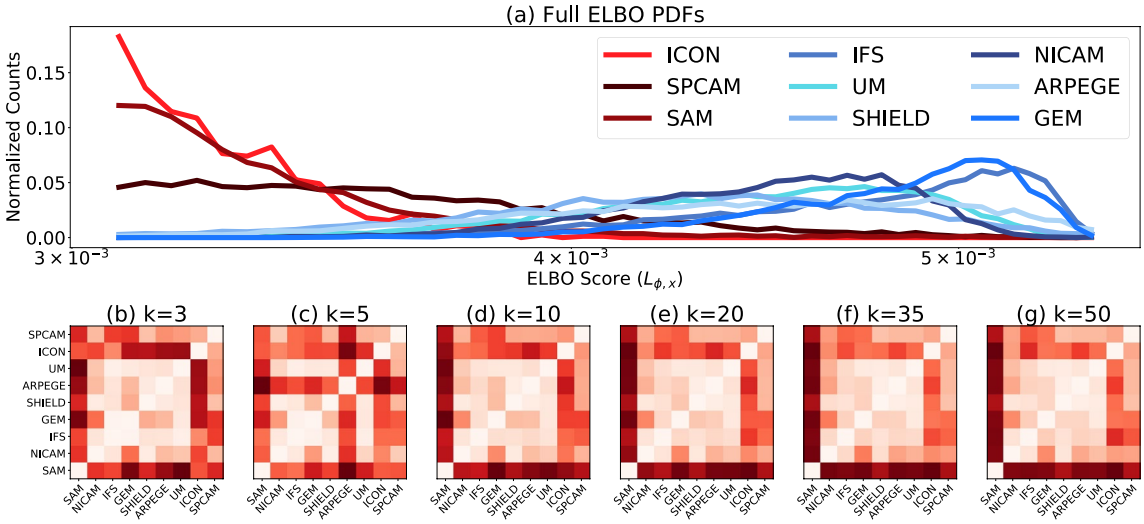

Figure 5. Unsupervised storm-resolving model (GSRM) inter-comparison. The top panel (**a**) shows the ELBO (Eq. 3) score distribution of data from different DYAMOND simulations. (The VAE encoder is trained on UM data before all nine different test datasets are applied.) We see that three model types (ICON, SPCAM, and SAM) have qualitatively different ELBO score distributions than the remaining models. Panels (**b**–**g**) show symmetrized KL divergences between DYAMOND models obtained through nonlinear dimensionality reduction and vector quantization (see main text). Panel (**b**) shows results obtained from $K = 3$ physically interpretable clusters while panel (**g**) shows the results from $K = 50$ in order to better approximate the true lower bound of the KL Divergence. Panels (**c**–**f**) are intermediate K values. To better highlight the structure, we apply agglomerative clustering to the columns[26] and symmetrize the rows. Regardless of the selected K value, the ultimate results are similar, particularly for $K > 20$. We find dynamical consistency between six of the nine GSRMs we examine (6×6 light red sub-region corresponding to NICAM, IFS, GEM, SHIELD, ARPEGE, UM), which is in agreement with panel (**a**).

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

We can use this information to colorize the vertical velocity samples in the latent space, as shown in Figs. 3 and S7.

Variational autoencoders

Variational autoencoders (VAEs) are widely-used latent-variable models for high-dimensional density estimation and non-linear dimensionality reduction[21]. VAEs differ from regular autoencoders in that (1) both encoders and decoders are conditional distributions (as opposed to deterministic functions), and (2) they combine the learning goal of data reconstruction with simultaneously matching a pre-specified "prior" in the latent space, enabling data generation.

In more detail, VAEs model the 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),

$$\mathcal{L}(\theta; \mathbf{x}) := \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\mathrm{KL}\left[\mathrm{q}_\theta(\mathbf{z}|\mathbf{x})||\mathrm{p}_\theta(\mathbf{z})\right]. \tag{3}$$

This involves a so-called variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, also called "encoder", and $p_\theta(\mathbf{x}|\mathbf{z})$ which is commonly referred to as the "decoder". Both the encoder and decoder are parameterized by neural network[21]. The β -parameter is usually set to 1 but can be tuned to larger or smaller values to trade off between data reconstruction ability and disentanglement of the latent space (the rate-distortion trade off), see[21,57,58] for details. To achieve a better model fit, one typically anneals β from zero to one over training epochs.

Our selected VAE architecture prioritizes representation learning over data reconstruction. For our experiments, we anneal β linearly over 1600 training epochs. We use 4 layers in the encoder and decoder with a stride of two) (Fig. 1). We use ReLUs as the activation function 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. For more details on the VAE design choices, see the Methods section of[59].

