communications
earth & environment

ARTICLE

https://doi.org/10.1038/s43247-023-01146-0 OPEN

Meta-learning to address diverse Earth observation problems across resolutions

Marc Rußwurm[1,2]✉, Sherrie Wang[3,4], Benjamin Kellenberger[1,5], Ribana Roscher[6,7] & Devis Tuia[1]

Earth scientists study a variety of problems with remote sensing data, but they most often consider them in isolation from each other, which limits information flows across disciplines. In this work, we present METEOR, a meta-learning methodology for Earth observation problems across different resolutions. METEOR is an adaptive deep meta-learning model with several modifications that allow it to ingest images with a variable number of spectral channels and to predict a varying number of classes per downstream task. It uses knowledge mined from land cover information worldwide to adapt to new unseen target problems with few training examples. METEOR outperforms competing self-supervised approaches on five downstream tasks, showing its relevance to addressing novel and impactful geospatial problems with only a handful of labels.

[1] Environmental Computational Science and Earth Observation Laboratory (ECEO), École Polytechnique Fédérale de Lausanne (EPFL), Route des Ronquos 86, Sion, 1951, Switzerland. [2] Laboratory of Geo-information Science and Remote Sensing (GRS), Wageningen University, Droevendaalsesteeg 3, Wageningen 6708 PB, The Netherlands. [3] Goldman School of Public Policy, University of California, Berkeley, 2607 Hearst Ave, Berkeley 94720 CA, USA. [4] Department of Mechanical Engineering and Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 55 Massachusetts Avenue, Cambridge 02139 MA, USA. [5] Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven 06520-8106 CT, USA. [6] Institute of Bio- and Geosciences, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, Jülich 52425, Germany. [7] Remote Sensing Group, University of Bonn, Niebuhrstr. 1a, Bonn 53113, Germany. ✉email: marc.russwurm@wur.nl

Earth science strives toward understanding and modeling processes occurring at the Earth's surface. Recently, more and more studies have appeared using increasingly growing amounts of satellite image data acquired at various spatial, spectral, and temporal resolutions. To extract actionable knowledge from this raw data, Earth scientists increasingly deploy deep learning models[1] that require large annotated datasets for performing at their best. Each dataset is annotated with a focus on a particular problem. Some, like building footprint segmentation[2] or land cover classification[3–5], are well-established, with large annotated datasets being available. Others, like marine debris detection[6–8], are less explored and typically scarce in labels. Furthermore, covariate shifts in satellite image data and concept shifts in annotations[9,10] are caused by various physical, environmental, social, and economic factors that differ between geographic regions, thus making it necessary to collect region-specific datasets. These under-explored or highly region-specific problems typically require researchers to invest a substantial amount of time and effort in annotating data to describe the specific problem in a machine-compatible way. Due to this heterogeneity in Earth observation problems, a wide and diverse landscape of Earth observation and remote sensing datasets has emerged. Each dataset is typically dedicated to describing one particular problem and viewed in isolation from the others. This is inefficient, as many Earth observation problems still share common knowledge that can inform different-but-related target problems. For instance, large-scale land cover classification is often done with different land cover categories and switching from one category scheme to another typically requires re-training entire models, even though the underlying problem remains very similar. Systematically utilizing this common knowledge between source and target problems with a specific learning algorithm is the objective of transfer learning[11], while meta-learning[12] extends this idea by learning this learning algorithm itself[13]. In particular model-based transfer learning, where the common knowledge is encoded in the weights of deep learning models, has gained popularity within deep learning[14]. Recently, remote sensing foundation models, such as RingMo[15] or contrastive self-supervised learning models like SSLTransformerRS[16] have been proposed that pretrain very large deep learning models on unified heterogeneous datasets, often made of collections of pre-existing smaller datasets. These models can then be fine-tuned on a particular downstream problem with fewer annotated data points. Meta-learning is also a transfer learning approach, in spirit similar to these contrastively learned foundation models, but it aims at pre-training a comparatively small model on a dataset of many source problems. A general problem is here explicitly expressed as a concrete task that contains a training (or support) dataset and a testing (or query) dataset that needs to be classified correctly. Three prominent taxonomies of meta-learning approaches exist[13], of which two have been used in remote sensing: metric-based and optimization-based. Metric-based meta-learning[17,18] pre-trains a deep feature extractor to create an expressive feature space where test samples are matched to class prototypes defined by each task training dataset. This feature extractor is frozen for test tasks, similar to the self-supervised learning approaches pre-trained on hand-designed source problems, which we compare to in the results section. architecturally similar to self-supervised learning strategies. Metric-based meta-learning has been particularly useful in remote sensing with high-resolution aerial RGB imagery[19–22]. Most research[19–21] addresses concept shift between problems, where a new set of classes needs to be classified from a few annotated images. Lunga et al.[22] target covariate shift in high-resolution imagery taken under different conditions, such as varying viewing angles[2], with a dedicated hashing and clustering framework. In optimization-based meta-learning, the entire pre-trained deep learning model is fine-tuned to a downstream task with stochastic gradient descent. The model-agnostic meta-learning algorithm (MAML)[23] and its variants[24–26] are prominent examples of models that are explicitly optimized to be fine-tuned to new problems with few annotated images. Optimization-based approaches with MAML have been used predominantly on Earth observation problems to mitigate concept and covariate shifts in medium-resolution multi-spectral imagery: applications cover global-scale cropland mapping[27] or land cover classification[28], where covariate shifts make it difficult to transfer models across different geographic regions. Despite their success, current meta-learning approaches are not yet used to their full potential since they focus on one family of problems for pre-training and fine-tuning. They cover urban scene classification[21], land cover classification[28], cropland classification and mapping[27], the most often taken in isolation from each other. It is common that studies use one data source, such as high-resolution aerial imagery[19–21,29], or medium-resolution multi-spectral imagery[27,28] and address the model transfer within this single homogeneous problem family. Only very recent approaches are starting to address these limitations and focus on integrating satellite imagery across heterogeneous problem families. For instance, MOSAIKS[29] uses a small single-layer small convolutional network[30] that extracts autocorrelation features from high-resolution aerial imagery. The features from this featurization approach proved effective when regressing socio-economic variables like housing prices, income, road length, nighttime lights, and environmental factors like forest cover and elevation. In this work, we address learning across different Earth observation problems systematically in METEOR: a meta-learning methodology for Earth observation problems across different resolutions. It is an optimization-based meta-learning approach that uses a small deep learning model with a single output. It is pre-trained with the model-agnostic meta-learning (MAML)[23] algorithm to distinguish different land cover categories on medium-resolution multi-spectral satellite data, as shown in Fig. 1. Extending previous works, we focus explicitly on fine-tuning this model to different heterogeneous real-world downstream classification problems involving a different number of classes, data with different spatial and spectral resolutions and few annotated samples. This heterogeneous transfer is enabled by three key methodologies as follows: first, we replace all batch normalization[31] layers with instance normalization[32] in the model, as we show experimentally that classical, transductive batch normalization[23] has detrimental effects on downstream problems with high-class imbalance (see "Designing Meteor"). Second, we dynamically change the convolutional kernels of the input channels to adapt to problems with different spectral bands, as detailed further in the methods section. Third, we address downstream problems with different numbers of classes by pre-training a binary meta-model, fine-tuning this model to each class separately, and ensembling a one-vs-all classifier.

These key modifications result in METEOR: a single pre-trained meta-model that can adapt to new problems of interest across geographies and sensors from limited label information. Using METEOR, domain experts can address these problems with satellite data of varying spatial and spectral resolutions, described by a few annotated images, and with a variable number of target classes.

Results

In this results section, we first experimentally highlight the importance of instance normalization in the METEOR model on realistic downstream problems, beyond an idealized class-balanced few-shot setting (Table 1). We then compare

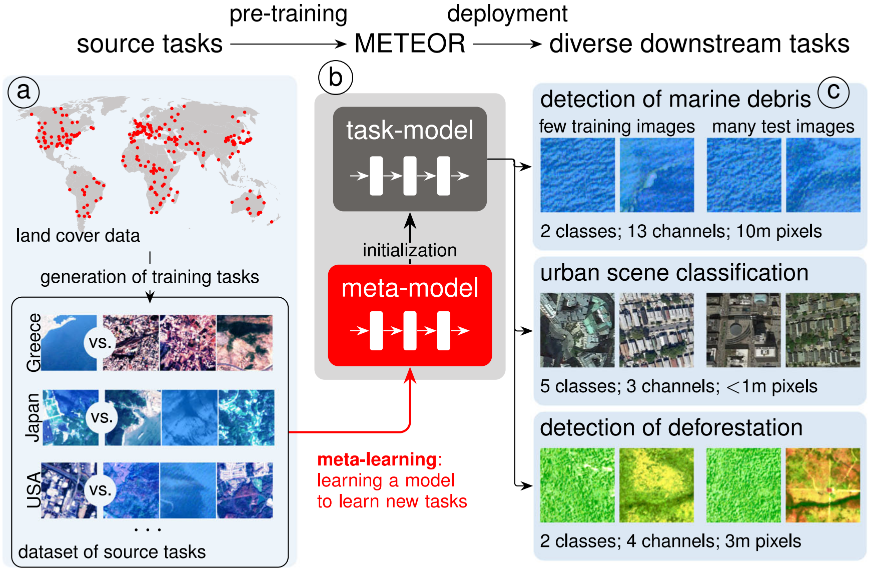
pre-training
deployment
source tasks
METEOR
diverse downstream tasks
a
land cover data
generation of training tasks
Greece
vs.
Japan
vs.
USA
vs.
. . .
dataset of source tasks
b
task-model
initialization
meta-model
meta-learning: learning a model to learn new tasks
c
detection of marine debris
few training images
many test images
2 classes; 13 channels; 10m pixels
urban scene classification
5 classes; 3 channels; <1m pixels
detection of deforestation
2 classes; 4 channels; 3m pixels

Fig. 1 Overview of METEOR in as a deep learning model pre-trained on land cover source tasks and deployed on diverse downstream tasks. A task is a (small) dataset containing few annotated images, divided into independent train and test sets. The task data describes a new problem in a format that a machine learning model can be optimized on. In METEOR, shown in (**b**), a randomly initialized meta-model is pre-trained with model-agnostic meta-learning (MAML)[23] to solve land cover classification source tasks, shown in (**a**). MAML[23] yields a deep meta-model that has explicitly learned to learn from different tasks with few labeled images. In each pre-training task, the model must distinguish one randomly chosen land cover type from others using satellite imagery of the same geographic area. A map of geographic regions from the Sen12MS[5] dataset with three examples of such pre-training tasks from Greece, Japan, and USA. The pre-trained meta-model can then be fine-tuned to diverse downstream problems shown in (**c**) with only few labeled images, thus leading to problem-specific task-models.

