Machine-Learned Cloud Classes From Satellite Data for Process-Oriented Climate Model Evaluation

Arndt Kaps, Axel Lauer, Gustau Camps-Valls, *Fellow, IEEE*, Pierre Gentine, Luis Gómez-Chova, *Senior Member, IEEE*, and Veronika Eyring

***Abstract*— Clouds play a key role in regulating climate change but are difficult to simulate within Earth system models (ESMs). Improving the representation of clouds is one of the key tasks toward more robust climate change projections. This study introduces a new machine-learning-based framework relying on satellite observations to improve understanding of the representation of clouds and their relevant processes in climate models. The proposed method is capable of assigning distributions of established cloud types to coarse data. It facilitates a more objective evaluation of clouds in ESMs and improves the consistency of cloud process analysis. The method is built on satellite data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument labeled by deep neural networks with cloud types defined by the World Meteorological Organization (WMO), using cloud-type labels from CloudSat as ground truth. The method is applicable to datasets with information about physical cloud variables comparable to MODIS satellite data and at sufficiently high temporal resolution. We apply the method to alternative satellite data from the Cloud_cci project (ESA Climate Change Initiative), coarse-grained to typical resolutions of climate models. The resulting cloud-type distributions are physically consistent and the horizontal resolutions typical of ESMs are sufficient to apply our method. We recommend outputting crucial variables required by our method for future ESM data evaluation. This will enable the use of labeled satellite data for a more systematic evaluation of clouds in climate models.**

Manuscript received 26 April 2022; revised 27 October 2022; accepted 21 December 2022. Date of publication 23 January 2023; date of current version 31 January 2023. This work was supported in part by the European Research Council (ERC) Synergy Grant "Understanding and Modeling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 Research and Innovation Programme under Grant 855187, in part by the ESA Climate Change Initiative Climate Model User Group (ESA CCI CMUG), in part by the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under Project ID bd1179, and in part by the Leibniz Supercomputing Centre. The work of Gustau Camps-Valls and Luis Gómez-Chova was supported in part by the Spanish Ministry of Science and Innovation under Project PID2019-109026RB-I00. The work of Pierre Gentine was supported by the National Science Foundation Science and Technology Center for Learning the Earth with Artificial intelligence and Physics (LEAP). *(Corresponding author: Arndt Kaps.)*

Arndt Kaps and Axel Lauer are with the Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), 82234 Oberpfaffenhofen, Germany (e-mail: arndt.kaps@dlr.de).

Gustau Camps-Valls and Luis Gómez-Chova are with the Image Processing Laboratory (IPL), University of Valencia, 46980 Valencia, Spain.

Pierre Gentine is with the Department of Earth and Environmental Engineering and the Center for Learning the Earth with Artificial intelligence and Physics (LEAP), Columbia University, New York, NY 10027 USA.

Veronika Eyring is with the Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, 82234 Oberpfaffenhofen, Germany, and also with the Institute of Environmental Physics (IUP), University of Bremen, 28359 Bremen, Germany.

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2023.3237008, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2023.3237008

***Index Terms*— Climate modeling, clouds, CloudSat, CUMULO dataset, ESA Cloud_cci, machine learning, Moderate Resolution Imaging Spectroradiometer (MODIS), process-oriented model evaluation.**

I. INTRODUCTION

EARTH system models (ESMs, also referred to as climate models) are important tools not only to improve our understanding of present-day climate but also to project climate change under different plausible future scenarios.

The simulation of clouds and their interactions with the climate system, however, remain a major challenge for ESMs [1]. The representation of clouds in these models has been identified as one of the primary sources of intermodel spread [2], [3]. An improved representation of cloud processes in ESMs is therefore an essential component in addressing these issues [4], [5], [6].

Observations frequently used to assess model performance are obtained from long-term satellite products providing near-global coverage, which have proven to be well suited for the evaluation of climate models [7], [8]. This conventional approach is, however, constrained in part due to limitations and uncertainties of observational products themselves [9], such as biases or varying spatial and temporal coverage.

We propose a new approach to ESM evaluation, designed to facilitate process-oriented evaluation of clouds in climate models and to address some of the apparent limitations of using conventional observational data. We use a priori knowledge about the characteristics of different cloud classes based on the cloud-type classification of the World Meteorological Organization (WMO). By exploiting this a priori knowledge, cloud processes can be highlighted in further evaluation. Our approach extends the recent development of machine learning based cloud classification methods for satellite data [10], [11], [12], [13], [14], [15] to climate models. Machine learning-based cloud classification is not a new idea (see [16]), but has only recently become feasible for large-scale applications due to the increase in available computing power and the different available methods have distinct properties. An important distinction between classification methods is whether they are supervised or unsupervised. The former relies on previously assigned classes and the latter aims at automatically finding distinct new classes. Supervised classification relies on the assumption that the assigned classes fit the purpose, whereas the user has limited control over the

makeup of the classes in unsupervised methods. Therefore, supervised methods allow for interpreting the final results without additional analysis steps but require a set of labeled data [11], [14]. If the goal is to find classes that are as distinct as possible, or if no previously labeled data are available, unsupervised methods are preferable [12], [13].

To our knowledge, no high-resolution [$\mathcal{O}$(1 km)] cloud-class-labeled satellite data have been used for analysis and evaluation of ESMs, so far. labeled datasets allow for a more detailed and more direct interpretation of cloud classes in the respective satellite data in contrast to comparatively coarse classifications as used for example in the D-Series of the International Cloud Climatology Project (ISCCP, [17]). Previous studies have used clustering for satellite products and the output of satellite simulators from models for unsupervised classification of clouds [6], [18], [19]. In these studies, morphological cloud regimes are then assigned to the identified clusters, according to the average physical properties of each cluster. Such a classification offers valuable insights into how individual models represent clouds in a more specific way than simple climatologies of physical variables would. However, in addition to being built on the rather low resolution of $(280\ \text{km})^2$ of the ISCCP-D1 [17] product, uncertainties and artifacts introduced by the satellite simulators can affect the results [20], [21]. In a recent study, a convolutional neural network was used on $(4000\ \text{km})^2$ grid cells to assign the amount of each of four cloud classes per cell [22]. In [22], the classes were derived from WMO classes detected from surface observations, and the method is applicable to climate model output. Other studies have classified satellite data by cloud regime to investigate specific cloud properties such as radiative effects or precipitation [23], [24]. Recently, clustering methods for cloud regimes have been applied to current-generation climate models using the $1^\circ \times 1^\circ$ resolution ISCCP-H product [25] for training, which has a much higher resolution than ISCCP-D1 [26]. Most of the clusters found this way are labeled with cloud regimes named after cloud types defined by the WMO.

Our method aims at establishing this connection between observations and models without the requirement to assign cloud classes a posteriori. We instead compute the relative amount of WMO cloud classes in coarse grid cells as explained in Section II. Statistical analysis can be conducted on these distributions in the same manner as for the traditionally used physical variables but in the phase space of cloud classes.

The proof-of-concept is outlined as follows. In Section II, we describe the satellite products used and introduce the two machine learning methods applied. In Section III, we use our results to establish: 1) that a neural network can be used to accurately assign cloud class labels to satellite data; 2) that this labeled satellite data provides a sufficient basis to train a regression model relating physical variable retrievals from satellites to cloud class distributions in coarse grid cells; and 3) that the application to a coarse-grained version of the alternative ESA Cloud_cci satellite product [27] is possible, showing the potential of the framework. Our findings are summarized in Section IV. A discussion and outlook are presented in Section V.

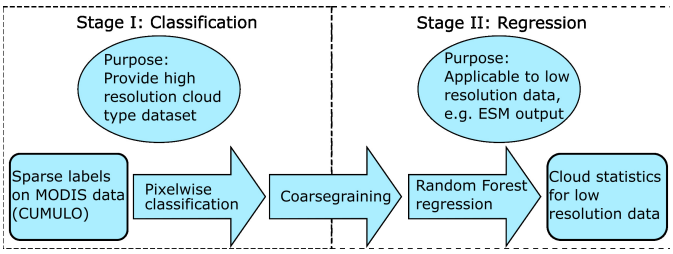

Fig. 1. Two stages of machine-learning—a classifier and a regression model—are required to obtain cloud-type predictions on datasets with low horizontal resolution.

II. Methods

A. Overview

Our goal is to evaluate clouds in climate models using observational data labeled with cloud types. For this, we need to: 1) obtain or create a cloud-labeled dataset and 2) enable a comparison to climate model output.

As a starting point, we use the partially labeled CUMULO dataset [11], which is then fully labeled using a neural network classifier. As this dataset has much higher horizontal resolution than typical climate model output this classifier can not be used for climate models directly. Instead, we train a regression model to predict the relative amounts of each cloud type that are present in larger areas (grid cells) of the fully labeled dataset. This concept is outlined in Fig. 1.

In Section II-C (Stage I in Fig. 1), we outline how the CUMULO dataset was created using data from the MODIS and cloud-type labels from CloudSat. Since classifying low-resolution data with individual cloud types is not appropriate, a regression model is trained on a coarse-grained version of this dataset (see Section II-D), Stage II in 1). Section II-E explains the steps applied to validate the regression model's performance (see Fig. 5) on coarse-grained data from ESA Cloud_cci, which are independent of the training data. In this step, we obtain cloud class distributions, i.e., the percentages of each cloud type in each coarse grid cell of the dataset. Model datasets can be evaluated by relating these distributions to the underlying processes driving formation and evolution of the specific cloud type.

The workflow of this framework is shown in Fig. 2, which illustrates how the different datasets and the two machine-learning models contribute to providing cloud-type distributions for low-resolution data.

B. Satellite Data Products

This work is based on data from three satellite products summarized in Table I.

