Combining crowd-sourcing and deep learning to understand meso-scale organization of shallow convection

Stephan Rasp* Hauke Schulz† Sandrine Bony‡ Bjorn Stevens†

The discovery of new phenomena and mechanisms often begins with a scientist's intuitive ability to recognize patterns, for example in satellite imagery or model output. Typically, however, such intuitive evidence turns out to be difficult to encode and reproduce. Here, we show how crowd-sourcing and deep learning can be combined to scale up the intuitive discovery of atmospheric phenomena. Specifically, we focus on the organization of shallow clouds in the trades, which play a disproportionately large role in the Earth's energy balance. Based on visual inspection four subjective patterns or organization were defined: Sugar, Flower, Fish and Gravel. On cloud labeling days at two institutes, 67 participants classified more than 30,000 satellite images on a crowd-sourcing platform. Physical analysis reveals that the four patterns are associated with distinct large-scale environmental conditions. We then used the classifications as a training set for deep learning algorithms, which learned to detect the cloud patterns with human accuracy. This enables analysis much beyond the human classifications. As an example, we created global climatologies of the four patterns. These reveal geographical hotspots that provide insight into the interaction of mesoscale cloud organization with the large-scale circulation. Our project shows that combining crowd-sourcing and deep learning opens new data-driven ways to explore cloud-circulation interactions and serves as a template for a wide range of possible studies in the geosciences.

Together, crowd-sourcing and deep learning offer a new way to discover knowledge from large datasets, which we illustrate on the example of shallow cloud organization.

The human visual system is exquisitely good at identifying patterns. A quick glance at a satellite image, for example, suffices to detect a multitude of interesting features, such as tropical cyclones, extra-tropical fronts or cloud clusters. While subjective, such intuitive pattern recognition can serve as a starting point for understanding new phenomena. Traditionally, however, this intuition is difficult to encode and scale up for statistical analysis.

Here, we combine two emerging tools to tackle this problem: crowd-sourcing and deep learning. Crowd-sourcing describes projects where a task is collaboratively solved by a group of people. This can be a small research group or a large group of internet users. One of the first examples of crowd-sourcing in the natural sciences is Galaxy Zoo[1], a project that has citizen scientists classify different galaxy types and has produced 60 peer-reviewed publications so far. An early meteorological example focused on estimating hurricane intensity (Hennon et al., 2015). Current climate projects on Zooniverse[2][3] ask volunteers to transcribe old, hand-written weather records. Thanks to the collaboration of many individuals such projects produce a wealth of data that would be unattainable for a single scientist.

Deep learning is a sub-field of machine learning based on multi-layered networks that has seen a surge in popularity in recent years. In particular, computer vision and natural language processing have been revolutionized by the switch from hard-coded, rule-based algorithms towards data-driven approaches (LeCun et al., 2015). Deep neural networks also have many potential applications in the Earth sciences, particularly where already existing deep learning techniques can be transferred to geoscientific problems (Reichstein et al., 2019). A perfect example of this is the

* Ludwig-Maximilian-University, Munich, Germany. Corresponding author: s.rasp@lmu.de
† Max Planck Institute for Meteorology, Hamburg, Germany
‡ Sorbonne Université, LMD/IPSL, CNRS, Paris, France

1 https://www.zooniverse.org/projects/zookeeper/galaxy-zoo
2 https://www.zooniverse.org/projects/edh/weather-rescue
3 https://www.zooniverse.org/projects/drewdeepsouth/southern-weather-discovery

