Explaining the Effects of Clouds on Remote Sensing Scene Classification

Jakob Gawlikowski, Patrick Ebel, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—Most of Earth is covered by haze or clouds, impeding the constant monitoring of our planet. Preceding works have documented the detrimental effects of cloud coverage on remote sensing applications and proposed ways to approach this issue. However, up to now, little effort has been spent on understanding how exactly atmospheric disturbances impede the application of modern machine learning methods to Earth observation data. Specifically, we consider the effects of haze and cloud coverage on a scene classification task. We provide a thorough investigation of how classifiers trained on cloud-free data fail once they encounter noisy imagery—a common scenario encountered when deploying pretrained models for remote sensing to real use cases. We show how and why remote sensing scene classification suffers from cloud coverage. Based on a multistage analysis, including explainability approaches applied to the predictions, we work out four different types of effects that clouds have on scene prediction. The contribution of our work is to deepen the understanding of the effects of clouds on common remote sensing applications and consequently guide the development of more robust methods.**

***Index Terms*—Classification, clouds, deep learning, explainability, remote sensing, robustness.**

I. Introduction

CLOUD coverage is detrimental to common remote sensing applications, such as remote sensing scene classification [1], [2], [3] and semantic segmentation [4], [5]. While clouds are characterized in great detail [6], [7] and different approaches for handling them have been investigated, less effort has been spent to investigate what exactly its effects on remote sensing applications are. The existing approaches range from learning cloud removal for preprocessing [8], [9], [10], [11], [12], [13] to familiarizing neural networks with clouds by including cloud-covered observations in the training dataset, such that the models learn to ignore clouds irrelevant to the task at hand [3], [4], [14]. Such approaches that include cloudy images in the training process are limited to samples with transparent clouds or samples where the crucial features for classification are not covered. Although recent work demonstrated that explicitly performing cloud removal may improve model robustness [15], the coverage of important features or the misinterpretation of features induced by clouds still poses a significant problem for remote sensing tasks sensitive to inter- and intraclass feature differences [16], [17]. Furthermore, the majority of curated optical satellite datasets are explicitly cleaned from clouds and remote sensing models are subsequently (pre-) trained on (predominantly) clear-view data [1], [2], [18]. This common practice, however, is in contrast to the application of networks typically trained on noncloudy datasets to data in the wild, which is to a large extent polluted by haze or clouds [6]. Fig. 1 illustrates the possible negative effects of cloud cover on scene classification. Fine-tuning such models on cloudy observations would require the post-hoc collection of new data plus task-related labels, which may thus be impracticable for the remote sensing practitioner. Hence, the issue of cloud-agnostic networks confronted with out-of-distribution data at test time commonly persists. That is, classifiers trained on cloud-free data may in practice still encounter samples significantly deviating from the distribution of data that the model has been trained on.

In order to understand the causes of the experienced drops in task performances [3], [14], we provide detailed insights into how clouds affect every single part of the remote sensing pipeline—from raw data to a model's predictions. To our knowledge, the only prior study explaining neural network's scene classifications focuses on clear data without taking the effects of clouds into account [19]. In our work, we explain the causes of overconfident miss-classifications resulting from scenes fully or partially covered by clouds. Specifically, we consider single-label scene classification on the SEN12MS

Manuscript received 19 January 2022; revised 7 August 2022 and 24 October 2022; accepted 26 October 2022. Date of publication 21 November 2022; date of current version 30 November 2022. This work was supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project "AI4Sentinels—Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of X. Zhu was jointly supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under Grant (ERC-2016-StG-714087, Acronym: *So2Sat*), in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Jakob Gawlikowski and Patrick Ebel contributed equally to this work.) (Corresponding author: Xiao Xiang Zhu.)*

Jakob Gawlikowski is with the Institute of Data Science, German Aerospace Centre (DLR), 07745 Jena, Germany, and also with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: jakob.gawlikowski@dlr.de).

Patrick Ebel is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Michael Schmitt was with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany. He is now with the Department of Aerospace Engineering, University of the Bundeswehr Munich, 85577 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: xiaoxiang.zhu.ieee@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3221788

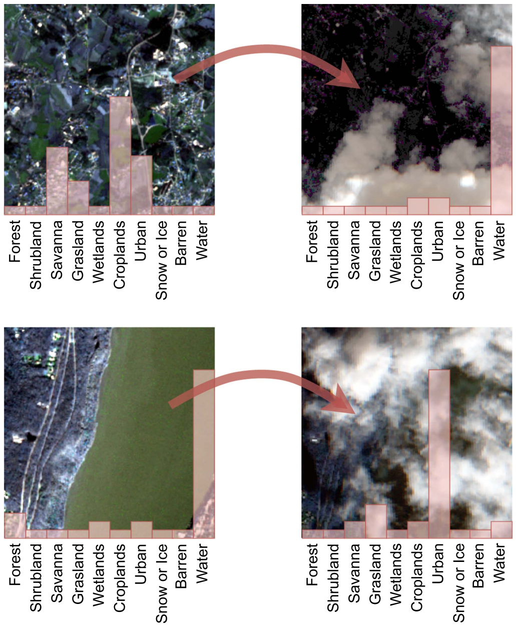


Fig. 1. Two examples of the effect of clouds on single-label scene classification. The visualization shows two examples of clear images, cloudy images, and the corresponding predicted class probabilities. While in both cases, the cloud-free image is classified correctly with respect to the ground truth, the cloudy version is misclassified. In the upper example, much of the croplands are obscured by cloud shadow, which causes the misclassification as a water body with a high soft-max probability. In the lower example, the clouds cover a large range of the water but keep a part of a city visible such that the sample containing clouds is misclassified as Urban with a high conviction. The cloud coverage of the samples is 19% and 77%, respectively. Although parts of the images are still visible, the classifier's predictions are misguided by the clouds and the resulting shadows.

dataset [2]. We use the Sentinel-2 images of the dataset, which have a resolution of 256×256 pixels and are assigned to one of 10 classes of land cover types. For cloud-covered samples, we utilize the corresponding and co-registered observations of the SEN12MS-CR dataset [20]. Our analysis is fourfold, as we consider the effects of clouds on the following.

1) *Data distribution*, by describing the effects of clouds on the statistics of the input dataset and how this affects individual land cover types.
2) *Classification performance*, by evaluating the impact of cloud coverage on a task performance level with respect to the considered single-label classification task, including individual class confusions.
3) *Effects on the network output*, by investigating the changes in the network predictions and the capability to separate cloudy samples from clear samples based on the network's output.
4) *Feature importance and network focus*, by analyzing which parts of an image drive a classifier's predictions and how this changes in the presence of clouds.

In sum, the contribution of this work is to provide a more thorough qualitative as well as quantitative analysis and interpretation of the effects of clouds on remote sensing applications, to subsequently allow further research to handle cloud-covered data more gracefully than currently feasible. The code base for the presented results and experiments can be found in our github repository: https://github.com/JakobCode/explaining_cloud_effects.