1) CloudSat: The cloud classes used in the CUMULO dataset are obtained from CloudSat. The product combines radar measurements from CloudSat and lidar data from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) of the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite [32]. These are used as the basis for a mixed threshold-based/fuzzy logic cloud classifier. The use of active sensors provides data that uniquely enable the labeling of multiple clouds along the vertical. A limitation

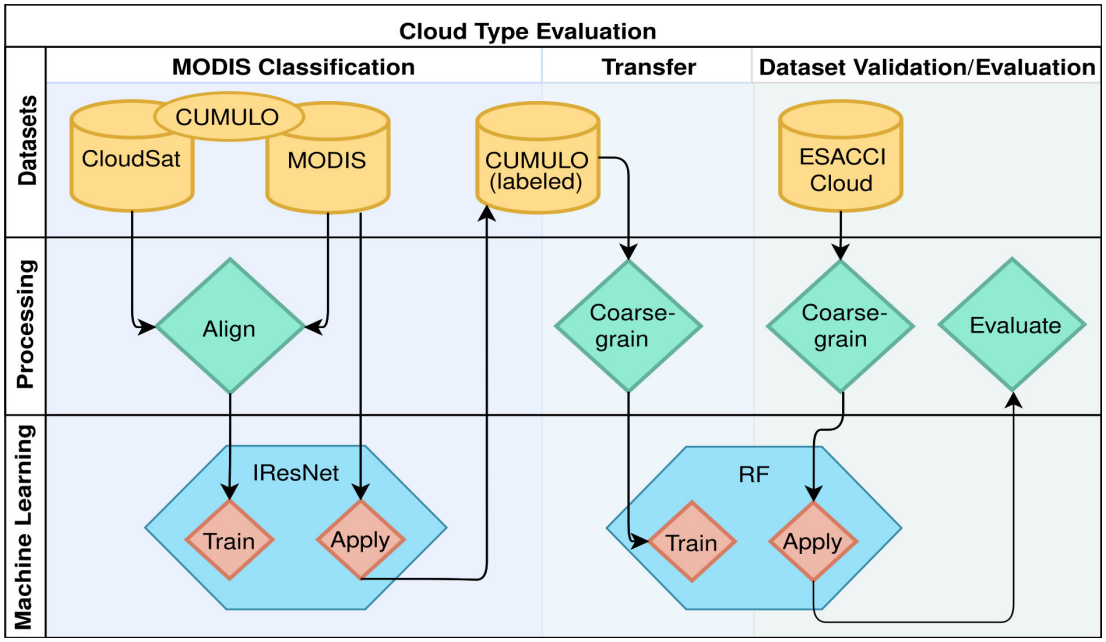

Fig. 2. Workflow schematic: 1) IResNet is trained on the CUMULO dataset and then applied on the unlabeled full-swath Moderate Resolution Imaging Spectroradiometer (MODIS) yielding the fully labeled CUMULO dataset; 2) RF regression model is trained on a coarse-grained version of this data to provide cloud class distributions; and 3) RF is applied to unseen data, allowing validation of the methods performance or evaluation of the target dataset.

TABLE I
THREE SATELLITE DATASETS USED AT DIFFERENT STAGES IN OUR FRAMEWORK

Name	Product / Version	Purpose	Resolution	Revisit period	Reference
CloudSat	2B-CLDCLASS-LIDAR / P1-R05	cloud type	(1.4×1.8) km^2	16 days	[28],[29]
MODIS	MYD06 / 6.1	physical variables	1 km	1-2 Days	[30]
ESA CCI	ESA Cloud_cci L3U / AVHRR-PMv3	physical variables for validation	0.05°	$\frac{1}{2}$ day	[31]

of this dataset is its sparse spatial and temporal coverage. Also, the small footprint size of the satellite instruments used makes distinction between stratocumulus and stratus clouds difficult.

2) MODIS: The inputs (called features in the following) to the machine-learning algorithms used here are also included in CUMULO and obtained from the Cloud Product of the MODIS instrument, which operates aboard the Terra and Aqua satellites. These are sun-synchronous polar-orbiting satellites like CloudSat and CALIPSO, which together with Aqua were part of NASAs afternoon constellation (A-Train) from 2006 to 2018, providing near simultaneous measurements. The MODIS Cloud Product data include nine pixel-level retrievals of physical variables (see Table II) as well as 13 radiance channels. These variables are provided for images of 1354 $\times$ 2030 1-km^2 pixels each covering 5 min.

Known limitations include large uncertainties in the detected cloud phase in high elevation regions, including Greenland and Antarctica. We use the MODIS data for the year 2008 provided with theCUMULO dataset.

3) ESA Cloud_cci: For validation of the method, we use the ESA Cloud_cci (ESA CCI) dataset. It is a long-term cloud product obtained from different observational sources. We use the dataset based on data from the Advanced Very High Resolution Radiometer (AVHRR) aboard the polar orbiting afternoon satellites from the National Oceanic and Atmospheric Administration (NOAA). Among others, the retrieved variables include a cloud mask and cloud physical variables that are also available from the MODIS Cloud Product (see Table II). The retrievals included in the ESA CCI data are obtained using the Community Cloud retrieval for Climate (CC4CL) algorithms [33]. CC4CL shows little dependence on the observational instrument or the surface properties of the measurement region, but still suffers from some particular issues for retrievals using passive sensor data. This includes difficulties in assessing the thermodynamic phase of mixed-phase clouds as well as the detection of thin cirrus clouds. We use the daily L3U dataset for June 2009 up to and including December 2011, with July 2010 being removed due to faulty temperature retrievals.

TABLE II
PHYSICAL VARIABLES AVAILABLE FROM THE MODIS CLOUD PRODUCT, USABLE AS FEATURES FOR THE MACHINE LEARNING MODELS

Abbv.	Physical variable
cwp	Cloud water path
cod	Cloud optical thickness
ptop	Cloud top pressure
htop	Cloud top height
ttop	Cloud top temperature
tsurf	Surface temperature
cer	Effective cloud particle radius
ceff	Effective Emissivity
phase	Cloud thermodynamical phase

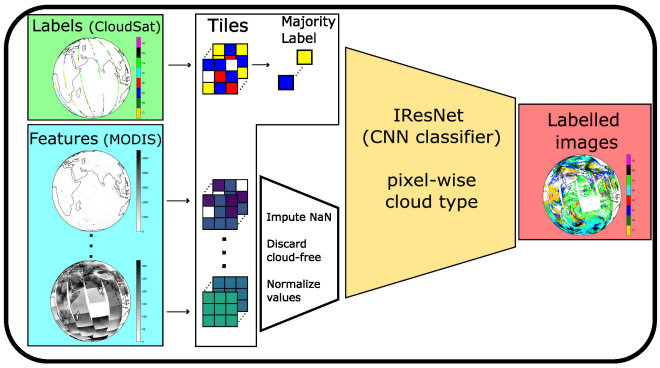

Fig. 3. Schematic of the pixelwise classifier, which is a convolutional neural network trained on features from MODIS and one of eight cloud-type labels from CloudSat per pixel.

TABLE III
CLOUD TYPES FROM THE CUMULO DATASET

Abbv.	Cloud type
Ci	Cirrus/Cirrostratus
As	Altostratus
Ac	Altocumulus
St	Stratus
Sc	Stratocumulus
Cu	Cumulus
Ns	Nimbostratus
Dc	Deep Convection

C. Pixelwise Classification

We create a fully labeled cloud-type dataset by applying a pixelwise classifier network to the sparsely labeled CUMULO datset [11]. The results is a high-resolution, high-coverage, cloud-labeled dataset (see Fig. 3). The classification scheme applied here is largely based classification algorithm used in [11]. The CUMULO dataset provides one of eight cloud-type labels (see Table III) per cloudy pixel along a narrow, vertically resolved path. To be compatible with the 2-D MODIS data, each vertical column is represented by the class which occurs most often in that column. The data from the two satellites are aligned such that, where available, the label track from CloudSat is superimposed on the corresponding MODIS data. This excludes nighttime measurements, as some MODIS retrievals are not available at night. Since the CloudSat swath is quite narrow, most of the pixels in the resulting CUMULO dataset are not assigned a label, which is why a neural network is trained using the labeled part of the dataset to predict cloud class labels for the unlabeled pixels.

The neural network used in [11] and here to classify clouds in the CUMULO dataset is a semisupervised convolutional network based on the invertible residual network (IResNet) [34]. Residual networks [35] have become the baseline for many image-related tasks and the IResNet additionally allows for semisupervised training. The training is termed semisupervised as both labeled and unlabeled samples are fed to the network. The model learns to minimize the cross-entropy for the labeled parts [see (2)] as well as the negative log-likelihood [see (1)] of the latent representation z of all (labeled and unlabeled) samples

$$\mathcal{L} = -\sum_{\substack{z_k = F(x_k),\\ x_k \in \mathcal{X}}} \log(p(z_k)) + \mathrm{Tr}(\mathbf{J}_F) \tag{1}$$

$$\mathcal{L}_l = \sum_{x_k, y_k \in \mathcal{X}_l} \log(x_k) y_k. \tag{2}$$

Here, $z_k = F(x_k)$ is the latent output of the IResNet F without its classifier head, and $\mathbf{J}_F$ is its Jacobian, with Tr denoting the trace operation. $\mathcal{X}$ contains all samples x and $\mathcal{X}_l$ contains only the samples with labels y.

The IResNet is applied to tiles consisting of 3×3 MODIS pixels to determine the cloud type of the central pixel. The training target is the cloud class that occurs most often in the tile or a random choice of classes that occur equally often in the tile. Note that this way, the class label is predicted such that it is representative of the whole tile, even though the label is only assigned to the central pixel. This is a design choice that possibly introduces a bias toward more frequent cloud classes but increases the number of usable tiles both for training and prediction by allowing for overlapping tiles. Tiles that contain less than six cloudy pixels according to the MODIS cloud mask are discarded. Therefore, the neural network is agnostic to such cases including clear sky situations. By applying the trained model, pixels in the CUMULO data that are yet unlabeled are assigned class labels, resulting in a set of fully labeled satellite data.

The CUMULO data contain MODIS radiance channels as well as retrieved physical cloud properties (see Table III). With potential application to climate models in mind, we decided to train the IResNet using the physical variables as features, these being more readily available from climate models than the radiances at the particular MODIS spectral channels. We found that the classes predicted by the model trained on the physical variable features were slightly more physically consistent. For example, we found that a number of high and thin clouds were given the Cumulus (Cu) label when using the radiances only, but this did not happen when using the physical cloud properties. However, the performance difference was marginal such that the classification step could be trained on either set of features. To be able to perform the training on physical variables, pixels containing missing values, e.g., from failed MODIS retrievals are imputed, using the mean value for each 3×3 pixel tile. As the tiles are small, this is not expected to skew the values in the individual tiles significantly as neighboring pixels are expected to have similar properties.

The IResNet is trained on all available CUMULO granules for the year 2008 ($\sim$48 000 multivariate images of 1354×2030 pixels) with standardized features. Instead of using a train/test split, we used fivefold cross-validation on the same data to assess generalization to unseen data. The model used for final predictions is then trained on the complete year. Due to the high temporal resolution of the data, there should be enough variance in the features for the training data to be representative of longer periods typical for climate models. This will compensate for the fact that only one year of training data is used.