Table 1 Different pre-training configurations tested on idealized (Sen12MS)[5] and realistic (DFC2020)[33] test tasks.

Task-datasets		**Sen12MS**	**DFC2020**
Number of tasks		**1000**	**7**
Task design		**Idealized**	**Realistic**
Label distribution		**Balanced**	**Imbalanced**
Exp. #1: fixed algorithm (MAML) vary normalization			
MAML	**Instance norm (IN)**[54]	0.78	**0.82 ± 0.08**
	Transductive BN[23]	**0.85**	0.26 ± 0.05
	Conventional BN[31]	0.84	0.60 ± 0.18
	Tasknorm-I[25]	0.83	0.59 ± 0.24
	Groupnorm[60]	0.72	0.54 ± 0.20
Exp. #2: vary algorithms fixed normalization (IN)			
IN	Fo-MAML[23]	0.66	0.77 ± 0.11
	SparseMAML[24]	0.74	0.79 ± 0.11
	SparseFoMAML[24]	0.63	0.74 ± 0.13

The meta-model is tested on Sen12MS test tasks with a similar structure to the pre-training source tasks (column Sen12MS) and on unbalanced land cover tasks from the DFC2020 dataset (column DFC2020), where the label distribution of the target task is unknown. In experiment #1, we fix the pre-training algorithm (model-agnostic meta-learning (MAML)[23]) and vary the normalization of the network. Highlighted by underscores, transductive batch normalization (BN)[31] achieved the highest accuracy on the idealized Sen12MS test tasks, but performed worst in the realistic use-case (DFC2020). We found that this finding also holds in experiment #2 where we fixed instance normalization and tested MAML against more recent meta-learning models like SparseMAML[24] or the first-order approximations of MAML (Fo-MAML) and SparseMAML (SparseFoMAML). Overall, a deep learning model trained with instance norm (IN) layers trained with the standard MAML algorithm performed best for all the results presented in this work. We highlight best scores by fold face and, for DFC2020, we report one standard deviation over five model runs with different query/support sets.

METEOR's meta-model to other state-of-the-art approaches within homogeneous land cover classification problems from different geographical regions (Table 2) and across different heterogeneous problem fields with different resolutions (Table 3). Finally, in the section "Interpreting and explaining METEOR's predictions across geospatial problems", we highlight the diversity of problems to which METEOR can be applied. We do so by a qualitative analysis of several example use-cases Earth scientists may encounter.

Designing METEOR. Table 1 shows the importance of instance normalization for class-imbalanced downstream problems (DFC2020) experimentally. This describes the central finding that enabled the deployment of this meta-learning approach across different realistic use-cases presented in this work. It shows the pre-training of METEOR with different configurations (rows in the table) tested on two datasets of test tasks (columns) of a very different nature regarding class-balancing. In all the experiments of this work, we used the same METEOR meta-model trained on globally distributed land cover tasks from the train regions of the Sen12MS dataset[5] (dataset details in the Methods section). The Sen12MS-column in Table 1 shows the model performance on 1000 class-balanced binary 4-way 2-shot tasks with 16 images per task from the Sen12MS test areas. This dataset presents an idealized class-balanced configuration that is common in few-shot meta-learning benchmarks[18,23] where always the same number of images per class has to be classified. The DFC2020-column shows the performance in the seven geographic areas from the Data Fusion Contest 2020 (DFC2020) dataset[33]. Here, a severe class imbalance is present where some land cover categories, such as water, are more frequent than others. In comparison to Sen12MS, the DFC2020 datasets present a more realistic land cover classification scenario in a class-imbalanced setting: depending on the region, the number of images available varies from 476 to 1439, and the number of classes from 5 to 7; consequently, the support images used for training the specific task-models also varies from 50 to 70 for a 10-shot training. Note that the accuracies between these two columns are not directly comparable due to the different difficulties of the respective datasets. Instead, we are interested in highlighting which pre-training configurations lead to the best results on idealized (Sen12MS) or realistic (DFC2020) downstream tasks.

Table 2 Model comparisons within land cover problems.

		Number of shots (training examples per class)				
Model	**Avg. rank**	**1**	**2**	**5**	**10**	**15**
SSL4EO	**2.51**	56.4 ± 9.7	**72.8** ± 11.8	**79.4** ± 10.2	80.5 ± 10.6	82.4 ± 10.3
METEOR	2.84	**61.5** ± 10.7	69.2 ± 11.9	78.6 ± 11.2	**81.5** ± 10.4	81.7 ± 11.9
MOSAIKS	2.86	61.3 ± 11.5	68.7 ± 14.8	77.3 ± 11.5	81.3 ± 10.3	**84.7** ± 9.2
BASELINE	2.99	58.1 ± 11.9	71.2 ± 9.9	**79.4** ± 7.9	81.0 ± 8.0	82.7 ± 7.9
SSLTRANSRS	5.70**	51.2 ± 10.5	61.4 ± 6.4	71.4 ± 8.4	74.2 ± 10.4	75.9 ± 11.0
SWAV	6.51**	46.5 ± 8.8	60.1 ± 13.0	67.6 ± 14.4	69.3 ± 14.4	72.1 ± 14.5
DINO	6.73**	45.4 ± 11.8	58.4 ± 13.7	66.9 ± 15.7	69.1 ± 14.9	71.6 ± 14.9
SECO	6.83**	49.1 ± 12.4	55.9 ± 13.2	66.5 ± 15.5	66.6 ± 16.8	67.5 ± 19.0
SCRATCH	9.00**	36.7 ± 16.8	46.7 ± 13.2	49.9 ± 12.9	51.6 ± 16.4	54.6 ± 15.3
IMAGENET	9.03**	42.3 ± 12.8	48.7 ± 12.1	59.0 ± 15.2	59.9 ± 14.4	62.8 ± 15.8

We report averaged accuracies obtained on the seven DFC2020 regions. Each model is fine-tuned to the 5–7 classes of each DFC region individually, using an increasingly large support set of 1, 2, 5, 10, and 15 training examples per class, i.e., shots. It is then tested on a query set containing all remaining images. We report the average rank (lower is better) to compare models across all shots simultaneously. We further test for the significance of the differences to METEOR with a Wilcoxon Signed Rank test and indicate signficiant deviations by **.

Table 3 Quantitative comparison of METEOR with several state-of-the-art methods (rows) across different heterogeneous Earth observation datasets (columns).

	5-Shot problem	**Human influence**	**Crop type mapping**	**Land cover classification**		**Marine debris**	**Urban scenes**
	Dataset	**AnthPr.**[43]	**DENETHOR**[42]	**DFC2020-KR**[39]	**EuroSAT**[40]	**fl. obj.**[6]	**NWPU-Urban**[41]
	Spatial res.	**10 m**	**3 m**	**10 m**	**10 m**	**10 m**	**<1 m**
	Spectral res.	**10 bands**	**4 bands**	**13 bands**	**13 bands**	**12 bands**	**3 bands**
	No. of classes	**2**	**3**	**5**	**10**	**2**	**5**
	No. of training imgs	**10**	**15**	**25**	**50**	**10**	**25**
	Model	**Rank (↓)**	**Accuracy (↑)**				
METEOR	**3.6**	83.7	75.6	**87.7**	60.9	**90.8**	57.4
SWAV[36]	4.2	**96.7**	69.8	54.2	**67.7**	65.4	70.4
MOSAIKS[29]	4.3	86.4	**76.4**	82.3	57.9	88.8	54.0
DINO[37]	5.0	91.2	66.2	56.6	61.3	65.1	**70.6**
SECO[35]	4.7	91.4	61.7	67.6	62.7	65.9	67.4
SSLTRANSRS[16]	5.3	90.7	65.5	76.3	59.7	78.9	52.1
SSL4EO[34]	5.5	96.2	58.0	80.2	59.1	82.4	49.9
BASELINE	6.8*	89.0	60.8	87.4	39.8	69.8	36.7
PROTO[17]	8.3**	59.7	56.2	76.9	46.1	67.3	39.1
IMAGENET	8.8*	83.7	59.7	50.8	42.7	64.1	60.5
SCRATCH	9.5**	64.8	61.1	66.5	25.7	64.4	32.3

This heterogeneous setting is most challenging, as each evaluated task is characterized by a different number of spectral bands, number of classes, and spatial resolution. Here, METEOR achieves the best average rank of 3.6 but is closely followed by SWAV with 4.2 and MOSAIKS with 4.3 across the evaluated datasets. Different models are optimal for different tasks, and no model dominates all tasks. This is reflected in the Wilcoxon Signed Rank test that shows that the performance of METEOR is only significantly different (indicated by * and **) from the BASELINE, PROTO, IMAGENET, SCRATCH models. Best values are highlighted by bold face.