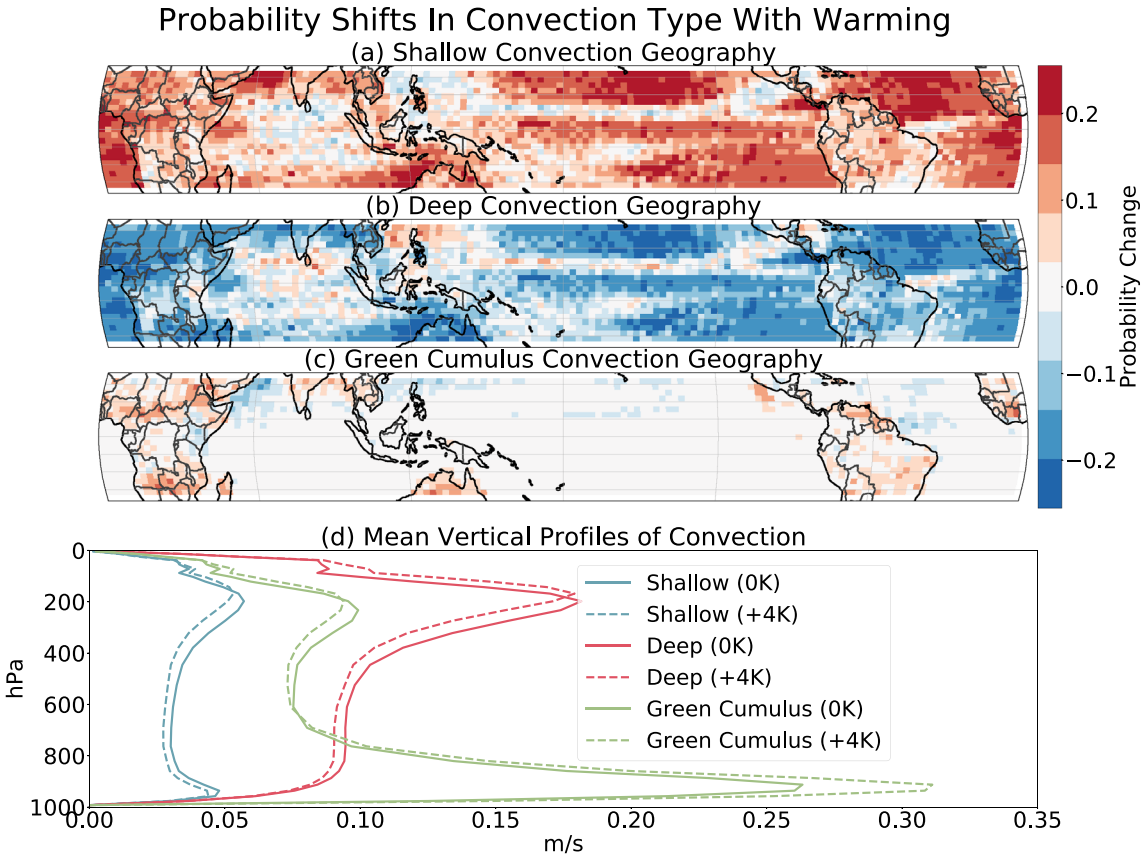

Figure 6. Convection type change induced by $+4K$ of simulated global warming (see main text) in the SPCAM model. Results are from a VAE trained on this SPCAM control ($+0K$) data. Panels (**a**–**c**) show differences in convection type proportion (see main text), where we stratified and plotted the 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 "Green 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**).

K-means clustering

A central element of our analysis pipeline is analyzing the distribution of the dimensionality-reduced, embedded data $\mathbf{z}_i$ using K-means clustering[60,61]. We use this algorithm both for small K (yielding interpretable convection types) and large K (for vector quantization, see below).

In a nutshell, K-means clustering alternates between assigning the (dimensionality-reduced) data points $\mathbf{z}_i$ to K cluster centers μ_k based on euclidean distance, and updating the cluster locations μ_k (setting them to the mean of the assigned data). To formalize the algorithm, one frequently defines the cluster assignment variables $m_i \in \{1, \ldots, K\}$, indicating which cluster data point $\mathbf{z}_i$ belongs to. A measure of convergence is the *inertia*, $\bar{\mathrm{I}} = \sum_{i=1}^{N} ||\mathbf{z}_i - \mu_{m_i}||^2$, measuring the intra-cluster variance of the data.

