Insight into cloud processes from unsupervised classification with a rotation-invariant autoencoder

Takuya Kurihana[*]
Dept. of Computer Science
University of Chicago
tkurihana@uchicago.edu

James Franke
Dept. of the Geophysical Sciences
University of Chicago
jfranke@uchicago.edu

Ian T. Foster
University of Chicago &
Argonne National Laboratory
foster@uchicago.edu

Ziwei Wang
Dept. of the Geophysical Sciences
University of Chicago
zwwang@uchicago.edu

Elisabeth J. Moyer
Dept. of the Geophysical Sciences
University of Chicago
moyer@uchicago.edu

Abstract

Clouds play a critical role in the Earth's energy budget and their potential changes are one of the largest uncertainties in future climate projections. However, the use of satellite observations to understand cloud feedbacks in a warming climate has been hampered by the simplicity of existing cloud classification schemes, which are based on single-pixel cloud properties rather than utilizing spatial structures and textures. Recent advances in computer vision enable the grouping of different patterns of images without using human-predefined labels, providing a novel means of automated cloud classification. This unsupervised learning approach allows discovery of unknown climate-relevant cloud patterns, and the automated processing of large datasets. We describe here the use of such methods to generate a new AI-driven Cloud Classification Atlas (AICCA), which leverages 22 years and 800 terabytes of MODIS satellite observations over the global ocean. We use a rotation-invariant cloud clustering (RICC) method to classify those observations into 42 AI-generated cloud class labels at $\sim$100 km spatial resolution. As a case study, we use AICCA to examine a recent finding of decreasing cloudiness in a critical part of the subtropical stratocumulus deck, and show that the change is accompanied by strong trends in cloud classes.

1 Motivation

Stratocumulus—decks of low clouds—are critical for the climate and represent the single biggest source of uncertainty in climate response to anthropogenic forcing [Zelinka et al., 2020]. The marine stratocumulus decks off the west coasts of North and South America and Africa in the sub-tropics cover only $\sim$6% of the Earth's surface but reflect a disproportionate amount of solar radiation, keeping the planet cool. Recent studies based on relatively small-scale numerical simulations suggest that these decks are vulnerable to increases in CO_2 and may break up after some threshold is reached,

---

[*]Globus Labs: https://labs.globus.org/

causing an additional 8K of global warming [Schneider et al., 2019, 2020]. However, the global models used in most climate studies are too coarse in resolution to capture the behavior of subtropical stratocumulus, or indeed marine low clouds in general. Current computing power does not allow simulating global clouds at the 10m resolution needed to realistically capture their behavior [Morrison et al., 2020], and will not be able to do so until at least 2060, even if computer power continues to grow at historical rates [Schneider et al., 2017]. Understanding the climate risk posed by stratocumulus transitions requires making use of satellite observations instead.

By now, satellite instruments have provided several decades of high-resolution multispectral images of the Earth. These datasets, now totaling many petabytes (PBs), should provide unprecedented opportunity to understand cloud processes. However, satellite imagery is underemployed in cloud studies because climate scientists have had no practical means of analyzing their spatial-temporal patterns. Traditional cloud classifications [Rossow and Schiffer, 1999, World Meteorological Organization] have therefore been based only on mean physical properties over some area, ignoring cloud morphologies and textures.

The difficulty of extracting insight from complex patterns in PB-scale datasets motivates the application of artificial intelligence (AI) algorithms to define and label cloud types that are relevant to climate research. However, supervised learning approaches [e.g., Lee et al., 1990, Rasp et al., 2019, Zantedeschi et al., 2019, Zhang et al., 2018] which rely on human-labeled training datasets, are not useful for research where the relevant classes themselves are not known. Unsupervised learning methods, on the other hand, can discover unknown cloud types relevant to climate change research, with definitions based on spatial distributions as well as means [Kurihana et al., 2019, 2021]. In this work, we leverage the unsupervised deep learning-based, rotation-invariant cloud clustering (RICC) approach of our previous studies [Kurihana et al., 2019, 2021] to create an atlas of cloud classes, and show that these classes can help in understanding observed trends in marine clouds.

