Understanding cloud systems structure and organization using a machine’s self-learning approach

Dwaipayan Chatterjee, a Claudia Acquistapace, a Hartwig Deneke, b Susanne Crewell, a

a Institute for Geophysics and Meteorology, University of Cologne, Cologne, Germany
b Leibniz Institute for Tropospheric Research, Leipzig, Germany

Corresponding author: Dwaipayan Chatterjee, dchatter@uni-koeln.de

Early Online Release: This preliminary version has been accepted for publication in Artificial Intelligence for the Earth Systems, may be fully cited, and has been assigned DOI 10.1175/AIES-D-22-0096.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2023 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
ABSTRACT: In this study, we introduce a self-supervised deep neural network approach to classify satellite images into independent classes of cloud systems. The driving question of the work is to understand whether our algorithm can capture cloud variability and identify distinct cloud regimes. Ultimately, we want to achieve generalization such that the algorithm can be applied to unseen data and thus help automatically extract relevant information important to atmospheric science and renewable energy applications from the ever-increasing satellite data stream. We use cloud optical depth (COD) retrieved from post-processed high-resolution Meteosat Second Generation (MSG) satellite data as input for the network. The network’s architecture is based on the DeepCluster version 2 and consists of a convolutional neural network and a multilayer perceptron, followed by a k-means algorithm. We explore the network’s training capabilities by analyzing the centroids and feature vectors found from progressive minimization of the cross entropy loss function. By making use of additional MSG retrieval products based on multi-channel information, we derive the optimum number of classes to determine independent cloud regimes. We test the network capabilities on COD data from 2013 and find that the trained neural network gives insights into the cloud systems’ persistence and transition probability. The generalization on the 2015 data shows good skills of our algorithm with unseen data, but results depend on the spatial scale of cloud systems.
SIGNIFICANCE STATEMENT: This study uses a self-supervised deep neural network to identify distinct cloud systems from cloud optical depth satellite images over central Europe. Satellite-retrieved products support the physical interpretation of the identified cloud classes and help optimize the number of identified classes. The trained neural network gives insights into cloud systems’ persistence and transition probability. The generalization capacity of the deep neural network with unseen data is promising but depends on the spatial scale of cloud systems.

1. Introduction

Cloud system appearance defined here as structure and distribution results from a complex interplay of physical processes and large-scale environmental conditions. Since the groundbreaking work by Howard (1803), there has been considerable interest in simplifying the complexity of cloud appearance and classifying cloud systems. With the availability of geostationary satellites providing images over large domains with ever-increasing temporal, spatial, and spectral resolution, the need for automatic classification becomes more and more imminent. A prominent pixel-based classification is the International Satellite Cloud Climatology Project (ISCCP Rossow and Schiffer 1999), which defines nine cloud types based on thresholds in cloud optical depth (COD) and cloud top pressure (CTP).

Extending pixel-based cloud classifications to a regional context, ISCCP has also established the concept of cloud regimes - sometimes also referred to as weather states – for analyzing cloudiness at regional scales. This approach was introduced by Jakob and Tselioudis (2003) and considers the joint 2-dimensional histograms of cloud optical thickness and cloud top pressure across a domain as a basis. Representative mean histograms for the different cloud regimes are obtained by k-means clustering (see MacQueen 1967). The method has subsequently been refined and extended to global scale analyses (Tselioudis et al. 2013), applied to other satellite instruments such as the Moderate Resolution Imaging Radiometer (MODIS) (Oreopoulos et al. 2014), and has been used for climate model evaluation (e.g. Williams and Webb 2008; Jin et al. 2016; Tselioudis et al. 2021). Recently Tzallas et al. (2022) has introduced a cloud regime dataset for Europe based on Meteosat Second Generation satellite observations and the CLAAS2 climate data record (Benas et al. 2017). In a more fitting variety of contexts, our work uses cloud regimes, cloud classes, and cloud patterns interchangeably.

Accepted for publication in Artificial Intelligence for the Earth Systems. DOI 10.1175/AIES-D-22-0096.1.
However, one major shortcoming of this approach is that using the joint 2-dimensional histograms of cloud properties completely neglects the spatial structure and organization of clouds within the considered domain. For cloud systems over oceans in the trade-wind regimes, some preliminary indications from recent research by Bony et al. (2020) showed how different mesoscale cloud patterns impact the top-of-atmosphere radiation budget differently and thus the local climate. Therefore, assessing the organization of cloud systems is important to understand regional climate (changes) but also has potential for other applications such as renewable energy.

Various authors also proposed to use of machine learning algorithms to perform the cloud classification task. Visa et al. (1998) suggested an unsupervised approach making use of Self-Organizing Maps (SOM) using five channels from Advanced Very High-Resolution Radiometer (AVHRR) in polar orbit, which were fine-tuned by the Learning Vector Quantization (LVQ). However, in that approach, the uncertainty in the cloud classification was high, and there was ambiguity in the physical understanding of the classification (Tian et al. 1999). Fabel et al. (2022) used a self-supervised approach to classify clouds from ground-based all-sky images and labeled the clouds according to low, medium, and high cloud base height. However, when prescribed to identify specific cloud features, the neural network does not fully utilize the rich representations present in the data as it focuses on a narrow objective. To overcome the issue of a missing “ground truth,” Stevens et al. (2020) tried to understand cloud systems through visual inspection from scientific peers labeling satellite images identifying four major patterns. Schulz et al. (2020), through a crowdsourcing effort, asked a group of scientists to label multiple independent and overlapping cloud organizations present in visible images of roughly 2,000 km domain into these predefined classes. Their study showed that the average participant needed 30 seconds to classify one image, and agreement over cloud patterns was only reached in 37 percent of cases, showing significant disagreement among scientists for the labeling task. Training a neural network with limited classes can also suppress the information on relevant other cloud regimes. Thus, Denby (2020) used unsupervised machine learning where input image combinations and loss function train the neural network’s parameters to learn the similarity between the same images again over the tropical North Atlantic. He showed that the identified classes are physically independent, supported by spectral bands representative of the physical properties. However, the approach involved human decisions in their learning stage, like the triplet selection of cloud images for training purposes.
Additionally, in contrast to crowdsourcing and Denby (2020) approach, Kurihana et al. (2022a,b) used learning of rotation-invariant dimensionally reduced features through unsupervised auto-encoders to generate a unique 22-year long record of global cloud classification with 42 clusters from Moderate Resolution Imaging Spectroradiometer (MODIS). However, their dataset covers ocean regions only.

