Combining Crowdsourcing and Deep Learning to Explore the Mesoscale Organization of Shallow Convection

Stephan Rasp, Hauke Schulz, Sandrine Bony, Bjorn Stevens

▶ **To cite this version:**

Stephan Rasp, Hauke Schulz, Sandrine Bony, Bjorn Stevens. Combining Crowdsourcing and Deep Learning to Explore the Mesoscale Organization of Shallow Convection. Bulletin of the American Meteorological Society, 2020, 10.1175/BAMS-D-19-0324.1 . hal-03022274

HAL Id: hal-03022274

https://hal.sorbonne-universite.fr/hal-03022274

Submitted on 24 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BAMS Article

Combining Crowdsourcing and Deep Learning to Explore the Mesoscale Organization of Shallow Convection

Stephan Rasp, Hauke Schulz, Sandrine Bony, and Bjorn Stevens

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

ABSTRACT: Humans excel at detecting interesting patterns in images, for example, those taken from satellites. This kind of anecdotal evidence can lead to the discovery of new phenomena. However, it is often difficult to gather enough data of subjective features for significant analysis. This paper presents an example of how two tools that have recently become accessible to a wide range of researchers, crowdsourcing and deep learning, can be combined to explore satellite imagery at scale. In particular, the focus is on the organization of shallow cumulus convection in the trade wind regions. Shallow clouds play a large role in the Earth's radiation balance yet are poorly represented in climate models. For this project four subjective patterns of organization were defined: Sugar, Flower, Fish, and Gravel. On cloud-labeling days at two institutes, 67 scientists screened 10,000 satellite images on a crowdsourcing platform and classified almost 50,000 mesoscale cloud clusters. This dataset is then used as a training dataset for deep learning algorithms that make it possible to automate the pattern detection and create global climatologies of the four patterns. Analysis of the geographical distribution and large-scale environmental conditions indicates that the four patterns have some overlap with established modes of organization, such as open and closed cellular convection, but also differ in important ways. The results and dataset from this project suggest promising research questions. Further, this study illustrates that crowdsourcing and deep learning complement each other well for the exploration of image datasets.

https://doi.org/10.1175/BAMS-D-19-0324.1
Corresponding author: Stephan Rasp, stephan.rasp@tum.de
Supplemental material: https://doi.org/10.1175/BAMS-D-19-0324.2
In final form 28 April 2020
©2020 American Meteorological Society
For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy.

AFFILIATIONS: Rasp—Technical University of Munich, Munich, Germany; **Schulz and Stevens**—Max Planck Institute for Meteorology, Hamburg, Germany; **Bony**—Sorbonne Université, LMD/IPSL, CNRS, Paris, France

A quick glance at an image, be it taken from a satellite or produced from model output, is often sufficient for a scientist to identify features of interest. Similarly arranged features across many images form the basis for identifying patterns. This human ability to identify patterns holds true also in situations where the features, let alone the patterns they build, are difficult to describe objectively—a situation which frustrates the development of explicit and objective methods of pattern identification. In these situations, machine learning techniques, particularly deep learning (see "Deep learning for vision tasks in the geosciences" sidebar), have demonstrated their ability to mimic the human capacity for identifying patterns, also from satellite cloud imagery (e.g., Wood and Hartmann 2006). However, the application and assessment of such techniques is often limited by the tedious task of obtaining sufficient training data, so much so that (in cloud studies at least) these approaches have not been widely adopted, let alone assessed.

Recently, Stevens et al. (2020) described a collective cloud classification activity by a team of 13 scientists supported by the International Space Science Institute (ISSI). This ISSI team aimed to identify mesoscale cloud patterns in visible satellite imagery taken over a trade wind region east of Barbados. Organization, or clustering, of clouds has been shown to have important implications for climate in the case of deep convection (Tobin et al. 2012), which raises the question to what extent this is the case in shallow clouds. The ISSI team's hand-labeling effort resulted in around 900 subjectively classified images. An initial application of machine learning to these images (by the first author) proved promising but also highlighted the need for more training data in order to obtain robust and interpretable results.

Based on these first insights, the authors organized a crowdsourced project (see "Crowdsourcing" sidebar) that would allow us to collect a substantially larger set of labeled images. This activity was designed to provide a better foundation for the application of machine learning to the classification of patterns of shallow clouds, as well as to explore methodological questions raised when attempting to marry crowdsourcing with machine learning to address problems in climate and atmospheric science. Specifically we sought to answer four questions:

Deep learning for vision tasks in the geosciences

Deep learning describes a branch of artificial intelligence based on multilayered artificial neural networks (Nielsen 2015). In recent years, this data-driven approach has revolutionized the field of computer vision, which, up to 2012, was to a large extent based on hard-coded feature engineering (LeCun et al. 2015). More specifically, the success of deep learning in vision tasks is based on convolutional neural networks which exploit the translational invariance of natural images (i.e., a dog is a dog whether it is in the top right or bottom left of the image) to greatly reduce the number of unknown parameters to be fitted. 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 detection of features in images, the topic of this study. 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. 2017; Liu et al. 2016; Hong et al. 2017; Kurth et al. 2018; Mudigonda et al. 2017). A notable exception is the aforementioned study by Wood and Hartmann (2006). They hand-labeled 1,000 images of shallow clouds and used a neural network to classify them into four cloud types, making it a predecessor to our study.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

Q1: How should a community-driven labeling exercise be set up to ensure 1) a good user experience for participants and 2) the usefulness of the gathered data for subsequent analysis?
Q2: Can a diverse set of scientist identify the subjective modes of cloud organization established by the ISSI team with satisfactory agreement to warrant further scientific analysis?
Q3: Can a deep learning algorithm learn to classify images as well as trained scientists?
Q4: To the extent that a machine can be trained to classify large numbers of images, what can be learned from applying this algorithm to global data?

In this paper, we present our findings. They suggest that, for suitable problems, the combination of crowdsourcing and deep learning allows scientists to analyze data on a scale beyond what would be possible with traditional methods. Though our main findings will be of particular interest to researchers interested in the mesoscale organization of shallow clouds, the methods used to obtain them may be of more general interest, and are presented with this in mind.

We begin by describing how the cloud patterns (or classes) we sought to classify were defined, followed by a summary of the crowdsourcing project. Then the results from the human data are presented before we explain how deep learning is used to extend the analysis. Finally, we summarize our findings as pertains to the research questions state above, from which inferences of potential relevance to future studies are drawn.

Crowdsourcing

Crowdsourcing 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 crowdsourcing 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). Recent climate projects on the crowdsourcing platform Zooniverse[2] asked volunteers to transcribe old, handwritten weather records. Thanks to the collaboration of many individuals, such projects produce a wealth of data that would be unattainable for a single scientist. Note that for this paper we understand the term crowdsourcing to indicate active labor by the participants rather than providing data through personal sensors or cameras. For a broader review of citizen science and crowdsourcing studies in the geosciences, see Zheng et al. (2018).

[1] www.zooniverse.org/projects/zookeeper/galaxy-zoo
[2] www.zooniverse.org/projects/edh/weather-rescue and www.zooniverse.org/projects/drewdeepsouth/southern-weather-discovery

Sugar, Flower, Fish, and Gravel

Mesoscale patterning of shallow cumulus is a common feature in satellite imagery. However, organization on these scales is largely ignored in modeling studies of clouds and climate. This applies to process studies with large-eddy simulations (e.g., Rieck et al. 2012; Bretherton 2015) as well as general circulation models, be it in traditional or superparameterizations (Arakawa and Schubert 1974; Parishani et al. 2018).