In all experiments, we perform the clustering ten times, each with a different, random initialization and finally select the result with the lowest inertia. This process enables us to derive the three data-driven convection regimes within an GSRM, which we highlight in Fig. 3h. Notably, we never find the clusters to be strictly spatially isolated; rather, our clustering can be thought of as a partitioning (or a Voronoi tessellation) of the latent space into semantically similar regions.

In order to identify the optimal number of cluster centroids in our analysis, we adopt a qualitative approach that takes into account our domain knowledge. Instead of relying on conventional methods such as the Silhouette Coefficient[62] or the Davies–Bouldin Index[63], we define a "unique cluster" as a group of convection in the latent space that exhibits physical properties (vertical structure, intensity, and geographic domain) that are distinct from those of other groups. By identifying the maximum number of unique clusters, we are able to create three distinct regimes of convection, as shown in Fig. 4. We have observed that increasing K above three usually results in sub-groups of "Deep Convection" that do not exhibit any discernible differences in either vertical mode, intensity, or geography. Therefore, for our purposes, we do not consider $K > 3$ to be physically meaningful.

Our method offers a significant advantage in creating directly comparable clusters of convection between different GSRMs. In recent works, clustering compressed representations of clouds from machine learning models often employs Agglomerative (hierarchical) clustering[64,65]. In contrast, our use of the K-means approach allows

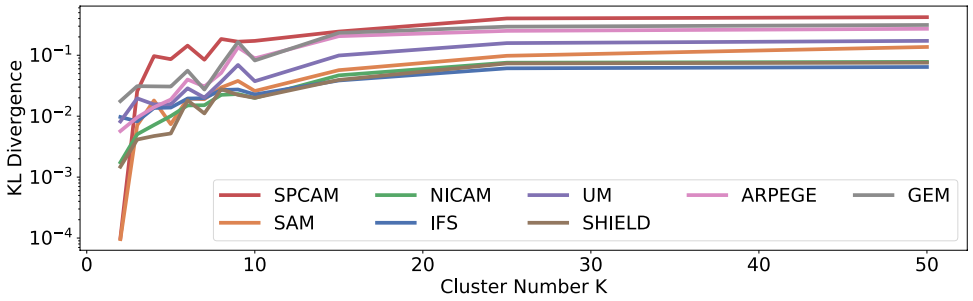

Figure 7. Approximating the KL divergence using vector quantization (VQ) based on K-means clustering, using a variable number of clusters. As discussed in the main paper, VQ lower-bounds the KL and becomes asymptotically exact for large K. We considered the distributional divergence between ICON and the eight other GSRMs. Empirically, the KL approximation seems to saturate at $K = 50$.

us to save the cluster centroids at the end of the algorithm, which provides a basis for cluster assignments for latent representations of out-of-sample test datasets when we use a common encoder as in "Latent space inquiry uncovers differences among storm-resolving models" of our results section. By only using the cluster centroids to get label assignments in other latent representations and not moving the cluster centroids themselves once they have been optimized on the original test dataset, we can objectively contrast cluster differences through the lens of the common latent space. Using this approach, we create interpretable regimes of convection across nine different GSRMs, as shown in Fig. 3d–l.

Vector quantization

We seek to approximate differences between data distributions by directly estimating their Kullback–Leibler (KL) divergence. The KL divergence is a measure of how one probability distribution differs or diverges from another. It quantifies the additional information needed to represent one distribution using another. In the context of our study, we utilize the KL divergence as a measure of distance between the distribution of convective features within our model and a reference distribution (Fig. 1).

The KL divergence is always non-negative and becomes zero only when two distributions match. 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 well approximated by using a technique called vector quantization[27]. This technique involves coarse-graining an empirical distribution into a discrete one obtained from clustering, allowing us to work in a tractable discrete space where the KL divergence can be computed.

In more detail, we perform a K-means clustering on the union of both data sets. We then define the *cluster frequencies* or *cluster proportions* as the fraction of the data claimed by each cluster k: $\pi_k = \frac{1}{N}\sum_{i=1}^{N}\delta(m_i, k)$, where δ denotes the Kronecker delta. By construction, $\sum_{k=1}^{K}\pi_k = 1$ are normalized probabilities.

By increasing the number of clusters (making enough bins), we can quantize continuous distributions into discrete ones with increasing confidence. 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

$$\mathrm{KL}\left(\mathrm{p}^{\mathrm{A}}(\mathbf{x})||\mathrm{p}^{\mathrm{B}}(\mathbf{x})\right) \geq \mathrm{KL}\left(\pi^{\mathrm{A}}||\pi^{\mathrm{B}}\right) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}}^{\mathrm{A}} \log \frac{\pi_{\mathrm{k}}^{\mathrm{A}}}{\pi_{\mathrm{k}}^{\mathrm{B}}}. \tag{4}$$

The inequality comes from the fact that any such discrete KL estimate lower-bounds the true KL divergence[66].