In this work, we want to investigate the suitability of a self-supervised deep neural network to identify distinct cloud systems over central Europe. In contrast to the tropical oceans, the central European domain is characterized by more synoptically driven cloud systems, strong land surface type, and topography variability that might hamper our endeavor. We use a resolution-enhanced cloud optical depth (COD) product following Deneke et al. (2021) derived from Meteosat Second Generation (MSG). This can act as a proxy for future improvements in the spatial resolution by Meteosat Third Generation (MTG) observations, launched at the end of 2022. While other approaches directly used RGB images (Denby 2020) or infrared brightness temperature images (Schulz et al. 2021) and thus relied directly on the level-1 satellite observations, we opt for COD and thus a level-2 product as input. The choice of COD is beneficial, as it is a cloud parameter that should be invariant to other influences, as it, e.g., avoids the influence of the diurnal cycle evident in the reflected radiances measured by the satellite. Furthermore, COD has the highest accuracy from all the retrieved parameters in the satellite data set (Deneke et al. 2021) and physically relates to the attenuation of clouds being the quantity that modulates solar energy yields most strongly. However, it needs to be noted that many of the boundary layer clouds driving the variability in solar energy might still not be fully resolved with the spatial resolution of our product.

The main challenge we tackle in the present study is whether deep neural networks can understand different cloud regimes based on the structure and distribution of COD and how this helps explore the cloud systems’ physical properties. We can frame four guiding research questions to detail our work better:

1. How can we use deep learning to capture cloud variability from COD images to identify different cloud regimes?
   Satellite images contain rich visual information, which we transfer into the feature space from which a given number of classes is extracted.
2. How can we physically interpret the identified classes?

In the traditional computer vision domain, self-supervision is often used as pre-training, followed by fine-tuning with some labels as a downstream task for vision applications, i.e., for medical image classification (Azizi et al. 2021). Without using labels to classify the scenes, we use information obtained by cloud and radiative properties retrieved from the multi-channel MSG images to interpret the physical meaning of the classes.

3. What are the measures to identify the correct number of classes?

Our principle for selecting the correct number of classes is that the right number of classes should be the best in identifying distinct properties for the classes in the physical parameter space avoiding correlation between classes.

4. Can the model generalize well on unseen data?

In deep learning, the true capability of a trained model is realized on unseen data, so testing the model on unseen data (i.e.), another year in our case, will ultimately reveal the network’s performance.

One of the inherent challenges of our study is the closeness of the cloud system patterns as they transit from one form to another. Cloud systems are generally chaotic, meaning their transformation from one type to another can be quick and transient. We want the neural network to understand this continuous relationship and to generate an understanding of this connection in the feature space, i.e., the n-dimensional space where the features of the cloud fields are collected.

An additional challenge is that the traditional datasets in computer vision, such as ImageNet (Russakovsky et al. 2015), are prepared in a controlled environment and have a well-balanced class distribution. In contrast, the satellite images are uncurated as their distribution space is not controlled. It is thus very likely that the classes are not well balanced and hence have a distribution sample imbalance among the categories, also known as the long-tailed effect. Under this effect, during training, the major cloud systems present in the satellite measurements drive the decision boundary of the neural network. Recent efforts have shown that Self-supervised learning (SSL) has a good tackle on the imbalanced learning problem (Liu et al. 2022). The learned features are more robust and easy to generalize as the SSL extracts more comprehensive information from the data space.
The present work has the following structure: section 2 describes the satellite dataset and our working domain. Section 3 provides details on the deep learning architecture, section 4 presents our results that aim at answering the questions posed above, and finally, section 5 concludes on our results and points at future work. In the following work, cloud regimes, cloud classes, and cloud patterns have been used interchangeably.

2. Satellite Dataset

Geostationary satellites provide high-spatial and temporal resolution monitoring of the Earth with ever-increasing capabilities and, thus, data volume. For Europe, the Meteosat Second Generation satellites located over the equator and the prime meridian offer full-disk multi-spectral images at 15 min temporal and at 3 x 3 km$^2$ nadir spatial resolution and in 11 narrow-band spectral channels since its launch in 2002 (Schmetz et al. 2002). To mimic the capabilities of the upcoming Meteosat Third Generation (MTG) mission, we use a Meteosat Second Generation (MSG) cloud dataset, which features an improved spatial resolution (Deneke et al. 2021). Herein, the Meteosat - 9, Spinning Enhanced Visible and InfraRed Imager (SEVIRI) is used, which is positioned at 9.6°E as backup satellite and provides the so-called rapid scan service at 5 min repeat cycle. Note that within the operational Meteosat program, Meteosat - 11 has taken over the rapid scanning at 9.6°E, and Meteosat - 9 has drifted over to 45.5°E. Channels 1 - 3 ($\lambda = 0.635, 0.810, 1.640 \mu m$) are used in combination with the broadband high resolution visible (HRV) channel ($\lambda = 0.4 - 1.1 \mu m$) to retrieve and downscale COD and effective cloud radius ($r_e$) from the reflected solar radiation observations. The retrieval is based on the Cloud Physical Properties algorithm developed at the Royal Meteorological Institute of the Netherlands and applied within the context of the Satellite Application Facility on Climate Monitoring (CMSAF) for the generation of the CLAAS-2 cloud data record (Roebeling et al. 2006). Together with COD and $r_e$, other cloud parameters are retrieved, such as liquid water path ($W_L$) and droplet number concentration ($N_D$). The data set also includes cloud top pressure (CTP) derived from thermal infrared channels via radiative transfer calculations, but this dataset does not benefit from the resolution enhancement. All products have 2 km x 1 km spatial and 5 min temporal resolution over central Europe.

We decided to use COD as input for our self-supervised approach because (i) the use of COD instead of reflectance(s) avoids the consideration of the solar zenith angle and surface characteristics,
which have a strong influence on the reflectance, (ii) COD is the most accurate retrieval product (Deneke et al. 2021) at the enhanced resolution, and (iii) COD is related to surface radiation and thus important for solar energy application which we aim to target at a later stage. Nevertheless, it is important to understand the magnitude and impact of retrieval uncertainties. The bi-spectral retrieval method used for estimating the COD relies on the principles described by Nakajima and King (1990) and is highly non-linear. A variety of uncertainty sources affect the overall accuracy, as is discussed in great detail in Grosvenor et al. (2018). Specifically, retrieval results are affected by uncertainties of the instrument itself and its calibration, atmospheric correction, and assumptions on surface albedo. Errors in COD become excessively large outside a value range of about 5 - 50 but are expected to lay below 8 - 10% within this range (see Platnick et al. 2017, their Figure 14). This estimate is valid for warm clouds as observed by the MODIS instrument. Given the efforts directed to validation and instrumental calibration within the context of the Satellite Application Facility on Climate Monitoring see (see Benas et al. 2017; Meirink et al. 2013), a similar accuracy is expected for the MSG-based retrievals of COD used here. It must, however, be noted that this uncertainty estimate does not take into account the effects of unresolved cloud variability, retrieval uncertainties arising for mixed-phase and ice clouds, as well as 3-dimensional radiative effects. Given the focus of the present paper to classify cloud scenes into characteristic regimes based on the COD, there is hope that such uncertainties are regime-dependent and implicitly taken into account by the classification.