Crucially, pre-training configurations of model-agnostic meta-learning (MAML) that achieve high accuracies on the idealized Sen12MS target tasks are not optimal for the more realistic DFC2020 tasks with gaps up to 60% in accuracy, as shown in Experiment 1 in Table 1. This performance gap is related to normalization layers in the network architecture and has been first identified and discussed by Bronskill et al.[25]. They show empirically that the running calculation of batch statistics in transductive batch normalization (BN) layers (used in the original MAML implementation[23]) at test time allows the model to exploit knowledge about the class balance to improve its accuracy. In our experiments, we confirm that this exploitation allows a MAML-trained model with transductive BN to achieve the highest accuracy on idealized balanced tasks (85%) but results in the worse accuracy (26%) in the realistic DFC2020 tasks (underlined row in Table 1). To alleviate this issue, Bronskill et al.[25] proposed to replace batch normalization with their proposed tasknorm-I[25] normalization layers. However, we found that simply replacing batch normalization with instance normalization[32] performed best on the realistically imbalanced DFC2020 data by a large margin 20%, as shown in the top row of Table 1. In experiment #2, we found that this configuration of MAML with instance normalization (IN) also outperformed more recently proposed meta-learning variants, such as SparseMAML[24] (in both the first-order SparseFoMAML and Second-order SparseMAML variants) that achieved state-of-the-art performance on machine learning benchmark datasets. Concluding from these experiments, we use instance normalization in a ResNet-12 deep neural network as the meta-model in the remainder of this paper.

Comparison of METEOR to other state-of-the-art models. This section compares METEOR with other state-of-the-art approaches, modified for few-shot classification, either within homogeneous land cover classification tasks (Table 2) or across different heterogeneous tasks (Table 3).

We compare METEOR to self-supervised approaches pre-trained on multi-spectral satellite data (SSLTRANSRS[16] and SSL4EO[34]), on RGB satellite data (SeCo[35]), and on natural RGB images SWAV[36] and DINO[37]. As further comparisons, we train a BASELINE to classify all 10 classes present in the training areas of the Sen12MS dataset in a supervised way, and add a ResNet-50 model initialized on IMAGENET weights, and with random initialization named SCRATCH. For these approaches, we load the respective feature extractors with pre-trained weights, encode the few training samples in the respective feature spaces, average them to class prototypes and assign the test imagery to the class of the nearest prototype, as done in Prototypical Networks[17]. In Supplementary Notes 2, we show empirically that this strategy leads to better results than fine-tuning a linear classifier when considering few-shot problems with less than 50 examples per class. We also generate MOSAIKS[29] features dynamically for each downstream task from the training data and predict the test data with a random forest classifier. Models pre-trained with RGB data, i.e., SECO, SWAV, DINO, and IMAGENET, are only able to process the three RGB channels, while the remaining models (METEOR, SSLRS-R50, SSL4EO, MOSAIKS, BASELINE) have access to all 13 Sentinel-2 bands present in the DFC2020 data.

In terms of metrics, we report the averaged accuracy on the test images (query set) that models achieved after being trained/fine-tuned on the support set of each downstream task. To compare across different datasets and configurations, we further report the average rank to compare the different methods across all tasks and configurations. A model that outperforms its competitors on all tasks would have an average rank of 1 (first). We further test statistically if the differences in accuracies across different datasets are significantly different to METEOR with a two-sided Wilcoxon signed rank test[38]. We indicate significance levels “*" if $p < 0.1$ and “**” if $p < 0.05$. No star indicates the difference between classifiers is not significant.

Comparison within land cover problems. We first compare METEOR on land cover classification target problems from seven DFC2020 regions that use the same imaging sensor (13-bands, Sentinel-2) and classes with similar semantics as the pre-training tasks (Sen12MS dataset) in Table 2.

Overall, METEOR compares well to their multi-spectral methods with an average rank of 2.84. Only SSL4EO[11] achieves a slightly better average rank of 2.51. Also, the supervised BASELINE achieves high few-shot accuracies with an average rank of 2.99, as the data and classes that this model was trained on (Sen12MS) align well with the DFC2020 data. MOSAIKS, SSL4EO, and the BASELINE are not significantly different from METEOR in accuracy, as shown by the Wilcoxon signed rank test[38], while METEOR was significantly better on these problems than SSLTRANSRS and all contrastive RGB approaches using only RGB imagers (SWAV, DINO, SECO) with p values < 0.05, as indicated by “**".

Comparison across diverse heterogeneous problems. In Table 3, we again compare METEOR to other approaches, but this time in heterogeneous tasks beyond land cover classification. The tasks involve Earth observation sensor data of different spatial and spectral resolutions, as indicated in the top rows. They cover common disciplines, such as land cover classification (DFC2020[39], EuroSAT[40]) and classification of urban scenes from the NWPU-RESISC[41] dataset, as well as more specific problems, such as a mono-temporal classification of crop types (DENETHOR[42]). Two niche applications are also studied to highlight even more the versatility of a single METEOR meta-model to address the diversity of Earth observation problems: AnthroProtect[43], which estimates the presence of human influence by classifying images of naturally protected areas from unprotected ones, and marine debris[6], which classifies the presence of floating objects, such as marine litter, in ocean scenes. The bottom row of Table 3 shows image examples of each of these downstream tasks. Further qualitative results from these problem fields are provided in the next results section. Analogously to the previous comparison, we use the average rank as the primary metric to compare the different methods across all tasks and assess whether the performance is significantly better than METEOR with the Wilcoxon signed rank test. All experiments in Table 3 are reported for 5-shots classification.

Here, METEOR compares well to the other approaches on these heterogeneous 5-shot problems and achieves the lowest average rank of 3.6. Still, it only achieves the best accuracy on two datasets (DFC2020-Kippa-Ring (KR) and floating objects[6]), while being among the best models for the other datasets. The main exception is NWPU, where DINO[37] andSWAV[36] achieved the best results. We hypothesize that the fine-grained features learned during contrastively pre-training from natural images in these approaches are here particularly helpful for very high-resolution urban scene imagery.

Interestingly, only METEOR and MOSAIKS[29] are among the best solutions for both few-shot problems within the land cover field (Table 2) and across heterogeneous problems (Table 3). For instance, the ResNet-50 trained contrastively with momentum contrast[44] from SSL4EO[34] only achieved rank 5.5 on the heterogeneous problems tested in this work. This is most prominently shown by the supervised BASELINE, trained on land cover Sen12MS data, which is competitive within land cover downstream tasks, but among the worst model in the heterogeneous tasks of Table 3.

Due to the high variance in accuracies across these diverse problems, only few comparison models (BASELINE, prototypical networks (PROTO), IMAGENET pre-training and a randomly initialized model SCRATCH) can be considered significantly worse than METEOR with the Wilcoxon signed rank test. Still, we can conclude from this experiment that METEOR compares well to the state-of-the-art in pre-training models on diverse few-shot Earth observation problems and can be applied to diverse heterogeneous problems successfully and that only METEOR and MOSAIKS achieve such consistency across tasks. After these quantitative comparisons, we explore METEOR prediction qualitatively on the diverse problems and provide some insight using interpretability methods. The iso-lines in the bottom panels show occlusion sensitivity. These indicate a decrease in prediction probability (in percentage) if this particular area is occluded.

Interpreting METEOR’s predictions across problems. This section studies the behavior of METEOR qualitatively in various heterogeneous environmental problems. The tasks differ in spectral and spatial resolution, demonstrating the broad usefulness of METEOR. The application of METEOR is as follows: a task-specific model (indicated in dark gray in Fig. 1, as well as in all figures in this section) is initialized in each task with the parameters of the same pre-trained meta-model (drawn in red in the figures). The task-specific model is then fine-tuned on a few training samples from the support set of the task under study. The following sections and figures illustrate the predictions quantitatively and qualitatively, and use explainable machine

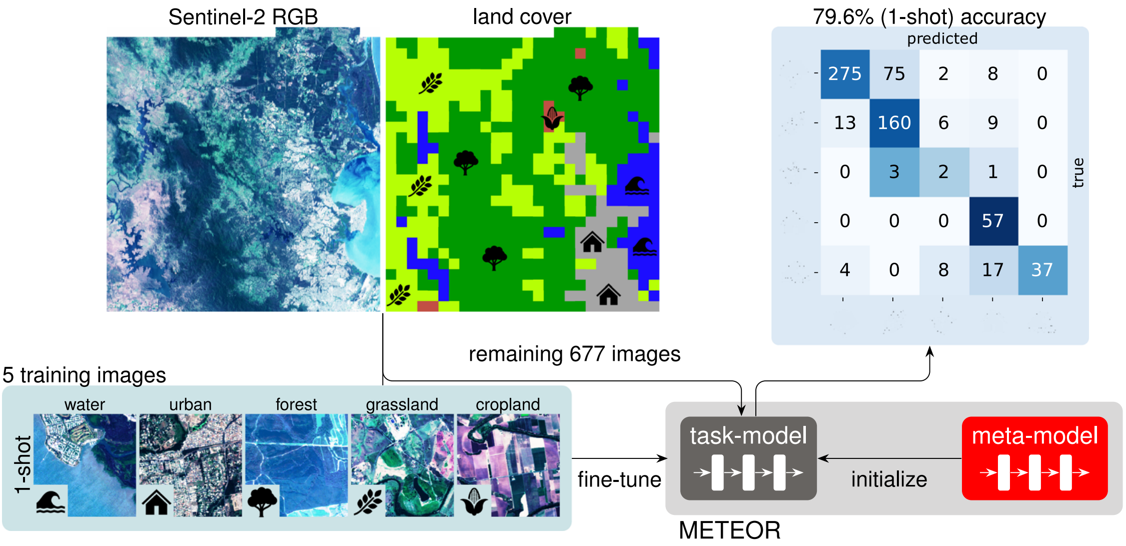

Fig. 2 Land cover classification in Kippa-Ring, Australia (one among the DFC2020 regions), when using one example per class (1-shot), so five examples in total. We show the confusion matrix of the predictions obtained on the 677 remaining images, which have been classified at a 79.6% accuracy in the first (of three) random splits. In 1-shot learning, the choice of training images is especially important, as the representation of classes are solely defined by these single training examples. The second random split is only slightly worse with 78.4%, while the third split is only classified with 47% accuracy. In that last case, several forest images were wrongly classified as grassland (not shown in the figure). To accommodate for this randomness, we average the accuracy of all three random splits leading to a 1-shot accuracy of 68%. The variance between splits decreases with more shots, as can be seen on the quantitative table in Supplementary Table 2.

learning[45,46] interpretations of the predictions combined with domain knowledge from the respective task to explain them.