The prevalence of mesoscale patterning in satellite cloud imagery led the ISSI team (Stevens et al. 2020) to identify four cloud patterns that frequent the lower trades of the North Atlantic. They named these patterns Sugar, Flower, Fish, and Gravel (Fig. 1). The choice of new and evocative names was motivated by the judgement that the patterns were different than those that have been previously described, for instance in studies of stratocumulus or cold-air outbreaks. Support for this judgement is provided by an application of the neural network from Wood and Hartmann (2006) and Muhlbauer et al. (2014), which was trained to distinguish between "No mesoscale cellular convection (MCC)," "Closed MCC," "Open MCC," and "Cellular, but disorganized." When applied to the scenes classified by the ISSI team the algorithm mostly resulted in the "disorganized" classification (I. L. McCoy 2017, personal communication). Despite the lack of a simple link between the patterns classified by the

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

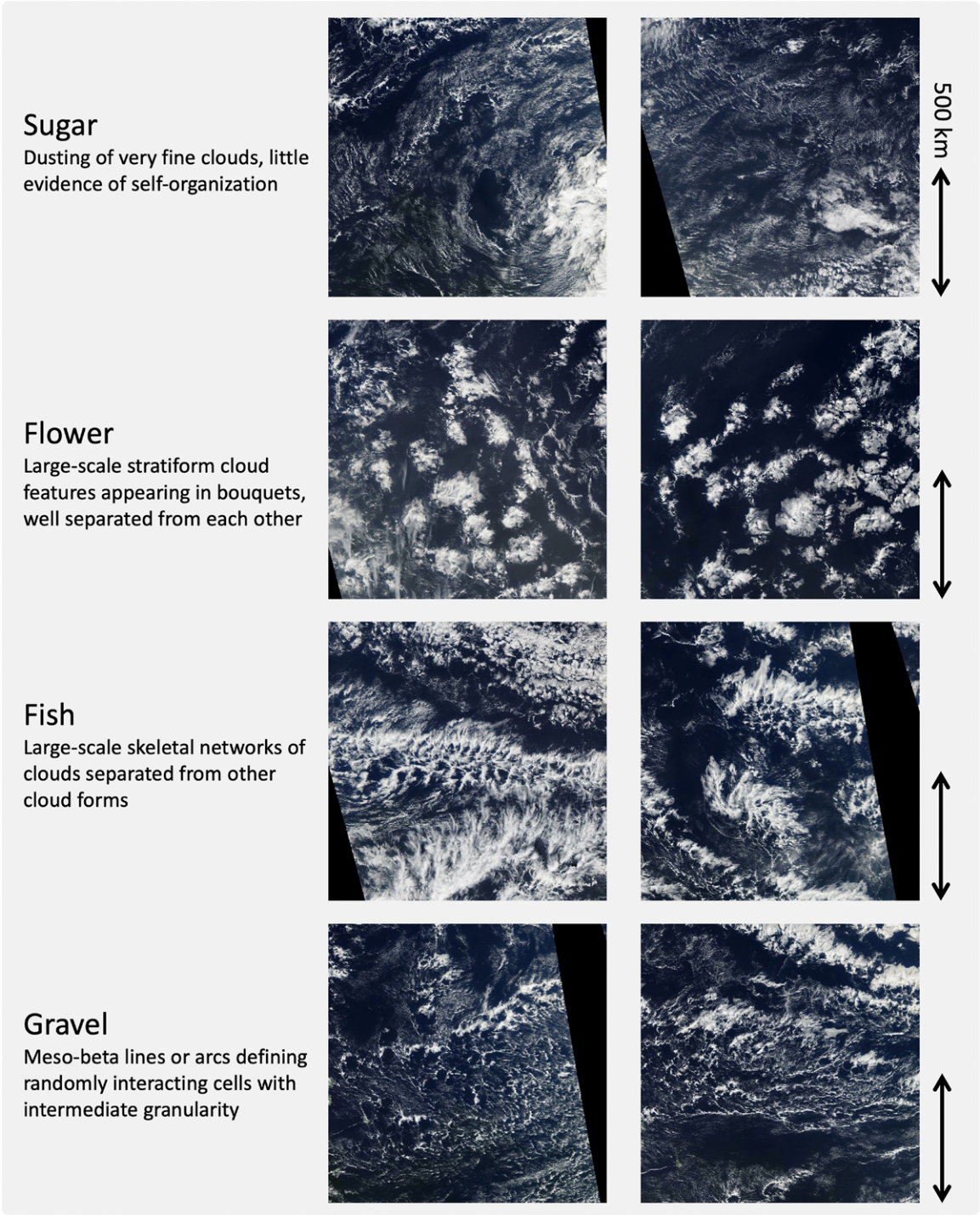


Fig. 1. Canonical examples of the four cloud organization patterns as selected by the ISSI team.

ISSI team and patterns previously described in the literature, below we point out previously identified patterns that may be related to the four patterns used here.

"Sugar" describes widespread areas of very fine cumulus clouds. Overall these fields are not very reflective, do not have large pockets of cloud-free regions, and, ideally, exhibit little evidence of mesoscale organization. Often, though, they are embedded within the larger-scale flow which gives them some structure. In strong flow, Sugar can form thin "veins," or feathers, which have been previously described as dendritic clouds (Nicholls and Young 2007).

"Flower" describes areas with isotropic cloud structures, each ranging from 50 to 200 km in diameter, with similarly wide cloud-free regions in between. This pattern overlaps to some

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

degree with canonical closed-cell MCC. Flowers, however, are often less densely packed than typical closed cells, which only have narrow cloud-free regions at the edges, and they are identified well outside of regions where stratocumulus are found (Norris 1998). One hypothesis is that they are successors of more closely packed closed-cell MCC which are in the process of breaking up.

"Fish" are elongated, skeletal structures that sometimes span up to 1,000 km, mostly longitudinally. As noted by Stevens et al. (2020), these features appear similar to what Garay et al. (2004) called actinoform clouds. They presented examples of these particularly well structured cloud forms taken from all ocean basins, near but typically downwind of regions where stratocumulus maximize. To the extent Fish are variants of the actinoform clouds found by Garay et al., they may be more common than previously thought.

Finally, "Gravel" describes fields of granular features marked by arcs or rings. The typical scale of these arcs is around 20 km. We suspect that these patterns are driven by cold pools caused by raining cumulus clouds (Rauber et al. 2007). In this regard, Gravel is fundamentally different from open-cell MCC, which has larger cells that are driven by overturning circulations in the boundary layer. However, the line between these two mechanisms can blur at times.

It is also interesting to compare our subjectively chosen labels to those of Denby (2020), who used an unsupervised learning algorithm to automatically detect different types of cloud organization (their Fig. 2). Some of their patterns bear resemblance to our classes, e.g., Sugar seems to most closely correspond to their patterns A and B, Gravel to G and H. However, their automatically detected classes appear less striking to the human eye.

Crowdsourced labels

To obtain a large pool of labeled images from the community, an accessible user interface is needed. Zooniverse[3] is an open web platform that enables researchers to organize and present research questions in ways that enable contributions from the broader public (see also "Crowdsourcing" sidebar). Its flexibility in serving and presenting images, choosing between different labeling tasks, and its ability to monitor and organize the information associated with the labeling activities made Zooniverse very well suited for our task.