2 Unsupervised classification of marine clouds in MODIS images

We use an unsupervised cloud clustering approach to reduce the dimensionality of more than two decades of multispectral satellite imagery from the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments.

2.1 MODIS dataset

The MODIS instruments on the Terra and Aqua satellites have been capturing images of the earth since 2000 and 2002 in 36 "moderate" resolution spectral bands, from visible to thermal range. Each instrument captures a ~2330 km × 2030 km *swath* every five minutes, covering much of the globe at a daily cadence. The total dataset of raw spectral imagery is now nearly 1 PB. We use selected fields from the MODIS Level 1B calibrated radiance product (**MOD02**), Level 1 geolocation fields (**MOD03**), and, for validation, Level 2 derived cloud properties (**MOD06**), all at 1 km resolution. We then extract ~331M non-overlapping ocean-cloud *patches* as the smaller geographical image unit, each of 128 pixels × 128 pixels (~100 km × 100km). We use only 6 of the 36 bands, selecting those most useful for diagnosing cloud optical properties, cloud phase, and cloud height. To avoid the complications of land reflectance, we further restrict sampling to ocean-only by eliminating any patches that include non-ocean pixels, using MOD06 land/water indicator. We also eliminate those with less than 30% cloud pixels, as indicated by the MOD06 cloud mask, yielding a set, `OC-Patches`, of 146M ocean-cloud patches.

2.2 Rotation-invariant cloud clustering: RICC

The goal of the RICC method that we employ in this work [Kurihana et al., 2021] is to generate latent representations from an autoencoder that can assign cloud types to images regardless of orientation when applying hierarchical agglomerative clustering (HAC). Rotation invariance is needed because any particular physically driven cloud pattern can occur in different orientations. A common l^2 loss function is insufficient to achieve this goal because it can generate different representations for an image x and a rotated version of the same image $\mathcal{R}(x)$ (where $\mathcal{R}$ is a set of scalar rotation operators), which may then be placed in different clusters. The RI autoencoder is inspired by a shifted transform invariant autoencoder [Matsuo et al., 2017]. It optimizes L_{RI} (see Equation 1), which combines both
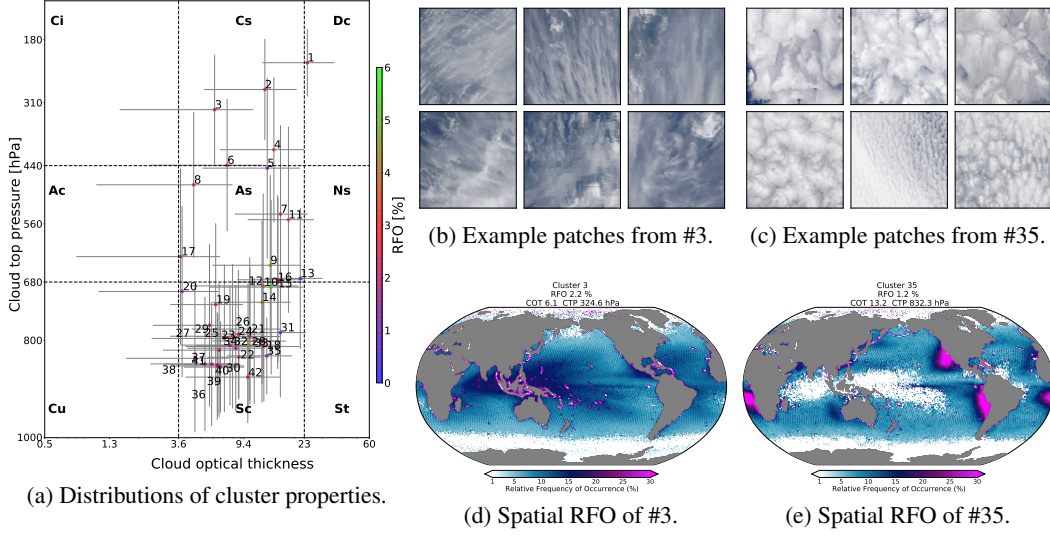
(a) Distributions of cluster properties.

(b) Example patches from #3.