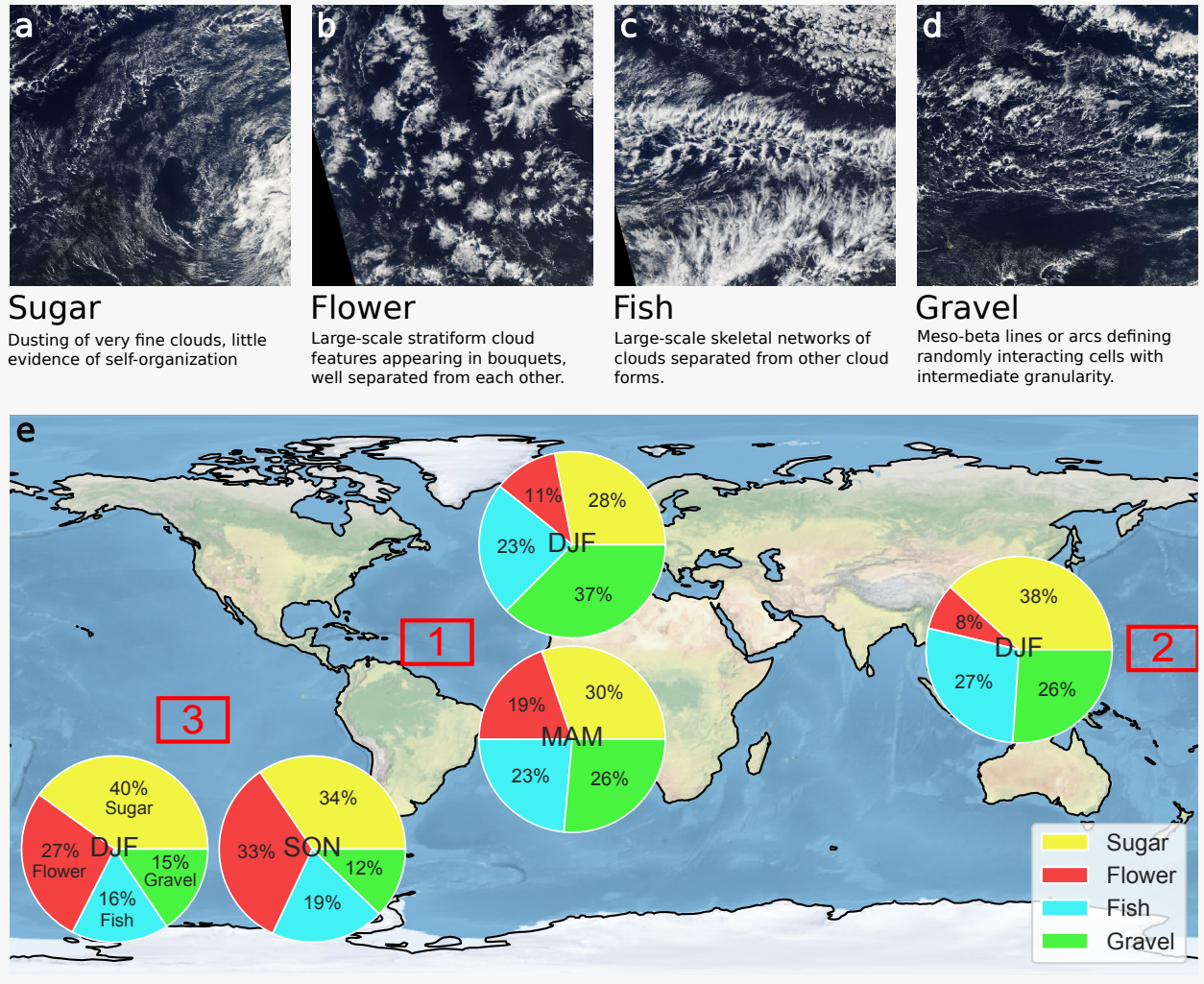

Figure 1: *(a–d) Examples of the four cloud organization patters. (e) Worldmap showing the three regions selected for the Zooniverse project. Pie charts show the area fractions of the human classifications for the regions and seasons.*

detection of features in images. One obstacle is that deep learning requires a large number, typically several thousands, of hand-labeled training samples. For Earth science problems, these are usually not available. For this reason, previous studies that used deep neural networks to detect atmospheric features relied on training data created by traditional, rule-based algorithms (Racah et al., 2016; Liu et al., 2016; Hong et al., 2017; Kurth et al., 2018).

Here, we present a community project that used a combination of crowd-sourcing and deep learning to tackle the question of mesoscale organization of shallow clouds, a topic of high relevance for the Earth's climate. In this paper, we will describe how we set up our project and what we learned from it, scientifically and organizationally. Further, we hope to convince fellow scientists that the approach presented here is a feasible way to tackle a number of research questions in the geosciences.

MESOSCALE ORGANIZATION OF SHALLOW CLOUDS

Shallow cumulus clouds might look innocent compared to their cumulonimbus counterparts but in terms of their importance for the global energy balance, they play a disproportionately large role. This has two reasons: first, they reflect a significant portion of the incoming solar radiation back to space while only contributing marginally to the greenhouse effect, thereby cooling our planet; and second, shallow cumulus cover large fractions of our planet's subtropical oceans (Bony et al., 2004). The global net

radiative effect of shallow cumulus is estimated to be around -20 $\mathrm{W\,m^{-2}}$ (Boucher et al., 2013), which illustrates that even small changes in cloud cover are important. Disagreement about these changes are also though to be the major cause for the uncertainty in model-based estimates of climate sensitivity (Bony and Dufresne, 2005; Vial et al., 2013; Stevens et al., 2016a). Understanding the mechanisms behind shallow cloud formation, therefore, is crucial.

Contrary to the textbook view of shallow cumulus, they are typically not horizontally uniform but exhibit a wide range of patterns on the mesoscale (20–200 km, meso-β; Young et al., 2002; Wood and Hartmann, 2006). So far the mechanisms driving many of these patterns are poorly understood. These modes of organization, however, could play a major role for the radiative effect of shallow clouds, a fact that has long been recognized for deep convection (Tobin et al., 2012). In today's climate models the typical assumption is that of a scale separation between the large and small scales, where the latter are slaved to the former. Mesoscale organization, therefore, is largely ignored, a potential cause of biases.

A first step towards better understanding shallow cumulus organization is to define it. While some patterns are easily detectable with traditional techniques (Muhlbauer et al., 2014), many other forms of organization are more ambiguous. Recently, twelve cloud experts browsed through hundreds of NASA Worldview[4] images and identified four frequently recurring cloud patterns, which they evocatively named Sugar, Flower, Fish and Gravel (Stevens et al., 2019b, Fig. 1a–d). The four categories were chosen entirely subjectively based on visual intuition without any climatological analysis. Flower and Gravel loosely resemble closed and open cell convection (Atkinson and Zhang, 1996), which occur behind mid-latitude cold fronts, but quite possibly involve different mechanisms.

In this first labeling exercise, 1000 images were labeled. This dataset already provided some clues about the physical mechanisms behind the cloud patterns and an initial machine learning model encouraged us to scale up our efforts. This is where crowd-sourcing and deep learning come in.

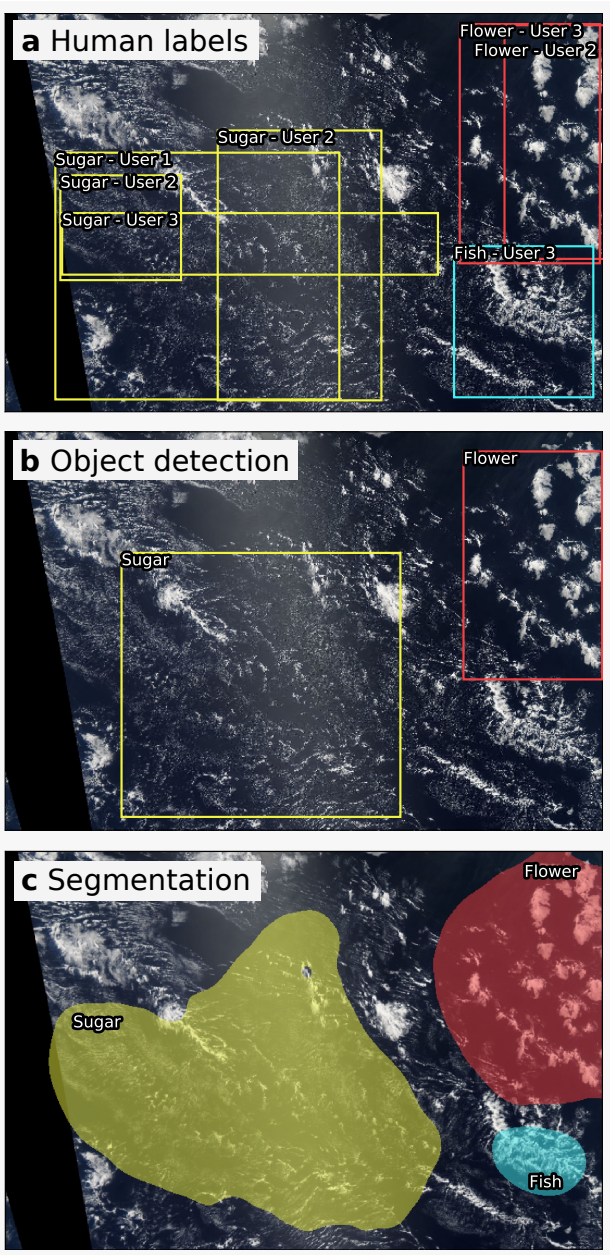

Figure 2: *(a) Crowd-sourced classifications. This validation image was labelled by three different users. Predictions of (b) the Retinanet object detection algorithm and (c) the image segmentation algorithm.*

