New Understanding of Cloud Processes via Unsupervised Cloud Classification in Satellite Images

Elisabeth J. Moyer[1,*], Ian T. Foster[2,3,*], James A. Franke[1], Rob Jacob[4], Rebecca Willett[3], and Takuya Kurihana [3]

[1]*University of Chicago, Department of the Geophysical Sciences*
[2]*Argonne National Laboratory, Data Science and Learning Division*
[3]*University of Chicago, Department of Computer Science*
[4]*Argonne National Laboratory, Environmental Science Division*
[*]*Correspondence: moyer@uchicago.edu | foster@anl.gov*

Focal Areas

This white paper addresses DOE AI4ESP focus areas #1 and #3. We propose to leverage advances in machine learning and artificial intelligence to create a new unsupervised cloud classification framework that will produce new datasets and cloud process understanding to reduce biases in Earth system models.

Science Challenge

Despite considerable improvement in climate model representation of many phenomena in recent years [e.g. 1], long-term climate projections remain uncertain due in particular to inadequate model representations of clouds [2]. Constraining cloud response to anthropogenic forcing requires running global simulations at the resolution needed to resolve the boundary layer eddies critical to convection [3]—a computationally infeasible feat for the foreseeable future. Deep learning data-driven approaches applied to decades of satellite imagery provide an unprecedented opportunity to fill the knowledge gap in cloud dynamics and feedbacks.

Rationale

Clouds are an obvious, critical part of the hydrological cycle and can have impacts far beyond the region where they are formed. For example, uncertainties in future changes in precipitation in the Sahel region (where a large portion of the population depends on rainfed agriculture for their livelihoods) are due in large part to the northern hemisphere temperature gradient which in turn depends on global cloud response to warming [4]. Cloud response to anthropogenic forcing is a root cause of hydrological cycle uncertainties in many regions. Thus improving cloud process understanding and the representation of clouds in models are critical steps to improving predictability of the hydrological cycle and the earth system as a whole [5].

Multispectral (hyperspectral) images from various Earth-observing satellites, such as MODIS [6], provide rich spatiotemporal information streams on cloud patterns and textures. These data represent an opportunity to help constrain cloud responses to warming. However, the volume (growing at ~100TB/day [7]) and complexity of these data pose a significant barrier to scientists. No automatic dimension reduction methodology exists that would enable researchers to understand the controls on cloud patterns and textures, identify trends, or compare to numerical simulations as a diagnostic test of their ability. Existing supervised deep learning classification approaches [8, 9, 10, 11] are insufficient, being limited by a lack of labeled images (stemming from the difficulty in consistently labeling 'blurry' cases and a general lack of agreement on what the classes ought to be) and, yet more concerning, being constrained to human-defined classes.

Recent developments in deep learning, especially unsupervised image classification, provide an excellent opportunity to overcome the paucity of labeled cloud data and capitalize on the wealth

DOI: https://doi.org/10.2172/1769754
DOE BER, AI4ESP 2021 #AI4ESP1094 (https://www.ai4esp.org)

of high-resolution data. Unsupervised classification techniques that automatically reduce satellite images would provide researchers with the opportunity to study distributions in the populations of cloud types, physical characteristics of novel classes of clouds, variability and trends in these quantities, and new Earth System Model evaluation methods which could be used to reduce biases and thus improve predictability.

Narrative

We propose the development of an unsupervised deep learning framework for the clustering and classification of satellite cloud images, with the goal of producing new understanding of the diverse and potentially changing roles of clouds in the climate system. Specifically, we propose to use deep learning methods to identify robust and meaningful grouping of diverse clouds into clusters, on the basis of which we hope to develop new classification schemes. Supervised cloud classification using deep learning was first attempted over 30 years ago [12]: perhaps the very first application of deep learning in Earth Science! Unsupervised cloud classification was first proposed a decade later [13, 14], but dismissed at the time as infeasible because of the limited accuracy achievable with the machine learning methods of that era.

In the intervening years, three critical developments have made it possible to revisit this problem. First, satellite data has vastly expanded since the late 90s. The MODIS satellite for example, was launched in 2000 and the number of earth observation satellites grows every year. Second, new deep learning classification methodologies, such as autoencoders [15, 16], promise to enable new approaches to clustering of complex cloud structures that were problematic to researchers in the late 90s. Finally, acceleration of deep learning performance via a combination of greatly enhanced computational power and improved numerical methods make scaling this problem feasible.