[3] www.zooniverse.org/projects/raspstephan/sugar-flower-fish-or-gravel

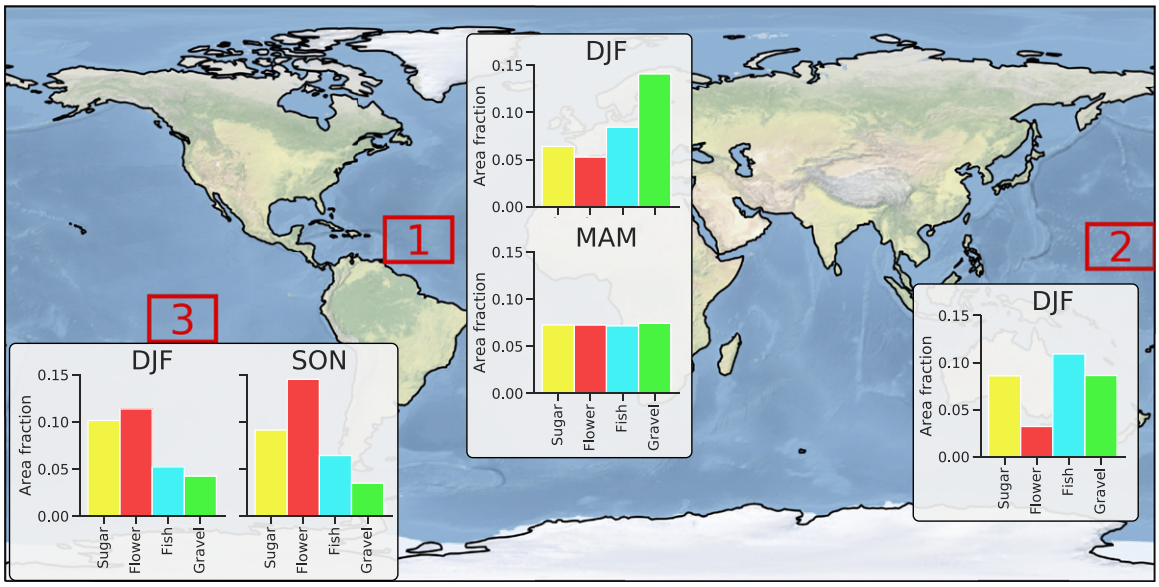


Fig. 2. World map showing the three regions selected for the Zooniverse project. Bar charts show which fraction of the image area was classified into one of the four regions by the human labelers. Note that the areas do not add up to one. The remaining fraction was not classified.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

For our project we downloaded roughly 10,000 (14° latitude × 21° longitude) *Terra* and *Aqua* MODIS visible images from NASA Worldview. To select the regions and seasons, we started with the boreal winter (DJF) east of Barbados as a reference. Barbados is home to the Barbados Cloud Observatory (Stevens et al. 2016). The clouds in its vicinity were not only the focus of the ISSI team's study, but have more generally come to serve as a laboratory for studies of shallow clouds and climate (Stevens et al. 2016; Medeiros and Nuijens 2016; Stevens et al. 2020; Bony et al. 2017). To obtain more images and sample a greater diversity of clouds, we subsequently added images from two further regions in the Pacific, which were chosen based on their climatological similarity to the original study region upwind of Barbados (Fig. 2; see supplemental material for details). Images were downloaded for an 11-yr period from 2007 to 2017.

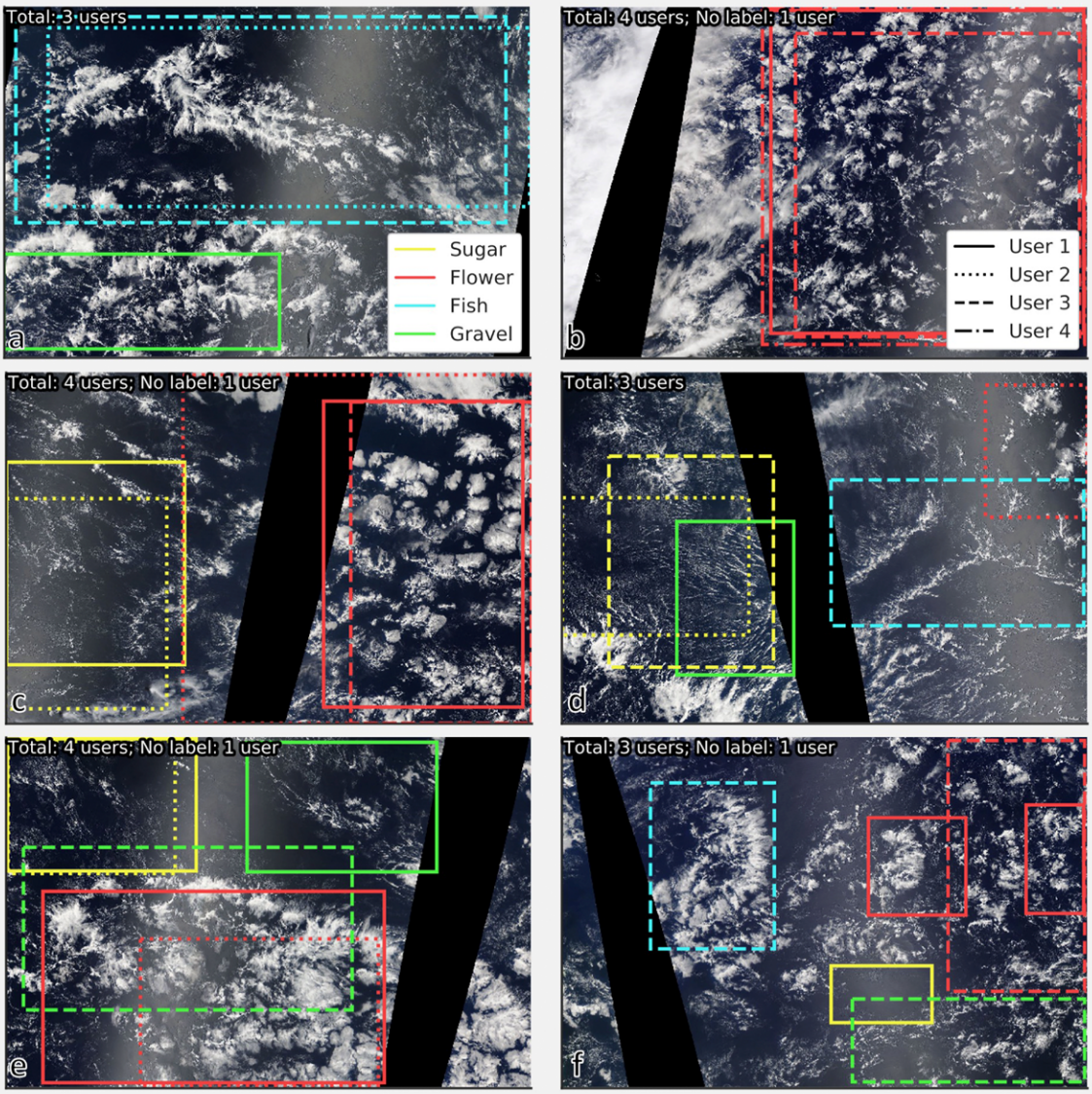
Stevens et al. (2020) speculated that their protocol of assigning a single label to the entire 10° × 20° image resulted in considerable ambiguity and disagreement between labelers. In an attempt to minimize this issue, we presented participants with slightly larger images and experimented with ways to allow the labeling of multiple, and possibly overlapping, subregions. This was accomplished by allowing users to draw rectangles around regions where they judged one of the four cloud patterns to dominate (see Fig. 3 for examples). 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. We arrived at this setup after experimenting with other options, such as labeling subsections of a predefined grid, or allowing users to label regions that they defined using polygons with an arbitrary number of sides. We opted for the rectangles to increase labeling speed and improve the user experience. Our thinking was that it would be better to have less accurate but more plentiful data, and that given the vague boundaries of the cloud structures, it was anyway doubtful that a more accurate labeling tool would add much information. As we will show later, this thinking paid off for the machine learning models we trained.