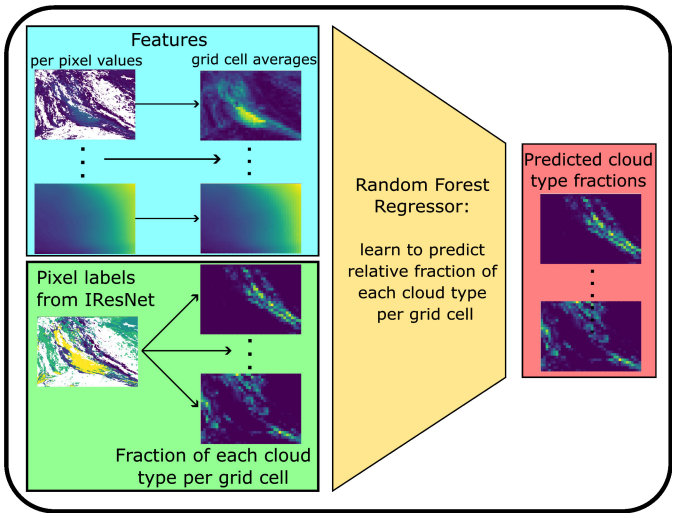

Fig. 4. Cloud type predictions for data with low horizontal resolution are obtained by coarse-graining high-resolution predictions as a basis to train a regression model predicting relative amounts of each cloud type for each coarse resolution grid cell.

D. Regression on Low Resolution Data

The second stage of machine learning (see Fig. 4) is designed to transfer the information contained in the (labeled) high-resolution satellite data to datasets of lower temporal and spatial resolution, like typical ESM output.

For this, the labeled satellite dataset obtained from the pixelwise classification is coarse-grained. All variables that are provided in both the CUMULO and the target dataset can be used as features.

The labeled data are provided on an evenly spaced metric grid, but many climate models are provided on evenly spaced angular grids. The area covered by individual pixels will not match between these two grids and scale differently depending on their geographic location. For simplicity and for the purpose of a proof-of-concept, we determine the grid cell size that on average is most representative of the target grid and use the averages of each variable over these grid cells as features for our model. We assume that the remaining differences between the grids are mitigated by averaging. The output is the relative cloud class occurrence in the grid cell, i.e., the fractional amount of each of the eight cloud classes plus an additional "undetermined" class. The "undetermined" class contains all pixels for which the prediction of a label was not possible due to failed MODIS retrievals, which often indicate clear sky. Missing values for pixels with no cloud are processed accordingly when computing the grid-cell averages (see Section II-F), such that cloud class fractions are predicted consistently for all properties including those that are not defined for clear sky (e.g., *ptop*). Grid cells containing only "undetermined" pixels are discarded. Thus, we obtain a multivariate regression problem with a 9-D output space, containing the eight classes plus "undetermined" pixels, and up to 11 features (i.e., the number of suitable physical variables provided by the CUMULO data, see Section II-F).

For our model, we choose the random forest (RF) [36] regression method for reasons of simplicity, computational efficiency, as well as its inherent normalization of the predicted

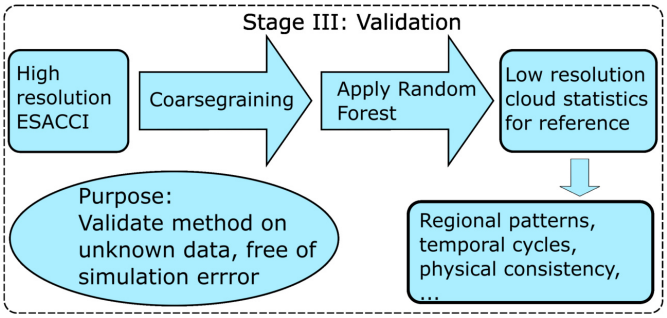

Fig. 5. Method is validated by applying the trained regression model to data the model has not seen before. The predictions are then analyzed for physical consistency.

fractions. After training the RF on the coarse-grained classified images, it can be directly applied to the target data, i.e., ESM output, providing cloud class fraction predictions for each grid cell. To investigate the sensitivity to the resolution and choice of features used, we trained multiple RFs for different respective choices. The individual training samples are weighted with weights w_i given by the L^1-norm $w_i = ||\mathbf{y}_i - \overline{\mathbf{y}}||_1$, where $\mathbf{y}_i$ denotes the cloud class fractions for the ith training sample and $\overline{\mathbf{y}}$ the average over all samples used in training. The weighting ensures that samples close to "the average sample" are given less weight in training, to reduce the effect of any bias in the data. We have about 48 000 labeled CUMULO granules (multivariate images of 1354×2030 pixels) available. To limit the amount of memory required, the RFs are trained on roughly $50 \cdot 10^6$ random samples drawn from a training split of 10 000 labeled data images. The amount of samples varies because grid cells containing only "undetermined" pixels are excluded. The models are then evaluated on a test split containing 8422 images. The hyperparameters of the RF models are chosen such that the depth of the individual regression trees is 17 or less. We apply a bagging subsampling fraction of 0.7 and a minimum leaf size of 2, with 400 individual trees per forest. These hyperparameters showed an optimal trade-off between the model's performance and size.

E. Application to ESA CCI Data

As a proof-of-concept, we apply our method to an independent satellite dataset, the ESA CCI data (see Section II-B). Application to the output of current ESMs (e.g., those contributing to CMIP6) is not yet possible due to too coarse horizontal and/or temporal resolution of the available output and/or key variables (e.g., *cer*, *cod*), required for sufficient performance, being unavailable. As illustrated in Fig. 5, we use this validation stage to show that the method generalizes to coarse data obtained from different sources and it is thus expected to also be applicable to other datasets such as suitable ESM output.

The ESA CCI data provide a similar representation of the observed cloud state and contain similar physical cloud variables as the MODIS product and are therefore comparable to the CUMULO data. In contrast to using ESM output as target data, this allows for a more direct assessment of uncertainties and limitations of this approach as possible model deficiencies do not play a role. We provide an analysis of predictions

TABLE IV

FRACTIONS OF THE CLOUD CLASSES FOR PIXELWISE CLASSIFICATION WITH PREDICTION ACCURACY AND F1 SCORE FOR THE SUPERVISED PART OF THE DATA. CLOUDSAT LABELS ARE FOR $21 \cdot 10^6$ LABELED PIXELS INCLUDED IN CUMULO, PREDICTIONS ARE FOR $800 \cdot 10^6$ PIXELS. SCORES ARE AVERAGES FOR THE FIVE VALIDATION SPLITS

	Ci	As	Ac	St	Sc	Cu	Ns	Dc
CloudSat fraction	0.259	0.132	0.112	0.021	0.313	0.065	0.082	0.015
Predicted fraction	0.154	0.10	0.180	0.027	0.353	0.014	0.134	0.041
Prediction accuracy	0.85	0.84	0.88	0.95	0.78	0.91	0.90	0.98
Prediction F1-Score	0.64	0.40	0.44	0.21	0.66	0.46	0.48	0.48

using different spatial and temporal resolutions in the Supplementary Material. For each validation experiment using the coarse-grained ESA CCI data, we randomly sample 20% of the available grid cells.

F. Features and Preprocessing

The RF regression model is trained on the features also available in the target dataset. For the ESA CCI data, these include cloud top temperature (*ttop*), cloud top height (*htop*), cloud top pressure (*ptop*), surface temperature (*tsurf*), cloud optical thickness (*cod*), cloud water path (*cwp*), and the effective radius of cloud particles (*cer*). For the cloud water path and the effective radius, the MODIS cloud product as well as the ESA CCI daily product do not distinguish between ice and liquid water. Instead, an additional flag is available to distinguish between liquid and ice cloud tops, which we use to separate *cwp* into liquid water path (*lwp*) and ice water path (*iwp*) as well as *cer* into the liquid and ice particle radii (*cerl/ceri*). This procedure is an approximation and only justified under the assumption that the phase flag provided by the satellite data is representative of the whole cloud column.

The grid box averages for the cloud liquid/ice water path, the radii, *cod* are computed over all pixels in each cell, i.e., replacing missing values with zero. This is useful as these values approach zero with decreasing cloud amount. In contrast, *ptop, htop, ttop* are only averaged over cloudy pixels ("in-cloud values"). The features used for the RF model should ideally complement each other. Information on the cloud thickness is implicitly contained in both *cwp* and *cod*. As the features *ptop, htop* and *ttop* effectively contain the same physical information, only one of them is used. Here, we use *ptop* as feature. In addition to the cloud variables, we also use surface temperature *tsurf* as it is readily available in many datasets. As a default, we therefore select *cwp, lwp, iwp, cerl, ceri, cod, ptop, tsurf*, which we call the *optimal set* of features in the following.

The ESA CCI dataset also provides a pixelwise uncertainty estimate for each variable, which we use to exclude pixels for which the uncertainty is larger than twice the actual value. We note that *cwp* in the ESA CCI data can take very large values. Such outliers with $lwp > 2000\ \mathrm{g/m^2}$ or $iwp > 6000\ \mathrm{g/m^2}$ are excluded from the prediction. The arithmetic means of the latitude and longitude coordinates of all pixels in a grid cell are used as representative geographical locations.

III. RESULTS

A. Predicted Cloud Classes at Pixel Level

To assess the performance of the IResNet, we use accuracy and F1-Score (see Table IV). A qualitative analysis of the physical properties of the predicted cloud classes is used to evaluate the consistency of the results. This is important because the physical properties of the classes predicted by the IResNet model need to be consistent with those from the WMO definitions.

The labels extracted from CloudSat that are available in CUMULO display a strong class imbalance (see Table IV), which we also find in the predicted classes. Most classes occur with a similar frequency in the source data and predictions, with deviations being small enough to be attributable to real differences in the data. We would like to highlight two key properties of the class distributions: 1) there are very few strati (St) and deep convective (Dc) clouds in both the source and the prediction and 2) Cumulus (Cu) and cirrus (Ci) clouds are strongly underestimated in the predictions compared to the source data. For example, Cu has the smallest amount of all predicted cloud classes while this class is more common than St and Dc in the source data. Further assessment of the representation of these four classes (St, Dc, Cu, Ci) is therefore of high importance. The mean accuracy of the classification in the validation splits of the cross-validation is larger than 0.8 for all classes but Sc, which suggests considerable skill in the classification method. However, for a multiclass problem with a large class imbalance such as the one we have here, the accuracy alone is not a suitable measure to fully assess the performance of the method. We therefore additionally look at the F1-Score, which is sensitive to the class imbalance and can help identify individual class biases in the predictions. The F1-Score is at least 0.4 for all classes except for St, and especially high for Sc and Cu with values larger than 0.6. When considering the accuracy scores for all classes, this suggests a good skill in representing the class imbalance. An exception is the St class with an F1-Score of 0.21. As noted in Section II-B, the CloudSat algorithm has trouble distinguishing between St and Sc, which is why this is also to be expected for the IResNet. To assess the uncertainty of the classification, we compute the mean metrics with standard deviation for each of the five validation splits in the crossvalidation and obtain an accuracy of 0.886 ± 0.003 and F1-score of 0.472 ± 0.005. We demonstrate in the Supplementary Material that the predicted classes are also physically consistent with the expectations of the respective WMO cloud types.

