Analyzing High-Resolution Clouds and Convection using Multi-Channel VAEs

Harshini Mangipudi
University of California Irvine
harshinimangipudi@gmail.com

Griffin Mooers
University of California Irvine
gmooers96@gmail.com

Mike Pritchard
University of California Irvine
mspritch@uci.edu

Tom Beucler
University of Lausanne
tom.beucler@unil.ch

Stephan Mandt
University of California Irvine
mandt@uci.edu

Abstract

Understanding the details of small-scale convection and storm formation is cru-
cial to accurately represent the larger-scale planetary dynamics. Presently, at-
mospheric scientists run high-resolution, storm-resolving simulations to capture
these kilometer-scale weather details. However, because they contain abundant
information, these simulations can be overwhelming to analyze using conventional
approaches. This paper takes a data-driven approach and jointly embeds spatial
arrays of vertical wind velocities, temperatures, and water vapor information as
three "channels" of a VAE architecture. Our "multi-channel VAE" results in more
interpretable and robust latent structures than earlier work analyzing vertical veloci-
ties in isolation. Analyzing and clustering the VAE's latent space identifies weather
patterns and their geographical manifestations in a fully unsupervised fashion. Our
approach shows that VAEs can play essential roles in analyzing high-dimensional
simulation data and extracting critical weather and climate characteristics.

1 Introduction

Long-term climate modeling is increasingly accurate, but any climate simulation high resolution
enough to explicitly resolve kilometer-scale processes controlling clouds and storms not only requires
high computational expense but also produces massive amounts of data. This large data volume
overwhelms physically-informed methods traditionally used to better understand storm organization
and how small-scale clouds connect to the large-scale circulation; even established dimensionality
reduction techniques such as Principle Component Analysis (PCA) [16] may fail to capture these
nonlinear relations. In contrast, unsupervised learning, more specifically deep generative modeling
[5, 3] can summarize explicitly-resolved storms without omitting fine-scale, nonlinear information.

Variational Autoencoders (VAEs) in particular are well-suited for dynamics discovery and under-
standing because of ability to learn meaningful latent representations of the data [11, 7, 17, 13]. VAE
applications on climate and weather data have thus far been rare. Physical processes like a 2D laminar
flow have been successfully represented by a VAE [4]. Some analysis has been conducted with VAEs
on spatio-temporal earth dynamics and with state variables like temperature and potential vorticity
data to cluster phenomena like the Polar Vortex [14, 15]. However, the training/test data applied to
the VAEs in all these cases was highly idealized and smaller in dimension than a storm-resolving
simulation output.

A *single-channel* VAE architecture (SVAE), modeling velocity fields in high-resolution climate
simulation has been implemented [12]; however, relationships between other variables known to be

important controls for convection were ignored. To bypass the limit of using a single physical state variable input in the context of deep generative modeling for climate simulation analysis, we upgrade our network architecture to that of a three-channel VAE - temperature and water vapor in addition to vertical velocity. We hypothesize that this multi-variate approach can enable increasingly satisfying physical clustering through a more thermodynamically-informed latent space.

2   Methods

2.1   Data and preprocessing

To provide our VAE with a multi-variate, high resolution simulation, we use output from the System for Atmospheric Modeling (SAM), a Storm-Resolving Model (SRM). We extract 1.5e6 30x128 snapshots or "images" of vertical velocity (m/s), temperature (K), and water vapor concentration (kg/kg) fields. The dimensionality of our images comes from the simulation design of an atmosphere with thirty vertical levels and 128 SRM columns embedded in each grid cell of a planetary-scale General Circulation Model (GCM) to explicitly resolve deep convection in thousands of embedded sub-domains [6]. The high-resolution convection data comes from all tropical areas ranging from latitudes of 20 degrees south to 20 degrees North during boreal winter. This geographically diverse output thus includes samples ranging from moist tropical rain forests to arid deserts, spanning both marine and continental zones. Each of the three variables is re-scaled separately from zero to one using standard normalization techniques.

2.2   Model architecture and upgrades

We expand on a baseline VAE architecture designed in [12]. Our VAE is fully convolutional and has four layers in the encoder and decoder. To address the inherent rate-distortion trade-off native to VAEs [1] and align the latent variables of the model more closely with the output, we implement a linear annealing [2] in the Kullback–Leibler (KL) Divergence term on our VAE loss function (Equation 1). We initially weight the KL Divergence as 0 in order to prioritize learning in the encoder for a superior latent space but increase β to 1 over 1600 epochs.

$$\text{ELBO} \left(\mathbf{x}; \theta, \phi, \mathbf{z} \right) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}) \right] - \beta \times D_{\text{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}) \right),$$ (1)

In contrast to earlier work, we interpret temperature, water vapor, and raw vertical velocity fields as three "channels" of an image, and adopt a VAE architecture to jointly embed them. This also increases the compression from the input vector to a latent space of 1024 from roughly 4x in [12] to 10x. We also continue to use a SVAE trained on vertical velocity fields as a baseline for comparisons.