CROWDSOURCING THE CLIMATE COMMUNITY

To obtain a large pool of classified images, we set up a cloud labeling interface on Zooniverse[5], an open web platform that enables researchers to organize and

[4] https://worldview.earthdata.nasa.gov/
[5] https://www.zooniverse.org/projects/raspstephan/sugar-flower-fish-or-gravel

present research questions in ways that enable contributions from the broader public. For our project we downloaded roughly 10,000 21° longitude by 14° latitude Terra and Aqua MODIS visible images from NASA Worldview. To select the regions and seasons, we started with the boreal winter east of Barbados as a reference. Barbados is home to the Barbados Cloud Observatory (Stevens et al., 2016b) and is the hub for field campaigns, both past (Stevens et al., 2019a) and upcoming (Bony et al., 2017), aiming to investigate shallow clouds. By identifying climatological factors thought to be important for shallow cloud formation we identified and subsequently added two further regions in the Pacific which are climatologically similar (Fig. 1e; see Methods for details). Images were downloaded for an eleven year period from 2007 to 2017.

On the web interface, participants are served an image randomly drawn from our library of 10,000 images. Users were then asked to draw rectangles around regions where one of the four cloud patterns dominates (Fig. 2a). Participants had the possibility to draw any number of boxes, including none, with the caveat that the box would cover at least 10% of the image. When an image was classified by four different users, it was retired, i.e. removed from the image library. No user was shown the same image twice. With the interface in place, cloud classification days were set up at the Max Planck Institute for Meteorology in Hamburg, Germany on Nov 2nd and at the Laboratoire de Météorologie Dynamique in Paris, France on Nov 29th 2018. After a brief instruction at the start of the day and a warm-up on a practice dataset, 67 participants, most of them researchers of the two institutes, labeled images for an entire day yielding roughly 30,000 classification, i.e. around three classifications per image. On average, participants needed around 30 seconds to classify one image, amounting to approximately 250 h of human classification. Overall, the four patterns occurred with roughly similar frequency but with notable differences depending on the geographic region and season (Fig. 1e).

HUMAN AGREEMENT AND PHYSICAL INTERPRETATION OF PATTERNS

The first key question of this labeling exercise is to which extent the human labelers even agreed on the subjectively chosen cloud patterns. Analysis of data from a small expert group suggested that there would be sufficient agreement (Stevens et al., 2019a). But could this be extended to a much larger group of people with less expertise in the subject? To find out we computed the following agreement metric: "In which percentage of cases, if one user drew a box of a certain class, did another user also draw a box of the same class, under the condition that the boxes overlap?" (see Methods for details). Overall, the agreement is 43% but there are notable differences between the four patterns (Fig. 3a). Humans agree most on Flowers (51%) while Fish (37%) are most controversial. Considering that some disagreement has to be expected for such subjective classes, these number and visual inspection of many classified images[6] lead us to conclude that the four patterns have some validity and could be communicated to and subsequently identified by a group of non-experts.

A second central question is whether the four patters, which were purely chosen based on their visual

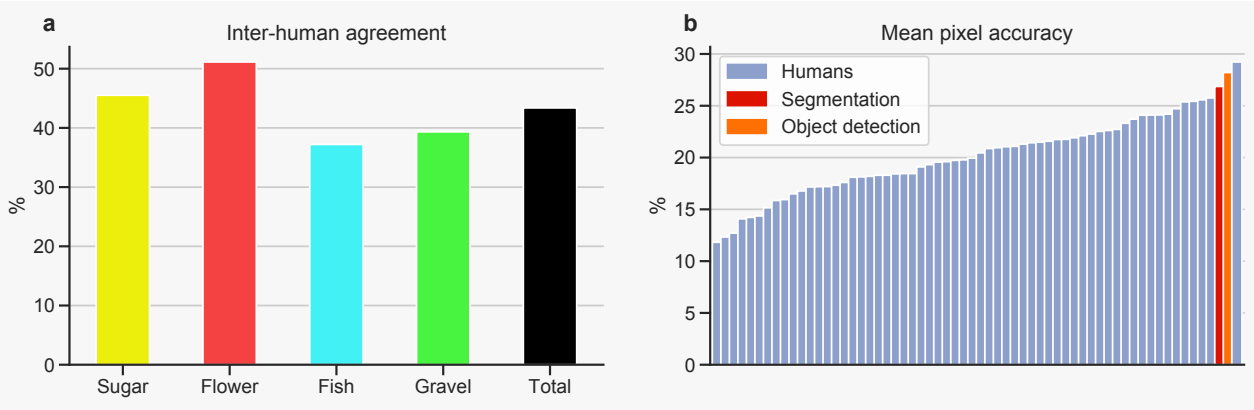

Figure 3: *(a) Mean agreement between humans. (b) Mean pixel accuracy for each human participant and the two deep learning algorithms. Definitions of the metrics can be found in the Methods.*

[6] `http://tiny.cc/w1th6y`

appearance on satellite imagery, actually correspond to physically meaningful cloud regimes. To investigate this, we created composites of the large-scale conditions from ERA-Interim reanalyses[7] corresponding to each pattern (Fig. 4). These composites suggest that Sugar, Flower, Fish and Gravel appear in climatologically distinct environments. Flowers are associated with a relatively dry and cold boundary layer with a very strong inversion. Sugar on the other hand appears in warm and humid boundary layers with strong downward motion in boundary layer. For Fish and particularly Gravel, on the other hand, the inversion and downward motion is rather weak.