B. Cloud Class Distributions at Coarse Resolution

The RF is expensive to train on large datasets such as the year-long, high-resolution CUMULO dataset. Because of these computational constraints, we train the RF on a subset of the labeled data of about 25% the size of the complete dataset. Tests have shown that the errors stabilize when using

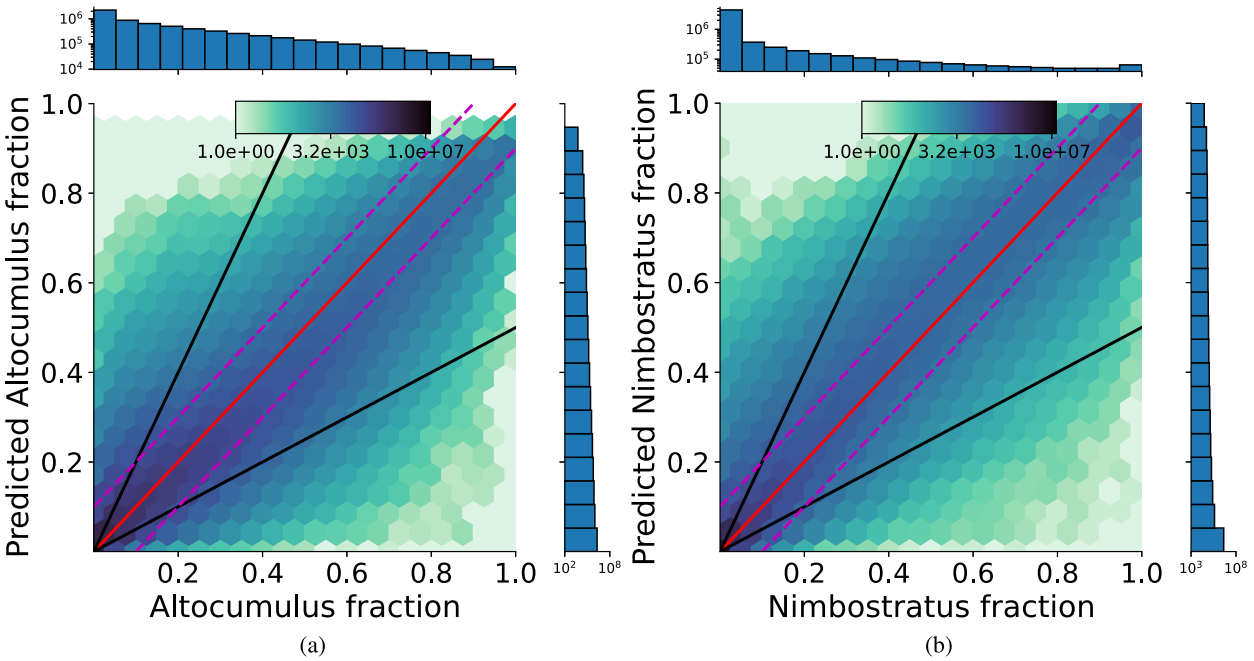

Fig. 6. Joint density of the predicted and true (a) Ac and (b) Ns fractions from the CUMULO test set for a grid cell size of $(100\ \mathrm{km})^2$, using the optimal set of features (see Section II-F). The color scale and the marginal histograms are logarithmic. The red line indicates the line of perfect correlation. The area between the dashed magenta lines indicates a deviation between ground truth and prediction of less than 0.1 and the area between the black lines indicates a deviation by less than a factor of two in either direction.

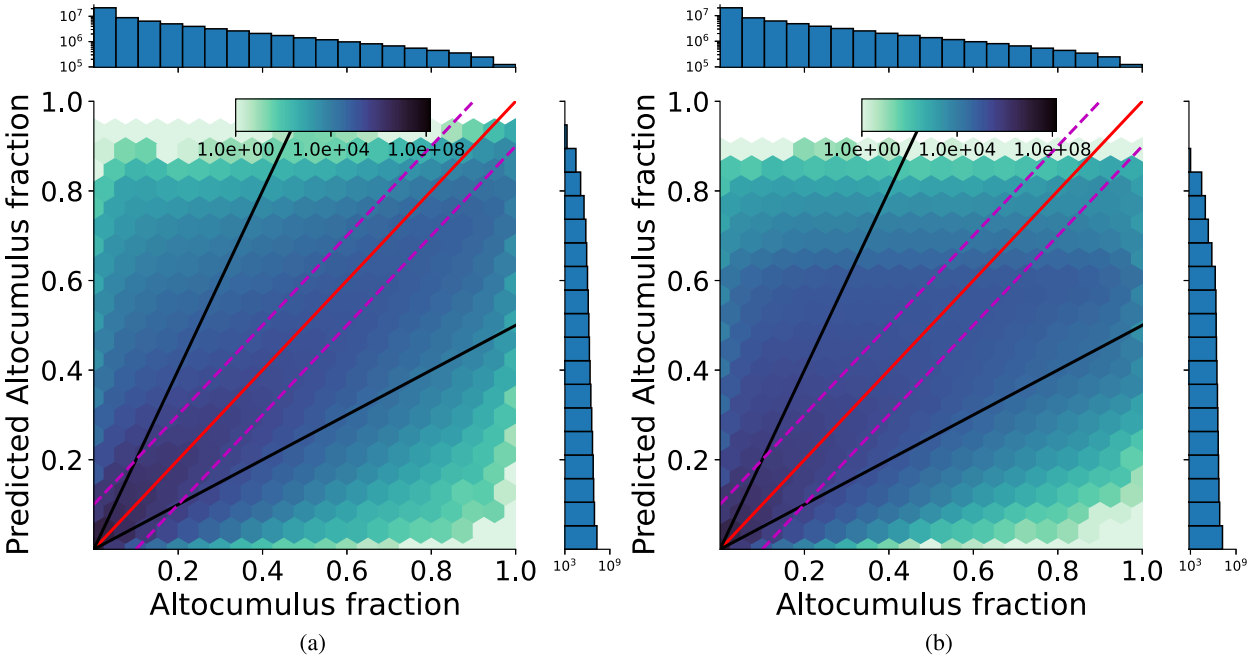

Fig. 7. Results for a model trained using features without (a) liquid/ice distinction (*cwp, cer, cod, ptop, tsurf*) and a model trained also without (b) *cod*, for comparison with Fig. 6.

even fewer training data. The mean errors and R2 scores for the different settings are summarized in Table V. Since with smaller grid cell sizes, more cells containing only "undetermined" pixels are excluded, and the relative amount of cloudy pixels increases, which is why we see larger mean absolute errors for small grid cells. Using the median, however, we see better performance for smaller cells. The R2-score increases with grid cell size, most likely due to the decreasing variance caused by averaging over more pixels. The performance is therefore judged not to be strongly dependent on the grid box size. We also use joint densities of predicted and ground truth cloud fractions of the test split as a performance indicator. Fig. 6 shows these for a grid cell size of $(100\ \mathrm{km})^2$ using the optimal set of features (see Section II-F). The joint density plot

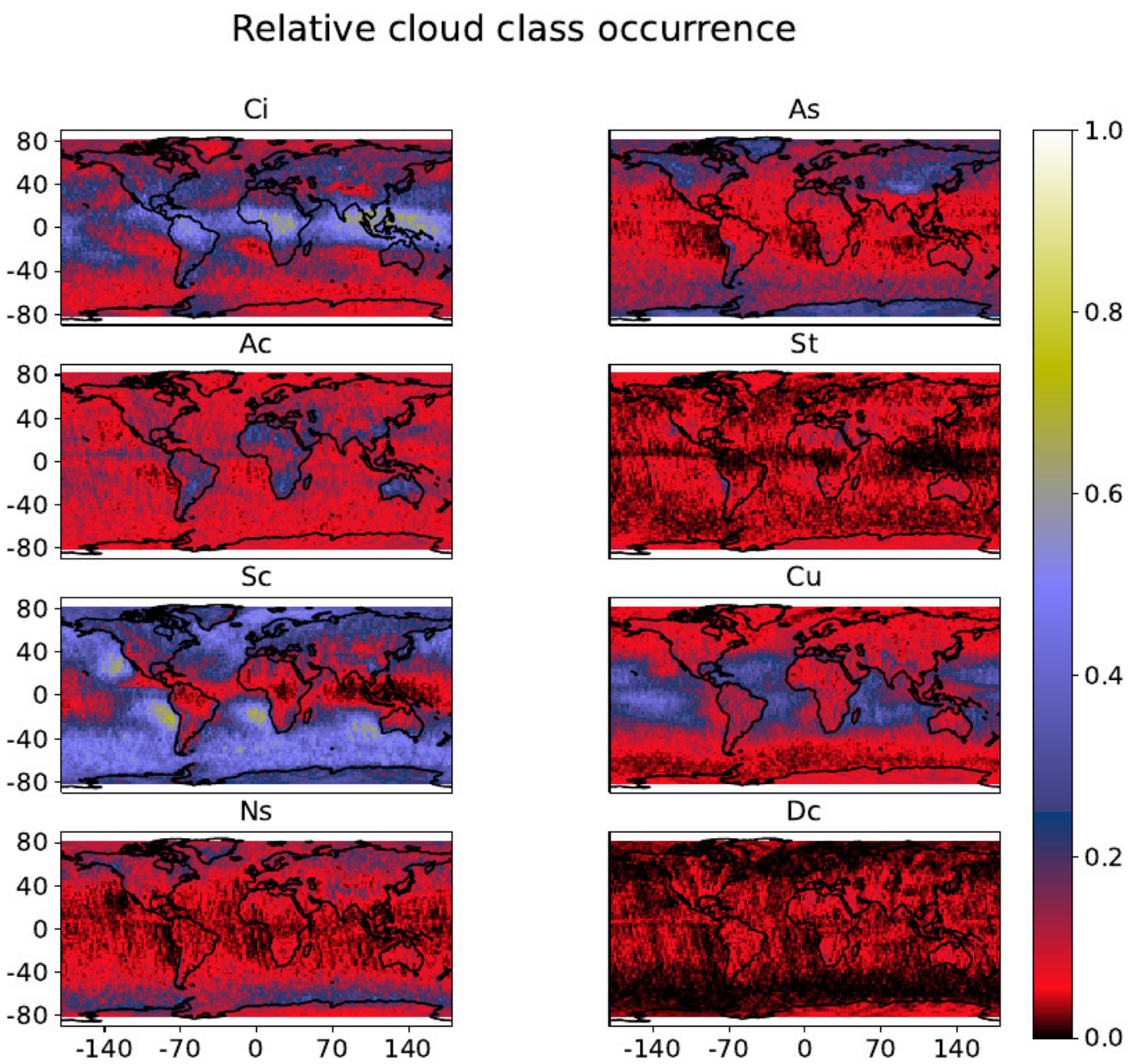

Fig. 8. Relative occurrence per class from CloudSat measurements (year 2008) from the 2B-CLDCLASS-LIDAR product per 2° × 2° grid cell.

displays the concentration of samples in the truth/prediction space, and along the x- and y-axis the marginal distributions of the true and predicted fractions, respectively.