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)). 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 (Fig. 7).

Computing pairwise GSRM distances

To quantify the similarities and dissimilarities among the data produced by different GSRMs (and hence measures of distance between models), we employ the vector quantization approach to compute 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 GSRM. Briefly, to compare Models A and B, we (1) save the K-means cluster centers from the latent vector of the VAE trained on Model A, (2) feed both models' outputs into Model A's encoder as test data, (3) obtain discrete distributions of cluster proportions for Model A and Model B, and (4) compute symmetrized KL divergences based on the discrete distributions using the right-hand side of Eq. (4).

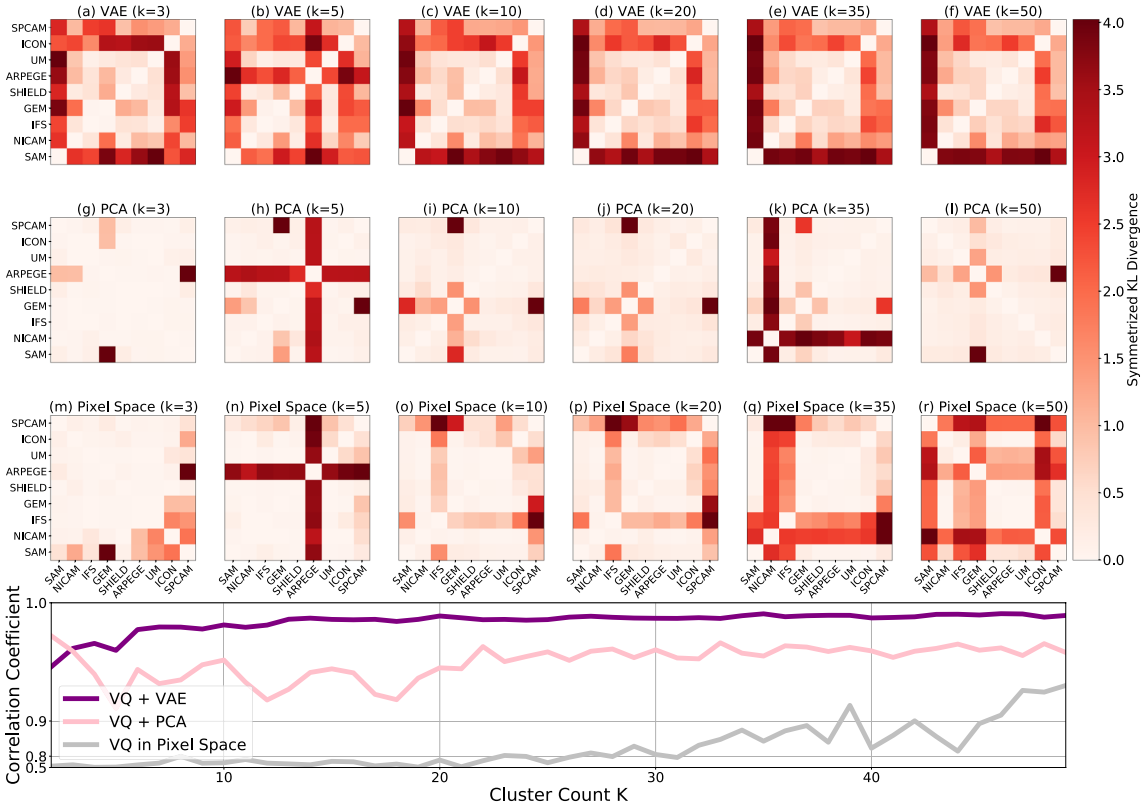

Figure 8. Comparing the robustness of our VAE-based approach with two baseline methods (clustering in PCA space and in pixel space), we assess symmetrized KL divergences across DYAMOND models. Across physically interpretable ($K = 3$), approximately converged ($K = 50$), and intermediate K values, only the VAE-based approach shows consistent performance. In the bottom plot, examining K from 2 to 50, our VAE approach exhibits increasing correlation coefficients close to one between symmetrized KL divergences at adjacent indices (K and $K + 1$), indicating robustness to clustering hyper-parameter variations. (We consider 15 different trials at each K and report the mean correlation coefficient.) This trend is not observed in the baseline approaches, where correlation coefficients are significantly less than one and do not trend upwards towards convergence as K approaches 50.