DEEP LEARNING SCALES UP HUMAN INTUITION

The 30,000 human classifications already provide a rich dataset which can be used to better understand the four patterns. However, even after 250 hours of labeling images, the classifications only cover a small fraction of the globe for a small fraction of the time. In fact, only around 0.6% of the data available during the selected eleven year period were labeled. Deep learning allows us to scale up this analysis by many orders of magnitude.

The cloud classification task presented here can be framed as one of two potential machine learning problems: object detection and semantic segmentation. Object detection algorithms draw boxes around features of interest, thereby exactly mirroring the human workflow for this task. In contrast, segmentation algorithms classify every pixel of the image. Fig. 2a,b shows examples of these two approaches for an image from a validation dataset that was not used during training (see `http://tiny.cc/w1th6y` for more randomly chosen examples). Details about the neural network architectures and preprocessing steps can be found in the Methods. Both types of algorithm accurately detect the most obvious patterns in the image and agree well with human labels. Interestingly, despite all training labels being rectangular, the segmentation algorithm learns to focus on the actual shape of the patterns.

To quantitatively compare the deep learning algorithms against the human labelers, we compute the mean accuracy, the percentage of correctly labeled pixels (see Methods), for each human and the two algorithms (Fig. 3b;). Both, the object detection algorithm and the segmentation algorithm, show a large consensus with the average human labels for a random validation set. The accuracy (Supp. Fig. S2) is higher for patterns where humans agreed. Further, the algorithms more frequently predicted patterns with a higher inter-human agreement, i.e. Flowers and Sugar. These results confirm that the deep learning models are able to detect the cloud patterns on par with human labelers.

Deep learning opens up many opportunities, primarily because the algorithms are very fast (less than one second to classify an image) and never tire. This makes it possible go much beyond the original, human dataset by applying the algorithm for the entire globe (Fig. 5a; see Methods for details). Caution is

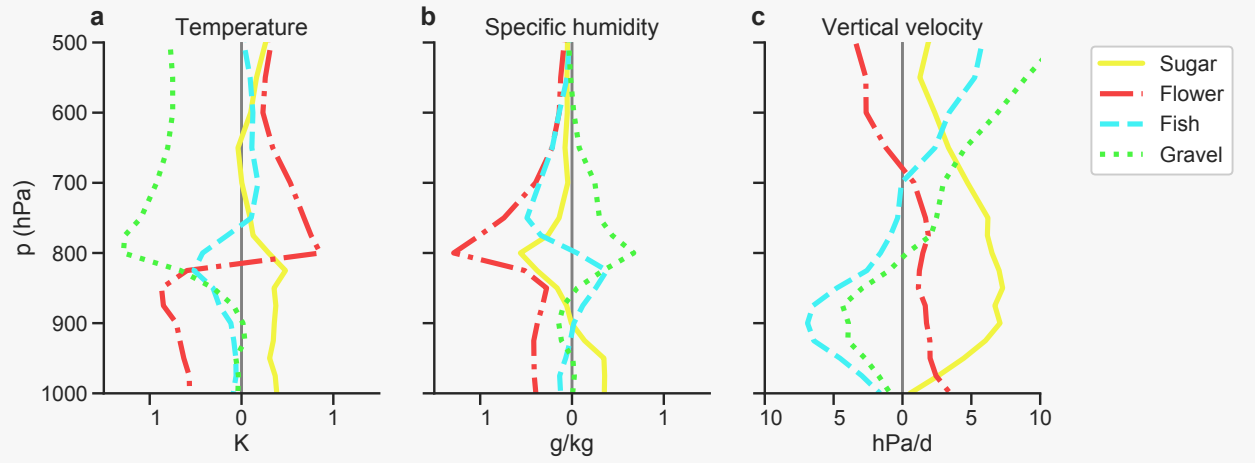

Figure 4: *Median of large-scale environmental conditions corresponding to the four patterns as identified by the human labelers. Figures show deviations of (a) temperature, (b) specific humidity and (c) vertical velocity relative to the climatological mean.*