For both cloud classes in Fig. 6, there is a clear correlation between the ground truth and the predictions with a Pearson correlation of $c_P = 0.96$ for Ns and $c_P = 0.89$ for Ac. Many predictions are, however, far off the target: Fig. 6(b) shows several hundred samples with a predicted Ns fraction of about 0.2 where the true fraction is close to 1. For this specific example, this is a small fraction [$\mathcal{O}(0.001\%)$] of the total number of grid cells, but it shows that the predictions can differ strongly from the true values in a nonnegligible number of cases. This deviation is a manifestation of ambiguity between different cloud states, likely caused by noise generated by the averaging of the features. Furthermore, this is an example of the predictions favoring low cloud fractions, as the "undetermined" class is prevalent in the training data. As shown in Fig. 6(a), large altocumulus (Ac) fractions (>0.9)

TABLE V

RESULTS OF THE REGRESSION MODELS FOR DIFFERENT GRID BOX SIZES. (1) TRAINED USING A DEFAULT SET OF FEATURES. (2) USING *cwp* AND *cer*, NOT SEPARATED INTO ICE AND LIQUID, IN ADDITION TO *cod, ptop, tsurf*. (3) USING (*cwp, cer, ptop, tsurf*) AS FEATURES

Grid cell size	Mean abs. error	Median error	Median relative error	R2-Score	Random MAE
3 km (1)	0.042	0.0018	18.9%	0.816	0.193
10 km (1)	0.036	0.002	33.9 %	0.836	0.178
20 km (1)	0.033	0.0025	41.3%	0.840	0.169
100 km (1)	0.027	0.0038	52.6%	0.845	0.151
200 km (1)	0.025	0.0045	54.2%	0.859	0.143
100 km (2)	0.028	0.0041	55.4%	0.837	0.151
100 km (3)	0.033	0.0043	60.4%	0.755	0.151

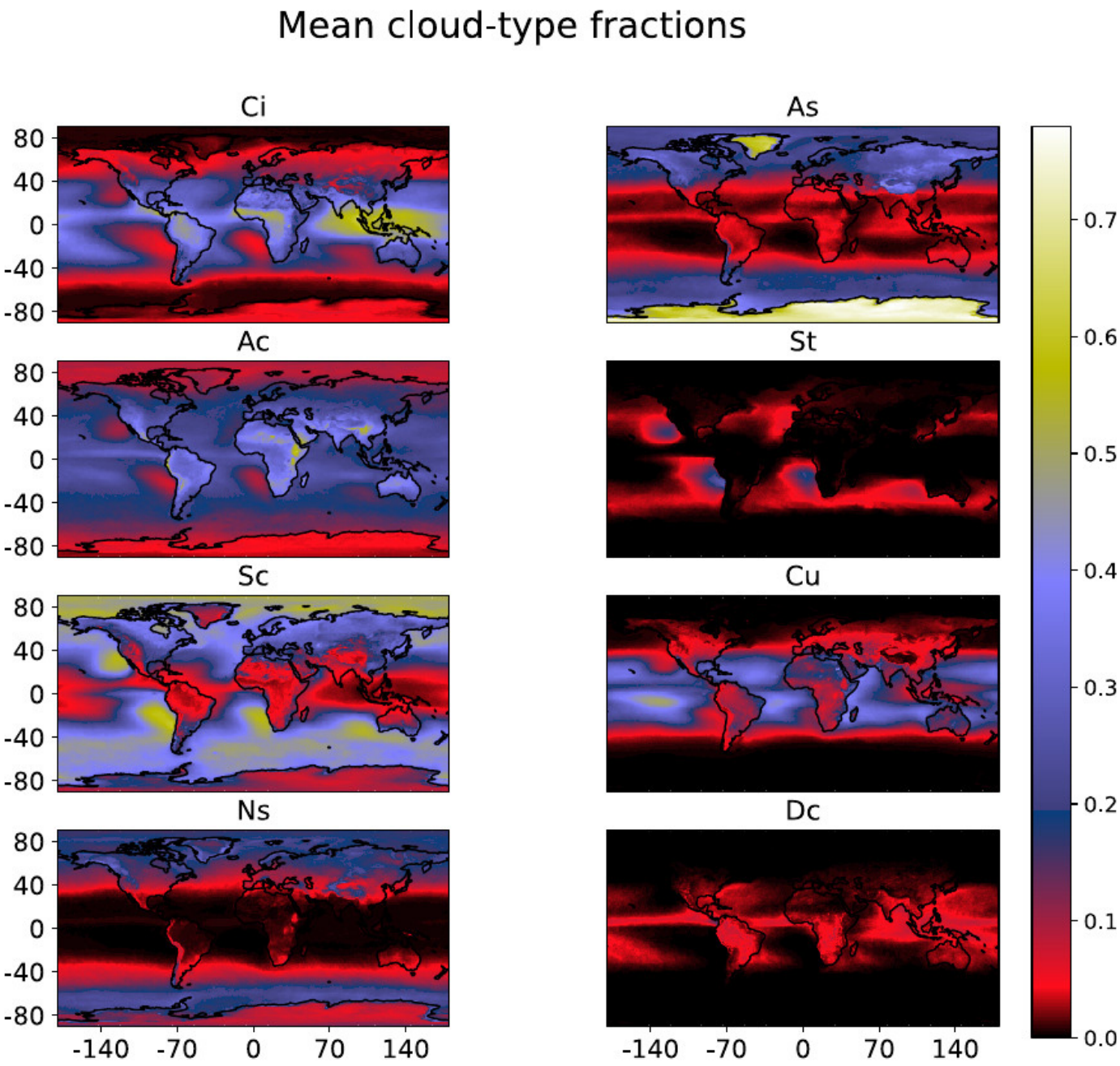

Fig. 9. Average class fractions for the predictions on coarse-grained ESA CCI data. The RF was trained on $(10\ \text{km})^2$ grid cells and applied to 10×10 pixel grid cells. The results are projected onto a $1^\circ \times 1^\circ$ grid. Many classes show similarities to the distributions in the CloudSat data (see Fig. 8), even though the location is not used as a feature.

are underestimated by the RF, but the deviation in this region remains largely below 0.1, as indicated by the dashed line. Most of the samples are, however, still contained within the $\Delta = 0.1$ range (magenta dashed lines). For fractions larger than 0.2, samples deviating by more than a factor of 2 (outside black lines) are rare. For fractions smaller than 0.2 (bottom left corner), deviations by a factor of more than 2 occur frequently, indicating difficulty in correctly predicting small fractions. Note that such predictions contribute significantly to the relative error, but have a negligible effect on the absolute error. We construct a random baseline by sampling from the class distributions in the IResNet predictions. We find that the mean absolute deviation is larger for the random baseline by roughly a factor of five, indicating that the regression model outperforms the random baseline.

In the Supplementary Material, we show that our method reproduces physically meaningful cloud class distributions. Specifically, we show that the features resulting in a prediction of a certain cloud class are in line with those expected from the meteorological definition of the respective class.

As the variables available in typical ESM output can vary, not always matching our optimal set of features, we also determine which of these features are essential to achieve good

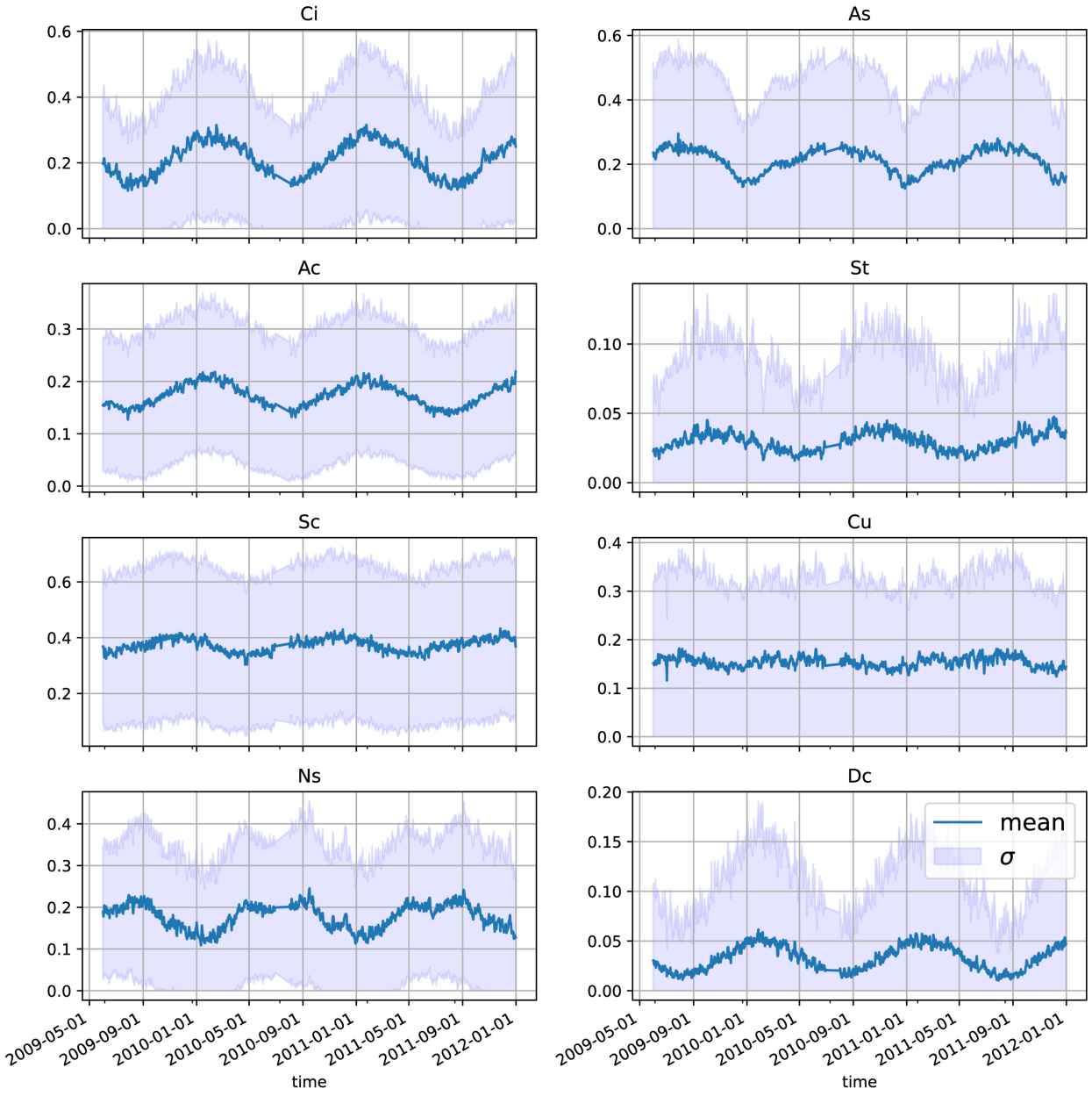

Fig. 10. Time series of mean predicted class fractions from 1 June 2009 to 31 December 2011. The per-class mean is computed for all locations in the southern hemisphere where at least one instance of the respective cloud class fraction is within the 90th percentile of all predictions of this class. Note that July 2010 has been excluded due to faulty data.