In the first example in Fig. 2, the meta-model is fine-tuned on five land cover classes with one example per class (1-shot setting) in Kippa-Ring near Brisbane, Australia, which is one of the seven DFC2020 regions. This leads to a task-model specialized in predicting land cover in Australia defined by the training classes. meta-model and fine-tuned with five training images. Each image defines the representations of the land cover classes in this downstream task, as shown at the bottom of the figure. The task-model then classifies the remaining 677 images from this geographic region, resulting in an average accuracy of 68%. This accuracy is achieved with only five training images (averaged over three runs with different train/test splits). With 40 training images, i.e., in a ten-shot setting, 88% of the remaining images are classified correctly, as shown quantitatively in Supplementary Table 2.

Second, we consider the monitoring of deforestation, which is a vital application to estimate drivers for climate change. We acquired PlanetScope imagery[47] with four spectral bands (RGB +NIR) at 3-m resolution between 2017 and 2022 from the Roraima state in northern Brazil (Fig. 3a). In this region, the clearing of tropical forests between 2017 and 2022 is visible. These scenes show the last 10 km of one of the multiple orthogonal access roads to the BR-174 highway, which provides infrastructure for deforestation in this region[48]. As a specific task, we distinguish between forest and no-forest classes and train the task-model with ten images from a deforested region (0°45’50"N 60°39’02"W) between 2017 and 2022 that is located north of the test scenes shown in Fig. 3 (top row). Once fine-tuned on these ten training images, the task-model then estimates a posterior probability for forest and no-forest in the test scenes by classifying 32 × 32 px image patches sampled on a regular grid. This results in a coarse segmentation map, as shown in Fig. 3c, at 96 × 96 m resolution. The estimated probability maps closely reflect the deforested areas visible in the satellite images.

Third, we address urban scene classifications from high-resolution aerial or satellite RGB imagery with less than 1-m ground sampling distance. We select the 3500 images pertaining to five urban classes (industrial, commercial, and dense, medium, sparse residential areas) from the NWPU-RESISC45[41] dataset. The meta-model of METEOR is fine-tuned to a task-model with 25 (five-shot) training images and achieves 65% accuracy on the remaining 3475 images, as shown in the last column of Table 3. Figure 4 presents one image of the test data, where the task-model has estimated high classification scores of multiple classes. The classification result is analyzed by an occlusion sensitivity analysis[46], which reveals that the irregularly spaced houses are responsible for the high classification score of commercial areas. At the same time, the regular rows of residential buildings are recognized as medium residential. Similarly, we can deduce that the 15% industrial score is caused by a single white building with roof installations, which are structurally similar to some buildings visible in the training set of the industrial class. This result on an ambiguous example shows that, thanks to task tuning, METEOR learns relevant features per each class: it correctly divided the prediction scores across classes thanks to the one-against-all learning.

Fourth is change detection, which can be realized by repeated image classification of the same area at different dates. We use METEOR on images of an explosion event in Beirut, Lebanon, on August 4, 2020, where a warehouse storing ammonium nitrate exploded after an initial fire. We acquired a sequence of 70 cloud-free Sentinel-2 images of Beirut spanning from September 3, 2019, until March 21, 2021 (Fig. 5a). Two classes, pre-event and post-event, are defined by taking the first and last five images from the sequence as training samples, covering periods between September 3 and 28, 2019, and February 19 and March 21, 2021, respectively. The meta-model is fine-tuned on these ten images to estimate probability scores for the two classes for each image in the sequence. As visible in the bottom plot of Fig. 5, the probability score for the class post-event remains low for all images before the explosion event on August 4, 2020. It increases sharply to 84.5% on the first image after the event on August 8 and remains high for the remaining images. An occlusion sensitivity analysis shows that this increase in score for the

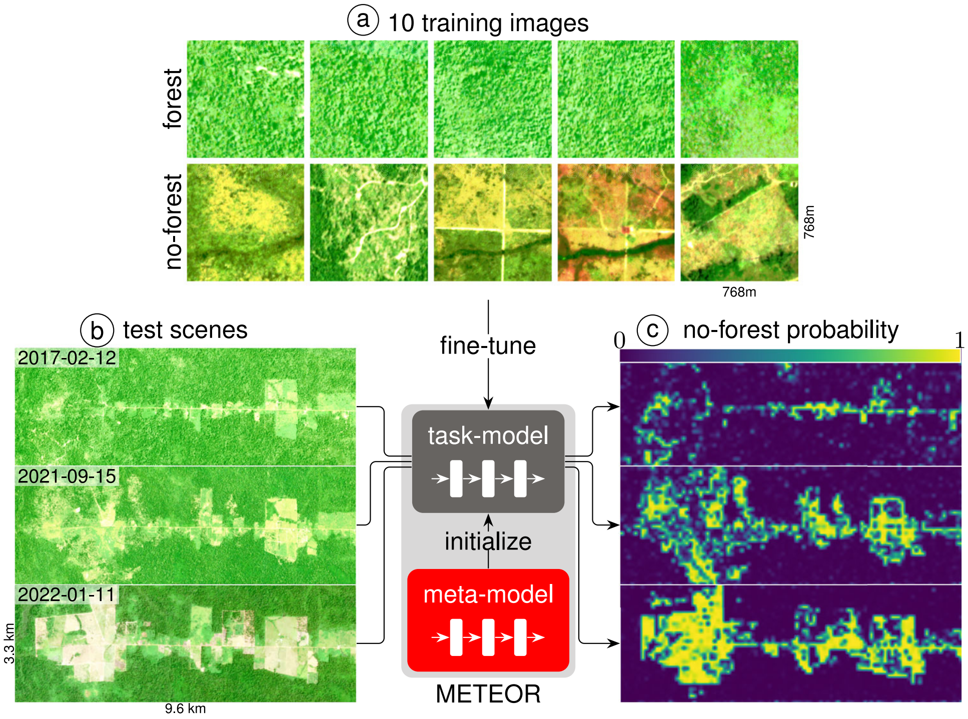

Fig. 3 Classification of deforested areas in the northern Brazilian Amazon forest. The task-model is initialized with the parameters of the meta-model and fine-tuned to classify forested from deforested areas in the Amazon rain forest (0°41'05"N 60°37'42"W) shown in (**a**) with 4-band PlanetScope imagery at 3-m ground resolution in (**b**). The ten training images, shown in the top panel, are taken from different areas (0°45'50"N 60°39'02"W) and were acquired on multiple dates between 2017 and 2022. **b** Three test scenes 9.6 × 3.3 km from 2017, 2021, and 2022, where deforestation is visible. We split these test scenes into 32 × 32 px tiles and predict the probability for the forest and no-forest classes to each tile. This tiling results in maps of deforestation at 96 × 96 m resolution in (**c**).

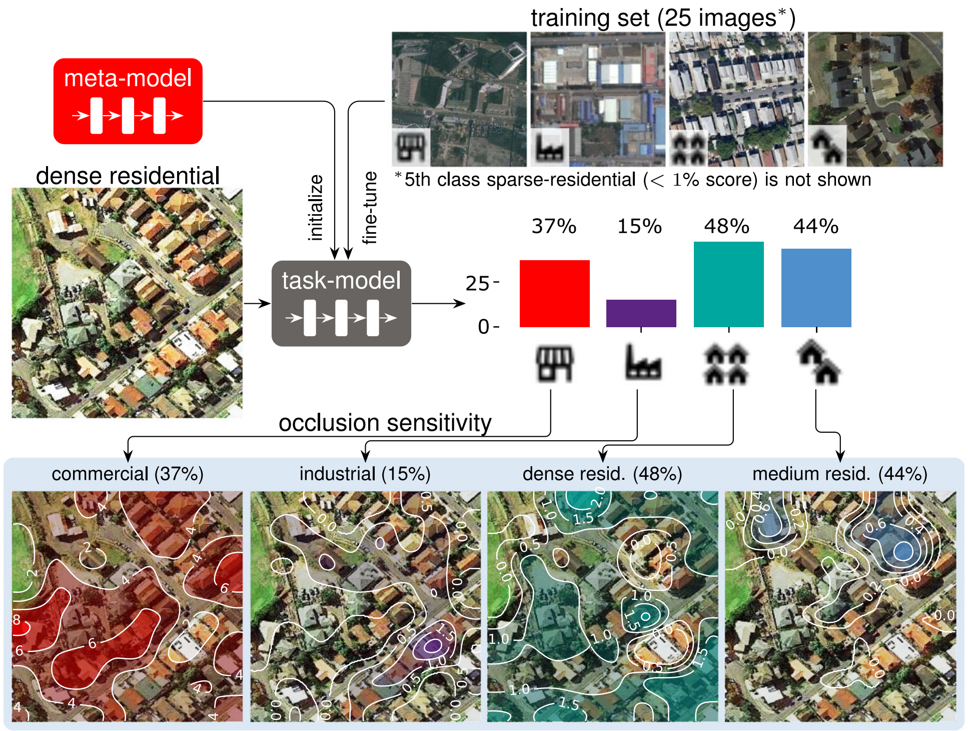

Fig. 4 Urban scene classification with high-resolution RGB imagery. High-resolution (less than 1 m) RGB satellite imagery is employed in this example, where we analyze the prediction of a challenging image where the model assigned multiple categories with a high classification probability. An occlusion sensitivity analysis[46] shows an irregular structure of houses, is partly recognized as a commercial area, as masking this part of the image decreases the score of class "commercial" by 6%. Regularly spaced houses are visible that the model associates with medium residential. A single flat-top building with roof installations causes the 15% probability for the industrial class. Note that similar white structures are present in the training set of the industrial class.