3   Results

Latent space credibility and exploration   By expanding the input to three high-resolution variables, we are asking substantially more of our encoder in terms of dimensionality reduction and feature extraction. There is no guarantee this architecture can create the same physically interpretable latent space as the SVAE in [12]. However, we can gain better understanding of the latent space by "colorizing" it by the magnitude of the corresponding vertical velocity field in the test dataset and its land fraction. In the tropics there are regions of intense convection, such as over the rain-forests and hot tropical waters of the Pacific warm pool; however, there will also be dry regions like deserts where deep convection will be suppressed. We want such known disparate convection types to be separated in some capacity on the latent space for it to be useful for physical analysis. When our latent space is colorized (Fig 1, a and b), we do see a clear sorting by magnitude, and convection over land preferentially appearing on different parts of the latent space than convection over the ocean. The full animated visual can be found at this link. This suggests that the added complexity is not overwhelming the VAE, and we now ask: *What are the advantages of the MVAE compared to the SVAE when it comes to anomaly detection and regime classification?*

Anomaly detection   We determine the similarity between the trained VAEs and their respective latent spaces by first looking at density estimation properties. To test whether the MVAE adds value

compared to the SVAE, we calculate the Evidence Lower Bound (ELBO) for both VAEs at each of the 1.5e6 samples in the test dataset. The higher this ELBO score the more anomalous the VAE finds the convection sample. We look at just the top five percent of ELBO anomalies for both the SVAE and MVAE. We then plot the frequency at which these anomalies occurred at each grid-cell in the tropics (Fig 2 b and c) and ask whether the results agree with domain knowledge of where geographically rare convective events tend to occur. Immediately we see different patterns lit up by the different VAEs. The SVAE hones in on continental zones, especially over more arid regions (Fig 2b). The MVAE highlights areas over both land and sea (Fig 2c), which is physically satisfying since rare storms are known to occur in both regions. Specifically, the MVAE finds deep convection over the tropical rain forests and the Inter-Tropical Convergence Zone to be most anomalous. But more significantly than just disagreeing about the geographic location of the most extreme anomalies, the two VAEs each find anomalies of very different vertical structures. In the median vertical profile in each anomaly group, the SVAE emphasizes strong convection near the surface of the earth, which is not typically recognized as an interesting phenomenon in the atmospheric sciences, whereas the MVAE focuses on convection deep in the upper Troposphere, a reassuringly familiar extreme tropical storm structure (Fig 2a).

Data driven convection regimes To analyze the latent space in an objective fashion, we bring in another unsupervised learning algorithm: K-Means Clustering. We cluster separately on both latent spaces to determine if the latent structures are different enough for the MVAE to highlight different types of convective organization. Our analysis turns up three distinct clusters in both the latent spaces (Figs 3). However, there are significant differences in the vertical profiles and geographic location of the regimes. The SVAE latent space successfully separates the deepest convection from shallower forms that occur in the subtropical ocean basins (Fig 3a and g, blue lines). But the third cluster in the SVAE group, similar to the anomalies it finds (Fig 2b), reflects the same puzzling mode of shallow convection over drier land areas (Fig 3d, blue line). The MVAE clusters deep convection similarly to the SVAE but attractively does not make a distinct regime for the mode of dry, continental shallow convection (Fig 3 d,g green vs. blue). Instead, unlike the SVAE, the subtropical shallow convective mode is satisfyingly separated into two distinct sub-regions – one in the Central Pacific/Atlantic, and another in the eastern parts of these ocean basins where especially distinct stratocumulus clouds are known to occur [8, 9, 10]. Since these two forms of low cloud have different characteristic horizontal eddy scales, we speculate that the inclusion of water vapor to the input of the MVAE down-weights the importance of the dry, shallow mode (Fig 3d; green vs. blue) and allows a more satisfying physical clustering of convective regimes.

4 Conclusion and broader impacts

We explore the ability of a Variational Autoencoder (VAE) to find a physically interpretable representation of a multi-variate, high-resolution climate simulation. We find that adding multiple variables as "channels" to our VAE highlights drastically different convection anomaly types and geographic locations compared to a VAE trained only on a single variable. The different regimes identified by clustering both latent spaces suggest there is value to a multi-variate approach using both dynamic and thermodynamic variables to cluster convection regimes successfully in a fully unsupervised fashion. It is possible that this multi-variate approach would also be valuable for global storm-resolving climate simulations that produce even more data and augment traditional physically-informed approaches to clustering the cloud regime-dependence of climate change projections.
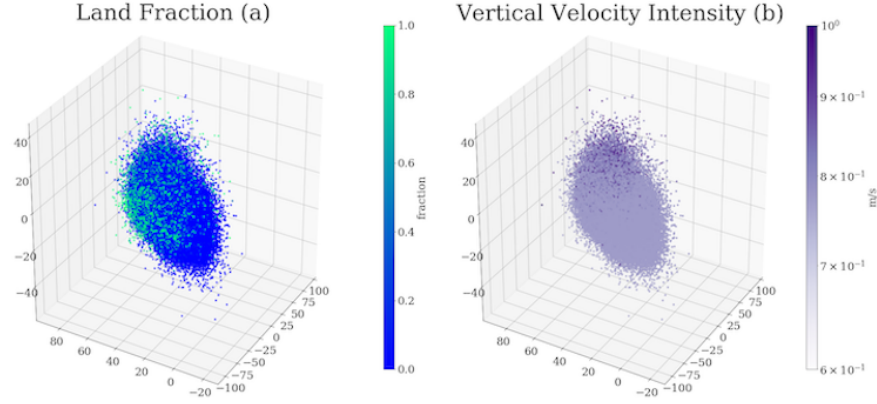
5 Acknowledgements