Most supervised learning attempts to date have focused on the 11 canonical classes familiar to meteorologists [8]. However, labeled cloud datasets are still small compared to other deep learning applications (2500 ground-based cloud images compared to 14M in ImageNET). A recent supervised deep learning project [9] employed a novel labeling approach by crowd-sourcing cloud labels. While this approach has promise as a means of delivering the large quantities of labeled data needed for deep learning, the manually defined classes had to be limited in number (here 4 classes) in order that non-experts could provide reasonably accurate labels. In this and other similar projects, the cloud classes do not necessarily capture all relevant distinctions, combining for example quite different types of low clouds.

Unsupervised learning relaxes the constraint of labeled data by automatically defining classes, thus vastly expanding the potential training set. The expanded data advantage comes at a cost however, as learned classes need to be validated with no obvious 'ground-truth.' An effective unsupervised classification scheme should generate classes/clusters that are (1) *separable* (cohesive and distinct in latent space), (2) *stable* (produce identical classes when trained on different subsets of the data), and (3) *rotationally invariant* (insensitive to the orientation of an image). Additionally for cloud classification, we propose that classes should (4) capture information on spatial distributions such as textures, and (5) be physically reasonable (embody scientifically relevant distinctions) [17].

Preliminary work has shown that unsupervised cloud classification using the MODIS instruments visible-to-thermal bands produces stable classes that are physically reasonable [17]. In this first unsupervised attempt, we produced 12 cloud classes which exhibit physical properties that are not randomly distributed and do not contradict existing physical knowledge. For example, low clouds are distinct in latent space from high clouds, and physical properties such as ice fraction and optical thickness occupy separate distributions across classes.

These preliminary applications were limited in geographic scope, with training and testing data focused on the east equatorial Pacific ocean. While clouds are varied and climatologically important in this region, all possible types of cloud are not uniformly sampled here. A classifier trained on equatorial cloud populations may break down when applied to ocean clouds in other regimes, such as high clouds in the southern ocean, and clouds over land pose additional challenges as the boundary conditions leading to convection are more heterogeneous. Background surface reluctance over land (or ice) will provide additional challenge to the classifier not present over the ocean. In more recent work, we have started to extend the work to a global scale [18], but more work is required to understand when and where the resulting classifiers are effective and what they tell us about cloud properties and changes in those properties over time.

Rotational-invariance proves to be a challenging aspect of unsupervised classification of clouds. Current ongoing research employing a rotationally invariant autoencoder [19] attempts to address this problem [17], with promising results via the use of convolutional network to learn spatial structure, a loss function that uses transform-invariant techniques, and a learning protocol that learns from satellite data without introducing biases [18]. These methods need to be evaluated at large scales on 30 years of satellite imagery.

Additional next step efforts include integration into earth system models (e.g., E3SM [20]). This integration may be performed first in the context of global high-resolution convection-permitting models such as SCREAM, and then (with modifications to deal with lower model resolutions) to ensembles of GCMs in CMIP as a tool for model validation. Scientific questions here include: do simulated clouds populations match historical data? Where and how do they fail?

Other questions include: Could high-resolution convection permitting models (and perhaps generative adversarial networks) be used in some novel ways here? For example, one could simulate clouds in quite different regimes (perhaps far from existing possibilities) and then apply the classifier to the resulting outputs to see what happens. For example, we might see evidence for the postulated breakup of stratocumulus decks under extreme warming [21].

Research efforts in this domain should address three major themes (1) the development of new deep learning techniques specific to the cloud problem; (2) operationalizing the prototype algorithms to global scale to produce classified datasets that can be disseminated alongside existing satellite image datasets; and (3) understanding what cloud classifiers tell us about cloud properties. The construction of a production-quality operational tool will require advances in methods for rapid processing of large quantities of climate simulation data and can make use of the advanced AI/HPC capabilities available at the national laboratories (e.g., Summit, Aurora, Cerebras). Any multispectral satellite image can in principle be used as training data, though making direct comparisons across cloud populations learned from different satellite products may prove challenging as resolutions and available spectral bands vary. Further steps can include the principled use of both observational data from multiple sources and the automated design and execution of climate simulation runs to expand input data into new regimes.

Advanced techniques developed here may have wide applications across many satellite image domains including beyond cloud classification: for example, improved cloud masking algorithms, land type classification, and monitoring of glaciers and surface water. Finally, this methodology could also be readily applied to *ARM: Total Sky Imager* ground based cloud data.