Robustness of results

Our unsupervised framework utilizes K-means clustering as part of the vector quantization process. However, the choice of the number of clusters (K) introduces variability in the results. To assess the generalizability of the results, we calculate symmetrized KL divergence between models from three different approaches: VAE, PCA, and pixel space (Pure K-means clustering on the full vertical velocity field) analysis. These tests involve comparing models generated from $K = 2$ to $K = 50$, with 15 unique trials conducted at each K.

To evaluate the variation in results for each K, we flatten the table of symmetrized KL divergences at a given K and calculate its Pearson correlation coefficient[67] with the table of KL divergences at $K + 1$. This process yields 15 unique Pearson correlation coefficients, which are then averaged. The summarized outcomes are presented in Fig. 8.

The analysis reveals that the VAE approach exhibits the highest level of robustness for an approximation of the true KL Divergence, showing a rapid convergence towards a correlation coefficient of nearly 1 as K increases. This suggests that, regardless of the selected K (when $K > 20$) value, the results remain consistent. Empirically we see this for the VAE approach in Fig. 8, where panels d, e, f show consistency in GRSM similarity but there are slight differences (particularly in ARPAGE) at lower K counts prior to convergence (a,b,c). In sharp contrast, the other approaches exhibit lower correlation coefficients and do not converge even at greater K counts (as shown in Fig. 8). Taken as a whole, these results suggest that for robustness of the measurement of model distance, a higher value of K is most appropriate.

However, it is important to note that we do not care solely about the approximation of the KL divergence when we consider the cluster count. We also desire for interpretability for our cluster's and for purposes of visualization we want each cluster to correspond to a unique regime of convection. Therefor, we still show results for lower values of K, in particular $K = 3$.

Data availability

Instructions for acquiring DYAMOND simulation data used to train our models can be found here. Compressed data used for main text and SI figures is publically available at 10.5281/zenodo.8024093.

Code availability

Training and postprocessing scripts, as well as saved model weights and python environments, are available on GitHub at 10.5281/zenodo.8024076. The geographic visualizations in Figs. 4, 6, S11, and S13 were rendered in Python[68] version 3.7.3 using cartopy[69] version 0.17.0 and matplotlib version 3.0.3.[70].

Received: 29 June 2023; Accepted: 8 December 2023
Published online: 15 December 2023