II. Data

A. Remote Sensing Data

To assess the effects of clouds on the scene classification task, both cloudy observations and patchwise land cover class annotations are required. For single-class labels and cloud-free observations, this work builds on the SEN12MS dataset of globally sampled Sentinel-1 and Sentinel-2 data [2], [21]. The Sentinel-2 data correspond to the Level-1 C top-of-atmosphere reflectance products. Semantic land cover annotations are given by the MODIS-derived [22] simplified IGBP scheme of [21], which consists of 10 different land cover types. For single-class labels, we use the provided target values in [2] which, for any sample, are given by the mode of its pixel-based simplified IGBP land cover type map. For every 252 globally distributed regions of interest, a large-scale observation is acquired within a given meteorologically defined season for each of the three sensors and collected semiautomatically via Google Earth Engine [23]. Each region on average covers an area of approximately 52×40 km^2 land surface, equating to images of about 5200×4000 pixels. All full-scene observations are translated into the Universal Transverse Mercator coordinate reference system. Afterward, the images are sliced into patches of sizes 256×256 pixels with a stride of 128 pixels, such that neighboring patches have an overlap of 25% to 50%. Patches that contain invalid pixels, either due to sensor noise or due to the coordinate transformation, are automatically removed from the dataset. For cloud-covered data, we utilize the compatible and co-registered SEN12MS-CR dataset of cloudy Sentinel-2 data [20].[1] The additional cloud-covered full-scene observations are acquired in the same year and season as their respective cloud-free counterparts to minimize surface changes and are preprocessed analogously. For training and testing data of this study, we use the intersection of both datasets' splits, respectively. That is, for each considered testing sample a cloud-free and a co-registered, potentially cloud-covered version exists.

In order to compute statistics on the extent of cloud coverage in the considered dataset, a pixelwise cloud map is required. We utilize s2cloudless [24] to compute binary cloud masks. The resulting distribution of cloud coverage on the considered test split is depicted in Fig. 2. The statistics indicate that the complete range of cloud coverage is present in the test split, from clear view to fully obscured. The distribution exhibits a concentration at high cloud coverage, implying an often impossible classification task. For hard or even impossible classification tasks, the predictions should be given with a larger entropy among the predicted soft-max probability vectors.

B. Data Distribution

The distribution of land cover types in the test split is reported in Fig. 3. The globally sampled land cover types are unbalanced,

[1] https://patrickTUM.github.io/cloud_removal

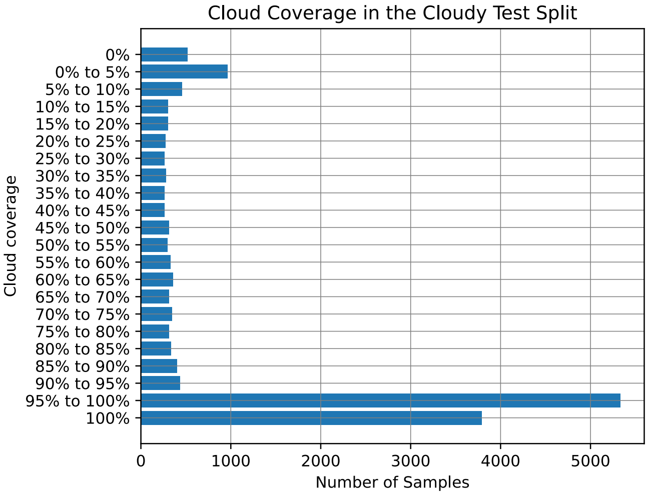


Fig. 2. Histogram of test split samples per percentage of cloud coverage. All extent of cloud coverage is present in the test split. The distribution exhibits a concentration at high cloud coverage, implying a challenging or even infeasible classification task.

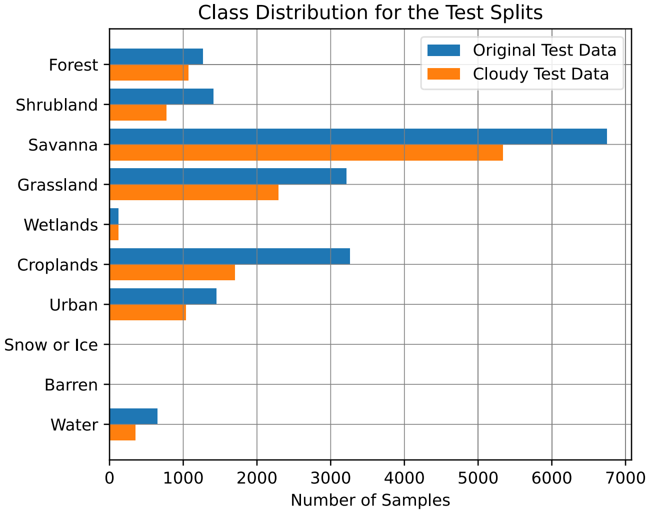
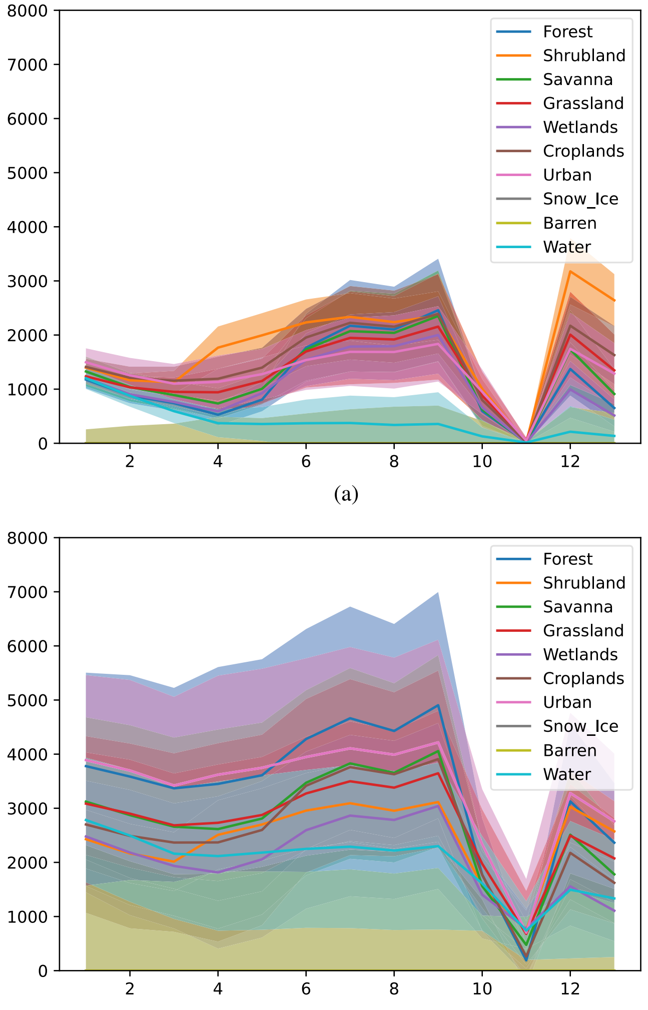


Fig. 3. Histogram of the land cover class distribution in the original test split of the SEN12MS dataset and the considered test split that is based on the intersection with the cloudy SEN12MSCR dataset. The globally sampled land cover types are unbalanced, with majority classes like *Savanna* while other classes hardly occur.



(a)

(b)

Fig. 4. Bandwise spectral fingerprint of each land cover class. The figures illustrate amplitude as a function of spectral bands and land cover type. Band intensities are computed as the grand mean across all samples, averaging across spatial dimensions for each class and band separately. The presence of clouds results in an average increase in band intensities and standard deviation. This indicates that, in the presence of clouds, land cover types become less separable on the basis of their spectral fingerprint. (a) Statistics of cloud-free data. (b) Statistics of 95% cloud-covered data.

with majority classes like *Savanna* while other classes (*Snow, Barren*) hardly occur. The distribution of land cover in the training split is comparable, which makes it representative of the holdout data.