Figure 1: The 3D projection of Principle Component Analysis of our multi-channel VAE latent space. We colorize by land/sea fraction and intensity of the vertical velocity field at each sample in the test dataset. The organization of this latent structure appears physically interpretable.
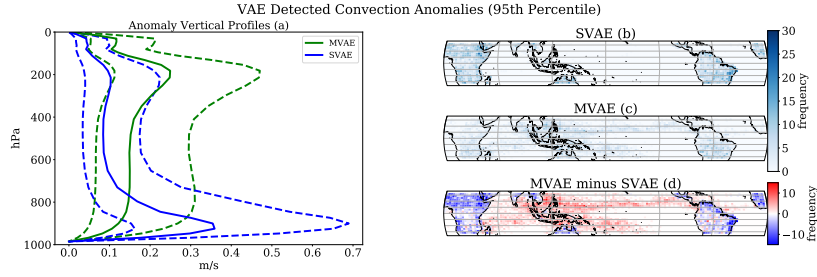

Figure 2: The vertical profiles (a, solid lines), 25th and 75th percentiles (a, dashed lines) and geographic distributions (b,c,d) of convection anomalies defined by our VAEs. We threshold the test data by the Evidence Lower Bound (ELBO) anomalies and examine only the top 5 percent of ELBO scores. The multi-channel VAE finds anomalies of different structure and geographic location than the single-channel VAE.
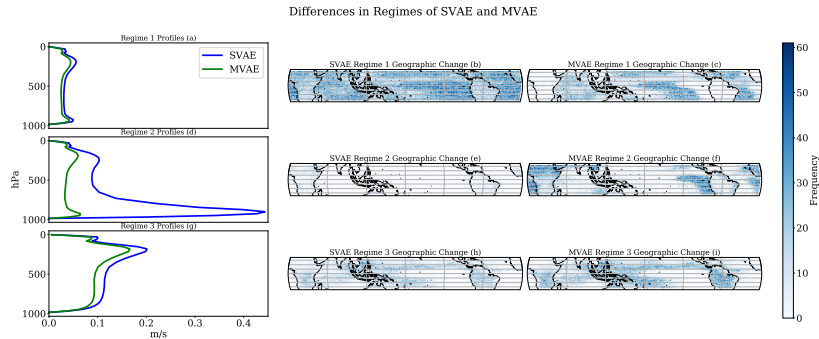

Figure 3: The results of K-Means clustering on the latent space of both VAEs separately. We find that three is the largest number of clusters in which each convective regime has both a distinct vertical profile (a,d,g) and distinct geographic pattern of frequency of occurrence (b,c,e,f,h,i). However, the three regimes found in each latent do not match suggesting unique organization in each VAE.

References

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

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

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

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

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[6] W. W. Grabowski and P. K. Smolarkiewicz. Crcp: a cloud resolving convection parameterization for modeling the tropical convecting atmosphere. *Physica D: Nonlinear Phenomena*, 133:171–178, 1999.

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

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

[9] B. Khouider and A. J. Majda. Multicloud convective parametrizations with crude vertical structure. *Theoretical and Computational Fluid Dynamics*, 20(5):351–375, 2006.

[10] B. Mapes, S. Tulich, J. Lin, and P. Zuidema. The mesoscale convection life cycle: Building block or prototype for large-scale tropical waves? *Dynamics of Atmospheres and Oceans*, 42(1):3–29, 2006. Atmospheric Convection and Wave Interactions (title for cover & spine) Atmospheric Convection and Wave Interactions: Convective Life Cycles and Scale Interactions in Tropical Waves (title for Special Issue title page).

[11] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks, 2018.

[12] G. Mooers, J. Tuyls, S. Mandt, M. Pritchard, and T. Beucler. Generative modeling for atmospheric convection, 2020.

[13] A. A. Pol, V. Berger, G. Cerminara, C. Germain, and M. Pierini. Anomaly detection with conditional variational autoencoders, 2020.

[14] X.-A. Tibau, C. Requena-Mesa, C. Reimers, J. Denzler, V. Eyring, M. Reichstein, and J. Runge. Supernovae: Vae based kernel pca for analysis of spatio-temporal earth data. In *Proceedings of the 10th International Conference on Climate Informatics*, 2018.

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

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

[17] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. On the quantitative analysis of decoder-based generative models, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mention the resolution of the data and the type of VAE and its associated strengths/weaknesses in Sections 2.1, 2.2
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See link to github repository in Section 5
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Model architecture described in Section 2.2, training data in Section 2.1, test data in Section 3. More details can be found on the model training script and config files. The link to the github repository containing these can be found in Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]