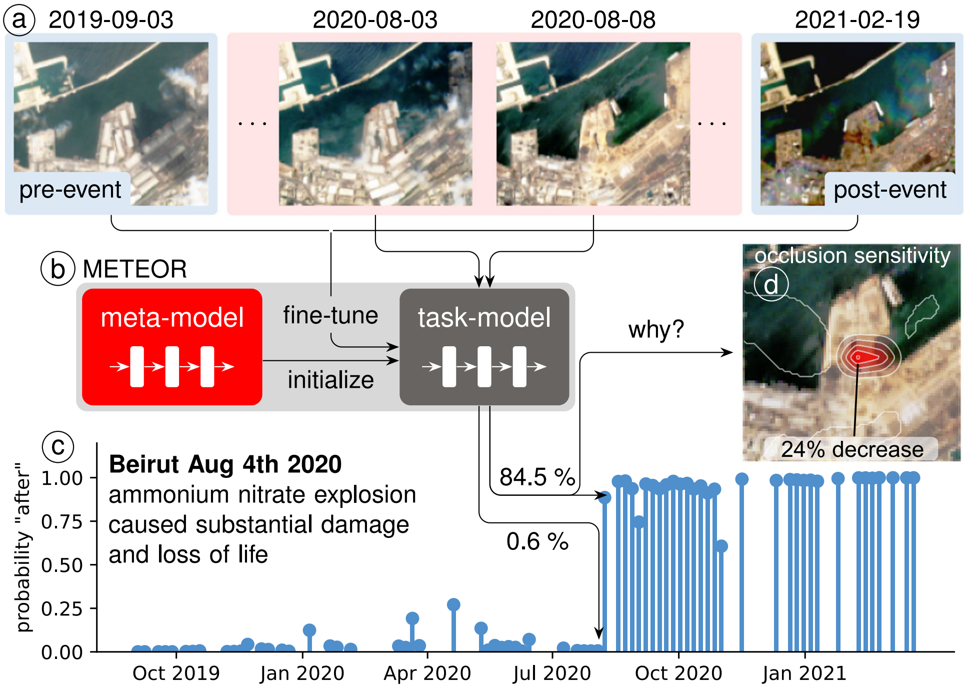

Fig. 5 Detection of changes with a sequence of multi-spectral medium-resolution imagery. This use-case shows the port of Beirut, Lebanon, where an explosion event caused substantial damage on August 4, 2020. **a** A total of 70 Sentinel-2 images where we use the first and last five images as support set to define the classes pre-event and post-event. task-model, which is fine-tuned on these examples. The meta-model in METEOR, shown in (**b**), is fine-tuned on these two classes and predicts all remaining images in the sequence. In (**c**), we show the resulting probability for the post-event class, where it remains low (0.6%) until August 3, 2020, and sharply rises probability of 84.5% on the following image of August 8, 2020. The crater and damaged buildings from the event caused the sudden increase in this probability score, as revealed by the occlusion sensitivity analysis drawn in (**d**). A further comparison to MOSAIKS and SSLTRANSRS is placed in Supplementary Fig. 5.

post-event class is predominantly caused by the explosion crater and the damages to adjacent buildings, as occluding these areas of the image leads to a decrease of the post-event score of up to 24%.

Fifth and last, we explore semantic segmentation of marine debris in satellite imagery, which is a vital requirement for quantifying marine litter on the world's oceans. We focus on coastal regions close to major rivers deltas and where notable plastic accumulation events were reported in the news[6]. The ResNet-12 task-model, used for image classification, can be modified for coarse semantic segmentation without changing the weights of the underlying model, as outlined in the Methods section. Thanks to this modification, it can now predict a score for each pixel in the image. Despite this modification, the task-model can still be initialized from the meta-model of METEOR. We fine-tune it on five examples of marine debris on RGB Sentinel-2 imagery from the coastal region of Accra, Ghana, from a dataset provided by Mifdal et al.[6]. These training images are shown in Fig. 6 alongside hand-annotated masks that serve as prediction targets. The bottom part of Fig. 6 presents one test image showing patches of liquid contaminants alongside the estimated probability map for the presence of marine debris. We provide a contour overlay of this probability map on the RGB image alongside the annotations for this image as a reference. The predictions show that the model has captured the nature of the floating object detection task and accurately predicted the shapes of floating marine litter using five training images only.

In summary, we demonstrate that with METEOR domain experts can fine-tune a single deep neural network on various downstream problems with few labeled examples. These downstream tasks can differ substantially from the land cover classification source tasks that the meta-model was pre-trained on in terms of problem scope, spatial, and spectral resolutions. The downstream problems can be diverse and span from land cover classification from different geographic regions over urban scene classification with high-resolution RGB imagery to segmenting marine debris at the sea surface. Few example images are needed to represent the downstream task classes, allowing us to define abstract categories, such as pre-event and post-event. This abstraction is also meaningful for other use-cases, such as land cover classification, where the representation of one land cover, e.g., cropland, can vary greatly from continent to continent. Crucially, leveraging a few labeled samples makes METEOR applicable to various use-cases that Earth scientists address every day.

Discussion and conclusion

This paper presents METEOR, a transfer learning methodology based on model-agnostic meta-learning[23] in which problem-specific neural networks are learned from a global (meta-)model using only few labeled examples describing such new problem. This is enabled by three simple-to-implement but important modifications (1) replacing transductive batch normalization with instance norm in the meta-model, (2) ensembling multiple binary classifiers to address problems with a varying number of classes, and (3) dynamically changing the input channels to account for problems with a different number of spectral bands. Thanks to these modifications, the generic meta-model, originally pre-trained to distinguish different land cover classes, can then address a diverse set of heterogeneous remote sensing tasks that vary in spatial and spectral resolution, as well as in the number of classes. Such as a model that can be trained on a collection of different-but-related tasks, and this can help in a variety of Earth science disciplines, where annotating samples is particularly difficult or costly. In this regard, we have shown quantitative and

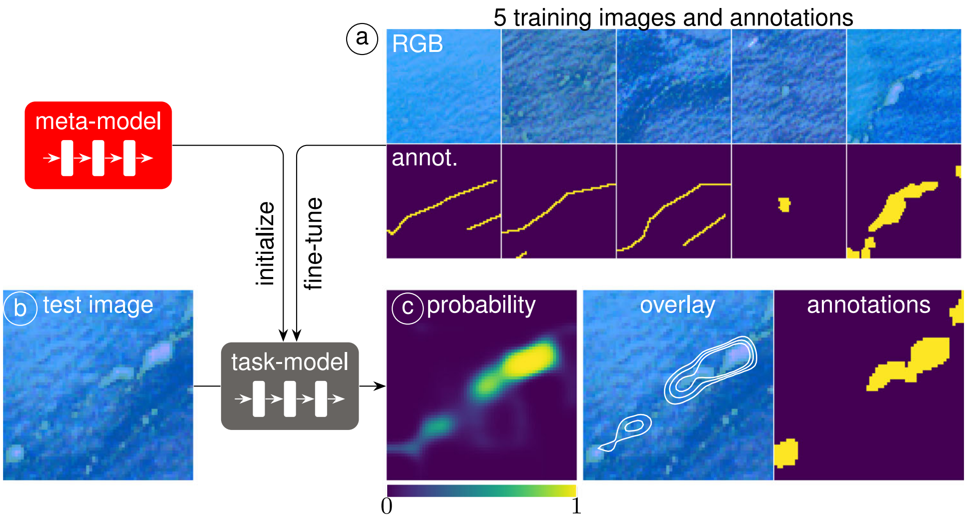

Fig. 6 Segmentation of marine debris with multi-spectral medium-resolution imagery. Marine debris on a 12-band, medium-resolution (10 m) Sentinel-2 scene near Accra, Ghana (October 31, 2018) have been visually detected and annotated by Mifdal et al.[6] in (**a**). This task-model is modified for segmentation and initialized with the weights of the meta-model. Five annotated images of (**a**) serve as a training dataset to fine-tune the task-model. **b** A test image and the corresponding prediction in (**c**) by the task-model to a map showing the probability of marine debris for each pixel. The same meta-model, as in the previous use-cases, achieves these predictions, highlighting the versatility of meta-learning for various Earth observation problems with few training samples.

qualitative experiments where a single pre-trained model can be of use across a variety of diverse Earth science disciplines from urban applications to deforestation mapping.

Our work extends related work based on meta-learning[19,27,28], self-supervised[16,35,49] or self-taught learning[29] from the homogeneous transfer learning setting (where all tasks share the same problem nature and input space dimensions) to a heterogeneous setting involving different satellite sensors and applications. In particular, this transfer across different resolutions has been identified as a major challenge toward a collective agenda on artificial intelligence for Earth science data analysis[50].