Suggested Partners/Experts

1. Bjorn Stephens, MPI
2. Climate Modeling Alliance (Caltech): next-gen earth system model framework project led by Tapio Schneider.
3. William Collins and Mr. Prabhat, Lawrence Berkeley Laboratory
4. Robert Wood, University of Wisconsin
5. Chris Bretherton, University of Washington

References

[1] T. J. Bracegirdle, C. R. Holmes, J. S. Hosking, G. J. Marshall, M. Osman, M. Patterson, and T. Rackow. Improvements in Circumpolar Southern Hemisphere Extratropical Atmospheric Circulation in CMIP6 Compared to CMIP5. *Earth and Space Science*, 7(6), June 2020.

[2] Mark D. Zelinka, Timothy A. Myers, Daniel T. McCoy, Stephen Po-Chedley, Peter M. Caldwell, Paulo Ceppi, Stephen A. Klein, and Karl E. Taylor. Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, 47(1), January 2020.

[3] Tapio Schneider, Shiwei Lan, Andrew Stuart, and João Teixeira. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44(24), December 2017.

[4] Paul-Arthur Monerie, Caroline M. Wainwright, Moussa Sidibe, and Akintomide Afolayan Akinsanola. Model uncertainties in climate change impacts on Sahel precipitation in ensembles of CMIP5 and CMIP6 simulations. *Climate Dynamics*, 55(5-6):1385–1401, September 2020.

[5] Graeme L. Stephens. Cloud Feedbacks in the Climate System: A Critical Review. *Journal of Climate*, 18(2):237–273, January 2005.

[6] Vincent V Salomonson, WL Barnes, Peter W Maymon, Harry E Montgomery, and Harvey Ostrow. MODIS: Advanced facility instrument for studies of the earth as a system. *IEEE Transactions on Geoscience and Remote Sensing*, 27(2):145–153, 1989.

[7] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019.

[8] Jinglin Zhang, Pu Liu, Feng Zhang, and Qianqian Song. CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network. *Geophysical Research Letters*, 45(16):8665–8672, August 2018.

[9] Stephan Rasp, Hauke Schulz, Sandrine Bony, and Bjorn Stevens. Combining crowd-sourcing and deep learning to explore the meso-scale organization of shallow convection. *arXiv:1906.01906 [physics]*, April 2020. arXiv: 1906.01906.

[10] Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt J. Kusner, and Duncan Watson-Parris. Cumulo: A Dataset for Learning Cloud Classes. *arXiv:1911.04227 [physics, stat]*, April 2020. arXiv: 1911.04227.

[11] Willem J. Marais, Robert E. Holz, Jeffrey S. Reid, and Rebecca M. Willett. Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations. *Atmospheric Measurement Techniques*, 13(10):5459–5480, October 2020.

[12] J. Lee, R. C. Weger, S. K. Sengupta, and R. M. Welch. A neural network approach to cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, 28(5):846–855, 1990.

[13] A. Visa, J. Iivarinen, K. Valkealahti, and O. Simula. Neural network based cloud classifier. *Industrial Applications of Neural Networks*, pages 303–309, 1998.

[14] Bin Tian, M. A. Shaikh, M. R. Azimi-Sadjadi, T. H. V. Haar, and D. L. Reinke. A study of cloud classification with neural networks using spectral and textural features. *IEEE Transactions on Neural Networks*, 10(1):138–151, 1999.

[15] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[16] Pascal Vincent, H. Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

[17] Takuya Kurihana, Ian Foster, Rebecca Willett, Sydney Jenkins, Kathryn Koenig, Ruby Werman, Ricardo Barros Lourenco, Casper Neo, and Elisabeth Moyer. Cloud classification with unsupervised deep learning. In *9th International Workshop on Climate Informatics*, 2019.

[18] Takuya Kurihana, Elisabeth Moyer, Rebecca Willett, Davis Gilton, and Ian Foster. A Data-driven Cloud Classification Framework Based on a Rotationally Invariant Autoencoder. page 19, 2021. in prep.

[19] Tadashi Matsuo, Hiroya Fukuhara, and Nobutaka Shimada. Transform invariant auto-encoder. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2359–2364, 2017.

[20] Jean-Christophe Golaz, Peter M Caldwell, Luke P Van Roekel, Mark R Petersen, Qi Tang, Jonathan D Wolfe, Guta Abeshu, Valentine Anantharaj, Xylar S Asay-Davis, David C Bader, et al. The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, 11(7):2089–2129, 2019.

[21] Tapio Schneider, Colleen M Kaul, and Kyle G Pressel. Possible climate transitions from breakup of stratocumulus decks under greenhouse warming. *Nature Geoscience*, 12(3):163–167, 2019.