The bandwise statistics of each class's spectral properties are illustrated in Fig. 4. The illustrated band intensities are computed by calculating the grand mean across all samples, averaging spatial dimensions for each class and band separately. The statistics show that the presence of clouds results in an average increase in band intensities as well as a considerable increase in standard deviations. That is, clouds result in land cover types being less separable based solely on their spectral properties. Furthermore, the considerable shift in the data distribution makes the behavior of neural networks unreliable and sensitive to misinterpretations caused by very confident but false predictions [25], [26].

III. Scene Classification Under Cloudy and Noncloudy Conditions

A. *Scene Classification Models*

We investigate the scene classification performance of a ResNet50 as well as a ResNet101 [27], a DenseNet121 [28], a VGG-16, and a VGG-19 model [29], which were already previously considered for this task [2]. Other than [2], we make use of all Sentinel-2 bands to include atmospheric information, which is of particular relevance in the presence of clouds. We trained on the cloud-free SEN12MS training data and randomly held out 10% of the training data for a validation set. The models were trained for 30 epochs and the models with the

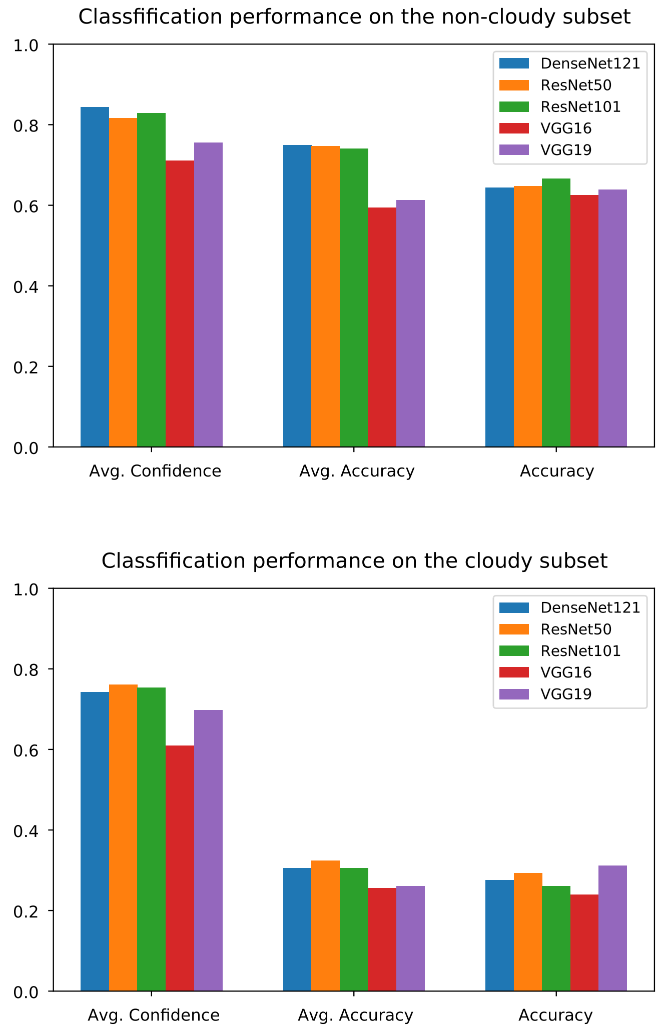


Fig. 5. Classification performances on cloudy and cloud-free data. The evaluated metrics demonstrate comparable performances for the different architectures considered in [2].

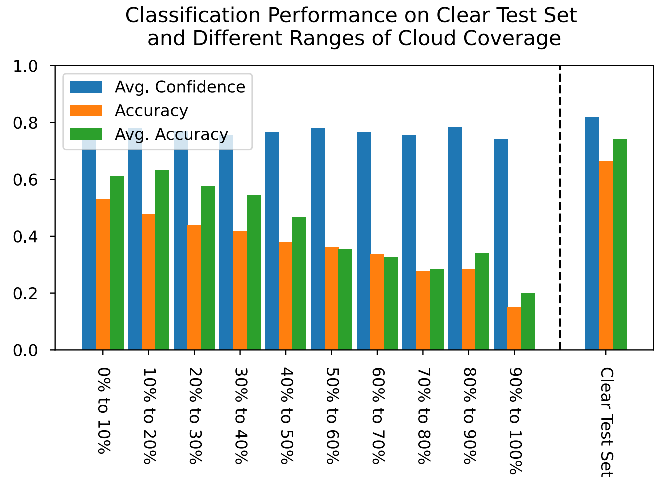


Fig. 6. Performance of the ResNet50 architecture as a function of varying ranges of cloud coverage. While accuracies are detriment with increasing cloud coverage, the network's confidence remains consistently high.

best performance on the validation set were saved during the training. For the optimization procedure, we utilized the ADAM optimizer [30] with a learning rate and weight decay of 10^{-5}. For the implementation, we extended the PyTorch [31] implementation provided by Schmitt and Wu.[2] The trained networks perform comparably to the baselines proposed in [2], which were trained on the 10 surface-relevant bands of Sentinel-2 only.

B. Classification Performance

Our trained networks achieve an average accuracy score between 0.61 and 0.75 (see also Fig. 5), which is comparable to the performance of the available networks pretrained on only 10 bands of Sentinel-2 [2]. In the following parts, we take the ResNet50 network as a representative use case for our further evaluations. The network can be seen as representative in a way that the presented findings based on the application of GradCam hold for all the trained networks. In contrast to the subset of clear images, the networks achieve only average accuracy scores between 0.26 and 0.32 on the cloudy test data. This denotes a considerably detrimental effect of clouds on the model's classification performance, in line with the high cloud coverage rates reported in Section II-A. In Fig. 5, the effects of the clouds on the accuracy, the average accuracy, and the confidence are illustrated. In general, the largest value within a network's soft-max output vector can be interpreted as the model confidence. Networks where the predicted probability represents the actual fraction of correct predictions are called calibrated while uncalibrated networks lead to over- or underconfident predictions [25]. We indicated the confidence by the average over the highest probabilities received from the network for the single samples. While there is a clear drop in classification performances, there is considerably less decrease in confidence.

Complementary, Fig. 6 details the performance of the ResNet50 network for different ranges of cloud coverage. The analysis shows that classification performances decrease with an increase in cloud coverage while confidence stays high.