The Zooniverse interface was further configured to serve participants an image randomly drawn from our library of 10,000 images. After being classified by four different users, images were retired, i.e., removed from the image library. In addition, 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 2 November and at the Laboratoire de Météorologie Dynamique in Paris, France, on 29 November 2018. After a brief instruction at the start of the day and a warm-up on the training dataset, 67 participants, most of them researchers, from the two institutes, labeled images for an entire day. The activity yielded roughly 30,000 classified images; i.e., each image was classified about three times on average. Because an image could have subregions with different classifications, the number of labels was somewhat larger, with 49,000. On average, participants needed around 30 s to classify one image, amounting to approximately 250 h of concentrated human labor. There were, however, considerable differences among users, as the interquartile range in classification times ranged from 20 to 38 s. Overall, the four patterns occupied similarly large areas, but notable differences occurred depending on the geographic region and season (Fig. 2).

Inferences from human labels

Given the subjective nature of labels assigned by visual inspection, our first research question was to what extent the human labelers agreed with each other. In the initial classification exercise of 900 images reported in Stevens et al. (2019), a majority of scientists agreed on one pattern in 37% of the cases, significantly more than random. In this project, in addition to choosing the category of the clouds, participants also had to choose the location. To explore the agreement we started by looking at many examples, six of which are reproduced in Fig. 3. Many more can be found at Rasp (2020a). The most notable conclusions from this visual inspection are that

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020



IoU values for each image

Panel	*Sugar*	*Flower*	*Fish*	*Gravel*	*Not classified*
a	N/A	N/A	0.28	0.00	0.44
b	N/A	0.44	N/A	N/A	0.59
c	0.14	0.31	N/A	N/A	0.41
d	0.19	0.00	0.00	0.00	0.67
e	0.18	0.13	N/A	0.01	0.55
f	0.00	0.09	0.00	0.00	0.68

Fig. 3. Six example images showing annotations drawn by human labelers. Different line styles correspond to different users. In addition the IoU values for each image and class are shown in the table.

users agreed to a high degree on features that closely resemble the canonical examples of the four classes but also that there was a lot of disagreement otherwise. Take Fig. 3a, where two out of three participants agreed on the presence of Fish in the top half of the image, or Fig. 3b, where three out of four participants recognized a region of Flower. On the other hand Fig. 3d shows an example of an image with plenty of ambiguity. Also note that users applied different methodologies when labeling, some labeling a single large region, others many small regions. Overall, we came to the conclusions that, while certainly noisy, clear examples of what was defined as Sugar, Flower, Fish, and Gravel could be robustly detected.

Next, we aimed to quantify the agreement. To our knowledge there is no standard way of evaluating subjective labels from multiple users. The most commonly used metric for

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

comparing a label prediction with a ground truth is the Intersect over Union (IoU) score, also called the Jaccard index. Given two sets, A and B, it is defined as the ratio of their intersection to their union, i.e., $I = A \cup B$ divided by $U = A \cup B$. An IoU score of one indicates perfect overlap, while zero indicates no overlap. We adapted the IoU score to this task by first iterating over every image and then computing the intersect I_{image} and union U_{image} for every user–user combination for this image. To compute the final "Mean IoU between humans" we computed the sum of the intersect and union over all images: $I = \sum_{\text{image}} I_{\text{image}}$ and $U = \sum_{\text{image}} U_{\text{image}}$. This was done for each cloud class separately. We also computed an IoU score for the "Not classified" area. Finally, the "All classes" IoU was computed by additionally taking the sum of I and U over all classes. The results for the interhuman mean IoU are shown in Fig. 4a.

At first glance, IoU values of around 0.2 seem low. In classical computer vision tasks such values would certainly indicate low agreement. However, as mentioned above, this dataset is different from classical object detection tasks in that there are more than two labelers for most images and there are many cases in which one or more participants did not label an image. In fact, the primary reason for the low mean IoU score are zero values, which arise from some users detecting a feature while others did not. Take Fig. 3b as an example. Here, three of four users agreed to a high degree of accuracy on the location of Flowers but the last user did not

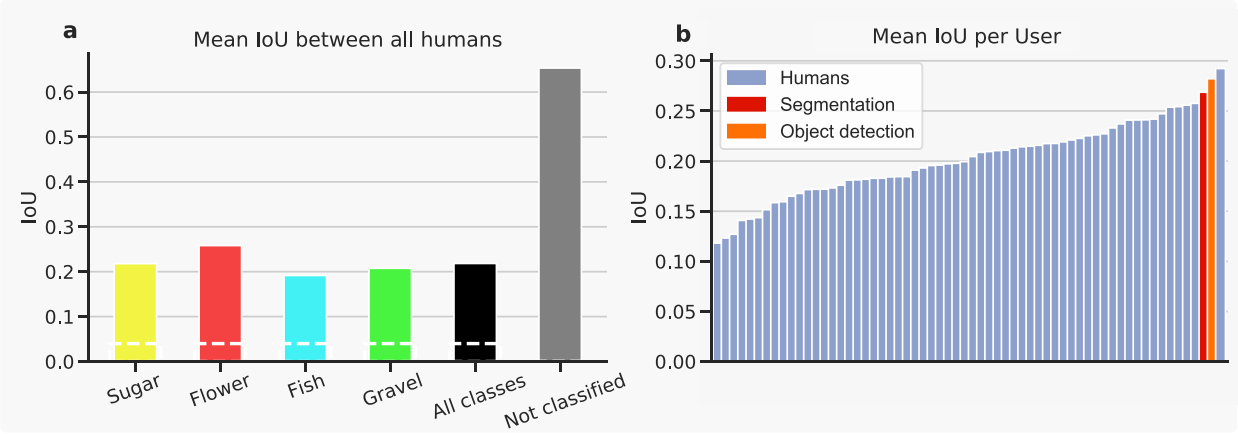


Fig. 4. (a) Mean IoU between humans. The dashed line represents random IoU; see text for details. (b) Mean IoU for each human participant and the two deep learning algorithms for a validation dataset.

submit a label. This results in three "no label"–"label" comparisons and three "label"–"label" comparisons. Even with perfect agreement between the three Flower labelers, the mean IoU would only be 0.5. In reality it is 0.44 for this example. These "no label"–"label" pairs with IoU = 0 make up 63% of all user–user comparisons (see Fig. ES2 in the supplemental material). Omitting these gives a mean IoU of 0.43. To get a feel for what this value means consider the two Sugar rectangles in Fig. 3d, which have an IoU of 0.46. The table at the bottom of Fig. 3 shows the mean IoU values for each of the example images. These numbers suggest that even for images where one would visually detect a high degree of agreement between the users, the IoU numbers are quite low. For this reason, the actual values should not be compared to other tasks where the IoU is used. Rather, for this paper they simply serve the purpose of comparing different classes and methods. To further illustrate this point we computed the IoU score for many randomly drawn labels from the same number and size distribution as the human labels, which gives an IoU of only 0.04. What the numbers do show is that there are noticeable differences between the four patters. People agreed most on Flower while Fish proved more controversial. With regard to Q2, we came to the conclusion that, despite the noise in the labels, there was sufficient consensus between the participants on clear features

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

to warrant further analysis, especially since, as we will see, the noise will largely disappear in the statistical average.