In contrast to the results on idealized Sen12MS test tasks (see Table 1; Sen12MS-column), which are similar to those usually obtained in meta-learning benchmarks, we found that using regular instance normalization in the deep meta-model neural network outperformed recently proposed meta-learning variants (e.g. SparseMAML[24], TaskNorm-I[25]) on realistic and class-imbalanced remote sensing problems. This finding enabled us to explore and compare METEOR with related work on tasks within the problem field of land cover classification (Table 2) and across challenging heterogeneous remote sensing problems (Table 3); in all comparisons, METEOR performed very well for few-shot classification methods, achieving the best average rank on heterogeneous problems and the second best within land cover problems. We studied in depth the behavior of METEOR in a series of potential use-cases, from change detection to marine debris segmentation, where domain experts can deploy METEOR to extract actionable information and insights into specific problems.

While this work presents a step toward deploying a learning-from-tasks framework on real-world Earth observation applications, it revealed some limitations that future research will need to address. First, implementing the task-model as an ensemble of one-versus-all classifiers (described in the Methods section) leads to poor performance when the number of classes is large. While predicting ten classes, as in the EuroSAT benchmark[40] was still accurate, performance was sub-par in the case of 45 classes as in the full NWPU-RESIC45[41]. Additionally, while fine-tuning the meta-model on a few annotated samples is computationally efficient, learning it in the first place with model-agnostic meta-learning is more memory expensive than pre-training with self-supervised algorithms. Training larger convolutional neural networks like ResNet-50 (23 million parameters) or ResNet-152 (60 million parameters) is considered out-of-reach for current meta-learning approaches.

Another point is the nature of the pre-training task. In this work, we limited the pre-training of the METEOR meta-model on the source tasks from land cover classification to demonstrate the versatility toward radically different target problems. Still, we believe that expanding the pre-training scope toward other labeled and unlabeled source tasks will likely further improve its performance on downstream Earth observation problems. Our work is valuable to several groups of domain experts, users, and scientists. First, the METEOR model can support researchers who aim to deploy state-of-the-art deep learning models in their particular field of expertise. METEOR is usable with the limited labeled data often available from specialized field campaigns, which often provide high-quality samples but are scarce in quantity. Moreover, we release METEOR as an open-source, simple, and ready-to-use package in Python. Second, we believe that our experiments across various remote sensing tasks can serve as a benchmark to compare future machine learning algorithms and measure the performance in a broader set of meaningful applications. In general, the results of this work indicate, along with other recent advances in meta-learning and self-supervised learning, that deep learning models trained on a learning-from-tasks framework can be employed for a versatile family of downstream problems. These models can provide increasingly intelligent solutions that, when deployed on impactful Earth science problems, help address some of the most pressing issues of our time.

Methods

General description of METEOR. The meta-learning methodology for Earth observation problems across different resolutions (METEOR) is a heterogeneous transfer learning approach. It is designed to capture knowledge across data modalities (different Earth observation sensor types) and efficiently adapt to different tasks (variable in terms of task semantics, number of classes, etc.). To do so, METEOR consists of a meta-model and a task-model (Fig. 1). The meta-model encodes knowledge from Earth observation source tasks and is pre-trained with the model-agnostic meta-learning algorithm[23] (detailed later) on a dataset of source tasks. The task-model is initialized with the parameters of the meta-model and fine-tuned on the particular problem with a few training samples describing the target task at hand. This framework falls into the family of model-based transfer learning[11,51] where knowledge from source tasks is encoded in model parameters to inform a target task. Other approaches from this category are self-supervised learning and pre-training, e.g., on ImageNet[52].

Meta-model implementation and pre-training on source tasks. The meta-model is a deep ResNet-12[53] neural network following the implementation of Oh et al.[26]. All batch normalization[31] layers are replaced with instance normalization[54] (see Table 1). The model used in all experiments has 15 input channels to accommodate the two radar bands of Sentinel-1 and the 13 spectral channels of the Sentinel-2 satellite and a single output dimension for binary one-versus-all classifications. We chose 15 channels to utilize all bands in the Sen12MS dataset as we observed no negative effect of including the two additional radar channels. Including these bands makes our pre-trained weights of METEOR applicable to radar downstream tasks, even though we did not show experiments based on radar data in this work. From the 13 optical channels, various satellite sensors, such as PlanetScope or Worldview, can be used for downstream tasks, as we demonstrate in the experiments. Pre-training METEOR with other channels is possible, as we did during development with only Sentinel-2 or RGB bands variants. the meta-model architecture. This meta-model is pre-trained on tasks of 16 images with four randomly selected land use and land cover classes, each task from one geographic area in the Sen12MS dataset[5]. We split the task images into a train and test partition with eight images each. This configuration corresponds to a 2-shot 4-way classification setting with two images per class. Note that we modify this multi-class task to binary one-versus-all classification by selecting one class randomly as a target. The task`s training objective is to learn the representation of this selected class from the two images against the other six negative examples (containing two examples of the three other classes).

Model-agnostic meta-learning. To obtain the meta-model, we use the model-agnostic meta-learning (MAML)[23] algorithm that optimizes the following objective:

$$\underbrace{\min_{\theta} \mathbb{E}_{\tau\sim p(\tau)}\left[L_{\tau}^{\text{test}}\left(\phi_{\tau,K}\left(\theta\right)\right)\right]}_{\text{outer loop/meta-learning}} \tag{1}$$

$$s.t.\ \underbrace{\phi_{\tau,k+1} \leftarrow \phi_{\tau,k} - \alpha\nabla L_{\tau}^{\text{train}}}_{\text{inner loop/fine-tuning}} \text{ and } \underbrace{\phi_{\tau,0} = \theta}_{\text{initialization}} \tag{2}$$

where a task-model ϕ_τ is initialized by the meta-model θ and iteratively fine-tuned with $k \le K$ steps based on gradients from a loss of training samples ∇L^{train} in an inner loop. The constant α denotes the inner learning rate. In the outer loop, the meta-model parameters θ are updated by minimizing the test loss L_{τ}^{test} over a batch of tasks $\mathbb{E}_{\tau\sim p(\tau)}$ with the fine-tuned parameters $\phi_{\tau,K}$. These fine-tuned parameters are a function of the initialization θ. This makes updating the meta-model parameters with second-order gradients (outer gradients through the inner loop gradients) possible. Over several thousand iterations, this yields a meta-model that is explicitly learned to learn differences between land cover categories from different geographic areas. In this work, we pre-trained the meta-model with the standard second-order MAML algorithm[23], as it achieved best results on the realistic use-cases in comparison to variants, such as SparseMAML[24] or tasknorm-I[25], as shown in Table 1.

Task-model implementation and fine-tuning on downstream tasks. The METEOR task-model tuning has three requirements: (i) it must have the same weight dimensions as the meta-model so that it can be initialized with the weights of the meta-model; (ii) it must consider a problem with more than one class; and (iii) it must apply to downstream tasks where the input channels involved are a subset of those of the meta-model.

- We fulfill the first and second requirements by implementing the task-model as an ensemble of multiple one-versus-all classifiers, each initialized with the parameters of the meta-model (first requirement), as shown in Supplementary Fig. 1b). Each classifier is responsible for predicting a single class when multiple classes are present in the task (second requirement). When fine-tuning this task-model on a downstream task with n classes, each classifier minimizes a binary cross-entropy loss with stochastic gradient descent concerning one positive class and $n-1$ negative classes. Given a new input sample for prediction (second row of Supplementary Fig. 1b), each ensemble member in the task-model predicts a score associated with its respective class. Classification probabilities for each class can be retrieved by either sigmoid-normalizing the prediction scores of each member separately or by combining the prediction scores by softmax. We used softmax normalization for all experiments except for the qualitative analysis in Fig. 4. Here, we obtained better occlusion sensitivity maps with sigmoid normalization, as the presence of one class did not influence the prediction of another.
- To address the third requirement, we select a subset of the learned convolutional filter banks in the first convolution in the input block of the ResNet-12 neural network, as shown schematically in Supplementary Fig. 1c. The meta-model's first layer normally convolves a 15-dimensional input image with 15 convolutional filter banks, as shown in the top row. When a downstream task provides data with, for instance, three RGB spectral bands (bottom row), we only copy the filter banks responsible for the RGB channels from the meta-model to initialize the task-model. This transfer is meaningful as long as the requested spectral bands form a subset of the spectral bands of the meta-model. In terms of downstream tasks with images of different spatial resolutions, no model modifications are necessary, as ResNets ingest images of different sizes natively.

Task-model modification for segmentation. Some downstream tasks, such as marine debris segmentation (Fig. 6), require the model to output segmentation maps rather than a single classification probability. For pixel-wise segmentation, the task-model needs to be adapted structurally, as shown in Supplementary Fig. 1a). We modify the network without changing the weight dimensions by removing the global average pooling in the penultimate layer and replacing the final linear layer with 1 × 1 convolutions. These convolutions are equivalent to linear layers applied to each feature-pixel separately and, thus, use identical weight dimensions. This modification yields a 9 px × 9 px segmentation map for a 64 px × 64 px image, which is then upscaled via bicubic interpolation to the original resolution.

Training details. During pre-training of the meta-model, we train in iterations containing batches of 16 tasks and aggregate metrics over cycles of 200 training iterations. We employ Adam[55] as an outer optimizer and set its learning rate to 0.001, which is further decreased by a factor of 0.1 if the validation loss has not decreased over 20 cycles. The training is stopped when the validation loss did not decrease over 40 cycles or at 40000 training iterations. The gradient update is done with standard second-order MAML in a single gradient step. The inner learning rate (i.e., step size α) is set 0.32 experimentally and the model weights are updated with stochastic gradient descent. We experimented with two gradient steps at the cost of a smaller batch size but found training on larger batches with a single gradient step to yield better performances. Fine-tuning on downstream tasks is realized through regular stochastic gradient descent with a step size between 0.32 and 0.4 and 20–60 gradient steps. We find that a comparatively wide range of step sizes and the number of gradient steps led to similar solutions in the classification experiments. The loss function is binary cross-entropy for both pre-training and fine-tuning on all downstream tasks. For the qualitative segmentation experiment in Fig. 6, we fine-tuned METEOR with a cross-entropy objective on each pixel. For the change detection experiment in Fig. 5, we define two classes, “pre-event" and “post-event", and similarly use cross-entropy.