(c) Example patches from #35.

(d) Spatial RFO of #3.

(e) Spatial RFO of #35.

Figure 1: Demonstration of physical reasonableness of of AICCA$_{42}$ classes. (a) Distributions of AICCA$_{42}$ classes in cloud optical thickness (COT) – cloud top pressure (CTP) space. Dots show mean values for each cluster and error bars show the standard deviation of properties. For comparison, dashed lines divide the nine regions corresponding to the International Satellite Cloud Climatology Project (ISCCP) cloud classes [Rossow and Schiffer, 1999, World Meteorological Organization]. Note that AICCA$_{42}$ allows far richer differentiation of cloud types, especially stratocumulus ("Sc"). Panels (b–e) show the texture and distribution of example classes #3 (a type of cirrostratus) and #35 (the dominant class in the stratocumulus decks off the N. and S. American W. coasts). (b–c) examples of MODIS true color images [Gumley et al., 2003] for representative patches from each example class. (d–e) all-season geographic distribution of each example class.

a rotation-invariant loss L_{inv} (which learns the rotation-invariance, to map different orientations of identical input images into an identical representation) and a restoration loss $L_{\text{res}}(\theta)$ (which learns the important spatial structures) to restore spatial patterns in inputs via RI autoencoder. The RI loss is formulated as:

$$L_{\text{RI}}(\theta) = \underbrace{\frac{1}{N} \sum_{x \in S} \sum_{R \in \mathcal{R}} ||D_\theta(E_\theta(x)) - D_\theta(E_\theta(R(x)))||_2^2}_{L_{\text{inv}}(\theta)} + \underbrace{\sum_{x \in S} \min_{R \in \mathcal{R}} ||R(x) - D_\theta(E_\theta(x))||_2^2}_{L_{\text{res}}(\theta)}, \quad (1)$$

where E and D represent the encoder and decoder, respectively; S are training images; and the scalars λ_{inv} and λ_{res} are the weights of the respective loss terms. (We use $(\lambda_{\text{inv}}, \lambda_{\text{res}}) = (32, 80)$.) Optimizing the values of θ produces a latent representation that preserves the spatial structure in x and transforms x and a set of $\mathcal{R}(x)$ to a canonical orientation. Kurihana et al. train the autoencoder at learning rate 0.01 for 100 epochs with 32 NVIDIA V100 GPUs. Our code is available in Github: `https://github.com/RDCEP/clouds`.

2.3 AICCA: An AI-driven Cloud Classification Atlas

We construct AICCA by assigning a cluster label to each of the 146M patches in `OC-Patches`. For each `OC-Patch` x_i, we calculate the cluster label assignment $c_{k,i}$ by identifying the cluster centroid μ_k with the smallest Euclidean distance to its latent representation, $z(x_i)$:

$$c_{k,i} = \underset{k=\{1,\cdots,k^*\}}{\arg\min} ||z(x_i) - \mu_k||_2, \quad (2)$$

where μ_k is a set of k^* cluster centroids by applying a trained RI autoencoder with Equation 1 on 1M patches selected at random from `OC-Patches` and applied HAC to the 74 911 ocean-cloud patches from the year 2003. We select $k^* = 42$ by following a five-criteria evaluation protocol based on quantitative and qualitative tests so that the dataset is formally "AICCA$_{42}$", but for convenience we use here simply "AICCA"; see [Kurihana et al., 2021, 2022] for further details on the protocol. The

3

evaluation protocol addresses the difficulty in unsupervised learning that there is no perfect ground truth against which output can be compared.

The resulting AICCA dataset displays strong explanatory power (Figure 1). It distributes classes in COT–CTP space in a manner of consistent with occurrence frequencies (panel a). The 42 AICCA classes allow far richer cloud information about stratocumulus than does the conventional 9-class scheme: 30 of these classes cover the space of the conventional "Sc" cloud type. The AICCA classes show consistent textures and physically reasonable properties (panels b–c) and physically reasonable geographic distributions (panels d–e). Examples shown here are tropical high-altitude clouds that would be broadly termed "cirrostratus" (class #3) and low-altitude stratocumulus decks that form along the West coast of N. and S. America (class #35).