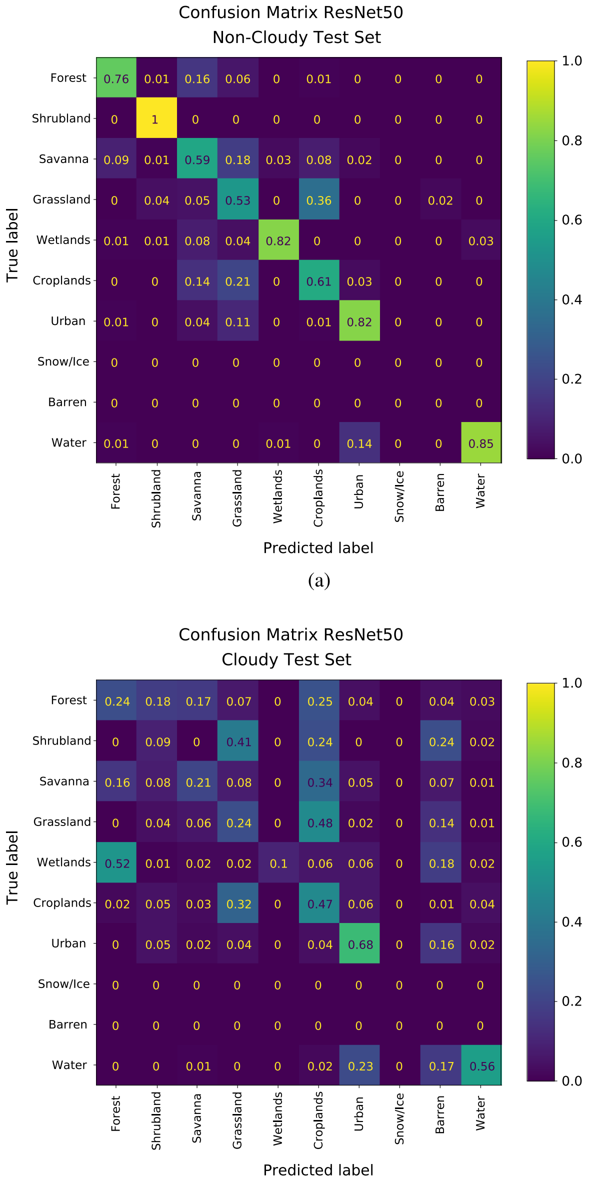
To attribute the decrease in performance to specific land types, we analyze the confusion matrices for clear and for cloudy observations shown in Fig. 7(a) and (b), respectively. For the cloud-free data, class 4 (*Grasslands*) is often confused with other types, specifically with class 6 (*Croplands*). The presence of clouds generally results in more misclassifications, but, in particular, reinforces the bias of predicting class 5 (*Wetlands*). Remarkably, especially the already harder-to-differentiate vegetation classes are much more distracted by the (partial) cloud cover with a clear bias toward class 4 and class 6.

IV. Analysis of Cloud Effects

A. Separability and Out-of-Distribution Analysis

The eventual occurrence of clouds poses the question of whether a given set of samples can be divided into cloudy and noncloudy images, solely based on a neural network's output.

[2] https://github.com/schmitt-muc/SEN12MS



(a)

(b)

Fig. 7. Confusion matrices of the cloudy and cloud-free test samples resulting from the intersection of SEN12MS and SEN12MS-CR. The true class labels are plotted versus the predicted class labels, with the row-normalized probabilities color-coded. Specifically, class 4 (*Grasslands*) is often confused with others, in particular with class 6 (*Croplands*). The presence of clouds generally results in more misclassifications, but in particular reinforces the bias of predicting class 6 (*Croplands*). Remarkably, the already harder-to-differentiate vegetation classes are much more distracted by the (partial) cloud cover with a clear bias toward classes 4 and 6. (a) Confusion matrix on cloud-free data. (b) Confusion matrix on cloud-covered data.

This can be seen as a case of Out-of-Distribution detection, which is a broadly studied topic in the field of machine learning [25], [32] and also applied in different remote sensing scenarios [26]. In order to evaluate the out-of-distribution detection performance of a classifier, one in general evaluates how well metrics can be used to separate a given test dataset into so-called in-distribution samples (in our case the noncloudy samples) and out-of-distribution samples (in our case the cloudy samples).

For every classification neural network, one can apply different metrics on the logit values as well as on the predicted probability vector. The motivation behind this analysis is driven by findings that predictions for data points from unknown data distributions might give a very confident prediction, but often differ considerably in the pure network output, the so-called logits [26]. An ideal model confronted with cloudy samples would express its uncertainty for example by a low confidence value or a high entropy in the resulting probability vector. Also, the features derived from a cloudy sample would fit relatively bad to the possible classes, and therefore, the predicted logit values should be small for all classes. Popular metrics are for example the *maximum probability* (or confidence), the *mutual information*, the *entropy*, the sum of the logit values (*log-sum*), and the *precision*. The precision is motivated by the Dirichlet distribution (a multivariate generalization of the Beta distribution) and can be interpreted as a description of the certainty on the predicted probability vector [25]. The precision is computed as the sum of the exponential of the logit values and the larger the precision value, the less variation in the prediction is assumed. In this article, we investigate the separability of cloudy and noncloudy samples based on the maximum probability, the entropy, the mutual information, the sum of the logit values, and the precision value.

B. *Grad-CAM for Saliency Map Computation*

Complementary to analyzing the effects of clouds on the scene classification performance via established statistics, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [33] to inspect the workings of the considered classifier when facing noisy optical data. Grad-CAM is a popular method to analyze which input region of an image contributed most to a given prediction. Grad-CAM can be applied post-hoc to a trained network to provide heat maps M_c of the models' attention on the image conditioned on a specific target class c, so-called saliency maps. To do so, the derivative $\frac{\delta y_c}{\delta A_k}$ of the output logit y_c for the conditioned class c with respect to the feature maps A_k is computed. The gradients are then global average pooled across the spatial dimensions H and W to obtain mapwise attention weightings

$$\alpha_{c,k} = \frac{1}{H \times W} \Sigma_{i=1,\ldots,H} \Sigma_{j=1,\ldots,W} \frac{\delta y_c}{\delta A_{k,i,j}}$$

which can be interpreted as the attribution of feature map A_k to drive the classification of c. The feature maps A_k at that layer are averaged across all output channels and the gradients for each channel are weighted by the respective layer's activations $\alpha_{c,k}$ in a simple linear combination. On the resulting pixelwise attribution of activations, a rectified linear unit σ is applied

$$M_c = \sigma(\Sigma_k \alpha_{c,k} A_k)$$

and the saliency map M_c is upsampled via bilinear interpolation to the dimensions of the input image. The resulting attention map specifies which areas in a given input to the network drive its classification as a scene of class c. We utilize Grad-CAM to analyze which regions of a land cover are salient in classifier's

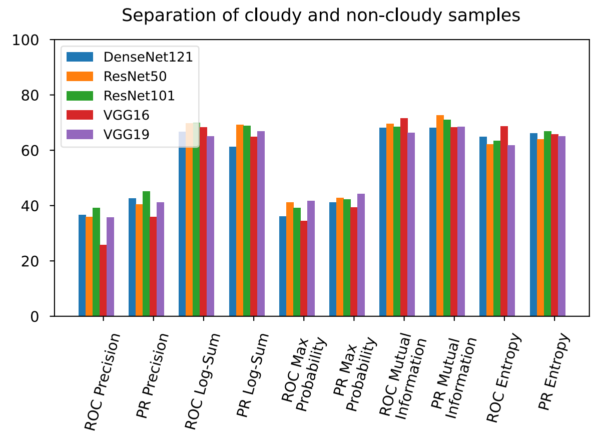


Fig. 8. Separability of samples from the cloudy and from the clear dataset, based on the PR and receiving-operating-characteristic (ROC) of different metrics applied to the output of different network architectures proposed in [2]. The evaluation shows that cloudy and noncloudy samples affect the output of different architectures differently. The best separability is reached with the VGG16 and the ResNet50 architecture and the mutual information metric, followed by the ResNet101 and the DenseNet121 architectures.

receptive fields, and how the presence of clouds affects these saliency maps.