We focus on a central European domain (Fig. 1), which exhibits a wide variety of clouds associated with synoptic events and local forcings. While we intentionally apply a machine-learning (ML) based cloud classification over land surfaces for the first time, we exclude the Alps’ topography, where the reflective snow can complicate cloud retrievals. The study period includes spring and summer, i.e., from April to July, to capture more variability in the cloud conditions, including locally triggered low clouds, which pose a challenge to numerical modeling and solar energy production due to their high variability. We avoid dusk and dawn by limiting the period to 6 am to 6 pm UTC. The year 2013 is used for training purposes, and the year 2015 to test the general applicability of the approach.

Ad-hoc, it is unknown which spatial scale is most meaningful for classification. Here we use images of 128 x 128 pixel size corresponding roughly to 260 km x 135 km, where cloud systems
Fig. 1. Top: Geographical map of Europe shown along with topography, thick black lines are coast, the thin black lines are political boundaries, and the black box shows the central European domain. Bottom left: MSG high-resolution cloud optical depth of the central European domain on May 22, 2013, at 13:30 UTC. Four images from 128 x 128 and 64 x 64 configurations are randomly cropped out at every time step. The boundary box inside the COD field represents a 128 x 128 area over Juelich, Germany; cropped at every time step is not used during training but is specifically used for the transfer learning task. Bottom right: The central European domain’s natural color RGB image at standard resolution is shown for better context. Note that cyan colors in the RGB indicate ice.

distributions up to the mesoscale are well represented. This choice is consistent with previous studies that used domain sizes around 250 km x 250 km (Denby 2020) or 1 x 1 to 2.5 x 2.5 degrees (Jakob and Tselioudis 2003; Oreopoulos et al. 2016). We also test a smaller image size of 64 x 64 pixels, roughly 130 km x 68 km, to study the influence of image size. At any instant, we crop four random images from our parent central European domain (Fig. 1). Therefore, with 144 parent images per day for about 120 days in 2013, we obtained 68,000 images per year. We keep 10,000 images unseen from the model for evaluation purposes. Finally, we also extract data over the specific location of Juelich to investigate the temporal variability and the transition from one class to another. The data over Juelich is not used during training and is specifically used for the transfer learning task. Also, data from 2015 are left out in training to investigate the generalization.
3. Deep Learning Architecture

Recent years have seen enormous progress in representation learning by the deep learning community (Caron et al. 2020; van den Oord et al. 2017; He et al. 2020). They facilitate the intelligent use of uncurated images with modest computational requirements. In our case, we want the neural network to learn distinctiveness in the distribution pattern of the cloud systems. The network should learn cloud systems’ global and local patterns and distribution properties from satellite images. As cloud systems change continuously, the neural network should learn the features of the satellite images such that it can discriminate at an instantaneous level.

To set up this architecture, we make use of the software package DeepCluster version 2 (named DCv2 in the following) from Facebook Artificial Intelligence Research (FAIR) (Caron et al., 2018, Caron et al., 2021). The open-source VISSL computer vision library (Goyal et al., 2021) was used to adapt the DCv2 neural network to our requirements. Based on sensitivity tests, we set the number of epochs always to 800, and depending on the available GPU type, batch size is set to 750 or 800.

For the classification task, we opted for a self-supervised approach because, during the representation learning phase, it learns more comprehensively and freely from the data distribution. In our self-supervised method, the network learns directly from the satellite COD observations from an initial dataset of 58,000 images from 2013. The network ingests data in batches, i.e., small groups of images, and for each epoch, the entire 58,000 image dataset makes the complete pass in 72 batches. For each batch, the network adjusts its parameters through stochastic gradient descent to achieve the minimization of the cross-entropy loss function. After 800 epochs (see above), the parameters are optimized. The network can now work with a dataset of COD images the model has not been trained on, and it will assign new images into k categories (or cloud classes). The user decides the number of categories k for clustering clouds; for this study, we varied k between 4 and 12. Our experiment is guided by the well-established ISCCP classification with nine cloud classes and the clear sky category. Thus, we started with k = 10. To investigate the sensitivity of cluster numbers, we run the experiment also with a higher class number, such as k = 12, and reduce them until k=4. The optimum number is investigated in section 4c.

Definition of the network input. \( N \) satellite images of COD built the input training data set \( X = \{x_1, x_2, x_3, \ldots, x_N\} \) of the deep learning architecture illustrated in Fig. 2. For robust and effective
Fig. 2. Schematic diagram of the deep learning network adapted in this work from VISSL library. From left to right: $X_{it}$ is the random crop from $X$ that is split into two similar subsamples $x_{it1/2}$. $F_x$ represents the convolutional neural network (CNN) and $h_x$ its output. For the CNN, the ResNet-50, a residual neural network architecture (He et al. 2015) 50 layers deep with 23.521 million parameters is used in our setup to learn essential features $h_x$ from the high dimensional satellite images. Then, $\text{Proj}(h_x)$ represents a multi-layer perceptron (MLP). The MLP attached to the CNN output, comprises three fully connected layers; the first two have 2048 neurons and the last layer with 128 neurons. To activate the neurons, the rectified linear activation function (ReLU) (Fukushima 1975) is used. The feature vector $z_x$ represents the output from MLP. The convolution of the functions $F_x$ and $\text{Proj}(h_x)$ is the function $f$, representing the non-linear transformations performed by the entire network’s architecture. The regularized architecture limits the latent space’s information capacity and learns the cloud systems’ representation, minimizing the multinomial cross-entropy loss.

Global mapping, image features are learned separately from two slightly differing sub-images. The physical reason behind this approach is that, since physical laws guide the appearance of the cloud systems, the neural network should know that the two random crops almost represent the same system. Thus, for creating two sub-images, each satellite image $x_i$ is transformed using random cropping to obtain two global views $x_{it}$ of the original image ($x_{it1}$ and $x_{it2}$ in Figure 2). Each $x_{it}$ covers 0.75 part of $x_i$.

**Network general architecture.** The neural network’s task is to learn visual features from each satellite image. A function $f$ describes the learning process via a nonlinear mapping $f(x_i) = z_j$ with
$i = 1 \ldots N, j = 1 \ldots M$ that maps the image $x_{it}$ into the array of features $Z = \{z_1, z_2, z_3, \ldots, z_M\}$, where $M$ is the dimension of the feature array $Z$ and in this study, the chosen dimension $M$ is equal to 128. The randomly cropped image $x_{it}$ has dimensions 96 x 96 (0.75 of 128 x 128) or 48 x 48 (0.75 of 64 x 64). The function $f$ represents the transformations performed by the network’s architecture components: a convolutional neural network (CNN) as the backbone and a multi-layer perceptron (MLP) as the head. The CNN can identify spatial dependencies in an image by learning relevant filters. Its output kernel values are then mapped by the MLP nonlinearly into the feature array $Z$ (Fig. 2 for more details). Figure 2 represents the network’s architecture, and the two branches graphically indicate the paths that the global views $x_{it}$ follow in the network. It is important to note that the parameters shared by CNN and MLP of the upper and lower branches are the same.