3   Trends in subtropical stratocumulus

For our case study, we consider whether the AICCA classes can provide insight into a recently discovered cloud trend: low cloud cover has been decreasing in the Pacific Ocean off the coast of Southern California and the Baja peninsula Andersen et al. [2022]. This finding was based on mean MODIS cloud properties alone, with no consideration of cloud patterns. Trends in AICCA classes suggest that the reduction is driven by losses in stratocumulus decks, but that other classes of sparse, optically thinner clouds actually increase in frequency.
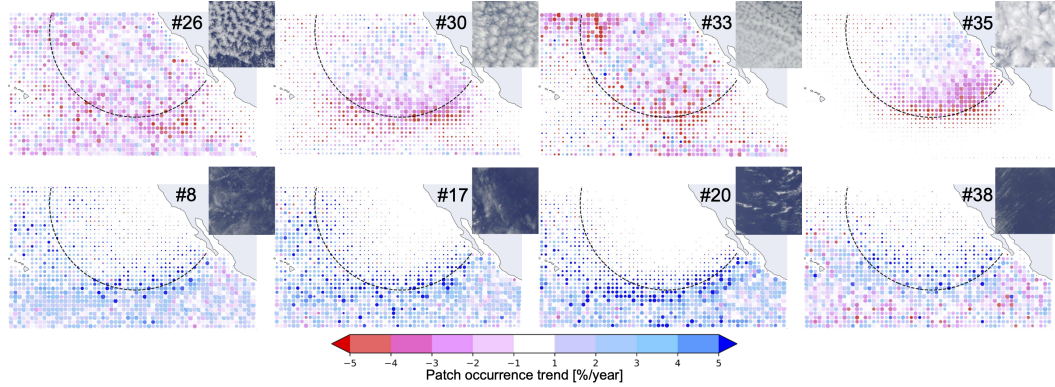

Figure 2: Trends in occurrence of selected AICCA cloud classes over the subtropical N. Pacific Ocean (100–160W and 5–40N). Colors show fitted linear trends over the 18 years 2003–2021, expressed in units of % of the mean value over this time. Dot size represents mean relative frequency of occurrence within a class. Dashed curve highlights the approximate edge of the deck.

Figure 2 shows the trends in 18 years (2003–2021) of MODIS observations off the S. California coast for eight cloud classes. We exclude 2000–2002 because the Aqua and Terra instruments were not operating simultaneously over an entire year during that period. Closed-cell (or transitional) stratocumulus are the most dominant cloud types in this part of the Pacific, making up a quarter of all cloud occurrences (top row, #26, 30, 33, 35, which are four of the five most predominant classes). All these classes decrease strongly, especially along the southern edge of the deck—by as much as 2/3 in some locations. However, several classes of optically thinner clouds increase, especially just south of the deck region (bottom row, #8, 17, 20, and 38). These classes represent very sparse cumulus or alto-cumulus, similar in visual texture but slightly higher in altitude. The combined effect is that sparse classes infiltrate the stratocumulus deck along its unstable edges. (The same analysis applied to ISCCP cloud classifications also shows a decrease in stratocumulus and increase in cumulus clouds, but AICCA cloud classes provide finer details on these transitions.) This trend may be temperature-driven: stratocumulus decks form only when atmospheric conditions are highly stable, and warmer sea surface temperatures tend to reduce stability. Sea-surface temperatures in the region from the ERA5 reanalysis [Hersbach et al., 2020] do increase by nearly 1K over this time period.

The AICCA dataset also lets us examine the temporal evolution of cloud patterns over shorter timescales. For now we consider evolution at fixed (Eulerian) locations, ignoring advection by winds. (Stratocumulus textures are driven by a mix of local environmental control and horizontal advection,

but the Eulerian perspective can provide insight at sub-daily to daily timescales.) Figure 3 shows, as a Sankey diagram, transitions from one time period to the next (~ daily) for the selected classes and region shown in Figure 2. We see significant evolution within closed-cell stratocumulus. The most frequently occurring class, #35, is unsurprisingly also the most stable, and evolves primarily into other closed-cell forms (#30 and #26), which in turn transition into a wide variety of classes. Less than 10% of transitions represent direct evolution from selected closed-cell to selected sparse cloud classes (though still more than the reverse direction). That is, transitions from closed cell classes to sparse cloud classes are largely modulated through a complicated network of mixed type classes.