V. Results

A. Effects on the Network Output

This section details the effects of clouds on the network output, including the predictions before applying the soft-max function to compute the categorical probability vectors. We utilize the metrics defined in Section IV-A and analyze the separability of the set of cloudy samples (with a coverage of at least 10%) to their cloud-free pairings and present in Fig. 8 the outcomes for the considered metrics in terms of the average under the curve of precision recall (PR) as well as the receiver operating characteristic curve (ROC). There is an effect of the different architectures on separability, dependent on the considered metric. Overall, separability works best for the mutual information and the entropy metric and the VGG16 architecture, followed by the ResNet50 and the DenseNet121 architecture. It is important to realize that a perfect separability, i.e., a value of 100, is unrealistic to reach in our setup, since several samples are only covered by clouds on a small fraction or do not contain any thick clouds at all (cmp. Fig. 2).

B. Feature Importance and Network Focus

To further analyze what drives misclassifications in the presence of clouds, we apply Grad-CAM to compute saliency maps as detailed in Section IV-B. Within our investigation, we encountered four different manners in which clouds affect the network's attention, presented in the following.[3]

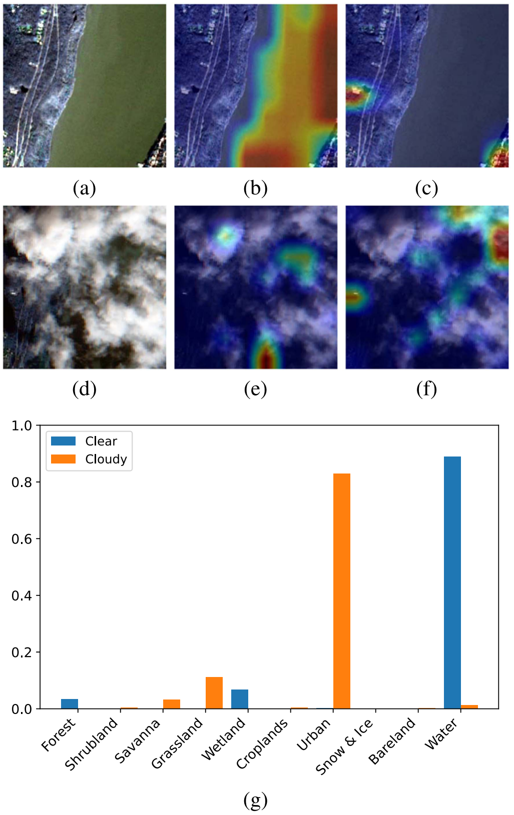


Fig. 9. (a) Clear and (d) 77% cloud-covered image with ground truth class water corresponding saliency maps with respect to the (b) and (e) classes water and (c) and (f) urban. In (g), the network's predictions are shown. This is an example of data where clouds partially cover the image such that homogeneous features are covered but "small feature classes" are still visible. Specifically, the few small buildings visible on the very edge of the image, and the small clouds, cause this confident misclassification.

1) Clouds Partially Cover the Image Such That Homogeneous Features are Covered But "Small Feature Classes" are Still Visible: Depending on the type of cloud coverage, a few clear features can already be enough to make the network predict a specific class with a high confidence value. Especially the urban class is an example of such behavior. Complementing Fig. 1 with the corresponding Grad-CAM results, Fig. 9 illustrates the saliency maps of a water-type land cover scene for both cloudy and cloud-free views. Evidently, the correct *Water* classification focuses on the whole water body, whereas the *Urban* misprediction is driven by the peripheral urban parts not covered by clouds. In both cases, the scenes are (in-)correctly classified at very high confidence, as shown in Fig. 1. Interestingly, the confidence of the network on the cloudy sample prediction is 86%, compared to 90% for the water prediction on the clean image.

[3] Please note that these chosen examples are exemplary in the sense that their class labels and the classifier's predictions are indeed representatives according to the land cover distribution of Fig. 3 and the confusion matrices of Fig. 7: The analyzed cases feature prominent land cover types such as Grassland, Croplands, Urban, and Forest—which, according to Fig. 2, make up a large proportion of the overall test data. Moreover, the considered cases are representative of salient changes to the network's performance. For instance, in the presence of clouds, the TPR of classifying Forest, the ground truth class in Fig. 13, drops drastically from 0.76 to 0.24. Meanwhile, the FPR to confuse Forest with croplands increases from 0.01 to 0.25, as shown in Fig. 7. As another example, Fig. 11 illustrates a confusion between the ground truth Grassland and the prediction of Cropland. In the presence of clouds, the FPR of this confusion is at 0.48, which is twice as large as the TPR of predicting Grassland correctly as shown in Fig. 7(b).

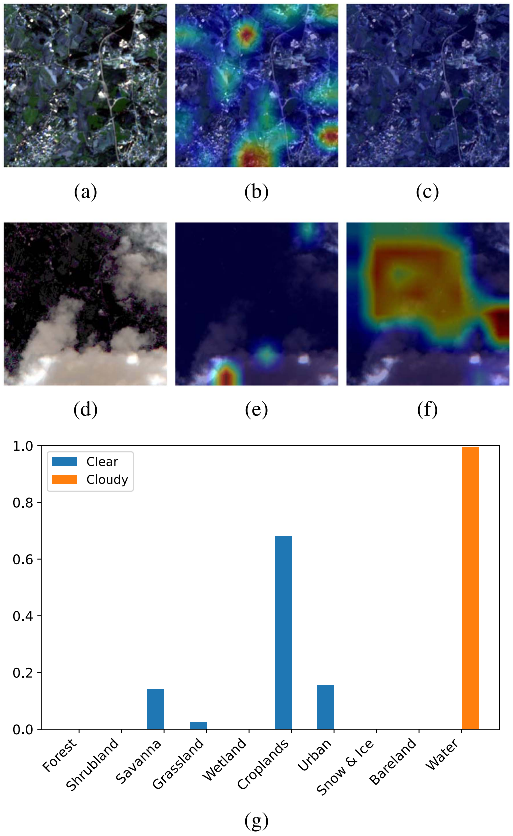


Fig. 10. (a) Clear and (d) 19% cloud-covered input images with ground truth class croplands and corresponding saliency maps for the (b) and (e) classes croplands and (c) and (f) water. In (g), the network's predictions on the clear and on the cloudy image are visualized. The illustrated example shows the case of large cloud shadow regions causing a confident misclassification. It is representative in featuring the majority class "croplands," constituting a large part of our dataset.

2) Structures are Hidden by Shadows: Even clouds that cover an image only partially on a small fraction can still have a considerable effect on the image caused by their shadow. Optical sensors are sensitive to illumination and large shadows impact the illumination. Based on this, shadows can hide structures and characteristics on the floor, leading to a more homogeneous-looking area. In Fig. 10, a very inhomogeneous side is visualized. As shown in Fig. 1, the confidence in the predictions is hence not very large. In contrast to this, the cloudy version covers most of the picture in a very dark monotonic-looking side. As a result, the network predicts the sample as a water body with high confidence. While the saliency map for the clear image shows several single regions that caused the correct prediction, the saliency map of the cloudy version clearly shows that the shadow caused the false prediction as a water body.