Another question that the new methodology of labeling allows us to answer is whether or not the patterns tend to span larger or smaller areas. Based on the Zooniverse labels, Flower boxes tended to be largest, covering around 25% (around 900,000 km^2) of the image. Fish and Gravel were somewhat smaller with a box size of around 20% (around 720,000 km^2). Sugar spanned regions smaller yet with boxes only taking up 15% (around 540,000 km^2) of the image on average. Because the initial classification by Stevens et al. (2020) required labelers to identify the entire scene, the relative infrequency with which they detected Sugar is likely due in part to the infrequency with which it covers large areas.

Further, we can ask whether the four patterns, which were purely chosen based on their visual 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[4] corresponding to each pattern (Fig. 5). To the extent ERA-Interim accurately represents the meteorological conditions in the region, the composites suggest that Sugar, Flower, Fish, and Gravel appear in climatologically distinct environments. This is supported by the standard error being smaller than the difference between the patterns. The standard error is a measure for how well the mean conditions of a given pattern can be estimated and is defined as $\sigma/\sqrt{N}$, where σ is the standard deviation and N is the sample size. At the same time, there is variability between individual profiles within a composite, as shown by the interquartile range. Hence, while the compositing suggests that the occurrence of a particular pattern is associated with significant changes in the large-scale environmental conditions, this is clearly not the only factor at play, and things like airmass history are likely also important.

[4] www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim

Flowers tend to be associated with a relatively dry and cold boundary layer with a very strong inversion (note that Fig. 5 shows deviations from the climatological mean). Sugar appears in warm and humid boundary layers with strong downward motion maximizing near the cloud base. For Fish and particularly Gravel, on the other hand, the inversion and downward motion is rather weak. The fact that Flower and Gravel are essentially opposites in terms of their environmental profiles suggests that they are not simply manifestations of

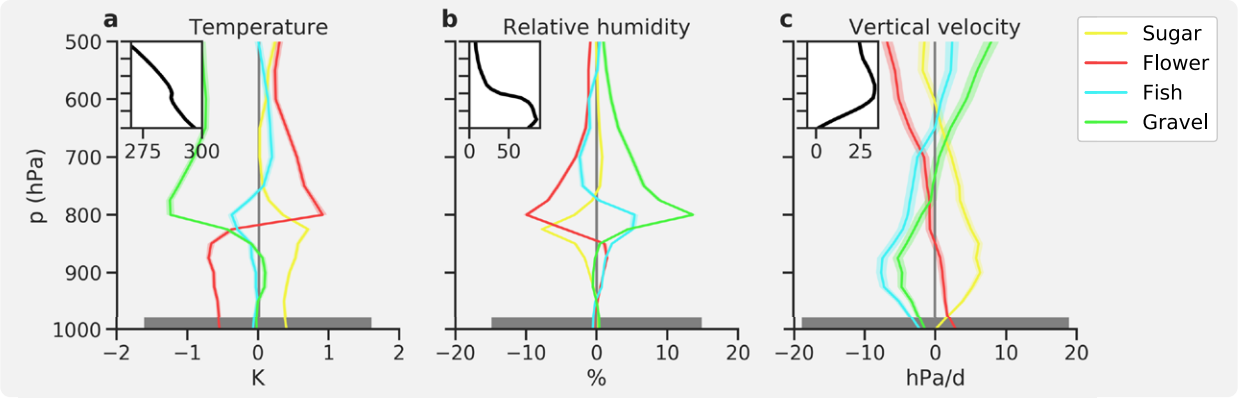


Fig. 5. Median of large-scale environmental conditions corresponding to the four patterns as identified by the human labelers: (a) temperature, (b) specific humidity, and (c) vertical velocity (shown in $\omega = dp/dt$) relative to the climatological mean, which is shown in the insets. The shading about the lines shows the standard error, and hence the statistical difference between the mean conditions associated with any particular pattern. The bar along the *x* axis shows the average interquartile spread (for the level where this spread maximizes, around 800 hPa) in the thermodynamic state associated with each pattern, indicating that the conditions associated with any given pattern can vary considerably.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

closed and open-cellular convection, which often transition smoothly into one another in similar large-scale environments (Muhlbauer et al. 2014).

Application of deep learning

While the 10,000 images labeled on Zooniverse already provide a useful dataset for further analysis, they only cover a small fraction of the globe for a small fraction of time. Only 0.6% of the data available during the selected 11-yr period were labeled. In this section, we explore whether deep learning (see "Deep learning for vision tasks in the geosciences" sidebar) can help to automate the detection of the four organization patterns and if so what can be learned from it.

The pattern recognition task can be framed as one of two machine learning problems: object detection and semantic segmentation. Object detection algorithms draw boxes around features of interest, essentially mirroring what the human labelers were doing. In contrast, segmentation algorithms classify every pixel of the image. Figure 6 shows examples of these two approaches for images from a validation dataset that was not used during training [see Rasp (2020b) for more randomly chosen examples]. Details about the neural network architectures and preprocessing steps can be found in the supplemental material. Both types of algorithm accurately detect the most obvious patterns in the image and agree well with human labels. Neither algorithm is perfect, however. The object detection algorithm sometimes