[7] `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim`

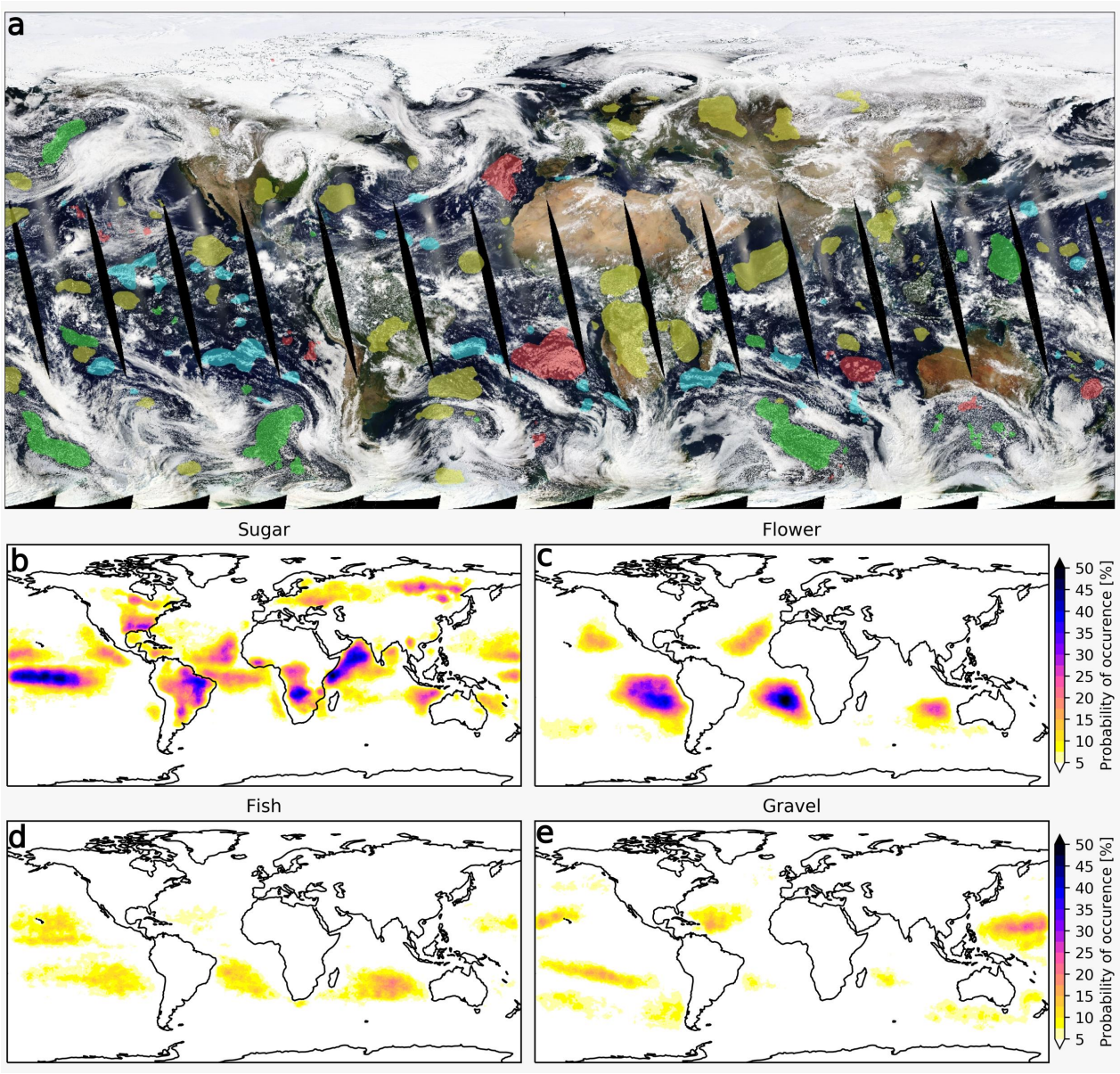

Figure 5: *(a) Global predictions of the image segmentation algorithm for May 1 2017. The colors are the same as in the previous figures. For more examples, see* `http://tiny.cc/fsth6y` *(b–e) Heatmaps of the four patterns for the year 2017.*

always advisable when applying machine learning algorithms outside of their training regime (Rasp et al., 2018). A visual inspection of the global maps (see `http://tiny.cc/fsth6y` for more examples), however, suggests that the algorithm's predictions are reasonable and physically interpretable as discussed below. Naturally, over land the predictions have to be assessed with greater care because no land was present in the training dataset. Nevertheless, the algorithm appears to correctly identify shallow cumuli over the tropical landmasses as sugar.

To obtain global climatologies of Sugar, Flower, Fish and Gravel we ran the algorithm on daily global images for the entire year of 2017 (Fig. 5b–e). This took only a few hours of computing on a single processor. As a comparison, the same feat would require more than 600 human hours. The heatmaps reveal coherent hotspots for the four cloud patterns. Sugar occurs predominantly north and south of the Inter-Tropical Convergence Zone (ITCZ) and in the Ara-

bian Sea. Flowers appear just west of the continents where the trade wind inversion is strong, which is in agreement with the large-scale composites (Fig. 4). Further downstream in the trade regions, Flowers transition to Gravel and Fish. These two patterns are geographically intertwined, which again confirms the physical analysis. Interestingly, Gravel seems to be relatively confined, in particular to our selected regions 1 and 2 (Fig. 1e). This contradicts our previous view of cold pool patterns dominating the trade wind regions globally (Rauber et al., 2007). Fish, which are linked to stronger convergence (Fig. 4c) are often associated with synoptic convergence lines, sometimes connected to trailing mid-latitude fronts. The physical coherence of the climatologies provide another piece of evidence that the four subjectively chosen patterns code for meteorologically meaningful regimes and that the deep learning algorithms are able to provide new insight. An unanswered question is whether important regimes of shallow cloud organization are missing, which could be tackled with emerging deep unsupervised learning approaches (Xie et al., 2016; Caron et al., 2018).

NEW OPPORTUNITIES

This project was an experiment for us. Without direct precedent it was hard to judge beforehand whether the results would turn out useful or not. Fortunately, they did. The combination of crowd-sourcing and deep learning allowed us to better understand the mesoscale organization of shallow clouds, a topic that has turned out to be quite elusive. The four patterns, that were defined subjectively based on their appearance–albeit, of course, by experienced cloud researchers–are physically meaningful and were identifiable by a large number of people, most of whom, while they share an atmospheric science background, are not versed in the study of low-clouds. The deep neural networks learned to classify the satellite images with human accuracy, despite considerable uncertainty in the training dataset, and were able to extrapolate beyond the human classifications. The physical insight gained from the first analyses presented in this paper revealed new information, some of which contradicts our preconceptions about these cloud patterns.

The results here are just a teaser of knowledge to be gained from this dataset. Because of the importance of shallow clouds for climate, ongoing and future analysis will focus on the radiative properties associated with mesoscale organization and how the environmental conditions in future climates might affect the frequency of occurrence of each pattern. For this, it might be helpful to use the existing classifications with different types of satellite images. Infrared imagery, for example, allows classification at all times, while geostationary satellites provide a much higher temporal resolution. This may help understand how shallow cloud organization develops over time. Ultimately, understanding better how clouds interact on the mesoscale and how these interactions affect the energy balance of our planet will hint at what current climate models, where the mesoscale is largely ignored, are missing.

In addition to the gain in scientific knowledge, we learned that such a project is feasible from an organizational point of view. Platforms like Zooniverse make setting up crowd-sourcing interfaces fast and easy. It is as simple as uploading the data and specifying which task users should complete (categorizing the entire image or drawing shapes). The results can then be downloaded as a tabular data file. Similarly, deep learning has become much more user-friendly over the last couple of years. Free online courses[8] and easy-to-use Python libraries such as Keras (Chollet and Others, 2015) and fastai[9] allow non-computer scientists to apply state-of-the-art machine learning models in a short amount of time. Further, for most common tasks in image processing, such as object detection and image segmentation in the paper, pre-existing and pre-trained neural network architectures are available, which make it convenient to transfer existing technologies to new tasks (see Methods for details on the models used in this paper). The computational demand is also manageable. For the networks used in this study, training took on the order of ten hours on a single graphics processing unit (GPU). GPUs are now available in most scientific computing centers and for rent on web computing services.