3) Small Clouds and Their Shadows Make the Ground Look Less Homogeneous: Clouds and their shadows cannot only cause homogeneity but also make images look more inhomogeneous. Especially many small clouds with many corresponding shadows make the image indicates more structure in the land side as their actually is. In Fig. 11, the cloud-free patch is accurately classified as "grassland." The cloudy patch of 40% cloud coverage is misclassified as "croplands." The corresponding saliency maps clearly show that while for the correct prediction on the clear image, most of the image is taken into account, the false prediction on the cloudy image is based mainly on cloudy and shadow parts of the image.

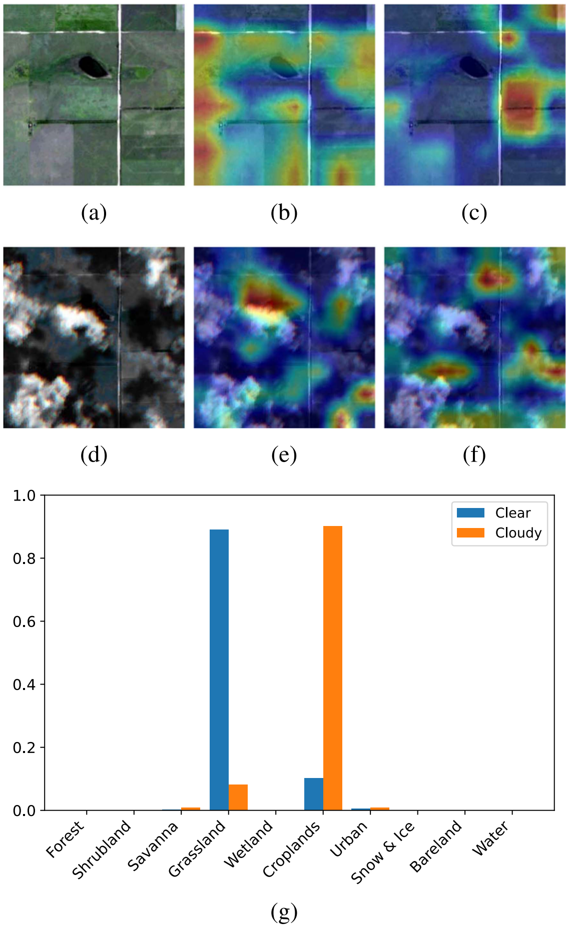


Fig. 11. (a) Clear and (d) 40% cloud-covered input images with ground truth class grasslands and corresponding saliency maps for the (b) and (e) classes grasslands and (c) and (f) croplands. In (g), the network's predictions are shown. This is a case of small clouds and their shadows making the ground look less homogeneous. Altogether, one can clearly see that the intensity of cloudy pixels and their high-contrast neighborhood capture the network's attention and result in misclassification. The shown misclassification is representative for many cases, as croplands are erroneously predicted twice as often as the correct class of grassland in the presence of clouds, according to Fig. 7(b).

4) Homogeneous and Semitransparent Clouds Make Ground Look More Homogeneous: Besides the above-considered non-transparent clouds with clear shapes and shadows, there also exist semitransparent and very homogeneous clouds. In Figs. 12 and 13, two examples are shown where these types of clouds lead to a wrong water and a wrong croplands prediction, respectively.

VI. Discussion

Following the four levels of analysis provided in Section V, this section communicates an interpretation of the observed results. The provided interpretations follow the preceding four stages of analysis to detail our views on the effects of clouds, from the raw data to network decisions and clarify how each step relates to one another.

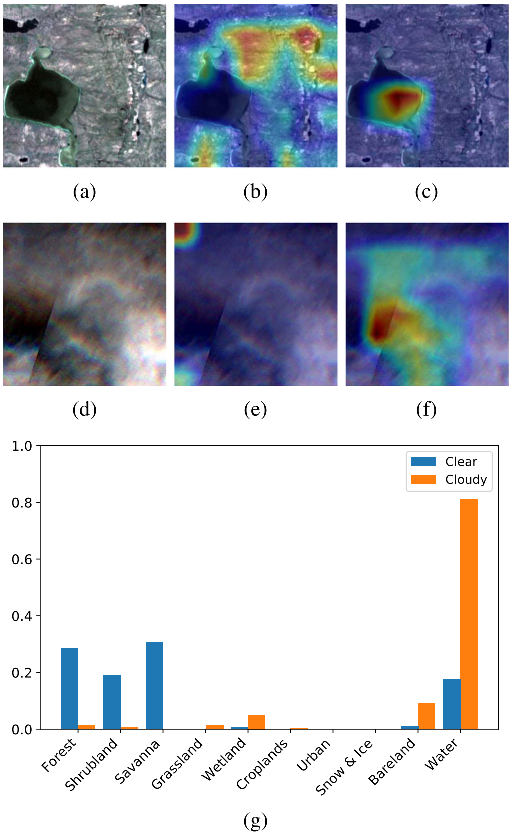


Fig. 12. (a) Clear and (d) 100% cloud-covered input images with ground truth class savanna and corresponding saliency maps for the (b) and (e) classes savanna and (c) and (f) water. In (g), the network's predictions are shown. The shown samples represent the case of homogeneous and semi-transparent clouds making ground appear more homogeneous. Altogether, one can clearly see that the lower contrast and the dark water shimmering through the clouds result in the water prediction.

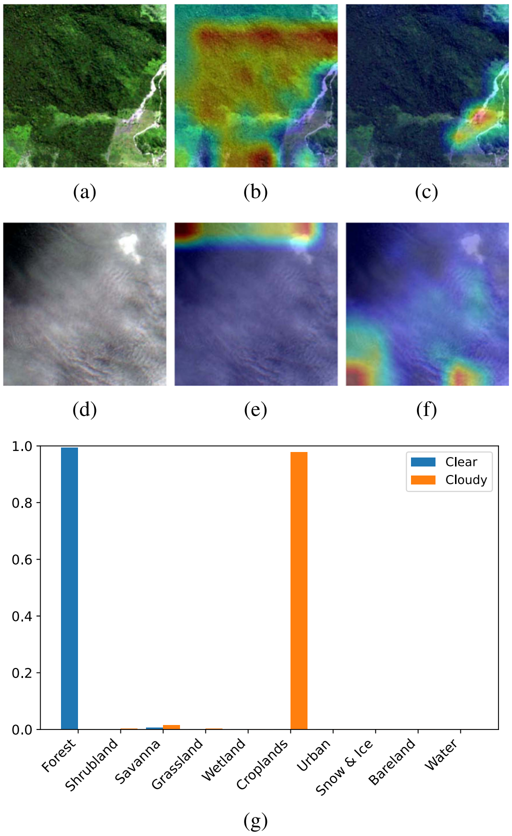


Fig. 13. (a) Clear and (d) 87% cloud-covered input images with ground truth class forest and corresponding saliency maps for the (b) and (e) classes forest and (c) and (f) croplands. In (g), the network's predictions are shown. This is a case of homogeneous and semitransparent clouds making ground appear more homogeneous. Specifically, some small regions with structured clouds result in the croplands prediction. This sample is represented as, in the presence of clouds, the correct classification of "forest" drops to a third of the original rate. Moreover, the probability of misclassifying "forest" as "croplands" outgrows the chance of a correct prediction, as analyzed in Fig. 7.