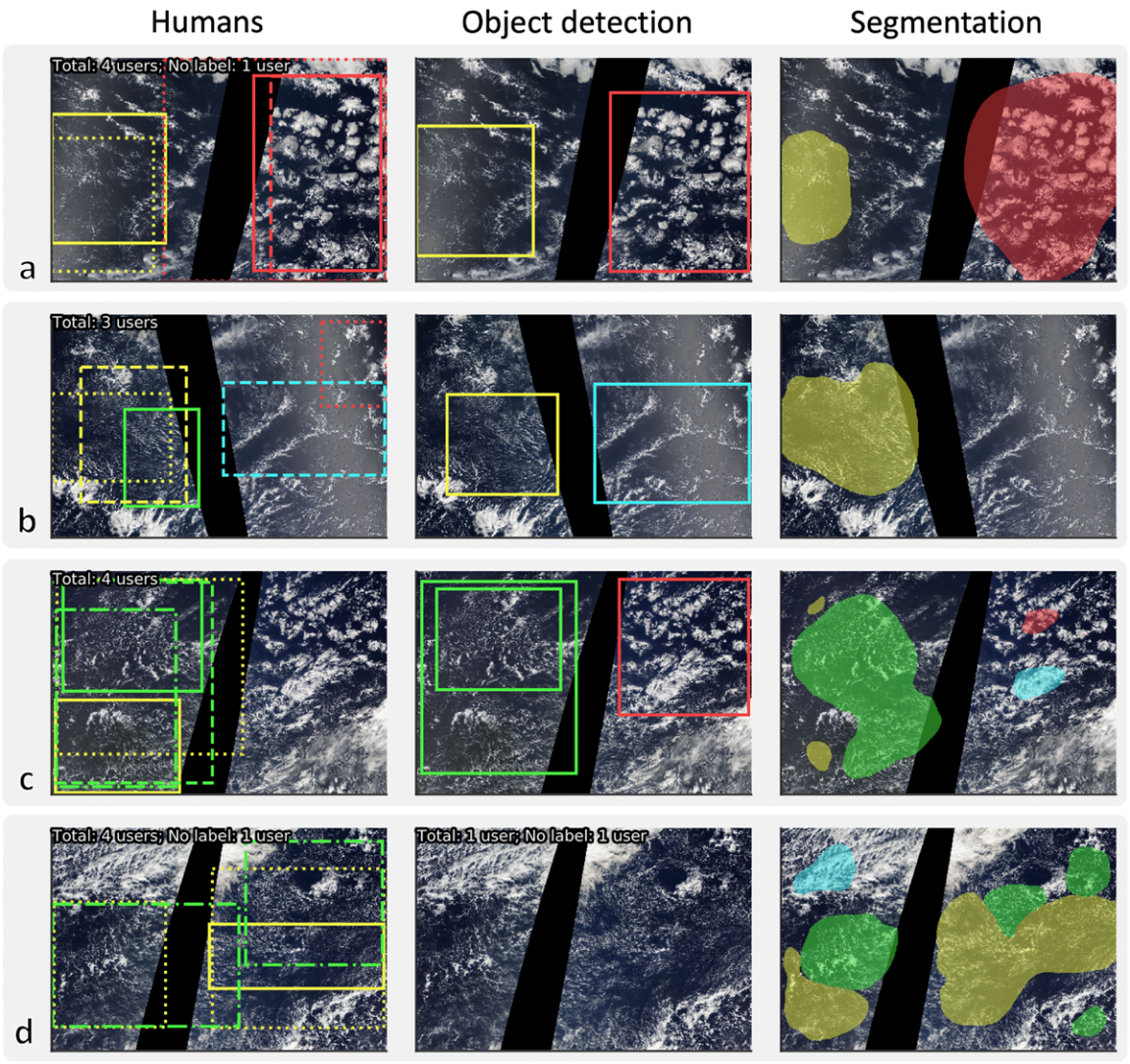


Fig. 6. Human and machine learning predictions for four images from the validation set. Note that (a) and (b) are also shown in Fig. 3.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

misses features, as is visible in Fig. 6d. The segmentation algorithm, on the other hand, tends to produce relatively small patches (Figs. 6c,d) because, other than humans and the object detection algorithm, in which the range of possible box sizes is an adjustable parameter, it has not been given instructions to only label larger patches. An interesting and advantageous feature of the segmentation algorithm is that, despite all training labels being rectangular, it appears to focus on the actual, underlying shape of the patterns, as visible by the rounded outlines of the predicted shapes. This suggests that despite the uncertainty in the human dataset, the deep learning algorithms are able to filter out a significant portion of this noise and manage to distill the underlying human consensus.

To quantitatively compare the deep learning algorithms against the human labelers, we compute the mean IoU for each human individually as well as for the two algorithms (Fig. 4b). Both algorithms show a large agreement with the human labels for a random validation dataset. The fact that the scores are higher than the mean interhuman IoU directly reflects the fact that the algorithms tend to produce less noisy predictions. Further analysis shows that the algorithms inherit some biases from the human training labels. The frequency and accuracy of the predicted labels is higher for patterns with a higher interhuman agreement, most notably Flower (Fig. ES3), which could slightly bias the deep learning predictions.

The main advantage of deep learning algorithms is that they are very fast at inference, 1 s per image compared to the 30 s a human needed on average, and they are more scalable. This allows us to apply the algorithm to the entire globe (Fig. 7a; see supplemental material for details). A healthy skepticism is warranted when applying machine learning algorithms outside of their training regime (Rasp et al. 2018; Scher and Messori 2019). A visual inspection of the global maps [see Rasp (2020c) 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, Fig. 7a suggests that the algorithm even 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 (Figs. 7b–e). The resulting heat maps reveal coherent hot spots for the four cloud patterns. The spatial distribution of these hot spots helps answer some further questions raised by the ISSI team's study. For instance, the heat maps indicate that organization is most common over the ocean. Only Sugar—the one pattern characterized by its lack of mesoscale organization—was identified over land (but keeping in mind the potential bias of the algorithm). Our results also indicate that Sugar, followed by Flower, are the most common forms of organization globally. This indicates a bias arising from the ISSI team's focus on a single study region, as large areas of Sugar are relatively rare near Barbados. A prevalence of Sugar in the trades adjacent to the deep tropics, and regions such as the Arabian sea, is consistent with its coincidence in association with strong low-level subsidence and a somewhat drier cloud layer (as seen by the large-scale composites, Fig. 5), indicating that it might be most favored in regions where convection is suppressed by strong subsidence from neighboring regions of active convection, or strong land–sea circulations.

Flower prevails slightly downstream of the main stratocumulus regions. Composites of the environmental conditions in which they form show them to be, on average, associated with large-scale environmental conditions characterized by more pronounced lower-tropospheric stability, and a somewhat drier free troposphere (Fig. 5). This lends credence to the idea that they are manifestations of closed-cell MCCs. Whereas the climatology of closed-cell MCC by Muhlbauer et al. (2014, their Fig. 5) shows similar hot spots in the subtropics, it also has strong maxima across the mid- and high-latitude oceans. The absence of such hot spots in our classification of Flower could indicate a bias of our algorithm toward the regions it was

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

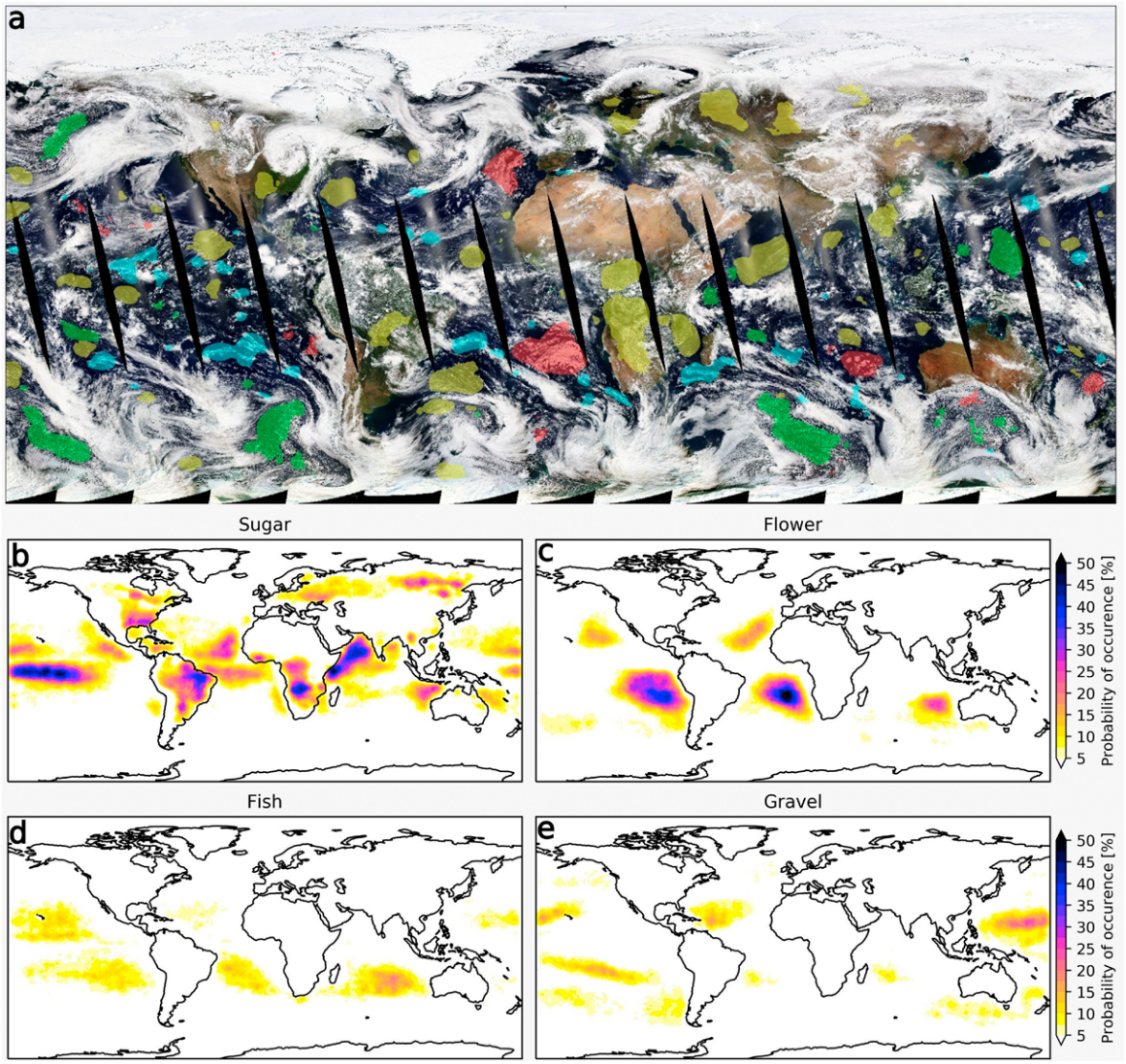


Fig. 7. (a) Global predictions of the image segmentation algorithm for 1 May 2017. The colors are as in the previous figures. See Rasp (2020c) for more examples. (b)–(e) Heat maps of the four patterns for the year 2017.

trained on. However, it could also suggest that "Flower" differs from typical closed cellular convection in their scale and spacing.