**Upper branch of the network.** In the upper branch (Fig. 2), applying the function $f$ to the image $x_{it2}$, we obtain the feature vector $Z = \{z_1, z_2, z_3, \ldots, z_M\}$ that then enters the Spherical k-means algorithm (Hornik et al. 2012), that clusters the data in the following way. The algorithm is initialized with random centroids $C = \{c_1, c_2, \ldots, c_k\}$ of dimension $M$. The number of centroids provided to the Spherical k-means corresponds to the number of classes in output $k$, decided by the user. Each centroid $C$ has a label $q_l$ with $l = 1 \ldots k$. The algorithm identifies the closest centroid to the feature vector $Z$ through cosine similarity and assigns $Z$ a pseudo-label $q_s$ corresponding to the selected closest centroid $C_s$. $q_s$ is a pseudo label because it is not the definitive label for the feature vector we will obtain at the end of the iterations as it changes with epochs to reduce the cross entropy loss. Thus, through the upper branch, the network assigns a preliminary label $q_s$ corresponding to a centroid $C_s$ to the feature vector $Z$ obtained from the image $x_{it2}$.

**Lower branch of the network and cost function minimization.** In the lower branch, the function $f$ is applied to the image $x_{it1}$, similar to the image $x_{it2}$ that went in the upper branch. The function $f$ outputs a feature vector $Z$, which is first compared with the centroids of upper branch $C$ by calculating a normalized dot scalar product. Then, $Z$ is provided as input to the cross entropy loss function, together with the output of the upper branch $C_s, q_s$. The network now minimizes the multinomial logistic loss function $L$ given by:

$$L(Z, q_s, C_s) = - \sum_{l=1}^{k} q_l \log(p_l)$$

(1)

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0096.1.
Where $p_l$ is the normalized exponential of the cosine score for the class $l$ divided by the sum of the cosine scores for all the classes and represents the probability distribution of the cosine score:

$$p_l = \frac{\exp(\frac{1}{\zeta}Z^TC_l)}{\sum_{m=0}^{k'}\exp(\frac{1}{\zeta}Z^TC_m)}$$

(2)

Note that the top and bottom branch between the Spherical K-means block and the cosine similarity is somehow connected as the centroids come from the k-means of the top branch. In equation 2, $\zeta$ is a normalization factor called temperature that controls the sharpening of the $p_l$, and $Z^TC_l$ represents the scalar product between the vectors $Z$ and $C_l$. We have taken it as $\zeta$ from the benchmark configuration in VISSL. When the feature vectors of the two branches capture similar information about the original image, the loss becomes lower and higher when they diverge. That’s how the network branches are encouraged to focus on the image characteristics, which progressively makes the feature vectors similar. The loss function minimization happens progressively layer by layer, derivating the loss function $L$ with respect to $Z$ parameters and adjusting parameter values in each layer by backpropagation. At the end of the minimization, we obtain a configuration $\theta$ of the parameters of the network that will be ready for a new iteration (batch) over the two branches. To investigate the suitability of the DCv2 model for cloud classification, we changed the loss code to obtain the distance of the feature vectors from the centroids, the features of the centroids, and the class category or the pseudo labels of each image sample. Furthermore, we made some changes in the nearest neighbor test algorithm in the VISSL library to find the nearest neighbors from the centroids.

**Network learning behavior.** To better understand how the computational model works internally and learns progressively from subsequent epochs, we extract the features at different levels of training to visualize and better comprehend the progress in the clustering around the k classes provided as input. Figure 3 shows the ability of a self-supervised approach to extract the rich representations in the dataset. In our case, we do not have labels to measure the accuracy of the scene classification task. Therefore we assess the quality of the learning algorithm by looking at the dimensionally reduced feature vector.

To represent the M-dimensional feature vector $Z$, in this case, we used the t-Distributed Stochastic Neighbor Embedding (tSNE) algorithm (van der Maaten and Hinton 2008), which tries to preserve
the relative position of points while mapping it to lower dimensional space. Here we map the 128 dimensions into two, and for the visualization, in Figure 3, we prescribe the number of classes to ten (k=10).

The neural network ingests the data at the first epoch and performs spherical clustering for the first time using the embeddings (synonym for feature vector Z). The model, in this way, generate the first idea of a few different class distributions (for instance, pink and yellow dots in Fig. 3a). At epoch 25, spirals of data points are clustering together (Fig. 3b). To interpret self-supervised learning
using an energy-based regularized framework (LeCun and Huang 2005), an energy function, the cross-entropy loss here, assigns low energies to training samples. It uses a regularizer, spherical k means here, and the loss function pushing the energy down along the data points that minimizes the volume of the low energy regions. In such context, Ranzato et al. (2007) showed that compressing the information to represent the data can help reduce the training points’ energy and maximize the unobserved points’ energy. We observe a spiraling behavior supporting the energy-based model theory. At epoch 250 (Fig. 3c), the model achieves convergence, and near epoch 800 (Fig. 3d), a clear segregation among the classes appears with maybe some outliers (like in yellow dots of Fig. 3d). The good performance of the network displayed convinced us to stop the training at 800 epochs.

**Optimization for the convergence.** We used stochastic gradient descent (SGD) (Ketkar 2017) to optimize the training process with momentum as compared to gradient descent which uses the average gradient of all training samples. SGD allows a slightly different loss curve with a randomly selected mini-batch, and this helps to reduce the computations enormously. Using momentum in the SGD lets the past and the current gradients guide the overall direction of the loss curve. In addition, DCv2 uses a learning rate scheduler that modifies the learning rate based on the progression in the training process. We use the learning rate scheduler of linear warm-up plus cosine with the linear start value of 0.3, end value of 4.8 and cosine start value of 4.8 end value of 0. VISSL supports mixed-precision training from NVIDIA’s Apex library to reduce the model memory requirement and better training speed. We have used mixed precision to save GPU memory and thus allow larger batch-size training. Therefore, NVIDIA’s Large Batch Training of Convolutional Networks (LARC) and SGD are used to improve large batch size training convergence. To avoid collapse, DCv2 applies batch normalization for stability apart from the clustering constraint. The task of the batch normalization layer is to take outputs from the last layer and normalize them before passing them as input to the next layer. For multi-GPU training, synchronized batch normalization normalizes the input within the mini-batch because it helps share the statistics across all the devices. For conversion to global batch normalization, VISSL supports NVIDIA’s Apex SyncBatchNorm. We have used four NVIDIA V100 GPU’S for training purposes. For training the neural network for 800 epochs on 4 V100 GPUs it took 16 hours or 64 core hours in total.
4. Results