References

1. Brient, F. & Bony, S. Interpretation of the positive low-cloud feedback predicted by a climate model under global warming. *Clim. Dyn.* **40**, 05. https://doi.org/10.1007/s00382-011-1279-7 (2012).
2. Blossey, P. N. *et al.* Cgils phase 2 les intercomparison of response of subtropical marine low cloud regimes to co2 quadrupling and a CMIP3 composite forcing change. *J. Adv. Model. Earth Syst.* **8**(4), 1714–1726. https://doi.org/10.1002/2016MS000765 (2016).
3. Schneider, T. *et al.* Climate goals and computing the future of clouds. *Nat. Clim. Change* **7**(1), 3–5. https://doi.org/10.1038/nclimate3190 (2017).
4. Stevens, B. *et al.* Dyamond: The dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Prog. Earth Planet Sci.* **6**(1), 61. https://doi.org/10.1186/s40645-019-0304-z (2019).
5. Randall, D., Khairoutdinov, M., Arakawa, A. & Grabowski, W. Breaking the cloud parameterization deadlock. *Bull. Am. Meteorol. Soc.* **84**(11), 1547–1564. https://doi.org/10.1175/BAMS-84-11-1547 (2003).
6. Christensen, H. M., Moroz, I. M. & Palmer, T. N. Simulating weather regimes: Impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Clim. Dyn.* **44**(7), 2195–2214. https://doi.org/10.1007/s00382-014-2239-9 (2015).
7. Daleu, C. L. *et al.* Intercomparison of methods of coupling between convection and large-scale circulation: 1 comparison over uniform surface conditions. *J. Adv. Model. Earth Syst.* **7**(4), 1576–1601. https://doi.org/10.1002/2015MS000468 (2015).
8. Li, Z. *et al.* Long-term impacts of aerosols on the vertical development of clouds and precipitation. *Nat. Geosci.* **4**(12), 888–894. https://doi.org/10.1038/ngeo1313 (2011).
9. Li, G. & Xie, S.-P. Origins of tropical-wide sst biases in cmip multi-model ensembles. *Geophys. Res. Lett.* **39**, 22. https://doi.org/10.1029/2012GL053777 (2012).
10. Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D. & Randall, D. A. Impacts of cloud superparameterization on projected daily rainfall intensity climate changes in multiple versions of the community earth system model. *J. Adv. Model. Earth Syst.* **8**(4), 1727–1750 (2016).
11. Judt, F. Insights into atmospheric predictability through global convection-permitting model simulations. *J. Atmos. Sci.* **75**(5), 1477–1497. https://doi.org/10.1175/JAS-D-17-0343.1 (2018).
12. Bretherton, C. S. & Khairoutdinov, M. F. Convective self-aggregation feedbacks in near-global cloud-resolving simulations of an aquaplanet. *J. Adv. Model. Earth Syst.* **7**(4), 1765–1787. https://doi.org/10.1002/2015MS000499 (2015).
13. Mapes, B., Tulich, S., Nasuno, T. & Satoh, M. Predictability aspects of global aqua-planet simulations with explicit convection. *J. Meteorol. Soc. Jpn. Ser. II* **II**(86), 175–185. https://doi.org/10.2151/jmsj.86A.175 (2008).
14. Blumenthal, M. B. Predictability of a coupled ocean-atmosphere model. *J. Clim.* **4**(8), 766–784. https://doi.org/10.1175/1520-0442(1991)004<0766:POACOM>2.0.CO (1991).
15. Yan Xue, M. A., Cane, S. E. Z. & Blumenthal, M. B. On the prediction of enso: A study with a low-order markov model. *Tellus A Dyn. Meteorol. Oceanogr.* **46**(4), 512–528. https://doi.org/10.3402/tellusa.v46i4.15641 (1994).
16. Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* (Elsevier, 2006).
17. Palmer, T. N. A personal perspective on modelling the climate system. *Proc. Math. Phys. Eng. Sci.* **472**(2188), 20150772. https://doi.org/10.1098/rspa.2015.0772 (2016).
18. Marat, K. & David, R. Cloud resolving modeling of the arm summer 1997 iop: Model formulation, results, uncertainties, and sensitivities. *J. Atmos. Sci.* **60**, 607–625. https://doi.org/10.1175/1520-0469(2003)060<0607:CRMOTA>2.0.CO;2 (2003).
19. Khairoutdinov, M. F. & Kogan, Y. L. A large eddy simulation model with explicit microphysics: Validation against aircraft observations of a stratocumulus-topped boundary layer. *J. Atmos. Sci.* **56**(13), 2115–2131. https://doi.org/10.1175/1520-0469(1999)056<2115:ALESMW>2.0.CO;2 (1999).
20. Khairoutdinov, M., Randall, D. & DeMott, C. Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *J. Atmos. Sci.* **62**(7), 2136–2154. https://doi.org/10.1175/JAS3453.1 (2005).
21. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. arXiv:abs/1312.6114 [CoRR] (2014).
22. Alemi, A. A., Poole, B., Fischer, I. S., Dillon, J. V., Saurous. R. A., & Murphy, K. Fixing a broken elbo. In *ICML* (2018).
23. Tulich, S. N., Randall, D. A. & Mapes, B. E. Vertical-mode and cloud decomposition of large-scale convectively coupled gravity waves in a two-dimensional cloud-resolving model. *J. Atmos. Sci.* **64**(4), 1210–1229. https://doi.org/10.1175/JAS3884.1 (2007).
24. Johnson, R. H., Rickenbach, T. M., Rutledge, S. A., Ciesielski, P. E. & Schubert, W. H. Trimodal characteristics of tropical convection. *J. Clim.* **12**(8), 2397–2418. https://doi.org/10.1175/1520-0442(1999)012<2397:TCOTC>2.0.CO;2 (1999).
25. Rabanser, S., Günnemann, S., & Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E. & Garnett, R. editors,), Vol. 32 (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf.
26. Schonlau, M. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *Stand. Genom. Sci.* **2**(4), 391–402. https://doi.org/10.1177/1536867X0200200405 (2002).
27. Gray, R. Vector quantization. *IEEE Assp Mag.* **1**(2), 4–29 (1984).
28. Yibo, Y., Stephan, M., & Lucas, T. An introduction to neural data compression. arxiv:2202.06533. (2022)
29. Knutti, R., Masson, D. & Gettelman, A. Climate model genealogy: Generation cmip5 and how we got there. *Geophys. Res. Lett.* **40**(6), 1194–1199. https://doi.org/10.1002/grl.50256 (2013).
30. Allan, R. P. *et al.* Physically consistent responses of the global atmospheric hydrological cycle in models and observations. *Surv. Geophys.* **35**(3), 533–552. https://doi.org/10.1007/s10712-012-9213-z (2014).
31. Neelin, J. D., Chou, C. & Su, H. Tropical drought regions in global warming and el nino teleconnections. *Geophys. Res. Lett.* **30**(24), 96. https://doi.org/10.1029/2003GL018625 (2003).
32. Chia, C. & Neelin, J. D. Mechanisms of global warming impacts on regional tropical precipitation. *J. Clim.* **17**(13), 2688–2701. https://doi.org/10.1175/1520-0442(2004)017<2688:MOGWIO>2.0.CO;2 (2004).
33. Parishani, H. *et al.* Insensitivity of the cloud response to surface warming under radical changes to boundary layer turbulence and cloud microphysics: Results from the ultraparameterized cam. *J. Adv. Model. Earth Syst.* **10**(12), 3139–3158. https://doi.org/10.1029/2018MS001409 (2018).
34. Zelinka, M. D., Klein, S. A. & Hartmann, D. L. Computing and partitioning cloud feedbacks using cloud property histograms part ii: Attribution to changes in cloud amount, altitude, and optical depth. *J. Clim.* **25**(11), 3736–3754. https://doi.org/10.1175/JCLI-D-11-00249.1 (2012).