Farther downstream in the trade regions, Flower makes way to Gravel and Fish. These two patterns are more geographically intertwined, which is in agreement with the similarity of the environmental profiles in Fig. 6. Interestingly, Gravel seems to be relatively confined to the Barbados region, the west of Hawaii, and the southern tropical Pacific near regions—like the South Pacific convergence zone—of climatological convergence (Fig. 2). Hence, the prevalence of Gravel in the more limited classification activity of Stevens et al. (2020), is not representative of the trade wind regions more broadly. There is also some coincidence of Gravel hot spots with regions of open-cell MCC regions as highlighted by the classification by Muhlbauer et al. (2014), specifically around Hawaii and in the southern subtropical Atlantic, but as with Flower their MCC algorithm picks up many more open cells in higher latitudes. This, again, suggests that there may be a fundamental difference between the classes, something we already suspected based on their physical driving mechanisms, i.e., cold pools versus boundary layer circulations. Fish, appears linked to stronger synoptic upward motion (Fig. 5c), which the image snapshot from 1 May 2017 (Fig. 6a) suggests is associated with synoptic convergence lines, often connected to trailing midlatitude fronts.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

Globally, the patterns are coherent, with hot spots for a given pattern appearing in a few spatially extensive and plausibly similar meteorological regimes. This coherence supports the hypothesis that the subjective patterns are associated with meaningful and distinct physical processes. Though the combination of crowdsourced labels and deep learning helped answering many of the questions raised at the outset of this study it also raises some new ones, for instance whether important cloud regimes are missing from our classification. Unsupervised classification algorithms like the one deployed by Denby (2020) can be a good starting point to explore this question.

The four questions

In this paper, we described a project to combine crowdsourcing, to detect and label four subjectively defined patterns of mesoscale shallow cloud organization from satellite images, with deep learning. The design and execution of the project raised a number of questions, four of which have been highlighted in this paper, and the answers to which we present as follows.

The first question (Q1) was concerned with how best to configure a crowdsourcing activity. We found that speed and ease of use for the participants is paramount. Drawing crude rectangles on the screen only took tens of seconds for each image, whereas more detailed shapes such as polygons would have taken significantly longer. Further, the quickness of drawing boxes on an image meant that less of an attention span was required from the participants. (Some even reported to have had fun.) For our task, which involves judgements with inherent uncertainty, the added noise introduced by crude labels turned out to be insignificant in the statistical average, as shown by the "consensus" found by the deep learning algorithm. Based on our experience, quantity trumps quality. This might, of course, be different for tasks where object boundaries are more clearly defined.

Q2 asked whether sufficient agreement exists between the human labelers to warrant scientific use of the labels. We believe that this is indeed the case. As discussed in the section titled "Inferences from human labels," there is a significant amount of disagreement between the participants, particularly because many cloud formations did not fit one of the four classes exactly. However, more importantly there was significant agreement on patterns that closely matched the canonical examples of "Sugar," "Flower," "Fish," and "Gravel." Taking a statistical average—training a deep learning model can be viewed as doing just that—removes some of the ambiguity from the labels and crystallizes the human consensus. Of course, the four classes chosen are not a complete description of all modes of organization, and others could have been defined. But the fact that the results are compatible with physical understanding suggest that the four classes do indeed capture important modes of cloud organization in the subtropics.

Q3 asked whether deep learning can be used to build an automated labeling system. The answer is a resounding yes. Both deep learning algorithms used in this paper show high agreement scores. Further, visual analysis of the deep learning predictions suggest that these are less noisy than the human predictions. In other words, the deep learning models have learned to disregard the noise of the human labels and instead extract the common underlying pattern behind points of agreement, i.e., the essence of the proposed patterns. In addition, the deep learning models are both many orders of magnitude faster than humans at labeling images, and less costly and difficult to maintain.

The application of deep learning enabled us to classify a significantly larger geographical and temporal set of data. This allowed us to look at global patterns of "Sugar," "Flower," "Fish," and "Gravel" thereby addressing our fourth research question (Q4). Here our main finding is that heat maps of pattern occurrence are distributed in a geographically coherent way across all the major ocean basins, and sample significantly different meteorological

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

conditions. Heat maps for two of the patterns (Flower and Gravel) show some overlap with closed-cell mesoscale cellular convection, but only over portions of the subtropics. As a rule the regions where patterns are identified (particularly for Fish and Flower) are not in regions familiar from past work on cloud classification.

Inferences and outlook

The coherence of the heat maps for individual patterns suggests the presence of physical drivers underpinning their occurrence; drivers that may change as the climate changes. Using the same classification categories but a different way of classifying the images, Bony et al. (2019) showed that differences in cloud radiative properties are associated with different forms of organization. Our study thus lends weight to the idea that quantifying the radiative effects of shallow convection, and potential changes with warming, may require an understanding of, or at least an ability to represent, the processes responsible for the mesoscale organization of fields of shallow clouds. This might seem to be a daunting task. However, if the occurrence of different modes of organization can be reliably linked to large-scale conditions, reanalysis data or historical climate model simulations could help reconstruct cloud fields. This could offer clues as to how mesoscale organization, and hence cloudiness, has changed in the past, and may change in the future.

This example helps highlight how crowdsourced and deep-learned datasets create new ways to study factors influencing shallow clouds and their radiative properties, and hopefully stimulates ideas for adapting the approach to other problems. The growing accessibility of these new research methodologies makes their application all the more attractive. Platforms like Zooniverse make it easy to set up a labeling interface free of charge. Plus, even if we did not do so, it is also possible to make the interface open to the public. Deep learning has also become much more accessible. Easy-to-use Python libraries[5] with pretrained models for many applications in computer vision as well as accessible online courses[6] make it possible even for non–computer scientists to apply state-of-the art deep learning techniques.

Our study also illustrates how crowdsourcing and deep learning effectively complement one another, also for problems in climate science. Deep learning algorithms typically need thousands of samples for training. These are not readily available for most problems in the geosciences. A key lesson from our project is that, even for the ambiguously defined images that characterize many problems in atmospheric and climate science, it is feasible to create sufficient training data with a moderate amount of effort. We found that 5,000 labels (i.e., one-sixth of what was collected here) were enough to obtain similarly good results to the ones shown here. This translates to a day of labeling for around 15 people.

This means that combining crowdsourcing and deep learning is a promising approach for many questions in atmospheric science where features are easily—albeit not unambiguously—detectable by eye but hard to quantify using traditional algorithms. In our case, the combination of the two tools allowed us to generate global heat maps, something that would have been impossible with traditional methods. Potential examples of similarly suited problems in the geosciences are detecting atmospheric rivers and tropical cyclones in satellite and model output,[7] classifying ice and snow particles images obtained from cloud probe imagery, or even large-scale weather regimes.