This result suggests that the replacement of closed-cell by sparse cloud types in Figure 2 does not simply result from a reduction in stratocumulus lifetime.

4   Summary

In summary, the unsupervised cloud clustering approach employed in this work enables reducing the dimensionality of multidecadal MODIS satellite imagery into a tractable set of physically meaningful cloud clusters. It permits the discovery of new classes based on both cloud morphology and physical properties that are unbiased by artificial assumptions and that capture the diversity of global cloud types. We have shown an illustrative example of how the AICCA dataset produced via this approach can be used to track both long-term trends and short-term transitions in cloud morphology, important for understanding cloud feedbacks. More broadly, this work shows the promise of unsupervised learning for unlocking new applications of large satellite datasets in Earth science.
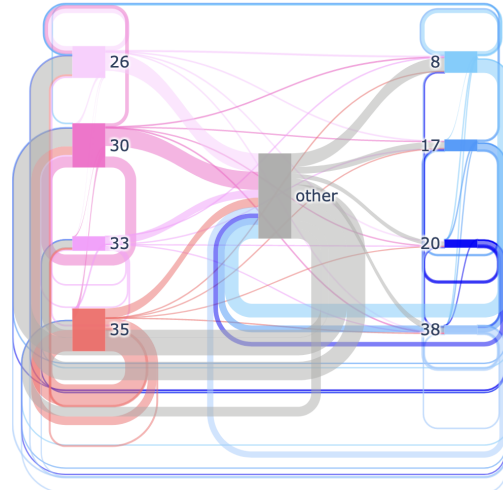

Figure 3: Sankey diagram of sub-daily and daily cloud class transitions within gridcells for the 2003–2022 period for the eight classes, and region, in Figure 2. Box widths represent occurrence frequency of each class, and colors are ordered by trend. Ribbon width represents the rate of transitions from one class type to another. Loops represent persistence of a class from one time period to the next.

Broader Impact

The AICCA dataset that we present here, generated via the unsupervised cloud clustering approach of Kurihana et al., has both immediate uses for cloud research and broader application in Earth science generally.

The AICCA dataset can be applied immediately to some of the most urgent questions in climate science: how clouds alter under conditions of higher CO_2 and warmer temperatures. The dataset is designed in particular to address questions about marine stratocumulus clouds, whose potential breakup is the largest uncertainty in climate projections. While observations will provide the most direct insight, the same approach can also be applied to analyzing the fidelity of cloud representations in Earth system models, which are producing increasingly realistic cloud distributions. Other promising lines for future analysis include the impacts aerosols or of large scale circulation pattern impacts such as ENSO.

More broadly, as the volume of data in Earth sciences explodes, unsupervised deep learning-based classification methods can find uses in many areas. Advances in measurement techniques have brought a deluge of new data, from PBs of satellite images or model output to location data from animal tracking to 3D scanning of fossil collections. Some means is needed to make these datasets tractable. Unsupervised deep learning methods can deliver these complex datasets in compact forms that allow interpretation. Finally, by facilitating access to core data, unsupervised methods can also help democratize climate research, reducing the barriers to entry in the field.

Unsupervised approaches, like any black-box interpretive tool, do offer some dangers. The proliferation of automated classifications not directly based on physical principles could negatively impact scientific fields by pulling researchers further from physical insight. As unsupervised classification grows in popularity, it will be important to produce community standards on how to validate the meaningfulness of the categories produced. Overall, however, the large potential benefits make its use worthwhile.

Acknowledgments and Disclosure of Funding

References

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. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL085782. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782.

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, March 2019. ISSN 1752-0894, 1752-0908. doi: 10.1038/s41561-019-0310-1. URL http://www.nature.com/articles/s41561-019-0310-1.