The pre-training of the meta-model was performed on two NVIDIA V100 GPUs within a computational SLURM cluster within 48 h. The estimated carbon footprint for pre-training one meta-model was 5 kg/eCO$_2$. Fine-tuning and prediction on the seven different DFC2020 regions took 4 min on a NVIDIA GeForce RTX 3090. The fine-tuning and prediction of 1-shot marine debris images took 10 s on a workstation with 32 CPU cores and 120GB RAM and 2 min 10 s on a MacBook Pro (2020) with Apple M1 CPU and 16GB RAM.

Comparison models. We compared METEOR across applications with prototypical networks and self-supervised contrastive learning algorithms in Tables 2 and 3.

Prototypical Networks (ProtoNets)[17] are a metric-based few-shot learning approach where a deep neural network (ProtoNet) is trained with the identical source tasks, as METEOR. In ProtoNets, a deep feature extractor maps all images of one task into a common feature space. The feature representations of the training images are then averaged into prototype vectors. The test images are associated with the class of the nearest prototype in Euclidean distance. The feature extractor is iteratively optimized via gradient descent to minimize the classification error of the test images over a batch of few-shot tasks. We implemented the prototypical network with a ResNet-18 model and trained it with a learning rate of 0.001 until the convergence of the validation error in the Sen12MS dataset.

Self-supervised contrastive learning provides an alternative way to obtain feature extractor representations. In contrast to prototypical networks, the optimization objective is to minimize the error on a hand-defined pretext task designed not to require annotations and to mimic some characteristic of interest of the data we want the model to be robust against. Usually, these methods are adapted to new downstream problems by freezing the feature extractor and fine-tuning a relatively small classifier network that can be a single linear layer or a multi-layer perceptron. However, we found in initial experiments (in Supplementary Fig. 2), where we compared different adaptation strategies, that this fine-tuning with a dedicated classifier network does not lead to accurate results for few-shot problems with less than 50 shots. We, therefore, decided to follow the strategy of prototypical networks and perform nearest neighbor classification with Euclidean distance in the feature space directly to obtain the results in Tables 2 and 3.

Specifically, we compared to the MoCo-trained ResNet-50 from SSL4EO[34] and the ResNet-50 from SSLTRANSRS[16] which were pre-trained on full multi-spectral data. In Supplementary Table 1, we compared to pre-trained variants provided in these code bases and choose models that achieved the best accuracy on the diverse downstream problems of Table 3. Furthermore, we compared METEOR to seasonal contrast (SeCo)[35], which uses Momentum Contrast v2 (MoCo v2)[44,56] on RGB representations of unlabeled Sentinel-2 images of the same scene at different dates, therefore learning robustness to image seasonality. We further use the weights of SWAV[36] and DINO[37], which are pre-trained on natural images and equally employ a ResNet-50. We also compare to a supervised BASELINE trained in a supervised way on ten classes of the Sen12MS dataset. We used a learning rate of 10^{-3} and a weight decay of 10^{-6} and take the model that achieved the best validation accuracy over 50 epochs. We trained a ResNet-12 (same architecture of METEOR), ResNet-18, and ResNet-50, and use the ResNet-18 in the comparison of Tables 2 and 3, as it achieves the best results across the different backbones (Supplementary Table 1). MOSAIKS[29] proposed a featurization strategy based on autocorrelation features from Random Kitchen Sinks[57] extracted by a small neural network[30]. As the most applicable featurization strategy of MOSAIKS is ambiguous, we test different configurations (Gaussian and Laplacian random features, empirical global features, and empirical local features and local features-supervised), as described further in the Supplementary Notes 2. We compared all variants to METEOR in Supplementary Table 1 where the “empirical local features" strategy led to the best accuracies across all problems which we use as MOSAIKS implementation in Tables 2 and 3.

Datasets. Nine datasets have been used throughout the experiments.

The Sentinel-12 Multi-Spectral (Sen12MS)[5] dataset is used for pre-training the meta-model. It contains Sentinel-1 and Sentinel-2 images with associated land cover labels in a coarse segmentation map in 125 globally distributed geographic regions. We use Sen12MS for classification by associating the image with the majority class observed in the patch[39]. The original dataset contains overlapping images of 256 px × 256 px. Following prior work[28], we remove the overlap in the images, which yields images of 128 px × 128 px in size. Nine different land use and land cover categories are present in this dataset that follow a simplified[5,39] International Geosphere Biosphere Program (IGBP)[58] classification scheme. These classes contain the general land cover categories forests, shrubland, savanna, grassland, wetlands, croplands, urban/built-up, snow/ice, barren, water. these classes throughout the globe is substantially different from each other. We split the data into distinct geographical regions for training (75), validation (25), and test (25) to prevent geographical autocorrelation and the potential positive biases of the training set leaking into the test regions. The test regions are used after training to evaluate final accuracy on Sen12MS tasks, as reported in Table 1. The meta-model is trained on tasks from the 75 training regions, while tasks from the validation regions are used for parameter tuning and early stopping of the pre-training process.

The public Data Fusion Contest 2020 (DFC2020) dataset[33] was designed to mirror Sen12MS with the same IGBP labels on seven different geographic regions. The annotations were semi-automatically refined and contain less label noise compared to Sen12MS, which makes these regions most suitable for qualitative and quantitative evaluation. Each DFC2020 region is partitioned into non-overlapping tiles of 256 px × 256 px and segmentation labels are provided alongside the optical Sentinel-2 and radar Sentinel-1 images. As for the previous dataset, we select the most frequent land cover in the segmentation map as a classification label for each tile following Schmitt and Wu[39]. The accuracy of all seven regions is used in Tables 1 and 2. We selected the Kippa-Ring region for the qualitative result in Fig. 2 and also compared it with other tasks in Table 3. In general, METEOR can achieve high performance on land cover classification problems (as in DFC2020 or Sen12MS), while accommodating the regional differences in the representation of land covers, as croplands or forests, which show a high intra-class variability, both spectrally and geographically. Concretely, this can increase the accuracy of land cover classification in general by fine-tuning many region-specific classifiers.

The EuroSAT benchmark dataset contains multi-spectral Sentinel-2 images of 64 px × 64 px with 13 spectral bands. It features nine land use and land cover classes: annual crop, herbaceous vegetation, industrial, permanent crop, river, forest, highway, pasture, residential, sea lake. The dataset is artificially balanced and contains 2500–3000 images per class. In this case, METEOR can learn the specific representation of land cover for southern Germany.

The NWPU-RESISC45 benchmark[41] is a broadly used benchmark dataset that contains RGB images of 256 px × 256 px size at different resolutions of 45 diverse classes. disciplines. Each class is represented by 700 images. To build an urban scene classification problem, we specifically select the classes commercial, residential, dense-, medium-, and sparse residential to create a downstream task for urban scene classification. We compare models on this dataset in Table 3 and show results qualitatively in Fig. 4. We use this dataset to test whether METEOR can learn more fine-grained distinctions of different urban scenes, even if it was pre-trained only on general land cover classes from Sen12MS.

The Floating Marine Objects dataset[6] contains Sentinel-2 images with hand-annotated labels of 26 coastal regions across the globe. Marine debris exhibits a strong heterogeneity[6]; hence, learning a single representation for marine debris is difficult due to different compositions of materials, different water transparency due to sediments, and different atmospheric conditions that vary between satellite scenes. Automated atmospheric correction cannot completely remove these latter effects[7]. We select images from the coastal region near Accra, Ghana, where liquid pollutants were visually detected and annotated on a Sentinel-2 scene on October 31, 2018. We use this dataset for segmentation in Fig. 6. In Table 3, we use this data in a binary classification setting where images of floating objects are assigned a positive class, and randomly sampled images from the entire Sentinel-2 scene are used as negatives. We use this dataset to test the limits of METEOR in distinguishing details within water scenes, even though the meta-model was pre-trained on generic Sen12MS land cover classes. Approaching this application field with meta-learning is particularly promising, as hardly any label data exists for marine debris detection. Moreover, the heterogeneity of the types of marine debris further complicates training a single dataset. Hence, METEOR can be fine-tuned in multiple debris and region-specific task-models calibrated for the specific area under study.

DENETHOR[42] is a crop type mapping dataset that provides PlanetScope and Sentinel-2 images of the year 2018 from nine crop categories. In Table 3, we use a single PlanetScope image from May 8, 2018, and obtain a cutout of the entire scene around each field parcel. This cutout is reshaped to a rectangular image of 128 px × 128 px. We select only field parcels that are larger than 30,000 m^2 to maintain a certain homogeneity after rescaling. We also select three classes wheat, corn, and meadow to obtain an annotated image dataset of 640 images, as we found that the classification of all crop types using a single image was too complex for all models with only few training samples. Here, the regional variability of croplands and the high temporal variability of growing crops makes it similarly difficult to construct a representative dataset. METEOR fine-tuned to few examples of cropland in one particular area and one particular growing phase can provide accurate predictions with little annotation effort.