35. Sherwood, S. C. *et al.* Relative humidity changes in a warmer climate. *J. Geophys. Res. Atmos.* **115**, D9. https://doi.org/10.1029/2009JD012585 (2010).
36. Romps, D. M. An analytical model for tropical relative humidity. *J. Clim.* **27**(19), 7432–7449. https://doi.org/10.1175/JCLI-D-14-00255.1 (2014).
37. Lauer, A., Hamilton, K., Wang, Y., Phillips, V. T. J. & Bennartz, R. The impact of global warming on marine boundary layer clouds over the eastern pacific—a regional model study. *J. Clim.* **23**(21), 5844–5863. https://doi.org/10.1175/2010JCLI3666.1 (2010).
38. Dror, T., Koren, I., Altaratz, O. & Heiblum, R. H. On the abundance and common properties of continental, organized shallow (green) clouds. *IEEE Trans. Geosci. Remote Sens.* **59**(6), 4570–4578. https://doi.org/10.1109/TGRS.2020.3023085 (2021).
39. Mapes, B. E. Convective inhibition, subgrid-scale triggering energy, and stratiform instability in a toy tropical wave model. *J. Atmos. Sci.* **57**(10), 1515–1535. https://doi.org/10.1175/1520-0469(2000)057<1515:CISSTE>2.0.CO;2 (2000).
40. Dror, T., Silverman, V., Altaratz, O., Chekroun, M. D. & Koren, I. Uncovering the large-scale meteorology that drives continental, shallow, green cumulus through supervised classification. *Geophys. Res. Lett.* **49**(8), e2021GL096684. https://doi.org/10.1029/2021GL096684 (2022).
41. Brenowitz, N. D., Beucler, T., Pritchard, M. & Bretherton, C. S. Interpreting and stabilizing machine-learning parametrizations of convection. *J. Atmos. Sci.* **77**(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1 (2020).
42. Dror, T., Chekroun, M. D., Altaratz, O. & Koren, I. Deciphering organization of goes-16 green cumulus through the empirical orthogonal function (eof) lens. *Atmos. Chem. Phys.* **21**(16), 12261–12272. https://doi.org/10.5194/acp-21-12261-2021 (2021).
43. Zhang, Y. & Klein, S. A. 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. *J. Atmos. Sci.* **70**(4), 1297–1315. https://doi.org/10.1175/JAS-D-12-0131.1 (2013).
44. Ahlgrimm, M. & Forbes, R. The impact of low clouds on surface shortwave radiation in the ECMWF model. *Mon. Weather Rev.* **140**(11), 3783–3794. https://doi.org/10.1175/MWR-D-11-00316.1 (2012).
45. Klocke, D., Brueck, M., Hohenegger, C. & Stevens, B. Rediscovery of the doldrums in storm-resolving simulations over the tropical Atlantic. *Nat. Geosci.* **10**(12), 891–896. https://doi.org/10.1038/s41561-017-0005-4 (2017).
46. Nugent, J. M., Turbeville, S. M., Bretherton, C. S., Blossey, P. N. & Ackerman, T. P. Tropical cirrus in global storm-resolving models: 1 role of deep convection. *Earth Sp. Sci.* **9**(2), e2021EA001965. https://doi.org/10.1029/2021EA001965 (2022).
47. Atlas, R. & Bretherton, C. Aircraft observations of gravity wave activity and turbulence in the tropical tropopause layer: Prevalence, influence on cirrus and comparison with global-storm resolving models. *Atmos. Chem. Phys. Discuss.* **1–30**, 2022. https://doi.org/10.5194/acp-2022-491 (2022).
48. Mangipudi, H., Mooers, G., Pritchard, M., Beucler, T., & Mandt, S. Analyzing high-resolution clouds and convection using multi-channel vaes (2021).
49. Eyring, V. *et al.* Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geosci. Model Dev.* **9**(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016 (2016).
50. Haarsma, R. J. *et al.* High resolution model intercomparison project (highresmip v1.0) for cmip6. *Geosci. Model Dev.* **9**(11), 4185–4208. https://doi.org/10.5194/gmd-9-4185-2016 (2016).
51. Norman, M. R. *et al.* Unprecedented cloud resolution in a gpu-enabled full-physics atmospheric climate simulation on olcf's summit supercomputer. *Int. J. High Perform. Comput. Appl.* **36**(1), 93–105. https://doi.org/10.1177/10943420211027539 (2022).
52. Hannah, W. M. *et al.* Initial results from the super-parameterized e3sm. *J. Adv. Model. Earth Syst.* **12**(1), e2019MS001863. https://doi.org/10.1029/2019MS001863 (2020).