performance. In addition to using the optimal set of features, we therefore also trained the model using different alternative sets, containing fewer features. Using the cloud top phase flag to distinguish between ice and liquid for some of our features (*lwp, iwp*, *cerl,ceri*) produces a small performance increase as the metrics indicate in Table V. Comparing Fig. 6(a) and Fig. 7(a) shows that the correlation between the true and predicted values becomes less pronounced when the information about the thermodynamic phase is removed. Further ablation studies reveal, that using *cod* and *ptop* is critical for the RF performance, but these variables are infrequently contained in ESM standard output. An example is shown using features without *cod*, where Table V shows a significant decrease in the R2-score. The effect on the joint density displayed in Fig. 7(b) is visible as predicted fractions being skewed toward smaller values. Predicted fractions above 0.8 are very rare and the joint density seems to be shifted toward the lower black line, corresponding to half the true value. We provide further information on the feature importance in the Supplementary Material.

C. Validation

To assess the generalization performance of the method, we compare predictions of the class distributions on ESA CCI data to the classes from the CloudSat CLDCLASS-LIDAR product. We use the class labels from CloudSat for the

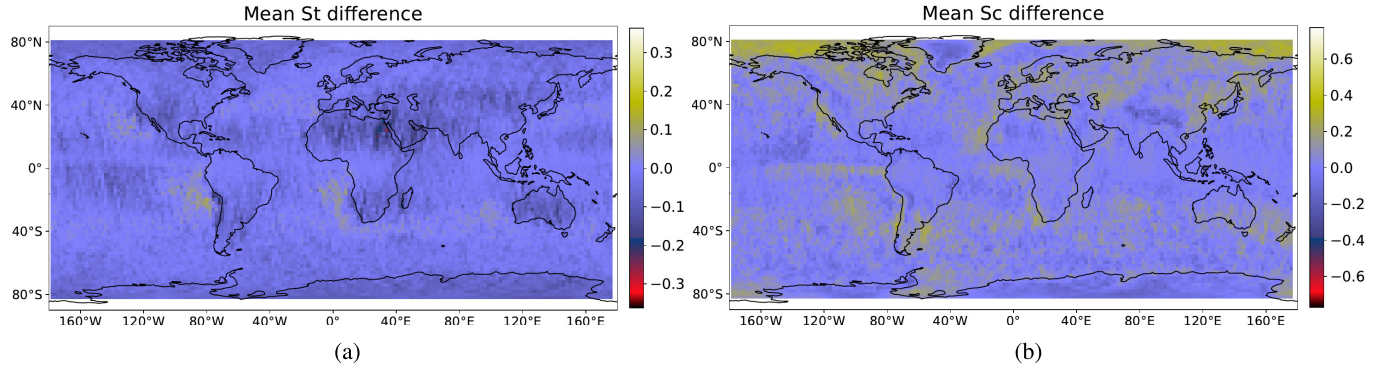

Fig. 11. Difference between the mean predicted fractions and CloudSat per $2^\circ \times 2^\circ$ grid cell for the relative amount of the classes with lowest/highest correlation, St [(a) $c_P = 0.18$]/Sc [(b) $c_P = 0.88$]. The color map is normalized to the range $[-m, m]$, where m is the maximum value for the class across both (CloudSat, predictions) distributions. Predictions obtained from RF trained on $(100\ \text{km})^2$ data and applied on 100×100 ESA CCI pixels.

year 2008. Again, the 3-D CloudSat data are aggregated into two dimensions using the most common cloud class within each vertical column as a representative cloud class. The labels provided by CloudSat for individual orbits are sparse, but using a whole year of measurements provides enough samples per location to compare to the predicted distributions. Fig. 8 shows the sparsity of CloudSat labels, even though the data have been aggregated to grid cells of $2^\circ \times 2^\circ$. Consequently, clear regional differences are not visible for all cloud classes, but Ci, Sc, and Cu show distinct areas of frequent occurrence. For example, Sc clouds are frequently detected in the subtropical subsidence regions off the west coasts of the continents, Ci clouds are frequent in the deep Tropics, and Cu is found frequently over the tropical and subtropical oceans away from the stratocumulus decks.

In the following, for a better comparison of the CloudSat ground truth and the predictions on the ESA CCI data, we exclude the "undetermined" predictions such that the cumulative fraction of all eight cloud classes equals one in each cell. The reported fractions are therefore a relative measure and independent of the total cloud amount in each grid cell.

The CUMULO and the ESA CCI data are provided on different grids. Because of averaging, same-size grid cells are not needed and thus no interpolative regridding is applied. We compare predictions obtained with the same model on differently sized ESA CCI grid cells in the Supplementary Material to this article. Our results show that for a model trained on large $(100\ \text{km})^2$ grid cells, there is no qualitative difference in the predictions for different grid sizes. However, some features become more pronounced when smaller grid cells are used during prediction. The maximum grid cell size for a reasonable application to ESM output, therefore, depends on the region or processes of interest.

Fig. 9 shows the predictions on the ESA CCI data using the RF trained on grid cells of $(10\ \text{km})^2$ and applied 10×10 pixel grid cells. Different classes occur in distinct patterns and the Sc class dominates in the predictions, while St and Dc occur very rarely.

Fig. 10 shows the time series of the class fractions averaged over grid cells in the Southern Hemisphere for which the respective cloud class can vary strongly. We define this by selecting grid cells for each class where at some time the class fraction is especially large (90th percentile). Using this method of analysis, almost all predicted classes show a seasonal cycle. Only for the classes Cu and Sc such a cycle is not visible. The Ns and As classes are predicted at higher fractions in the cold months. In contrast to Ns and As, Ac and St have higher fractions in summer. The CloudSat ground truth is too sparse to similarly assess seasonal cycles and enable a direct comparison.

D. Uncertainty Estimate

For the ESM evaluation, deviations introduced by the data need to be separable from those caused by the evaluation method.

As an uncertainty estimate for the consecutive application of both machine-learning methods, we compute the Pearson correlation and difference between the predictions and the CloudSat labels. For this purpose, we bin the relative amount of each class to grid cells of $2^\circ \times 2^\circ$ size for both datasets. As an example, differences for the cloud types with highest (Sc) and lowest (St) correlation are shown in Fig. 11. The difference increases with the fraction of occurrence of each class (as displayed in Fig. 9). Note that this is only a rough measure of accuracy as the two datasets differ in temporal and spatial resolution. Additionally, an exact match cannot be expected as the CloudSat data cover the year 2008, while the ESA CCI data cover the period June 2009 to the end of 2011. The mean within-class correlation is 0.65. Table VI shows the mean fractions of the classes in the predictions and the CloudSat data. The predictions here are comparable to the pixelwise predictions obtained using the IResNet (see Table IV). The most notable difference in the distribution is again the under-representation of Ci in the predictions relative to the CloudSat labels, which is caused by the under-representation in the predicted pixelwise labels. Table VI also shows the relative difference between the two distributions for grid cells showing a large class fraction in the predictions (90th percentile). For all classes, the magnitude of this deviation is below 50%. This is also the range of relative deviation we

TABLE VI

MEAN FRACTION OF THE PREDICTED CLASSES COMPARED WITH THE RELATIVE AMOUNTS OF THE CLASSES IN CLOUDSAT. THE LAST ROW SHOWS THE MEAN DIFFERENCE FOR PIXELS WITH PREDICTIONS IN THE 90TH PERCENTILE Δ_{90} RELATIVE TO THE MEAN μ_{90} OF THESE PREDICTIONS. PREDICTIONS ARE TAKEN FROM A MODEL TRAINED ON THE DEFAULT SET OF FEATURES USING $(100\ \text{km})^2$ AND APPLIED ON 100×100 PIXEL ESA CCI GRID CELLS

	Ci	As	Ac	St	Sc	Cu	Ns	Dc
Predict.	0.13	0.14	0.19	0.01	0.31	0.10	0.10	0.02
CloudSat	0.20	0.13	0.11	0.05	0.27	0.12	0.09	0.03
c_P	0.87	0.80	0.60	0.18	0.88	0.84	0.83	0.36
Δ_{90}/μ_{90}	-29%	49%	49%	1%	18%	14%	30%	39%

found on the test split, leading to an overall estimate of the uncertainty of 50%.

IV. SUMMARY

We presented a method for the evaluation of clouds in coarse resolution data, employing the consecutive application of machine-learned classification and regression models. Using this method, information on clouds from high-resolution, 3-D CloudSat and CALIPSO lidar products is first added to passive sensor data from MODIS using the CUMULO framework and then transferred to coarse resolution data. This approach offers a new perspective on statistical and process-oriented assessment of the performance of climate models by being able to analyze the model output in terms of different cloud classes and thereby distinguishing the driving mechanisms for the formation and evolution of different cloud types more clearly. This provides the potential to better understand and ultimately improve existing model deficiencies.

The pixelwise classification has a high accuracy of at least 0.8 for each class, with little variation across the validation splits. The relative amounts of predicted Cu and St, however, can differ by more than a factor of 2 compared to the CloudSat data used as ground truth. The predicted cloud classes show distinct physical properties that are consistent with the expected properties of the corresponding WMO cloud classes.

The subsequent regression can reproduce consistent cloud class distributions on regional scales with mean errors being at least one order of magnitude smaller than the random baseline. Furthermore, the RF regression successfully generalizes to different data as could be shown using the ESA CCI data. The predicted global distributions of the individual cloud classes compare well with the CloudSat ground truth. This is evident when qualitatively comparing the distributions for each class as well as in the correlations and differences in areas of a high-class fraction. The correlation is larger than 0.6 for all but two classes (St/Dc) and the relative difference in areas of a high-class fraction is smaller than 50% for all classes. The spatio-temporal location of a sample is not used as an input for the machine-learning algorithms. Therefore, any predictions are solely based on the physical properties represented by the features. Yet, even small-scale regional characteristics of the CloudSat ground truth are similarly represented in the predictions using the ESA CCI data. Notable examples are a peak of Ci in the tropical warm pool region or an increased As fraction in the Himalayas. Additionally, the geographical means for all classes correlate positively with the respective relative occurrence in the CloudSat ground truth, with higher correlations for the classes with many available samples. We further showed that the regression model associates each class with specific feature values (see the Supplementary Material). These values are consistent with the expected properties of the different WMO cloud types. Analysis of the effects of temporal averaging of the target data showed that the method works well with near-instantaneous data but cannot be applied to monthly averaged data (see the Supplementary Material).