AnthroProtect[43] was gathered to measure the presence of human influence from Sentinel-2 imagery in Fennoscandia. It consists of Sentinel-2 images of areas that are designated as naturally protected areas and are, thus, minimally influenced by humans. These images are classified against Sentinel-2 scenes of non-protected areas within the same countries. This dataset contains 990 annotated images and results are reported in Table 3. This dataset tests the applicability of METEOR to distinguish fine-grained differences between human-influenced and natural (protected) areas.

The datasets for the qualitative deforestation and change detection problems have been created by the authors, and details are provided in Figs. 3 and 5, respectively. They will be available in the provided repository. Fine-tuning METEOR on deforestation detection with few training examples can support remote sensing and environmental research in quickly exploring and identifying newly deforested areas with few training examples. These deforestation events are often time-critical, especially when emerging in novel areas that have not been mapped before. Hence, fine-tuning a deep learning model with a few training examples can accelerate the identification of new emerging deforestation hot spots. Similarly, the change detection application in Beirut highlights the need to quickly identify affected areas in natural disaster cases from a few training examples.

Data availability

The self-created datasets for change detection are acquirable from the provided repository https://github.com/marccoru/meteor/blob/master/doc/beirutdata.md, where we provide a direct download link for the image data. The Sentinel-2 images were obtained through Google Earth Engine[59] and are free of charge in a Creative Commons CC BY-SA 3.0 IGO licence. The PlanetScope data for the deforestation use-case were obtained through the PlanetScope Education and Research Plan. We provide a file in the repository under https://github.com/marccoru/meteor/blob/master/doc/deforestationdata.md that lists the PlanetScope explorer query and the image ids used in this study. To acquire the actual images, one must re-download either through their own PlanetScope Education and Research Plan or by purchasing the data from PlanetScope directly. All other datasets (Sen12MS[5], EuroSAT[40], NWPU-RESISC45[41], DFC2020[33], Floating Marine Objects[6], Denethor[42], AnthroProtect[43]) are publically available from the respective authors. We provide an overview file with download links under https://github.com/marccoru/meteor/blob/master/doc/public_datasets.md.

Code availability

The source code, model weights of the meta-model are available at https://github.com/marccoru/meteor alongside scripts to reproduce the experiments.

Received: 10 June 2023; Accepted: 27 November 2023;
Published online: 12 January 2024

References

1. Camps-Valls, G., Tuia, D., Zhu, X. X. & Reichstein, M. (eds) *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing* (Wiley & Sons, 2021).
2. Weir, N. et al. Spacenet mvoi: a multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 992–1001 (IEEE, 2019).
3. Sumbul, G., Charfuelan, M., Demir, B. & Markl, V. Bigearthnet: a large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-*

2019 IEEE International Geoscience and Remote Sensing Symposium, 5901-5904 (IEEE, 2019).

4. Sumbul, G. et al. Bigearthnet-mm: a large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geosci. Remote Sens. Mag.* **9**, 174–180 (2021).
5. Schmitt, M., Hughes, L. H., Qiu, C. & Zhu, X. X. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7, 153–160 (2019).
6. Mifdal, J., Longépé, N. & Rußwurm, M. Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **3**, 285–293 (2021).
7. Biermann, L., Clewley, D., Martinez-Vicente, V. & Topouzelis, K. Finding plastic patches in coastal waters using optical satellite data. *Sci. Rep.* **10**, 1–10 (2020).
8. Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitsos, D. E. & Karantzalos, K. Marida: a benchmark for marine debris detection from sentinel-2 remote sensing data. *PLoS One* **17**, e0262247 (2022).
9. Murphy, K. P. Beyond the iid assumption. In *Probabilistic Machine Learning: Advanced Topics*, Ch. 19, 727–762 (MIT Press). http://probml.github.io/book2. Version 2023-08-15 (2023).
10. Lemberger, P. & Panico, I. A primer on domain adaptation. Preprint at *arXiv:2001.09994* (2020).
11. Yang, Q., Zhang, Y., Dai, W. & Pan, S. J. *Transfer Learning* (Cambridge University Press, 2020).
12. Schmidhuber, J. *Evolutionary Principles in Self-referential Learning, or on Learning How to Learn: The Meta-meta-... hook.* PhD thesis, Technische Universität München (1987). https://mediatum.ub.tum.de/813181?show_id=813180.
13. Hospedales, T., Antoniou, A., Micaelli, P. & Storkey, A. Meta-learning in neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5149–5169 (2021).
14. Tan, C. et al. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III 27*, 270–279 (Springer, 2018).
15. Sun, X. et al. Ringmo: a remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing* (2022).
16. Scheibenreif, L., Hanna, J., Mommert, M. & Borth, D. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1422–1431 (2022).
17. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).
18. Vinyals, O. et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems* 29 (2016).
19. Zhang, P., Bai, Y., Wang, D., Bai, B. & Li, Y. Few-shot classification of aerial scene images via meta-learning. *Remote Sensing* **13**, 108 (2021).
20. Sharma, S., Roscher, R., Riedel, M., Memon, S. & Cavallaro, G. Improving generalization for few-shot remote sensing classification with meta-learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 5061–5064 (IEEE, 2022).
21. Tang, X. et al. Multi-scale meta-learning-based networks for high-resolution remote sensing scene classification. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 4928–4931 (IEEE, 2021).
22. Lunga, D., Arndt, J., Gerrand, J. & Stewart, R. Resflow: a remote sensing imagery data-flow for improved model generalization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14**, 10468–10483 (2021).
23. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1126–1135 (PMLR, 2017).
24. Von Oswald, J. et al. Learning where to learn: gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021).
25. Bronskill, J., Gordon, J., Requeima, J., Nowozin, S. & Turner, R. Tasknorm: rethinking batch normalization for meta-learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1153–1164 (PMLR, 2020).
26. Oh, J., Yoo, H., Kim, C. & Yun, S.-Y. Boil: towards representation change for few-shot learning. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).
27. Tseng, G., Kerner, H., Nakalembe, C. & Becker-Reshef, I. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1111–1120 (2021).
28. Rußwurm, M., Wang, S., Korner, M. & Lobell, D. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 200–201 (2020).
29. Rolf, E. et al. A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 1–11 (2021).
30. Coates, A. & Ng, A. Y. Learning feature representations with k-means. In G. Montavon, G. B. Orr, K.-R. Müller (Eds.) *Neural Networks: Tricks of the Trade*, 561–580. Second Edition (pp. 561-580). Springer Berlin Heidelberg 2nd Edition LNCS 7700 (2012).
31. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, 448–456 (PMLR, 2015).
32. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: the missing ingredient for fast stylization. Preprint at *arXiv:1607.08022* (2016).
33. Schmitt, M., Hughes, L., Ghamisi, P., Yokoya, N. & Hänsch, R. IEEE GRSS Data Fusion Contest. *IEEE Dataport*. https://doi.org/10.21227/rha7-m332 (2020).
34. Wang, Y. et al. Ssl4eo-s12: a large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. Preprint at *arXiv:2211.07044* (2022).
35. Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D. & Rodriguez, P. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 9414–9423 (2021).
36. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Proc. Syst.* **33**, 9912–9924 (2020).
37. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 9650–9660 (2021).
38. Wilcoxon, F. *Individual Comparisons by Ranking Methods* (Springer, 1992).
39. Schmitt, M. & Wu, Y.-L. Remote sensing image classification with the sen12ms dataset. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. V-2-2021, 101–106 (2021).
40. Helber, P., Bischke, B., Dengel, A. & Borth, D. Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**, 2217–2226 (2019).
41. Cheng, G., Han, J. & Lu, X. Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE* **105**, 1865–1883 (2017).
42. Kondmann, L. et al. Denethor: the DynamicEarthNET dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (Round 2)* (2021).
43. Stomberg TT, Leonhardt J, Weber I and Roscher R. Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery. *Front. Artif. Intell.* **6**, 1278118 (2023).
44. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738 (2020).
45. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explain it to me–facing remote sensing challenges in the bio-and geosciences with explainable machine learning. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **V-3-2020**, 817–824 (2020).
46. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833 (Springer, 2014).
47. Team, P. Planet application program interface: In *Space for Life on Earth*. https://api.planet.comx (2017).
48. Barni, P. E., Fearnside, P. M. & Graça, P. M. L. d. A. Simulating deforestation and carbon loss in Amazonia: impacts in Brazil's Roraima state from reconstructing highway br-319 (Manaus-Porto velho). *Environ. Manage.* **55**, 259–278 (2015).
49. Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou and X. X. Zhu. Self-Supervised Learning in Remote Sensing: A review. In *IEEE Geoscience and Remote Sensing Magazine*, **10**, 213–247, (2022).
50. Tuia, D. et al. Toward a collective agenda on AI for earth science data analysis. *IEEE Geosci. Remote Sens. Mag.* **9**, 88–104 (2021).
51. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
52. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255 (IEEE, 2009).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 770–778 (2016).
54. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1501–1510 (2017).

55. Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.
56. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738 (2020).
57. Rahimi, A. & Recht, B. Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems* 21 (2008).
58. Loveland, T. & Belward, A. The international geosphere biosphere programme data and information system global land cover data set (discover). *Acta Astronaut.* **41**, 681–689 (1997).
59. Gorelick, N. et al. Google Earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
60. Wu, Y. & He, K. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19 (2018).

Author contributions

M.R. envisioned and implemented the method, the experiments, wrote the first draft the manuscript, and integrated the suggestions of the co-authors. S.W. contributed by suggesting additional experiments and overall revision of the manuscript. B.K. contributed by experimental design and manuscript writing. R.R. provided expertise on explainability methods in experimental design and manuscript refinement. D.T. contributed in advising in model and experimental design and manuscript revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-023-01146-0.

Correspondence and requests for materials should be addressed to Marc Rußwurm.

Peer review information *Communications Earth & Environment* thanks Sun Xian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Martina Grecequet. A peer review file is available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.