a. Analysis of the clustering capability

Once the network has sorted the individual images into different classes, we want to check the consistency within a single class. Thus, we quantify how close to each other the evolving cloud patterns are in the feature space. We can check how images from the same cluster that are spatially close or far in the feature space are visually similar to gain confidence in the quality of the representation of the different cloud patterns in the feature space. For this analysis, for each centroid, we select the closest image to a centroid (called centroid in the following) over Central Europe (Fig. 4) and compare them with their two closest and their two farthest images. The seven centroids from Figure 4 show distinct COD patterns; however, the physical separation of identified classes and their interpretation will become clearer in the later sections. Centroid 2 is associated mainly with clear sky conditions, and centroid 3 has optically thin clouds. Centroids 4 - 6 reveal extended cloud fields with a mean cloud fraction higher than 94% though their distribution differs, with centroid 6 having the highest COD values representing intensive convective cells. The images closest to the centroids resemble their behavior, while the ones farthest away sometime deviate to a larger extent. That indicates the distance to a centroid can be used to measure the classification’s quality. It is worth noting that Figure 4 visually demonstrates the distinctiveness and separation of the features learned within a single class from the cloud systems found in other classes. Additionally, to quantitatively assess the quality of our embeddings in seven clusters, we calculated the silhouette score (Rousseeuw 1987) for our high-dimensional and complex dataset, yielding a score of 0.54. This score ranges between -1 to +1, and low or negative values indicate that the clustering configuration may have too many or too few clusters. Thus, the mean score of 0.54 for all the embeddings shows that clusters are considerably compact and non-overlapping, supporting our qualitative assessment. Therefore, scores like silhouette can also be used to find the optimal number of clusters, as discussed in section 4b. Further analogies in the cloud patterns will be investigated in section 4d, where we aim at a physical interpretation.

b. Determination of the optimal number of classes

As stated in section 3, users provide the number of classes k utilized by the network to cluster the data. We want to understand, for each configuration, what is the optimum number of classes.
Fig. 4. Each row shows five 128 x 128 COD images belonging to a certain cloud regime (CR) over central Europe. From right to left: centroid, closest 1\textsuperscript{st}, closest 2\textsuperscript{nd}, 2\textsuperscript{nd} farthest, farthest are shown along with the color scale for COD reference purposes.

This number should be as low as possible, but at the same time, it should separate the radiative and physical properties of different clouds. We use for this purpose the spatial distribution of COD and CTP as retrieved from MSG. In this way, we follow Oreopoulos et al. (2016) who distinguished cloud regimes based on COD - CTP using k-means clustering methods, starting from a core ensemble of six cloud regimes and then verifying if each regime can be split into sub-regimes. To split a regime into sub-regimes, they evaluate the differences in the spatial pattern correlation coefficients between their centroids (mean joint histograms) to some threshold values.
Here, we use a bottom-up approach, starting from varying amounts of cloud regimes (from 12 to lower cluster numbers). We check the correlation in the joint histograms of COD and CTP and also look at sensible distributions of the mean cloud fraction and the relative frequency of occurrence (shown in Fig. 6) over the central European domain (more details on the joint histograms can be found in section 4d). Based on this, we define the following thresholds for the correlations: 0.65 for the 128 x 128 and 0.7 for the 64 x 64 sample images. A correlation between two clusters higher than a threshold value would indicate that clusters are not sufficiently independent. In fixing the threshold, we got inspired by Oreopoulos et al. (2016), who selected a threshold of 0.6, but in this work, the thresholds are slightly more restrictive. Additionally, for 64 x 64, since the domain is comparatively smaller than the 128 x 128 sample images, it has more variability in terms of cloud system pattern and distribution. Here we restrict with a relatively higher threshold of 0.7.

For the 128 x 128 configuration, Figure 5 shows that the configuration with seven classes is the first one in which the correlation among the classes is never larger than the prescribed threshold. The number of classes whose correlation is higher than the threshold (red squares in Fig. 5) gets reduced if we decrease the number of cloud regimes from 12 to 7. The network can distinguish seven classes, thus the largest number of classes, without producing some superposition of the cloud radiative properties. In conclusion, we find the optimal number of classes to be 7 for the 128 x 128 configuration and 8 for the 64 x 64 configuration.

c. Optimum classification

Now we want to investigate how the clustered satellite images arrange themselves in the representation space. Our goal is to see if, giving k classes (as determined in section 4b) as a clustering constraint, the features from the satellite images are strong enough to be captured by the deep neural network for a qualitative classification of distinct types of clouds. Since visualizing the feature vector Z is impossible because it has 128 dimensions, we reduce the dimension of the feature array to a 2D space using the tSNE algorithm by van der Maaten and Hinton (2008) and visualize the classes in a 2-dimensional space in Figure 6.

Figure 6 shows the representation learning capability of the neural architecture using the self-supervised algorithm. When looking at the 2-dimensional reduced feature space, distinct cloud patterns occupy distant areas. As already noted in section 4a, class 2 relates to clear sky conditions.
Fig. 5. Correlation patterns among classes for the number of cloud regimes varying from 12 to 7 for 128 x 128 configuration. Red squares with hatches indicate the classes with a correlation of higher than 0.65 in COD-CTP space. Note that for smaller numbers of cloud regimes, no significant correlation appears.

and is located clearly on the edge of the 2D space (Fig. 6a left). In contrast, on the opposite edge (Fig. 6a right), we find the classes with the strongest convective cloudiness (classes 5 and 6) as identified by small areas of enhanced COD. Similar distinct behavior can be found in other regions of the 2D space and will be discussed in more detail in the next section. Comparing the configurations with 128 x 128 and 64 x 64 pixels, one can initially notice that the latter captures more variability in the cloud systems because smaller images can detail more information on cloud structures. Section 4d will investigate this in more detail.

d. Physical significance of the identified classes

In this section, we analyze the physical properties of the identified cloud classes. The goal is to verify the independency of the identified classes in the physical spaces given by a combination of radiative and cloud properties. An overview of different characteristics in different classes is given in Table 1. We first use the surface downwelling solar radiation (SDS), and the top of the atmosphere (TOA) reflected solar radiation (TRS) (Fig. 7a). Note that these products have been
Fig. 6. Classification by the deep neural network at a) $k = 7$ for 128 x 128 configuration and b) $k = 8$ for 64 x 64 configuration. Thirty random samples were selected over central Europe and visualized for each class. Each image is assigned a unique class through a colored frame. The bar charts in the lower part of a) and b) represent the number of images in relative percentage and cloud fraction in each regime. To better associate the class number with the cloud regime, the centroid image is shown as well.

derived with radiative transfer calculations (see section 2). Moreover, we analyze CTP and COD (Fig. 7b). As in Denby (2020), we randomly select 1000 images for each of the seven optimal classes of the 128 x 128 configuration. Then, we derive their mean position in the space of the SDS-TRS (Fig. 7a) and similarly for COD and CTP (Fig. 7b).