Tapio Schneider, Colleen M. Kaul, and Kyle G. Pressel. Solar geoengineering may not prevent strong warming from direct effects of CO2 on stratocumulus cloud cover. *Proceedings of the National Academy of Sciences*, page 202003730, November 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2003730117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.2003730117.

Hugh Morrison, Marcus van Lier-Walqui, Ann M. Fridlind, Wojciech W. Grabowski, Jerry Y. Harrington, Corinna Hoose, Alexei Korolev, Matthew R. Kumjian, Jason A. Milbrandt, Hanna Pawlowska, Derek J. Posselt, Olivier P. Prat, Karly J. Reimel, Shin-Ichiro Shima, Bastiaan van Diedenhoven, and Lulin Xue. Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8), August 2020. ISSN 1942-2466, 1942-2466. doi: 10.1029/2019MS001689. URL https://onlinelibrary.wiley.com/doi/10.1029/2019MS001689.

Tapio Schneider, João Teixeira, Christopher S. Bretherton, Florent Brient, Kyle G. Pressel, Christoph Schär, and A. Pier Siebesma. Climate goals and computing the future of clouds. *Nature Climate Change*, 7(1):3–5, January 2017. ISSN 1758-678X, 1758-6798. doi: 10.1038/nclimate3190. URL http://www.nature.com/articles/nclimate3190.

William B Rossow and Robert A Schiffer. Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, 80(11):2261–2288, 1999. URL https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2.

World Meteorological Organization. International Cloud Atlas. https://cloudatlas.wmo.int/.

Jonathan Lee, Ronald C Weger, Sailes K Sengupta, and Ronald M Welch. A neural network approach to cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, 28(5):846–855, 1990. URL https://doi.org/10.1109/36.58972.

S. Rasp, H. Schulz, S. Bony, and B. Stevens. Combining crowd-sourcing and deep learning to understand meso-scale organization of shallow convection, 2019. URL https://doi.org/10.48550/arXiv.1906.01906.

Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt Kusner, and Duncan Watson-Parris. Cumulo: A dataset for learning cloud classes. In *NeurIPS Workshop on Tackling Climate Change with Machine Learning*, 2019. URL https://www.climatechange.ai/papers/neurips2019/11.

J. Zhang, Pu Liu, Feng Zhang, and Qianqian Song. CloudNet: Ground-based cloud classification with deep convolutional neural network. *Geophysical Research Letters*, 45:8665–8672, 2018. URL https://doi.org/10.1029/2018GL077787.

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

Takuya Kurihana, Elisabeth Moyer, Rebecca Willett, Davis Gilton, and Ian Foster. Data-driven cloud clustering via a rotationally invariant autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–25, 2021. URL https://doi.org/10.1109/TGRS.2021.3098008.

Liam Gumley, Jacques Descloitres, and Jeffrey Schmaltz. Creating reprojected true color MODIS images: A tutorial. *University of Wisconsin–Madison*, 19, 2003.

Tadashi Matsuo, Hiroya Fukuhara, and Nobutaka Shimada. Transform invariant auto-encoder. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2359–2364, 2017. URL https://doi.org/10.1109/IROS.2017.8206047.

Takuya Kurihana, Elisabeth J. Moyer, and Ian T. Foster. AICCA: AI-driven cloud classification atlas. *Remote Sensing*, 14(22), 2022. ISSN 2072-4292. doi: 10.3390/rs14225690. URL https://www.mdpi.com/2072-4292/14/22/5690.

Hendrik Andersen, Jan Cermak, Lukas Zipfel, and Timothy A. Myers. Attribution of observed recent decrease in low clouds over the northeastern Pacific to cloud-controlling factors. *Geophysical Research Letters*, 49(3), February 2022. ISSN 0094-8276, 1944-8007. doi: 10.1029/2021GL096498. URL https://onlinelibrary.wiley.com/doi/10.1029/2021GL096498.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot,

Massimo Bonavita, Giovanna Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, July 2020. ISSN 0035-9009, 1477-870X. doi: 10.1002/qj.3803. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803`.

Checklist

If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

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]
   (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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] because we did not need regression tasks.
   (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)? [Yes]

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? [Yes]
   (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]