Tests with multiple sets of input features have shown that information about the cloud height, cloud water content and optical thickness are essential for good performance, with information on the thermodynamic phase of the cloud providing additional robustness. In contrast, the horizontal resolution of the data, the model is trained on, seems to be less critical (see the Supplementary Material). Models trained on different grid cell sizes show differences but no clear optimal resolution can be defined from these initial results. It is recommended, however, that predictions be performed using data at their highest available horizontal resolution as more features can be resolved.

Predictions on test data show clear correlations of ground truth and predictions (average $\bar{c}_p = 0.92$). Predictions on the ESA CCI data provide enough information to isolate individual features and processes. This suggests that this method can be successfully applied to any dataset of sufficient length and horizontal and temporal resolution to allow for statistically robust predictions. When applied to the test data, the median relative deviation was about 50%. Comparing the predictions with the raw labels from CloudSat we find similar values. Especially in regions where specific cloud types are predicted with a high frequency of occurrence, we find relative deviations mostly below 50%. Only the St class is consistently underestimated for which the pixelwise classification already showed poor performance.

V. DISCUSSION

The results from testing the regression model with unseen data show physical and temporal consistency of the results across all analyses. This is the primary goal of this method aiming to evaluate physical processes. We can therefore be confident that the results are meaningful, even though the results of the classification do not exactly reproduce the label distribution in the source data. The deviations in the amount of St predicted by the classifier can be at least partly explained by the relatively small amount of training samples and the similarity to the physical properties between St and Sc. We conclude that pixelwise labeled data are therefore suitable as a basis for training a regression model which learns cloud class distributions on datasets with a horizontal resolution typical for climate model scales. Generally, our results suggest that the method is, therefore, suitable for a process-oriented assessment of clouds simulated by climate models. Using the predicted distributions this can be performed in the space of

cloud classes, providing several advantages. First, a layer of subjective interpretation is removed by being able to analyze the results in terms of cloud classes clearly defined by the underlying classification algorithm (CloudSat, [28]). Because we are using the commonly used WMO classes, the resulting cloud class distributions can then be analyzed and interpreted in a process-based manner as the key processes driving formation and evolution differ between the cloud classes. This greatly simplifies the analysis and evaluation of selected physical properties related to cloud processes in the climate models. Second, as the deep learning algorithm learns from high-resolution 3-D data, the climate models are implicitly analyzed in a horizontally super-resolved manner which also takes into account information about the vertical structure of clouds, i.e., learning from a combination of 2-D and 3-D data can potentially take advantage of information from the vertical that would not be included in the cloud top view. That the method can resolve phenomena on regional and seasonal scales provides the opportunity to identify spatio-temporal areas in which clouds are not correctly represented. This could for example be done for the low-level clouds found in the subtropics west of the continents, investigating their horizontal extent, their dependence on feature values, and their temporal evolution. However, due to the nature of the multistage process, some limitations apply: by building the regression on 2-D, spatially averaged source data it is hard to make correct predictions on individual grid cells. This results in several samples for which the predicted cloud fraction differs by a factor of 2 or more. Additionally, the under-representation of the Cu class and the limited accuracy of the St class show that this method can still be improved. This probably stems at least partly from the CloudSat ground truth itself, as the CloudSat algorithm has trouble distinguishing between St and Sc. We, therefore, recommend combining these classes when applying these or similar methods. Some features of the predicted cloud distribution such as, for example, the high fractions of Ns along the Antarctic coast, might be amplified or hidden by noisy satellite retrievals. Especially in high latitudes, clouds can be challenging to characterize with passive sensors like MODIS. Our ML models do not provide satisfactory results when applied to temporally averaged data because they are trained on instantaneous measurements. Using geostationary data available, e.g., every 30 min (GOES satellite [37]) for the pixelwise classification instead of MODIS data available only twice a day might improve the results. Such an approach has been applied to other atmospheric variables like convection and rainfall [38], [39]. The physical properties of the predicted clouds could then be safely averaged over time due to the high and consistent temporal resolution of the data allowing the regression model to train on data more comparable to typical ESMs output. However, the processes to be evaluated with our approach are not resolved at large temporal scales. This contributes to the poor regression performance for monthly mean data and will still be an issue when the RF is trained on temporally averaged data. In turn, this means that this method is suitable to detect model deficiencies relatively quickly in contrast to using climatological means from long-term simulations. This is because we would expect an inaccurate representation of the global and regional cloud distributions to be already detectable with model output available for less than a year.

The consecutive application of two machine learning steps makes it difficult to quantitatively estimate the propagated errors. Even though the error of the classification and regression can be individually estimated using test splits, the combined impact of these errors is not clear. The small variations of the different IResNets used in the crossvalidation suggest high confidence of the networks in their predictions, but errors or uncertainties can not be propagated through an RF. Error estimation is further complicated by possible inconsistencies in the CloudSat classification algorithm, i.e., clouds not necessarily being labeled the same way a human expert would. An example of this is the difficulties in differentiating between St and Sc. We do not, however, see any specific inconsistencies in the physical properties or the regional distributions of the predictions, suggesting that the propagated uncertainties are reasonable. We estimate the uncertainty of at least 50% to what would be reported by CloudSat for individual predictions can be expected. However, in large datasets, the method can identify individual regions of high occurrence for a class. The absence or underestimation of such phenomena in the global cloud distribution would be signs of possible ESM deficiencies. Even for classes for which limited training data are available (Cu, St, Dc), we find that the predictions are self-consistent. This is apparent in the characteristic feature values, which are distinct for each class and do not vary regionally (see the Supplementary Material). The regional distributions of the classes are attributable to the predominant atmospheric conditions. For example, the Dc class occurs more frequently near the equator, Cu is found predominantly in tropical and subtropical regions over the ocean, St is mostly west off the continents in the subtropical subsidence regions. Both, Cu and St are as expected low-level clouds with low cloud top heights.

In terms of implementation, this method can be applied to new datasets quickly and does not require individual implementation for each model. We would like to note that many of the variables needed for this method are typically part of the standard output of climate models so the main requirement would be to provide instantaneous or near-instantaneous values, i.e., model output not averaged over longer time intervals. We would therefore like to encourage the modeling community to provide such an output, e.g., in future model intercomparison projects such as CMIP7. We especially recommend adding the cloud optical thickness, and the effective particle radii for liquid and ice particles as instantaneous 2-D variables to the CMIP7 data request. Future improvement of the method could include replacing the RF as a regression model. The most significant advantage of the RF is the use of the bagging process during training, which helps to generalize well to unseen data. However, the size of the RF scaling with the size of the dataset and batched training is not straightforward requiring a high degree of subsampling to make training computationally feasible. Therefore, as noted previously, this required us to discard data. However, once trained, applying the RF to batched data for predictions is possible and fast, making it suitable for practical application.

A CNN could be a reasonable replacement for the RF due to the resemblance of the data to images. First attempts to replace the RF with a CNN, however, did not yield satisfactory results independent of specific architecture, with the network's loss not converging. Additionally, implementation and training of the RF are much simpler than that of a CNN, which makes the RF more suitable in practice. Also, in this study, the CloudSat cloud classes in the source data are aggregated in the vertical dimension by assigning the most common class in the cloud column to the respective pixel. Even though this provides an implicit resolution of vertical features, a full classification in three dimensions would be a clear improvement. An improved representation of the vertical cloud structure might be obtained by using a more sophisticated aggregation algorithm. Taking into account the physical properties of the observed pixel in each vertical column might lead to more representative ground truth.

ACKNOWLEDGMENT

The authors would like to thank Valentina Zantedeschi, Fabrizio Falasca and their coauthors for kindly providing the CUMULO dataset and code. Our own code related to the random forest and processing can be found at `github.com/EyringMLClimateGroup/kaps22tgrs_ml_cloud_eval` or via `DOI:10.5281/zenodo.7248773`.

REFERENCES

[1] P. P. Vignesh et al., "Assessment of CMIP6 Cloud Fraction and Comparison with Satellite Observations," *Earth Space Sci.*, vol. 7, no. 2, pp. 1–21, 2020.

[2] J.-L. Dufresne and S. Bony, "An assessment of the primary sources of spread of global warming estimates from coupled Atmosphere–Ocean models," *J. Climate*, vol. 21, no. 19, pp. 5135–5144, Oct. 2008. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/21/19/2008jcli2239.1.xm%l

[3] M. D. Zelinka et al., "Causes of higher climate sensitivity in CMIP6 models," *Geophys. Res. Lett.*, vol. 47, no. 1, pp. 1–12, 2020.

[4] S. Bony et al., "Clouds, circulation and climate sensitivity," *Nature Geosci.*, vol. 8, no. 4, pp. 261–268, 2015.

[5] T. Schneider et al., "Climate goals and computing the future of clouds," *Nature Publishing Group*, vol. 7, no. 1, pp. 3–5, 2017.

[6] K. D. Williams and M. J. Webb, "A quantitative performance assessment of cloud regimes in climate models," *Climate Dyn.*, vol. 33, no. 1, pp. 141–157, Jul. 2009.

[7] A. Lauer et al., "Benchmarking CMIP5 models with a subset of ESA CCI phase 2 data using the ESMValTool," *Remote Sens. Environ.*, vol. 203, pp. 9–39, Dec. 2017.

[8] H. Kawai et al., "Significant improvement of cloud representation in the global climate model mri-esm2," *Geosci. Model Develop.*, vol. 12, no. 7, pp. 2875–2897, 2019. [Online]. Available: https://gmd.copernicus.org/articles/12/2875/2019/

[9] C. Jakob, "An improved strategy for the evaluation of cloud parameterizations in GCMS," *Bull. Amer. Meteorol. Soc.*, vol. 84, no. 10, pp. 1387–1402, Oct. 2003.

[10] S. Rasp, H. Schulz, S. Bony, and B. Stevens, "Combining crowdsourcing and deep learning to explore the mesoscale organization of shallow convection," *Bull. Amer. Meteorological Soc.*, vol. 101, no. 11, pp. E1980–E1995, 2020.

[11] V. Zantedeschi, F. Falasca, A. Douglas, R. Strange, M. Kusner, and D. Watson-Parris, "Cumulo: A dataset for learning cloud classes," 2019, *arXiv:1911.04227*.