The collocation of the classes in the radiative and cloud properties space permits the identification of the types of clouds the network distinguishes (Table 1, Fig. 7). Class 1 (light blue dots) probably corresponds to cirrus clouds with a COD mean value smaller than ten and cloud top pressure around 400 hPa, approximately 7500 m. Cirrus show a large variability in albedo and surface downwelling.
Table 1. The relative frequency of occurrence (RFO), Mean and standard deviation (std) of cloud mask (CMASK), cloud optical depth (COD), cloud top height (CTH), cloud top pressure (CTP), cloud top temperature (CTT), shortwave downwelling radiation (SDS), TOA reflected solar radiation (TRS), for all the classes identified by the network in the 128 x 128 configuration.

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFO</td>
<td>13.2</td>
<td>19.6</td>
<td>15</td>
<td>16.3</td>
<td>16.8</td>
<td>12.4</td>
<td>6.6</td>
</tr>
<tr>
<td>CMASK mean (%)</td>
<td>92.2</td>
<td>27.3</td>
<td>64.5</td>
<td>96.5</td>
<td>99.1</td>
<td>94.8</td>
<td>88.2</td>
</tr>
<tr>
<td>CMASK std (%)</td>
<td>10.7</td>
<td>21.6</td>
<td>17.6</td>
<td>5.5</td>
<td>3.6</td>
<td>8.5</td>
<td>12</td>
</tr>
<tr>
<td>COD mean</td>
<td>5.3</td>
<td>1.2</td>
<td>3.3</td>
<td>15.3</td>
<td>24.8</td>
<td>23.4</td>
<td>7.4</td>
</tr>
<tr>
<td>COD std</td>
<td>4.6</td>
<td>2.7</td>
<td>3.3</td>
<td>6.9</td>
<td>11.5</td>
<td>13.9</td>
<td>4.4</td>
</tr>
<tr>
<td>CTH mean (m)</td>
<td>8212</td>
<td>5152</td>
<td>3198</td>
<td>3866</td>
<td>6001</td>
<td>5642</td>
<td>3529</td>
</tr>
<tr>
<td>CTH std (m)</td>
<td>1734</td>
<td>3606</td>
<td>2226</td>
<td>1625</td>
<td>2160</td>
<td>2245</td>
<td>2259</td>
</tr>
<tr>
<td>CTP mean (hPa)</td>
<td>374</td>
<td>613</td>
<td>736</td>
<td>657</td>
<td>498</td>
<td>530</td>
<td>697</td>
</tr>
<tr>
<td>CTP std (hPa)</td>
<td>101</td>
<td>245</td>
<td>164</td>
<td>118</td>
<td>142</td>
<td>146</td>
<td>157</td>
</tr>
<tr>
<td>CTT mean (k)</td>
<td>236</td>
<td>260</td>
<td>271</td>
<td>264</td>
<td>248</td>
<td>251</td>
<td>265</td>
</tr>
<tr>
<td>CTT std (k)</td>
<td>10</td>
<td>24</td>
<td>15</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>SDS mean (Wm(^{-2}))</td>
<td>442</td>
<td>546</td>
<td>582</td>
<td>408</td>
<td>253</td>
<td>194</td>
<td>488</td>
</tr>
<tr>
<td>SDS std (Wm(^{-2}))</td>
<td>183</td>
<td>232</td>
<td>196</td>
<td>143</td>
<td>104</td>
<td>123</td>
<td>171</td>
</tr>
<tr>
<td>TRS mean (Wm(^{-2}))</td>
<td>298</td>
<td>155</td>
<td>224</td>
<td>401</td>
<td>530</td>
<td>284</td>
<td>328</td>
</tr>
<tr>
<td>TRS std (Wm(^{-2}))</td>
<td>183</td>
<td>232</td>
<td>196</td>
<td>143</td>
<td>104</td>
<td>123</td>
<td>171</td>
</tr>
</tbody>
</table>

Solar radiation that depends on the natural variability of cirrus clouds: they can be thicker or semitransparent and contain larger or smaller ice crystals, profoundly impacting their radiative properties. The cloud-free class 2 has the lowest cloud fraction (27%), with most solar radiation reaching the surface. Due to the lower albedo of the surface compared to clouds, less radiation is reflected back to space. Note that SDS and TRS experience a strong diurnal cycle, and only at noon SDS values > 700 Wm\(^{-2}\) are reached. Class 3 and 7 correspond to relatively low-level clouds (orange-shaded dots in Fig. 7) with a medium cloud fraction of 65 and 88%, respectively. These clouds have relatively small cloud optical depth values, with their mean values displaying cloud top altitude below 600 hPa (around 3,000 m). Note that frequently low-level clouds are topped by overlaying (semi-transparent) ice clouds (as shown in Fig. 1), which explains the relatively high standard deviation in CTH of more than two kilometers. However, there are differences between class 3 and class 7: class 3 has only half of the COD as class 7, but its relative standard deviation is much higher, indicating high variability, which is demanding for solar energy production. Class 7,
Fig. 7. a) Mean TOA reflected solar radiation (TRS) versus mean surface downwelling solar radiation (SDS) given by dots of 1000 randomly selected images of the 128 x 128 configuration with color indicating the class they belong to. Triangles indicate the mean position of the classes. The images associated with the triangles are the closest to the mean position of the 1000 images. The crosses in the triangles represent the error bars of each cluster group for the radiative properties on x and y. b) Same as a), but in the 2d space given by COD (abscissa) and CTP (ordinate).

possibly associable with deeper low-level clouds, displays larger COD values and smaller surface downwelling solar radiation than the shallower low-level clouds of class 3.

Classes 4, 5, and 6 (blue-shaded dots in Fig. 7) all describe overcast situations with cloud fractions of more than 94%. For these classes, average COD values are larger than 10, and they span a broad region of CTP values above 800 hPa, indicating various stages and intensities of convective development. There are more interesting different features among the classes. Class 5 and 6 have very similar mean values of COD and CTP around 25 and 500 hPa, which especially differ from the mean COD of 15 and the CTP of 650 hPa of class 4. These differences indicate that classes 5 and 6 correspond to deeper and thicker clouds than class 4. However, class 6 has the highest relative standard deviation of COD (55%) of the three classes, which aligns with the impression of the strongest convective activity as discussed with respect to Figure 6. Still, there is a distinct difference, as class 5 reflects twice as much radiation into space as class 6. Thus, despite
of similar cloud fraction, the network is able to identify different structural features within these three "overcast" classes.

Finally, we investigate the joint histograms of cloud optical depth and cloud top pressure (Fig. 8) for distinctness. Clearly, each class occupies its characteristic areas in the 2-dimensional space; however, some overlap exists: From the three overcast classes, both classes 5 and 6 include high clouds though class 6 is more limited to the highest tops and has a broader spectrum of COD. The third overcast class (class 4) also includes lower clouds and thus slightly overlaps with the low-cloud class 7, but the occurrence of high COD values (23 and above) becomes more prominent for class 4 at the mid-cloud level. The overlap between classes 2 and 3, clear sky and shallow low-level clouds, respectively, can be explained by the rapidly changing nature of low-level clouds, which can easily form or dissolve and show large variability in space.