1) Distribution Shift: As presented in Section II-B, the presence of clouds changes the bandwise data statistics. That is, an overall shift in the data distribution is observable. Distribution shifts have previously been shown to make the behavior of neural networks unreliable and sensitive to misinterpretations caused by very confident but false predictions [25], [26]. Moreover, the bandwise standard deviations increased considerably. This, in return, causes the individual land cover classes to be less separable on their spectral statistics alone. While convolutional neural networks do also incorporate spatial information via local context, the spectral statistics of a sample become less indicative of its class belongings. Finally, preprocessing pipelines based on statistics priorly computed on the cloud-free training data (as in [2]), are no longer appropriate as they do not match the cloudy data distribution and thus do not normalize the cloud-covered data.

2) Classification Performance and Overconfidence: The performance and confidence metrics presented in Section III-B indicate that the classifier is oblivious to the presence of previously unencountered clouds and their effects caused by the shift in the input data distribution as described in Section II-B. Interestingly, the drop in the accuracy is not uniformly distributed, but the confusion matrix in Fig. 7(b) shows a bias toward particular classes. Moreover, this bias is not toward the class with the most training samples (savanna). In addition to the biased decrease in classification performance, the classifier's high overconfidence in the cloudy samples is an undesirable effect caused by clouds. Even though the data are very different from the data known from the training (as seen in the band statistics), the network still gives predictions with high confidence. This behavior is in line with prior observations that neural networks are overly confident in their predictions even in the presence of noise and on changing data domains and distributions [25], [34].

3) Cloudy Noncloudy Separability: Even though the clouds have such a strong effect on the classification performance, the results in Section V-A showed that the separation between cloudy and noncloudy images based on different metrics on the network output is only possible to a certain extent. Even the most discriminative network architectures and measures can only separate in-distribution from out-of-distribution samples in roughly two-thirds of the considered cases. This behavior was also observed when the threshold for the cloud coverage was increased from 10% to a larger value or even to 100%. Besides this, the classifiers and metrics also differ in the extent to which

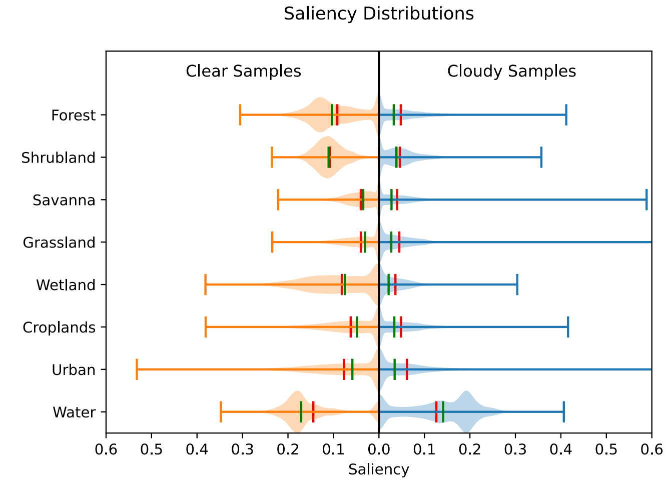


Fig. 14. Violin plot visualization of the pixelwise saliencies with respect to the correct class over the noncloudy (left) and cloudy (right) test data. The violin bodies indicate a smoothed empirical probability distribution, per class. The mean and the median intensities for each class are given by the red (mean) and the green (median) line, respectively. For the clear case, the plots clearly show classwise differences in the average number of pixels and intensities contributing to the prediction. For the cloudy case, they show a reduction of saliency for all classes but the water class.

they can grasp differences between representations of cloudy and cloud-free images. Especially the performances based on the precision value and the maximum probability underline the findings that networks are overly confident and further support the interpretation that the scene classifier is oblivious to the presence of previously unencountered clouds.

4) Outliers as Distractors: As evidenced by the Grad-CAM analysis in Section V-B, cloud coverage poses an obstacle to land cover classification in four different kinds: First, clouds partially cover the image such that large areas are covered but relatively irrelevant “small feature classes” may still be visible. Second, otherwise apparent structures may be hidden by cloud shadows. Third, small clouds and their shadows make the ground look less homogeneous. Fourth, transparent clouds lead to a different representation of the (often already on clear images hard to differentiate) classes of land cover. These four cases can be directly related to the shift in the confusion matrix represented in Fig. 7(b), as, for example, the large shift from water to urban classes can be explained and the interplay of houses and water was presented as shown in Fig. 9. Moreover, samples across all four cases highlight that the network’s spatial attention often shifts toward clouds, their shadows, or the transition between both. That is, outstandingly bright, very dark, or high-contrast areas often coincide with a focus of attention. As these are oftentimes entailed by the presence of clouds, we interpret that out-of-distribution image intensities function as a distractor. In sum, clouds and their shadows distract classifiers on a macroscale by obscuring large areas—but also on a per-pixel level, as cloud or cloud-shadow induced intensity changes equally distract the classifier from the actual land cover. Moreover, the evaluation of the pixelwise saliencies in Fig. 14 shows that all areas of the water land cover type contribute to a relatively high relevance. In contrast, for the more feature-based urban class, the majority of the class is not that relevant for the prediction. At the same time, the values for the urban saliency sometimes reach larger values than for all other classes. Interestingly, an equivalent but less significant trend of larger areas of an image into account also appears for the forest and the shrubland class. For the Wetland class, the relevance is not that concentrated on single values but seems to also take a variety of areas into account. Those classes also have the largest relative drop in the true positives, indicating that the coverage of clouds harms these types of classes more than those, which focus on smaller areas.

Hence, clouds and shadows covering parts of an image and hiding information for specific areas affect the scenewise classification of regions differently. When comparing the pixelwise saliency of cloud-free data to one of cloudy samples, a clear decrease in saliency is visible while the outliers become more extreme. That is, on average, a smaller fraction of a scene’s pixels contributes to its classification in the presence of clouds, except for a few extrema. This finding validates the hypothesis that less information covering multiple pixels leads to class prediction but mainly local information. This is also what the presented saliency maps, except of the false water prediction in Fig. 10, indicate. Fewer pixels driving a classification are in contrast to the majority principle that the most prominent class (i.e., the one covering the largest area) defines a scene’s label. The only exception from the trend of shrinking saliencies is the Water class, for which larger areas of cloud shadows in other scene types tend to be misclassified as water. Altogether, the presented analysis clearly shows that conventionally (pre-)trained networks are not fit for domain shifts in data common in remote sensing. Specifically, the derived features cannot be used to give a strong idea of the underlying class, even if only parts of the image are covered by clouds.

Overall, our multistage analysis reveals that the effects of clouds on remote sensing applications manifest in many different aspects of the pipeline, from the raw data to the information a trained network extracts from these images. As the visualizations and evaluations of the Grad-CAM images underlined, the structure caused by clouds and their shadows contain misleading information leading to very confident but false predictions.

While our analyzed data comprise a large cohort of globally distributed regions acquired through several seasons that should be sufficiently heterogeneous and representative, our analysis may nonetheless be dependent on, e.g., the choice of datasets and cloud detection algorithms. For instance, future work may conduct our analysis focused on a single-country level, e.g., on the dataset in [35]. Moreover, recent publications have provided novel large-scale datasets for cloud detection or removal in time series [12], [13], [36], which may serve as an extended version of our analysis. With respect to the cloud detector algorithm, s2cloudless was chosen for being commonly deployed, easily applicable, and performing well [37], [38]. However, many alternative approaches exist [35], [39], [40], [41], [42], whose variable sensitivity thresholds may result in qualitatively different cloud masks and thus different downstream analysis results. The chosen s2cloudless algorithm is reported to show a fair “balance (within 10%) between commission and omission errors” [38], which may avoid any one-sided biases to either false alarms or misses of clouds in our subsequent analysis.