[12] L. Denby, "Discovering the importance of mesoscale cloud organization through unsupervised classification," *Geophys. Res. Lett.*, vol. 47, no. 1, pp. 1–10, Jan. 2020.

[13] T. Kurihana, E. Moyer, R. Willett, D. Gilton, and I. Foster, "Data-driven cloud clustering via a rotationally invariant autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–25, 2021.

[14] C. Zhang, X. Zhuge, and F. Yu, "Development of a high spatiotemporal resolution cloud-type classification approach using Himawari-8 and CloudSat," *Int. J. Remote Sens.*, vol. 40, no. 16, pp. 6464–6481, Aug. 2019.

[15] W. J. Marais, R. E. Holz, J. S. Reid, and R. M. Willett, "Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations," *Atmos. Meas. Techn.*, vol. 13, no. 10, pp. 5459–5480, Oct. 2020. [Online]. Available: https://amt.copernicus.org/articles/13/5459/2020/

[16] J. Lee, R. C. Weger, S. K. Sengupta, and R. M. Welch, "A neural network approach to cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 5, pp. 846–855, Sep. 1990.

[17] W. B. Rossow and R. A. Schiffer, "Advances in understanding clouds from ISCCP," *Bull. Amer. Meteorological Soc.*, vol. 80, no. 11, pp. 2261–2287, Nov. 1999.

[18] C. Jakob, "Objective identification of cloud regimes in the tropical western Pacific," *Geophys. Res. Lett.*, vol. 30, no. 21, pp. 1–4, 2003.

[19] N. D. Gordon, J. R. Norris, C. P. Weaver, and S. A. Klein, "Cluster analysis of cloud regimes and characteristics dynamics of midlatitude synoptic systems in observations and a model," *J. Geophys. Res. D, Atmos.*, vol. 110, no. 15, pp. 1–13, 2005.

[20] R. Pincus, S. Platnick, S. A. Ackerman, R. S. Hemler, and R. J. P. Hofmann, "Reconciling simulated and observed views of clouds: MODIS, ISCCP, and the limits of instrument simulators," *J. Climate*, vol. 25, no. 13, pp. 4699–4720, Jul. 2012.

[21] D. J. Swales, R. Pincus, and A. Bodas-Salcedo, "The cloud feedback model intercomparison project observational simulator package: Version 2," *Geoscientific Model Develop.*, vol. 11, no. 1, pp. 77–81, Jan. 2018.

[22] P. Kuma, F. A. Bender, A. Schuddeboom, A. J. Mcdonald, and O. Seland, "Machine learning of cloud types shows higher climate sensitivity is associated with lower cloud biases," *Atmos. Chem. Phys. Discuss.*, Mar. 2022.

[23] L. Oreopoulos, N. Cho, D. Lee, and S. Kato, "Radiative effects of global MODIS cloud regimes," *J. Geophys. Res., Atmos.*, vol. 121, no. 5, pp. 2299–2317, Mar. 2016.

[24] J. Tan, C. Jakob, W. B. Rossow, and G. Tselioudis, "Increases in tropical rainfall driven by changes in frequency of organized deep convection," *Nature*, vol. 519, no. 7544, pp. 451–454, Mar. 2015.

[25] A. H. Young, K. R. Knapp, A. Inamdar, W. Hankins, and W. B. Rossow, "The international satellite cloud climatology project H-Series climate data record product," *Earth Syst. Sci. Data*, vol. 10, no. 1, pp. 583–593, 2018.

[26] G. Tselioudis, W. B. Rossow, C. Jakob, J. Remillard, D. Tropf, and Y. Zhang, "Evaluation of clouds, radiation, and precipitation in CMIP6 models using global weather states derived from ISCCP-H cloud property data," *J. Climate*, vol. 2021, pp. 1–42, Jun. 2021.

[27] M. Stengel et al., "Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the cloud-CCI project," *Earth Syst. Sci. Data*, vol. 9, no. 2, pp. 881–904, 2017.

[28] Z. Wang. (Apr. 2019). *CloudSat 2B-CLDCLASS-LIDAR Product Process Description and Interface Control Document*. [Online]. Available: http://www.cloudsat.cira.colostate.edu/sites/default/files/products/files/2B-CLDCLASS-LIDAR_PDICD.P1_R05.rev0_.pdf

[29] K. Sassen, Z. Wang, and D. Liu, "Global distribution of cirrus clouds from CloudSat/Cloud-aerosol lidar and infrared pathfinder satellite observations (CALIPSO) measurements," *J. Geophys. Res.*, vol. 113, pp. 1–12, Oct. 2008.

[30] S. Platnick et al., "The MODIS cloud products: Algorithms and examples from terra," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 2, pp. 459–472, Apr. 2003.

[31] M. Stengel et al., "Cloud_CCI advanced very high resolution radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties," *Earth Syst. Sci. Data*, vol. 12, no. 1, pp. 41–60, 2020.

[32] D. M. Winker, C. A. Hostetler, M. A. Vaughan, and A. H. Omar, "CALIOP algorithm theoretical basis document Part 1: CALIOP instrument, and algorithms overview," Release 2, NASA Langley Research Center, Hampton, VA, USA, 2006, pp. 1–29.

[33] O. Sus et al., "The Community Cloud retrieval for CLimate (CC4CL)—Part 1: A framework applied to multiple satellite imaging sensors," *Atmos. Meas. Techn.*, vol. 11, no. 6, pp. 3373–3396, 2018.

[34] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J. H. Jacobsen, "Invertible residual networks," *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 894–910.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[37] A. Walther, W. Straka, and A. K. Heidinger. (Jun. 2013). *NOAA NESDIS Center for Satellite Applications and Research ABI Algorithm Theoretical Basis Document For Daytime Cloud Optical and Microphysical Properties (DCOMP)*. [Online]. Available: https://www.goes-r.gov/products/baseline-cloud-opt-depth.html
[38] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff, "Applying machine learning methods to detect convection using geostationary operational environmental satellite-16 (GOES-16) advanced baseline imager (ABI) data," *Atmos. Meas. Techn.*, vol. 14, no. 4, pp. 2699–2716, Apr. 2021.
[39] V. A. Gorooh et al., "Deep neural network cloud-type classification (DeepCTC) model and its application in evaluating PERSIANN-CCS," *Remote Sens.*, vol. 12, no. 2, p. 316, 2020.

Arndt Kaps received the M.Sc. degree in physics from Georg-August University, Göttingen, Germany, in 2019. He is currently pursuing the Ph.D. degree with the University of Bremen, Bremen, Germany, under the supervision of V. Eyring.

Since 2020, he has been with the Institute for Atmospheric Physics, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. His research interests include machine learning, satellite products, and climate model data.

Axel Lauer received the Ph.D. degree in meteorology from the Free University of Berlin, Berlin, Germany, in 2005.

Before joining DLR, he spent five years as a Researcher at the International Pacific Research Center, Honolulu, HI, USA, studying clouds and aerosols and two years at the Institute of Advanced Sustainability Studies (IASS, now RIFS), Potsdam, Germany, where he was heading a small group investigating the impact of black carbon on air quality and climate. He is currently a Research Scientist with the Institute of Atmospheric Physics, German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and is a Core Developer of the Earth System Model Evaluation Tool (ESMValTool). His main research interests include aerosols, clouds, and cloud-climate feedbacks as well as their interactions.

Gustau Camps-Valls (Fellow, IEEE) is a Physicist and a Full Professor in electrical engineering with the Universitat de València, Valencia, Spain, where lectures on machine learning, remote sensing, and signal processing. He is the Head of the Image and Signal Processing (ISP) Group, an interdisciplinary group of 40 researchers working at the intersection of AI for Earth and Climate sciences. He has published over 250+ peer-reviewed international journal articles, 350+ international conference papers, 25 book chapters, and five international books on remote sensing, image processing, and machine learning.

Dr. Camps-Valls has been a Program Committee Member of international conferences, such as IEEE, SPIE, EGU, and AGU, and the Technical Program Chair for the IEEE IGARSS 2018 and the General Chair for AISTATS 2022. He served in the technical committees of the IEEE GRSS and IEEE SPS, as an Associate Editor for five top IEEE journals, and in the prestigious IEEE Distinguished Lecturer program of the GRSS from 2017 to 2019 to promote "AI in Earth sciences" globally.

Pierre Gentine received the Ph.D. degree in electronics engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2010.

He is the Director of the National Science Foundation's Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center (STC), Columbia University, New York, NY, USA. He has authored or coauthored more than 180 international journal articles and several international book chapters.

Dr. Gentine is a fellow of the American Geophysical Union and a recipient of the American Geophysical Union Macelwane Medal.

Luis Gómez-Chova (Senior Member, IEEE) received the Ph.D. degree in electronics engineering from the University of Valencia, Valencia, Spain, in 2008.

He is currently a Full Professor with the Department of Electronic Engineering, University of Valencia. He is also a Researcher at the Image and Signal Processing (ISP, isp.uv.es) Group, where his work is mainly related to pattern recognition and machine learning applied to remote sensing multispectral images and cloud masking. He conducts and supervises research on these topics within the frameworks of several national and international projects. He has authored or coauthored more than 55 international journal articles (JCR), more than 140 international conference papers, and several international book chapters.

Dr. Gómez-Chova was awarded by the Spanish Ministry of Education the National Award for Electronic Engineering and has completed different doctoral and post-doctoral stays at prestigious research centers. He is a referee in many international journals and conferences. For more information, visit www.uv.es/chovago for more information.

Veronika Eyring is the Head of the Earth System Model Evaluation and Analysis Department, Institute of Atmospheric Physics, German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and also a Professor of climate modeling with the University of Bremen, Bremen, Germany. She has authored many peer-reviewed journal articles and has contributed to the Intergovernmental Panel on Climate Change (IPCC) climate assessments since many years, including her role as a Coordinating Lead Author for Chapter 3 "Human influence on the climate system" in the IPCC Sixth Assessment Report of Working Group I published in 2021. Her research interests include understanding and modeling the Earth system with machine learning to improve climate projections and technology assessments for actionable climate science.

Dr. Eyring was a PI of the Earth System Model Evaluation Tool (ESMValTool) until 2020, has been a fellow of the European Laboratory for Learning and Intelligent Systems (ELLIS) since 2019, and a member of the Scientific Advisory Board of Worldfund. She received the German Research Foundation (DFG) Gottfried Wilhelm Leibniz Prize in 2021 for her significant contributions to improving the understanding and accuracy of climate projections through process-oriented modeling and model evaluation. She is involved in the World Climate Research Program (WCRP) since many years, for example through her roles as the Chair for the Coupled Model Intercomparison Project (CMIP) Panel from 2014 to 2020 and a member of scientific steering committees, including the Working Group on Coupled Modeling (WGCM) from 2008 to 2018.