53. David, A. R. Beyond deadlock. *Geophys. Res. Lett.* **40**(22), 5970–5976. https://doi.org/10.1002/2013GL057998 (2013).
54. Duras, J., Ziemen, F., & Klocke, D. The dyamond winter data collection. In EGU General Assembly Conference Abstracts, EGU21-4687 (2021).
55. Deardorff, J. W. Closure of second-and third-moment rate equations for diffusion in homogeneous turbulence. Phys. Fluids, **21**, 525–530 (1978). https://api.semanticscholar.org/CorpusID:121223716.
56. Beucler, T. & Cronin, T. A budget for the size of convective self-aggregation. *Q. J. R. Meteorol. Soc.* **145**(720), 947–966. https://doi.org/10.1002/qj.3468 (2019).
57. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M. M, Mohamed, S., & Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In ICLR (2017).
58. Alemi, A., Fischer, I., Dillon, J., & Murphy, K. Deep variational information bottleneck. In ICLR (2017). arXiv:1612.00410.
59. Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. G. Generative modeling of atmospheric convection. In *Proceedings of the 10th International Conference on Climate Informatics, CI2020, New York, NY, USA*, 98–105 (Association for Computing Machinery, 2020). ISBN 9781450388481. https://doi.org/10.1145/3429309.3429324.
60. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28**(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489 (1982).
61. Macqueen, J. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297 (1967).
62. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
63. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909 (1979).
64. Denby, L. Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophys. Res. Lett.* **47**(1), e2019GL085190. https://doi.org/10.1029/2019GL085190 (2020).
65. Kurihana, T., Moyer, E., Willett, R., Gilton, D., & Foster, I. Data-driven cloud clustering via a rotationally invariant autoencoder (2021).
66. Duchi, J. Lecture notes for statistics 311/electrical engineering 377. *Stanford* **2**, 23 (2016).
67. Student. Probable error of a correlation coefficient. *Biometrika*, **6**(2/3):302–310 (1908). http://www.jstor.org/stable/2331474.
68. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, 2009).
69. Met Office. Cartopy: A cartographic python library with a Matplotlib interface. Exeter, Devon, 2010 (2015). https://scitools.org.uk/cartopy.
70. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).

Acknowledgements

The authors acknowledge funding by the National Science Foundation (NSF) Machine Learning and Physical Sciences (MAPS) program and NSF Grant 1633631, the Department of Energy, Office of Science under Grant number DE-SC0022331, the 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, 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. This work was also supported by gifts from Intel, Disney, and Qualcomm. We further acknowledge funding from NSF Science and Technology Center LEAP (Learning the Earth with Artificial Intelligence and Physics) 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. We are grateful to Scientific Reports Editor Ryan Sriver and our two anonymous editors for their constructive feedback. The authors express their gratitude to Jens Tuyls for helping with the initial model repository and also thank Yibo Yang, Veronika Eyring, Gunnar Behrens, Ilan Koren, Tom Dror, Peter Blossey, Peter Caldwell, Claire Monteleoni, David Rolnick, Imme Ebert-Uphoff, and Maike Sonnewald for helpful conversations that advanced this work.

Author contributions

G.M., S.M., M.P., and T.B. designed the research. G.M., M.P., L.P., and T.B. performed numerical simulations. G.M., S.M., M.P., T.B., P.G., L.P., P.S., and H.M. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-49455-w.

Correspondence and requests for materials should be addressed to G.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.