[5] Keras (Chollet et al. 2015) and fastai (https://docs.fast.ai/) were used for this study; see supplemental material.

[6] https://course.fast.ai/ and https://deeplearning.ai

[7] See www.nersc.gov/research-and-development/data-analytics/big-data-center/climatenet/ for a similar project.

Acknowledgments. First and foremost, we 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

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

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).

Data availability statement. All data and code are available at https://github.com/raspstephan/sugar-flower-fish-or-gravel.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020

References

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *J. Atmos. Sci.*, **31**, 674–701, https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2.

Bony, S., and Coauthors, 2017: EUREC4A: A field campaign to elucidate the couplings between clouds, convection and circulation. *Surv. Geophys.*, **38**, 1529–1568, https://doi.org/10.1007/s10712-017-9428-0.

——, H. Schulz, J. Vial, and B. Stevens, 2020: Sugar, gravel, fish, and flowers: Dependence of mesoscale patterns of trade-wind clouds on environmental conditions. *Geophys. Res. Lett.*, **48**, e2019GL085988, https://doi.org/10.1029/2019gl085988.

Bretherton, C. S., 2015: Insights into low-latitude cloud feedbacks from high-resolution models. *Philos. Trans. Roy. Soc.*, **373A**, 20140415, https://doi.org/10.1098/rsta.2014.0415.

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

Denby, L., 2020: Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophys. Res. Lett.*, **47**, e2019GL085190, https://doi.org/10.1029/2019gl085190.

Garay, M. J., R. Davies, C. Averill, and J. A. Westphal, 2004: Actinoform clouds: Overlooked examples of cloud self-organization at the mesoscale. *Bull. Amer. Meteor. Soc.*, **85**, 1585–1594, https://doi.org/10.1175/BAMS-85-10-1585.

Hennon, C. C., and Coauthors, 2015: Cyclone center: Can citizen scientists improve tropical cyclone intensity records? *Bull. Amer. Meteor. Soc.*, **96**, 591–607, https://doi.org/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. arXiv, 4 pp., http://arxiv.org/abs/1708.03417.

Kurth, T., and Coauthors, 2018: Exascale deep learning for climate analytics. *SC18: Int. Conf. for High Performance Computing, Networking, Storage and Analysis, Dallas*, TX, IEEE, 649–660, https://doi.org/10.1109/SC.2018.00054.

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

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. *Advances in Big Data Analytics*, H. R. Arabnia and F. G. Tinetti, Eds., CSREA Press, 81–88, http://worldcomp-proceedings.com/proc/p2016/ABD6152.pdf.

Medeiros, B., and L. Nuijens, 2016: Clouds at Barbados are representative of clouds across the trade wind regions in observations and climate models. *Proc. Natl. Acad. Sci. USA*, **113**, E3062–E3070, https://doi.org/10.1073/pnas.1521494113.

Mudigonda, M., and Coauthors, 2017: Segmenting and tracking extreme climate events using neural networks. *31st Conf. on Neural Information Processing Systems*, Long Beach, CA, NIPS, 5 pp., https://dl4physicalsciences.github.io/files/nips_dlps_2017_20.pdf.

Muhlbauer, A., I. L. McCoy, and R. Wood, 2014: Climatology of stratocumulus cloud morphologies: Microphysical properties and radiative effects. *Atmos. Chem. Phys.*, **14**, 6695–6716, https://doi.org/10.5194/acp-14-6695-2014.

Nicholls, S. D., and G. S. Young, 2007: Dendritic patterns in tropical cumulus: An observational analysis. *Mon. Wea. Rev.*, **135**, 1994–2005, https://doi.org/10.1175/MWR3379.1.

Nielsen, M. A., 2015: *Neural Networks and Deep Learning*. Determination Press, http://neuralnetworksanddeeplearning.com.

Norris, J. R., 1998: Low cloud type over the ocean from surface observations. Part II: Geographical and seasonal variations. *J. Climate*, **11**, 383–403, https://doi.org/10.1175/1520-0442(1998)011<0383:LCTOTO>2.0.CO;2.

Parishani, H., M. S. Pritchard, C. S. Bretherton, C. R. Terai, M. C. Wyant, M. Khairoutdinov, and B. Singh, 2018: Insensitivity of the cloud response to surface warming under radical changes to boundary layer turbulence and cloud microphysics: Results from the ultraparameterized CAM. *J. Adv. Model. Earth Syst.*, **10**, 3139–3158, https://doi.org/10.1029/2018MS001409.

Racah, E., C. Beckham, T. Maharaj, S. E. Kahou, Prabhat, and C. Pal, 2017: ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *31st Conf. on Neural Information Processing Systems*, Long Beach, CA, NIPS, 12 pp., http://arxiv.org/abs/1612.02095.

Rasp, S., 2020a: Sugar, flower, fish or gravel - Example human labels. figshare, https://doi.org/10.6084/m9.figshare.12141264.v1.

——, 2020b: Sugar, flower, fish or gravel - Example ml predictions. figshare, https://doi.org/10.6084/m9.figshare.8236289.v2.

——, 2020c: Sugar, flower, fish or gravel - Global predictions. figshare, https://doi.org/10.6084/m9.figshare.8236298.v2.

——, M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Rauber, R. M., and Coauthors, 2007: Rain in shallow cumulus over the ocean: The RICO campaign. *Bull. Amer. Meteor. Soc.*, **88**, 1912–1928, https://doi.org/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**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Rieck, M., L. Nuijens, and B. Stevens, 2012: Marine boundary layer cloud feedbacks in a constant relative humidity atmosphere. *J. Atmos. Sci.*, **69**, 2538–2550, https://doi.org/10.1175/JAS-D-11-0203.1.

Scher, S., and G. Messori, 2019: Weather and climate forecasting with neural networks: Using general circulation models (GCMs) with different complexity as a study ground. *Geosci. Model Dev.*, **12**, 2797–2809, https://doi.org/10.5194/gmd-12-2797-2019.

Stevens, B., and Coauthors, 2016: The Barbados Cloud Observatory: Anchoring investigations of clouds and circulation on the edge of the ITCZ. *Bull. Amer. Meteor. Soc.*, **97**, 787–801, https://doi.org/10.1175/BAMS-D-14-00247.1.

——, and Coauthors, 2019: A high-altitude long-range aircraft configured as a cloud observatory: The NARVAL expeditions. *Bull. Amer. Meteor. Soc.*, **100**, 1061–1077, https://doi.org/10.1175/BAMS-D-18-0198.1.

——, and Coauthors, 2020: Sugar, gravel, fish, and flowers: Mesoscale cloud patterns in the tradewinds. *Quart. J. Roy. Meteor. Soc.*, **146**, 141–152, https://doi.org/10.1002/qj.3662.

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. *J. Climate*, **25**, 6885–6904, https://doi.org/10.1175/JCLI-D-11-00258.1.

Wood, R., and D. L. Hartmann, 2006: Spatial variability of liquid water path in marine low cloud: The importance of mesoscale cellular convection. *J. Climate*, **19**, 1748–1764, https://doi.org/10.1175/JCLI3702.1.

Zheng, F., and Coauthors, 2018: Crowdsourcing methods for data collection in geophysics: State of the art, issues, and future directions. *Rev. Geophys.*, **56**, 698–740, https://doi.org/10.1029/2018RG000616.

Downloaded from http://journals.ametsoc.org/bams/article-pdf/101/11/E1980/5018932/bamsd190324.pdf by guest on 24 November 2020