For the fine resolution of 64 x 64 also, no strong overlap exists but it appears that the high occurrence is bound to smaller areas in the COD-CTP space except for classes 6 and 8, likely due to the stronger separation into eight classes. In fact, this might be helpful to identify low-level clouds, which seemed to be more confined in class 5 as in the 128 x 128 configuration. Overall, we find that the classes identified and the properties represented through these classes are exclusive to a large extent though some overlaps occur. Therefore using satellite-retrieved products supports the physical interpretation of the identified cloud classes. Note that a detailed investigation of the different classes’ characteristics is beyond this study’s scope.

Fig. 8. Joint histogram of COD (abscissa) and CTP (ordinate) for 128 x 128 (top) and 64 x 64 pixel (bottom) configurations.
Fig. 9. Correlation of centroids obtained from 2013 and 2015 datasets for a) 128 x 128 configuration and b) 64 x 64 configurations. Below each joint histogram are the images closest to the centroid from 2015 (indicated with *) and 2013, respectively, for each class. Note that the class number is resorted as indicated by * according to matching in the correlation of centroids.

**e. Generalization**

This section will test how the model behaves when ingested with satellite data from a different year. The research question we want to answer is whether the model can identify the same centroids detected with the data from 2013 using data from 2015. The goal is thus to discover how the centroids will arrange themselves to represent the cloud regimes of unseen cloud system distributions. The only information the network retains from the run on 2013 data are the values of the internal parameters of the neural network (CNN and MLP). We also classify the 2015 data set into the same number of optimal classes as discussed in section 4c.

To answer this question, we then correlate the 2013 and 2015 centroid embeddings and compare them around other respective centroids for both configurations (Fig. 9). This type of analysis should reveal the inter-cluster spread in both years. The 128 x 128 configuration generally performs better on unseen data than the 64 x 64 configuration. In the 128 x 128 configuration, all centroids from unseen (2015) data coincide with the centroids from seen (2013) data (Fig. 9a). The highest off-diagonal column-wise centroid correlation is 0.53, which suggests that in a 128 x 128 configuration,
the newly identified classes from the 2015 data set fit nicely with the pre-trained network from 2013.

In the 64 x 64 configuration, it is more challenging for the network to classify scenes similar in both years. Here the off-diagonal column-wise centroid correlations are much higher, with values up to 0.93. This especially holds for two classes: 1) a mixture of deep convections along with a spread out of low conviction (class 3) and 2) varying clear sky conditions (class 1). For example, both centroids of classes 3 and 6 from 2015 display a correlation higher than 0.9 (both main and off-diagonal) with class 3 of 2013, while other centroids get discriminated more clearly (Fig. 9b). These classes have almost the same type of mixed convection, but rather than being in one class as in 2013; they belong to two classes in 2015. Furthermore, from 2015 class 8 represents very low convection, whereas class 4 groups almost clear sky cases. Both these classes correlate for 0.84 (main diagonal) and 0.74 (off-diagonal) with class 1 of 2013, representing an almost clear sky. Similar to the mixed convection case, rather than being in one class of almost clear sky type, it is present in two different cloud classes in 2015.

While the generalization works well for 128 x 128 configuration, problems appear for the smaller scale. We think this might be related to two probable reasons. 1) The spatial scale of cloud systems, i.e., the scale at which cloud systems occur over a region. For the network to properly represent the clouds, such a scale should be well represented in the input dimensions (128 x 128 representing well over central Europe). Still, in the 64 x 64 case, the scale captured might be small to understand the structure and distribution of cloud systems, which is especially challenging when scales include a mixture of clear sky, convective, and stratiform regimes. 2) We get more classes of variability on a smaller scale and thus likely need more training data. However, letting the network run on 2015 data for a few more epochs might also improve its performance and decrease the centroid correlations in complex cloud systems.

f. Persistency of cloud regimes and probability of transition

Finally, we move towards applications and use the network’s ability to identify cloud regimes over a fixed location (Juelich, Germany, in our case). We can study the temporal development of specific cloud regimes and exploit the 5-minute temporal resolution of the dataset to observe the transformation of cloud systems with time at a higher resolution. Note that the network trained
with random crop settings over central Europe has not previously ingested the temporal continuous data over Juelich. In other words, the neural network has no idea of the temporal cohesivity of the satellite images. This experiment also gives us insights into the application of transfer learning for the neural network to study cloud systems over a fixed geospatial location. Looking at images from Juelich (see Fig. 1) for an entire day, we can investigate how one class transitions towards another in time. The goal is to understand how close classes should be in the feature space to make a transition and derive a transition probability from one cloud regime to another. Figure 10 illustrates the temporal evolution of cloud regimes for two selected days. In this way, the computational model can provide information on the persistency of the cloud classes over a specific region. Furthermore, it also reveals that sometimes a mixed regime appears, as can be seen for 20 June 2013 (Fig. 10a) in the morning when the alteration between two deep convections (class 5 and 6 is observed). The network also shows how deep convection decreases, the lower level cloud appears (class 3), and the clear sky conditions prevail in the afternoon. Figure 10 also illustrates how similar the four closest images (in terms of cosine distance) look to a randomly selected test image for both days.

More specifically, we can detect how long a specific cloud regime stays on average over the location or what is the most probable cloud regime it can evolve into. In the 128 x 128 configuration, the clear sky regime (class 2) displays the longest persistency over the site (Fig. 11a)). It is also the cloud regime that most probably evolves into the shallow low-level type (class 3) (Fig. 11b)). Similarly, classes 4, 5, and 6, representing different stages of convective cloud structures and distributions, display a persistency of around 75 minutes and often evolve into one another. Class 7 is a class that stays in one fixed location for the shortest time, an average of 30 minutes. These clouds are located near the boundary layer and could represent deeper low-level clouds. They often evolve into class 5, i.e., grow deeper to a full convective cloud, or with slightly less probability, can also evolve into class 3 (shallower low-level clouds), maybe after a precipitation event. Note that the low persistence time results from the occurrence of frequent jumps between two classes (see Fig. 11) during the transition time and that temporal smoothing might be beneficial for the analysis of the long-term time series.
Fig. 10. Temporal evolution and transition of two example days for the cloud systems over Juelich a) on the 20 June 2013 as identified by the neural network. Here $t = 0$ represents a COD image that is given as a test case (identified with the vertical line + symbol), and the neural network is asked for the four nearest images sampling over Juelich. The time stamps represent the order of their closeness from the test image ($t = 0$) in the feature space. b) same, for 25 July 2013.