VII. Conclusion

With over 50% of our planet's surface covered by clouds at any point [6], haze and clouds pose a considerable obstacle to the continuous monitoring of Earth. In this work, we investigated in detail the effects of clouds on a deep neural network performing remote sensing scene classification. To start with, clouds considerably alter the spectral characteristics of data and make individual land cover types less separable from one another. In terms of performance, we observed a considerable drop in overall classification accuracy to almost half of the rates at clear views. A confusion matrix analysis revealed that existing biases toward predicting certain classes are reinforced in the presence of clouds. Even though the network remains highly confident in its predictions, it cannot separate between cloud-free and cloud-covered observations—indicating the classifier's unawareness of clouds. Finally, we complemented the reported statistics with a qualitative analysis of the classifier's attention maps. The saliency maps highlighted that clouds distract the network from the actual land cover surface. That is, rather than focusing on the actual land cover, previously unseen noise is so salient that it becomes the focus of the classifier's attention. These insights contribute to a better understanding of the effects of clouds on remote sensing applications and may consequently guide the future development of more robust models. We plan to continue our research and develop a methodology that is more robust to the effects of outliers and noise detailed in this contribution. For future approaches, evaluating the distribution of image regions relevant to the prediction is an interesting way to identify misconceptions and misclassifications. In addition, training methods that incorporate clouds and shadowy regions and can express the uncertainty and the lack of knowledge due to obscured parts of the image are a promising route to more robust approaches in the future.

Acknowledgment

The authors would like to thank Dharani Deivasihamani for the support in implementing the evaluation pipeline and visualizing preliminary results.

References

[1] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[2] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the SEN12MS dataset," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. V- 2-2021, pp. 101–106, 2021.

[3] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, "Convolutional neural networks based remote sensing scene classification under clear and cloudy environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Oct. 2021, pp. 713–720, doi: 10.1109/ICCVW54120.2021.00085.

[4] M. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1479–1486.

[5] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A cross fusion network for joint land cover classification using optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1562–1574, Jan. 2022, doi: 10.1109/JSTARS.2022.3144587.

[6] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[7] D. Spänkuch, O. Hellmuth, and U. Görsdorf, "What is a cloud? Toward a more precise definition?," *Bull. Amer. Meteorol. Soc.*, vol. 103, pp. E1894–E1929, Mar. 2022, doi: 10.1175/BAMS-D-21-0032.1.

[8] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li, "Thick clouds removal from multitemporal ZY-3 satellite images using deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 143–153, Dec. 2020, doi: 10.1109/JSTARS.2019.2954130.

[9] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 333–346, 2020.

[10] Y. Zi, F. Xie, N. Zhang, Z. Jiang, W. Zhu, and H. Zhang, "Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3811–3823, Mar. 2021, doi: 10.1109/JSTARS.2021.3068166.

[11] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, "Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for sentinel-2a imagery," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 157.

[12] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5222414, doi: 10.1109/TGRS.2022.3146246.

[13] A. Sebastianelli et al., "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412216, doi: 10.1109/TGRS.2022.3208694.

[14] M. Rußwurm and M. Körner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 11–19.

[15] Z. Gu, P. Ebel, Q. Yuan, M. Schmitt, and X. X. Zhu, "Explicit haze & cloud removal for global land cover classification," in *Proc. Comput. Vis. Pattern Recognit. Conf. Workshop Multimodal Learn. Earth Environ.*, Jul. 2022, pp. 1–6. [Online]. Available: https://elib.dlr.de/186738/

[16] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[17] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020, doi: 10.1109/JSTARS.2020.3005403.

[18] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.

[19] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102520.

[20] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2021.

[21] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.

[22] M. A. Friedl et al., "Global land cover mapping from MODIS: Algorithms and early results," *Remote Sens. Environ.*, vol. 83, no. 1/2, pp. 287–302, 2002.

[23] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017.

[24] A. Zupanc, "Improving cloud detection with machine learning," *Sentinel-Hub*, 2017, Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[25] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2021, *arXiv:2107.03342*.

[26] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616819, doi: 10.1109/TGRS.2022.3140324.

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

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:main.html

[30] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, May 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:main.html

[31] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[32] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 7047–7058.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[34] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2016.

[35] J. Li et al., "A lightweight deep learning-based cloud detection method for Sentinel-2A imagery fusing multiscale spectral and spatial features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5401219, doi: 10.1109/TGRS.2021.3069641.

[36] C. Aybar et al., "CloudSEN12–A global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Zenodo*, Aug. 2022.

[37] J. Braaten, K. Schwehr, and S. Ilyushchenko, "More accurate and flexible cloud masking for Sentinel-2 images," *Medium*, 2020, Accessed: Oct. 16, 2022. [Online]. Available: https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f

[38] S. Skakun et al., "Cloud mask intercomparison exercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2," *Remote Sens. Environ.*, vol. 274, 2022, Art. no. 112990.

[39] Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, and L. Zhang, "Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 89–108, 2022.

[40] L. Sun et al., "A new cloud detection method supported by GlobeLand30 data set," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3628–3645, Oct. 2018.

[41] H. Guo, H. Bai, and W. Qin, "ClouDet: A dilated separable CNN-based cloud detection framework for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9743–9755, Sep. 2021, doi: 10.1109/JSTARS.2021.3114171.

[42] D. López-Puigdollers, G. Mateo-García, and L. Gómez-Chova, "Benchmarking deep learning models for cloud detection in Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 992.

Jakob Gawlikowski received the B.Sc. and M.Sc. degrees in mathematics in 2015 and 2019 from the Technical University of Munich, Munich, Germany, where he is currently working toward the Ph.D. degree in robust data fusion with the Chair of Data Science in Earth Observation, Department of Aerospace and Geodesy.

He is currently a Researcher with the German Aerospace Center's (DLR) Institute of Data Science, Jena, Germany. His research focuses on data fusion machine learning approaches with a special focus on uncertainty quantification and robustness in deep learning models.

Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrü, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is currently working toward the Ph.D. degree in optical satellite image reconstruction with the SiPEO Lab, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany.

His research interests include machine learning and its applications in computer vision and to remote sensing data.

Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the habilitation in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2021, he has held the Chair for Earth Observation with the Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany. Before that, he was a Professor in applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences. From 2015 to 2020, he was a Senior Researcher and Deputy Head with the Professorship for Signal Processing in Earth Observation, TUM; in 2019, he was additionally appointed as an Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist with the University of Massachusetts, Amherst. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations, particularly interested in remote sensing data fusion with a focus on SAR and optical data.

Dr. Schmitt is a Co-Chair of the Working Group "Active Microwave Remote Sensing" of the International Society for Photogrammetry and Remote Sensing, and also of the Working Group "Benchmarking" of the IEEE-GRSS Image Analysis and Data Fusion Technical Committee. He frequently serves as a Reviewer for a number of renowned international journals and conferences and has received several Best Reviewer awards.

Xiao Xiang Zhu (Fellow, IEEE) received the master's (M.Sc.), Doctor of Engineering (Dr.-Ing.), and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Chair Professor for Data Science in Earth Observation with TUM and was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de). Since 2019, she has also been the head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been a Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and the University of California, Los Angeles, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor with ESA's Phi-lab. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is on the scientific advisory board of several research organizations, among others the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.