Crowd-sourcing is a solution for one of the big problems when applying deep learning in many scientific disciplines: the lack of labeled training data. With this study we hope to convince fellow researchers that the effort required to create enough training data is manageable. As a rough estimate of how much data is required, we trained our networks with less data and found that useful results can still be obtained with 5,000–10,000 classifications. This translates to a day of classification for around 15 people, which is within the capabilities of even small research groups. Of course, the amount of required training images

[8] `https://course.fast.ai/`, `https://deeplearning.ai`
[9] `https://docs.fast.ai/`

depends strongly on the complexity of the task. For our project, most participants were climate scientists. Another interesting question is how good the classifications would be if they were done by the general public, as has been common for most previous crowdsourced science projects.

Similar projects could be useful for a wide range of research questions in the geosciences. Typically, if a feature is easy to identify by eye but hard to objectively define, a subjective crowd-sourced approach could be a feasible way to harness human intuition on a statistically significant scale.

Data availability

This dataset will be used for a Kaggle[10] competition. To ensure a fair competition, the raw data will stay private for now. Interested researchers are encouraged to contact us directly to obtain the data, deep learning models and Jupyter notebooks for analysis. After the Kaggle competition has finished the repository will be made public. The Zooniverse project is still online[11], where readers can try labeling clouds themselves.

Acknowledgements

First and foremost, we would like to thank all the participants of the cloud labeling days. Special thanks go to Ann-Kristin Naumann and Julia Windmiller for initiating this collaboration and to Katherine Fodor for suggesting Zooniverse. SR acknowledges funding from the German Research Foundation Project SFB/TRR 165 "Waves to Weather". This paper arises from the activity of an International Space Science Institute (ISSI) International Team researching "The Role of Shallow Circulations in Organising Convection and Cloudiness in the Tropics". Additional support was provided by the European Research Council (ERC) project EUREC4A (Grant Agreement 694768) of the European Union's Horizon 2020 Research and Innovation Programme and by the Max Planck Society. We acknowledge the use of imagery from NASA Worldview, part of the NASA Earth Observing System Data and Information System (EOSDIS).

References

Atkinson, B. W., and J. W. Zhang, 1996: Mesoscale Shallow Convection in the Atmosphere. *Reviews of Geophysics*, doi:10.1029/96RG02623.

Bony, S., and J. Dufresne, 2005: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophysical Research Letters*, **32 (20)**, L20 806, doi:10.1029/2005GL023851.

Bony, S., J.-L. Dufresne, H. Le Treut, J.-J. Morcrette, and C. Senior, 2004: On dynamic and thermodynamic components of cloud changes. *Climate Dynamics*, **22 (2)**, 71–86, doi:10.1007/s00382-003-0369-6.

Bony, S., and Coauthors, 2017: EUREC4A: A Field Campaign to Elucidate the Couplings Between Clouds, Convection and Circulation. *Surveys in Geophysics*, **38 (6)**, 1529–1568, doi:10.1007/s10712-017-9428-0.

Boucher, O., and Coauthors, 2013: *Clouds and aerosols*, 571–657. Cambridge University Press, Cambridge, UK, doi:10.1017/CBO9781107415324.016.

Caron, M., P. Bojanowski, A. Joulin, and M. Douze, 2018: Deep Clustering for Unsupervised Learning of Visual Features. https://arxiv.org/abs/1807.05520.

Chollet, F., and Others, 2015: Keras. https://keras.io/.

Gneiting, T., and A. E. Raftery, 2007: Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378, doi:10.1198/016214506000001437.

He, K., X. Zhang, S. Ren, and J. Sun, 2015: Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385.

Hennon, C. C., and Coauthors, 2015: Cyclone Center: Can Citizen Scientists Improve Tropical Cyclone Intensity Records? *Bulletin of the American Meteorological Society*, **96 (4)**, 591–607, doi:10.1175/BAMS-D-13-00152.1.

Hong, S., S. Kim, M. Joh, and S.-k. Song, 2017: GlobeNet: Convolutional Neural Networks for Typhoon Eye Tracking from Remote Sensing Imagery. http://arxiv.org/abs/1708.03417.

Kurth, T., and Coauthors, 2018: Exascale Deep Learning for Climate Analytics. http://arxiv.org/abs/1810.01993.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521 (7553)**, 436–444, doi:10.1038/nature14539.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár, 2017: Focal Loss for Dense Object Detection. http://arxiv.org/abs/1708.02002.

Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, 2016: Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets. https://arxiv.org/abs/1605.01156.

Muhlbauer, A., I. L. McCoy, and R. Wood, 2014: Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmospheric Chemistry and Physics*, **14 (13)**, 6695–6716, doi:10.5194/acp-14-6695-2014.

Racah, E., C. Beckham, T. Maharaj, S. E. Kahou, Prabhat, and C. Pal, 2016: ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. http://arxiv.org/abs/1612.02095.

[10] `https://www.kaggle.com/`

[11] `https://www.zooniverse.org/projects/raspstephan/sugar-flower-fish-or-gravel`

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, **115 (39)**, 9684–9689, doi:10.1073/pnas.1810286115.

Rauber, R. M., and Coauthors, 2007: Rain in Shallow Cumulus Over the Ocean: The RICO Campaign. *Bulletin of the American Meteorological Society*, **88 (12)**, 1912–1928, doi:10.1175/BAMS-88-12-1912.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566 (7743)**, 195–204, doi:10.1038/s41586-019-0912-1.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. http://arxiv.org/abs/1505.04597.

Stevens, B., S. C. Sherwood, S. Bony, and M. J. Webb, 2016a: Prospects for narrowing bounds on Earth's equilibrium climate sensitivity. *Earth's Future*, **4 (11)**, 512–522, doi:10.1002/2016EF000376.

Stevens, B., and Coauthors, 2016b: The Barbados Cloud Observatory: Anchoring Investigations of Clouds and Circulation on the Edge of the ITCZ. *Bulletin of the American Meteorological Society*, **97 (5)**, 787–801, doi:10.1175/BAMS-D-14-00247.1.