Fig. 11. a) Cloud systems persistency over Juelich, Germany, given by the average interval of occurrence of cloud regimes over the site. b) Matrix representing the probability of transition of each cloud regime to another class derived over Juelich, Germany. We used one year of data for both analyses.
5. Conclusions and outlook

This work uses a deep neural network architecture and a spherical k means algorithm to detect cloud regimes from cloud optical depth images over Europe for the first time. Classifying cloud regimes is a long-standing problem in Earth system sciences. Central Europe’s mid-latitude geographical location diurnally gives rise to majorly two cloud families: 1. a mix of stratiform and convective clouds for night time and 2. convective type in the day time. Keeping our long-term vision on solar energy applications, we wanted first to disentangle them and focus on convective cloud systems. We used COD measurements, available in the daytime only, for summer periods. In this way, we were able to investigate a wide variety of convective and stratiform clouds. In the future, with many more years of data, we intend to analyze the diurnal behavior of the different classes. We propose this approach to exploit the greater freedom that the deep neural network offers in understanding the complexity of the cloud patterns. We then verify the results by analyzing the cloud physical and radiative properties of the cloud regimes obtained from the network. The cloud optical depth values and all the other radiative properties are retrieved as a post-processed satellite product by Deneke et al. (2021) to mimic the Meteosat Third Generation (MTG). In this way, our approach can also be seen as an opportunity to reduce future satellite missions’ complexity and data volume. We investigated two different configurations of 128 x 128 and 64 x 64 pixels to study the effect of scale. At first, we tried to understand how close classes are in the feature space to better capture how cloud regimes can transform into one another. A high similarity is found among the closest and farthest images to a randomly selected one from each class. We then determined the optimal number of classes representing the cloud pattern variability. The cloud regime number that adequately represents independent and mutually exclusive cloud patterns is seven for the 128 x 128 configuration and eight for the 64 x 64 configuration. Adopting those number of classes, in fact, the cloud regimes appear visually distinct and positioned in different areas of the feature space.

Exploiting cloud and radiative satellite products, namely cloud optical depth, cloud top pressure, surface downwelling, and TOA reflected solar radiation, we associated each cloud regime with specific physical properties for the 128 x 128 configuration. Class 1 represents cirrus clouds, while class 2 groups clear sky cases; classes 3 and 7 represent two types of low-level clouds, namely shallower and deeper. Finally, classes 4, 5, and 6 comprehend various convective clouds.
To test the applicability of the network, which did not get any temporal information during training, we investigate temporal variability of the cloud regimes over a specific location (Juelich, Germany). In the 128 x 128 configuration, we identified the most persistent cloud regime as the clear sky (155 minutes), followed by convective clouds (70 - 75 minutes). Low-level clouds, as well as cirrus clouds, last around 30 - 40 minutes. These results can help better understand each cloud class’s diurnal cycle over the area of interest and are relevant to describe the highly variable conditions for solar energy production.

Finally, we also tested how the network was clustering data on a new unseen dataset by running it with data from 2015. We found that all centroids from unseen data coincide with those from 2013 in the 128 x 128 configuration. In the 64 x 64 configuration, the network had difficulties discriminating the same centroids, possibly also due to the smaller dataset available. However, the generally good agreement between 2013 and 2015 shows that the data distribution between the two years has not changed much. Nevertheless, this highlights the importance of choosing the right image size according to the problem at hand.

This study followed a purely data-driven approach and did not investigate the properties of detected cloud regimes in detail. However, in the future, we aim to investigate which classes correspond to situations most determining for production and distribution of solar energy. Herein also, the transition probability between different classes offers the potential for short-term prediction capability. For this application, additional training information will also be explored.

Despite using a resolution-enhanced product, geostationary measurements are limited in the detection of boundary layer clouds because of their lower spatial resolution, the question that still can be further explored is whether the lower image resolution offered by geostationary satellites eliminates valuable information for the classification of boundary layer clouds. Sabottke and Spieler (2020) find a case-to-case dependence of image resolutions in radiography. Therefore, determining the optimal satellite resolution for identifying different cloud regimes by machine learning applications remains an open problem and will depend on the application. COD retrieval is also prone to uncertainties (see Sec. 2) Tan et al. (2015) find that cloud regimes are less susceptible to minor, random, or statistically insignificant changes in cloud properties. The highest uncertainties occur for COD retrievals at large values outside the range of 50. However, at these
high values, downward solar radiation is already heavily extinguished so that the application for solar energy, which we aim to pursue next, will not be affected strongly.

This work shows the potential inductive bias of convolutional architecture to handle complex patterns during the first few initial epochs of self-supervised training. At each epoch, the representation learning of the feature improves. However, future work can focus on seeing whether the network can also learn some hidden properties of the satellite images, such as the geo-localization of cloud systems or environmental parameters associated with a given cloud pattern. The network in the feature space might be able to reveal the specific combination of environmental parameters leading to convection.

In this work, one of the main difficulties connected with the training of the DCv2 on satellite images is the absence of control over the distribution of images. For example, specific cloud types like shallow clouds could be overrepresented within the datasets compared to deep convection. In this case, the model could saturate, and it would not be able to extract more information when fed with additional images. Moreover, by design, DCv2 favors a well-balanced class dataset; therefore, the network’s performance is expected to decrease when the algorithm works with satellite images because naturally occurring cloud systems do not conserve class balance properties. One solution to attempt in future work could be to attach a weight (1D Tensor) to the cross-entropy loss function that contains the prior knowledge about the class distribution of the dataset.

Finally, satellite images have more factors of variation than typical benchmark computer vision data sets such as ImageNet. For example, clouds usually spread over the image, generating a different distribution from a usual computer vision data set, which is normally highly object-oriented and centered. Next, it is also likely that two or more very different cloud patterns may be present in an image. Thus have multiple concepts in a single image. However, the spherical k-means algorithm uses the dominant concept (pattern and distribution) present in the images and attributes the whole image to a single label. We thus think it would be beneficial if we could extract multi concepts from the features and use them for training. Present-day transformers pre-trained using self-supervision and different pre-text tasks (Atito et al. 2021; Caron et al. 2021) have shown good promise in understanding the contextual relation of semantics present in the image. Their suitability will be investigated in the future.
Acknowledgments. Dwaipayan Chatterjee’s research has been supported by Federal Ministry for the Environment, Nature Conservation, Nuclear Safety, and Consumer Protection (Grant agreement: 67KI2043, https://kiste-project.de/). This work was performed as part of the Helmholtz School for Data Science in Life, Earth, and Energy (HDS-LEE). Dwaipayan would also like to thank Quentin Duval and others from Facebook’s artificial intelligence research for maintaining the VISSL computer vision library. Dwaipayan would like to thank the three anonymous reviewers for taking the time to provide constructive comments and for making the manuscript better.

Claudia Acquistapace’s research has been supported by the Deutsche Forschungsgemeinschaft (Individual proposal, project number 437320342), project “Precipitation life cycle in trade wind cumuli” (https://gepris.dfg.de/gepris/projekt/437320342, last access: 23 December 2021).

Data availability statement. The data and code to produce this work can be accessed at DOI:10.5281/zenodo.7437949 or via web link https://zenodo.org/record/7437949

References


Howard, L., 1803: I. on the modifications of clouds, and on the principles of their production, suspension, and destruction; being the substance of an essay read before the askesian society in the session 1802–3. *The Philosophical Magazine, 17 (65)*, 5–11.


Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0096.1.


Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0096.1.