Stevens, B., and Coauthors, 2019a: A high-altitude long-range aircraft configured as a cloud observatory–the NARVAL expeditions. *Bulletin of the American Meteorological Society*, BAMS–D–18–0198.1, doi:10.1175/BAMS-D-18-0198.1.

Stevens, B., and Coauthors, 2019b: Sugar, Gravel, Fish, and Flowers:Mesoscale cloud patterns in the Tradewinds. *Quart. J. Roy. Meteor. Soc.*, **submitted**.

Tobin, I., S. Bony, and R. Roca, 2012: Observational Evidence for Relationships between the Degree of Aggregation of Deep Convection, Water Vapor, Surface Fluxes, and Radiation. *Journal of Climate*, **25 (20)**, 6885–6904, doi:10.1175/JCLI-D-11-00258.1.

Vial, J., J.-L. Dufresne, and S. Bony, 2013: On the interpretation of inter-model spread in cmip5 climate sensitivity estimates. *Climate Dynamics*, **41 (11)**, 3339–3362, doi:10.1007/s00382-013-1725-9.

Wood, R., and D. L. Hartmann, 2006: Spatial Variability of Liquid Water Path in Marine Low Cloud: The Importance of Mesoscale Cellular Convection. *Journal of Climate*, **19 (9)**, 1748–1764, doi:10.1175/JCLI3702.1.

Xie, J., R. Girshick, and A. Farhadi, 2016: Unsupervised Deep Embedding for Clustering Analysis. http://arxiv.org/abs/1511.06335v2.

Young, G. S., D. A. R. Kristovich, M. R. Hjelmfelt, and R. C. Foster, 2002: Rolls, Streets, Waves, and More: A Review of Quasi-Two-Dimensional Structures in the Atmospheric Boundary Layer. *Bulletin of the American Meteorological Society*, **83 (7)**, 997–1001, doi:10.1175/1520-0477(2002)083⟨0997:RSWAMA⟩2.3.CO;2.

Supplemental Methods

Region selection criteria

The regions were selected ahead of the classification days according to a similarity analysis of atmospheric conditions that resemble the conditions encountered during the DJF season east of Barbados where these patterns were first found (Stevens et al., 2019b).

Because the mesoscale organization of shallow cumulus is a relatively new research topic, the meteorological conditions influencing it are primarily an educated guess. Lower tropospheric stability (LTS), surface wind speed (FF) and total integrated column water vapour (TCWV) are three parameters one could naively imagine to describe the meteorological setting to a sufficient degree. Starting with the inter-annual seasonal mean of these atmospheric properties at the region east of Barbados, we searched for climatologically similar regions and seasons within a 120°-wide latitudinal belt (60°N to 60°S) around the globe. We used a k-means clustering with eight clusters to find similar patterns within our search perimeter. As input to the algorithms we used the climatological means of LTS, FF10 and TCWV for each of the four seasons. The eight clusters explain more than 90% of the variance in the dataset and provide large enough regions to fit 21° longitude by 14° latitude boxes reasonably well.

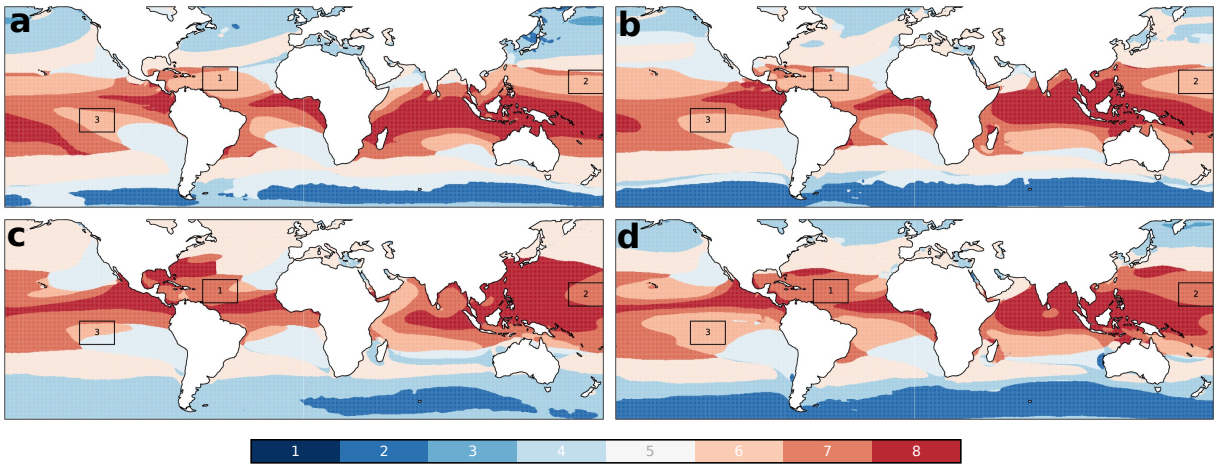

Figure S1: *Cluster analysis of LTS, FF10, TCWV separated by season (DJF, MAM, JJA, SON). The colors identify the 8 clusters as a result of the k-means algorithm. For a better visual impression the clusters are sorted by cluster mean column integrated moisture with cluster 1 being the driest. Black boxes indicate regions chosen for human-classifications.*

Fig. S1 shows the clusters for the four seasons. Our analysis indicates that the meteorological conditions over the Northwestern Atlantic change with season. This is not surprising due to the migration of the ITCZ, but it illustrates that we shouldn't expect to see the same cloud patterns or at least the same distribution throughout the year. The final choice of seasons and regions was made to match the climate of region 1 in DJF (Table S1)

Table S1: *Selected domains used for human-classification of cloud patterns.*

Domain	Bounds	Seasons used
1	-61°E -40°E; 10°N 24°N	DJF, MAM
2	159°E 180°E; 8°N 22°N	DJF
3	-135°E -114°E; -1°N -15°N	DJF, SON

Agreement metrics

In the paper we use two different metrics for agreement. First, the agreement score, used to compare the inter-human agreement, defined as follows: "In which percentage of cases, if one user drew a box of a certain class, did another user also draw a box of the same class, under the condition that the boxes overlap." The overlap is measured using the Intersection-over-Union (IoU) metric. For above metric an IoU of larger than 0.1 is required. While this value might seem low, for two equally sized boxes this actually indicates an overlap of 20%, almost one quarter of the box. Changing the threshold changes the absolute values but not the relative agreement for each of the patterns. To measure inter-human agreement, for each image all combinations of two users are compared against each other and subsequently averaged.

The second metric is the pixel accuracy used to compare the machine learning models to the human predictions. Here, for each pixel, the accuracy of one user (or a machine learning prediction) compared to another user is computed for each pattern. Pixels where both users predict no pattern are omitted for this score.

The reason for using two different metrics is that while the first metric is easily understandable and interpretable, it is not a proper metric (Gneiting and Raftery, 2007). This means that predicting the truth does not necessarily give the best score. For example, because of the IoU threshold, predicting larger boxes would result in a higher agreement score. The pixel agreement, in contrast, is a proper score and is therefore suited to compare inter-human agreement with the two deep learning algorithms.

Deep learning models

Two deep learning models are used, one for object detection and one for semantic segmentation. For object detection, an algorithm called Retinanet (Lin et al., 2017) is used. Here we used the following implementation in Keras (Chollet and Others, 2015): `https://github.com/fizyr/keras-retinanet`, which uses a Resnet50 (He et al., 2015) backbone. The original images had a resolution of 2100 by 1400 pixels. For Retinanet the images were downscaled to 1050 by 700 pixels. This is necessary to fit the batch (batch size = 4) into GPU RAM.

For semantic segmentation, we first converted each human classification, i.e. all boxes by one user for an image, to a mask. Sometimes boxes for different patterns overlap. In this case, the mask is chosen to represent the value of the smaller box. Overall, the amount of overlapping boxes is small, however, so that the resulting error is most likely negligible. To create a segmentation model, we used the fastai Python library[12]. The network architecture has a U-Net (Ronneberger et al., 2015) structure with a Resnet50 backbone. For the segmentation model the images were downscaled to 700 by 466 pixels (batch size = 6).

To create the prediction masks, first a Gaussian filter with a half-width of 10 pixels was applied to smooth the predicted field. Then, for each pixel the highest probability for each of the four patterns was used, if this probability exceeded 30%. This last step counteracts the tendency to predict background, which is by far the most common class in the training set.

Global heatmaps

To create the heatmaps, the segmentation algorithms was used. Predictions were created for a 21° longitude by 14° latitude region at a time, with a windows sliding in 10.5° and 7° increments over the globe. The highest pattern probability for the overlapping images was then taken to create the global mask. This was necessary because the algorithm tends to predict background at the edges of the image, a consequence of the human labelers not drawing boxes that extend all the way to the edge of the image. The climatology was created from one year of Aqua data.

[12] `https://docs.fast.ai/`

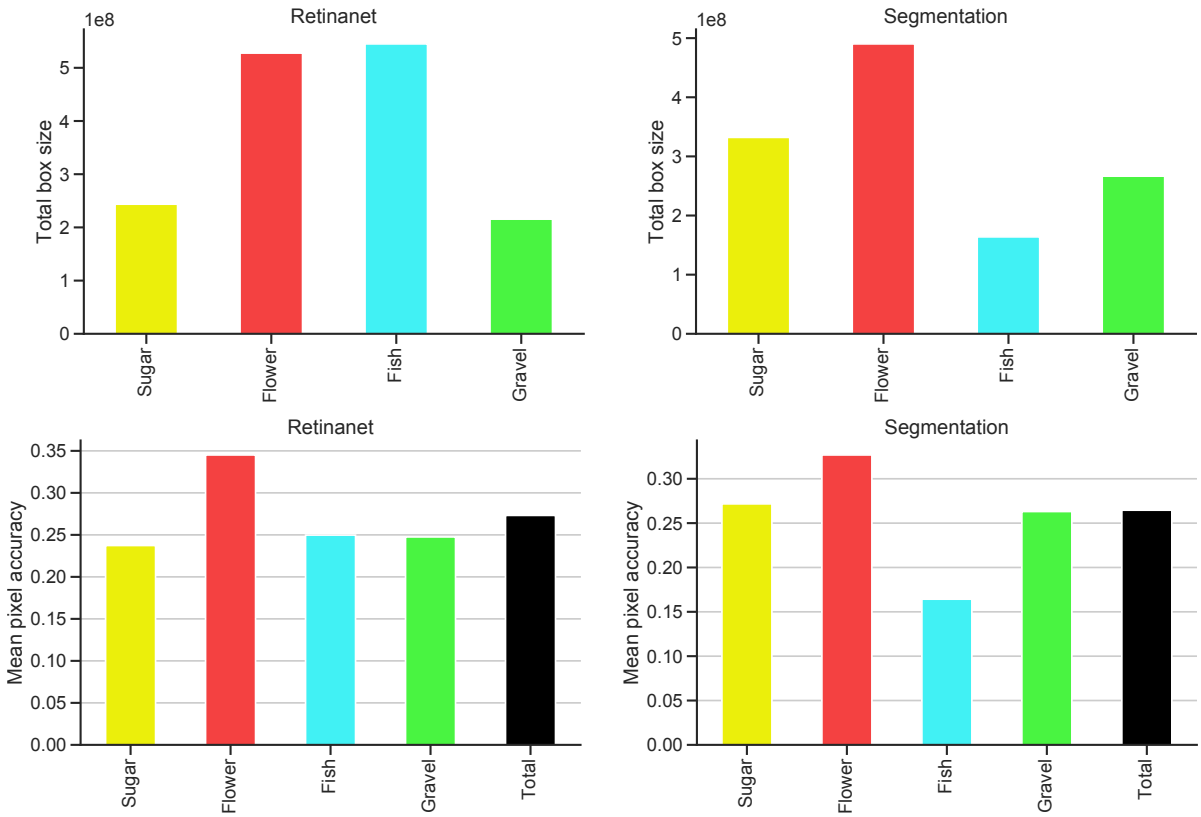

Figure S2: *(Top row) Total size of classifications for the two deep learning algorithms for a random validation dataset. (Bottom row) Mean pixel accuracy for the two algorithms stratified by pattern, also for a random validation set.*