# METEOR: Meta-learning connecting Earth problems observed from space


Marc Russwurm
EPFL
Sion, Switzerland
marc.russwurm@epfl.ch

Ribana Roscher
Forschungszentrum Jülich
Jülich, Germany
r.roscher@fz-juelich.de

Benjamin Kellenberger
Yale University
New Haven, CT
benjamin.kellenberger@yale.edu

Sherrie Wang
MIT
Boston, MA
sherwang@mit.edu

Devis Tuia
EPFL
Sion, Switzerland
devis.tuia@epfl.ch



## Abstract

*Satellite remote sensing has become a key technology for monitoring Earth and the processes occurring at its surface. It relies on state-of-the-art machine learning models that require large annotated datasets to capture the extreme diversity of the problems of interest to achieve effective monitoring. While datasets for established problems like land cover classification exist, niche applications such as marine debris detection, deforestation, or glacier dynamics monitoring still miss datasets of sufficient size and variety to train successful deep learning models. Despite some advances in transfer learning, current approaches remain problem-specific and perform poorly out of domain. In this work, we propose METEOR, a meta-learning model providing a holistic, fine-grained classification setup capable of adapting to new problems with limited labels. We demonstrate the performance and versatility of METEOR on a series of remote sensing benchmark tasks from different disciplines.*


## 1. Introduction

Satellite remote sensing is an emerging technology to monitor the pulse of planet Earth and is becoming a prime sensor data source for studying the effects of climate change [26] and human activities [8]. Thanks to deep learning algorithms [1, 12], the accuracy of products derived from satellite images is steadily increasing, and researchers are considering increasingly multi-sensor, hybrid, and explainable models [22].

But despite its promises, deep learning for satellite remote sensing still cannot live to its full potential, since many problems of interest such as marine debris detection, deforestation mapping, or glacier dynamics monitoring still lack the large annotated dataset needed for model training and evaluation. This is especially problematic as remote sensing models are commonly learned in isolation for specific problems of interest and geographical regions, limiting the effectiveness of deep learning for less explored problems. Moreover, this is further exacerbated by the differences in sensors used, preventing straightforward model adaptation.

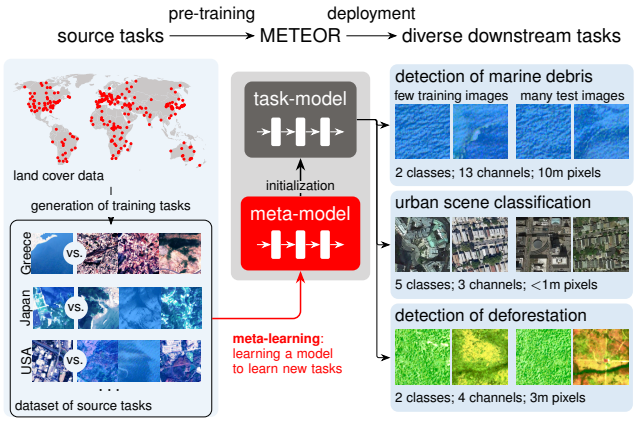



Figure 1. **Concept of METEOR (center), which is pre-trained on land cover source tasks (left) and deployed on diverse downstream tasks (right)**. A task is a dataset containing few annotated images, divided into independent train and test sets. The task data describe a new problem in a format that a machine learning model can be optimized on. The meta-model is pre-trained with model-agnostic meta-learning (MAML) [5] to solve land cover classification source tasks in different geographic regions (left). MAML yields a deep meta-model that has explicitly *learned to learn* from different tasks with few labeled images. The pre-trained meta-model can then be fine-tuned to diverse downstream problems (right) with only few labeled images, thus leading to problem-specific task models.

Therefore, models able to process data from multiple sensors and to learn new problems from a handful of examples (i.e., *few-shot*) are in high demand.

In this work, we address learning across different Earth observation problems systematically: we propose METEOR, a *met*a-learning methodology for *E*arth *o*bservation problems across different *r*esolutions. METEOR is an optimization-based meta-learning approach that uses a small ResNet-12 deep learning model, which outputs a one-against-all classification score. METEOR is pre-trained with the model-agnostic meta-learning (MAML) [5] algorithm to distinguish different land cover categories on medium-resolution multi-spectral satellite data, as shown in Fig. 1. Once trained on a dataset of land cover tasks, it can then meta-learn downstream tasks of interest with limited labeled samples. In our experiments, we focus explicitly on fine-tuning this meta-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*. Our experimental results show that METEOR improves over the current state of the art and is a competitive transfer learning strategy to learn across problems with limited examples.

## 2. Methods

### 2.1. The starting point: MAML

To train the meta-model, we use the model-agnostic meta-learning (MAML) [5] algorithm that optimizes the following objective:

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

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

A task-model $\phi_\tau$ is initialized from the meta-model $\theta$ and iteratively fine-tuned with $k \leq K$ steps based on gradients from a loss of training samples $\nabla L^{\text{train}}$ in an inner loop. The constant $\alpha$ denotes the inner learning rate. In the outer loop, the meta-model parameters $\theta$ are updated by minimizing the test loss $L_{\tau}^{\text{test}}$ 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 $\theta$. 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. We chose the standard second-order MAML algorithm [5] over more recent variants like SparseMAML [24], as it achieved better results on the realistic use-cases in our initial experiments. Once the meta-model $\theta$ is trained, we use it as a starting point to learn specific downstream task models with limited labels.

### 2.2. METEOR

METEOR is designed to facilitate heterogeneous transfer across remote sensing problems that involve data of different sensors at different resolutions. This heterogeneous transfer is enabled by three modifications of the original MAML model:

- First, we replace all batch normalization [7] layers with instance normalization [23]. It was shown that classical, transductive batch normalization, usually used in models trained with MAML [5], has detrimental effects on downstream problems with high class imbalance [14]).
- Second, when learning downstream task models, we dynamically select only the kernels in the first convolution layer of the model that correspond to the input channels (*e.g.*, spectral bands) of the downstream task. This allows us to fine-tune the model for tasks containing imagery with fewer spectral bands. This selection is meaningful as long as the spectral bands form a subset of those used to train the meta-model. Here, pre-training was done on 2 radar and 13 optical channels, which enables downstream tasks using various satellite sensors, such as PlanetScope or Worldview.
- 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 easy-to-implement, but important methodological 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.

## 3. Data and competing methods

We first present the dataset used to train the meta-model (Sen12MS, Sec. 3.1) and then the six datasets used as downstream tasks Sec. 3.2. In Sec. 3.3, the competing methods are briefly presented.

### 3.1. Training the meta-model: Sen12MS

The *Sentinel-12 Multi-Spectral (Sen12MS)* [18] dataset contains Sentinel-1 (synthetic aperture radar) and Sentinel-2 (multispectral) images with associated land cover la-

bels in a coarse segmentation map in 125 globally distributed geographic regions, shown as red dots on the map of Fig. 1. We use Sen12MS for classification by associating the image with the majority class observed in the patch [19]. The original dataset contains overlapping images of 256 px by 256 px. Following prior work [15], we remove the overlap in the images yielding 128 px by 128 px images. Nine different land use and land cover categories are present [18, 19]: *forests*, *shrubland*, *savanna*, *grassland*, *wetlands*, *croplands*, *urban/built-up*, *snow/ice*, *barren*, *water*. We split the data into distinct geographical regions. The meta-model is trained on tasks from 75 training regions, while tasks from the 25 validation regions are used for parameter tuning and early stopping of the pre-training.

### 3.2. The six downstream tasks

We assess the downstream adaptability of METEOR on a variety of downstream tasks, including:

- *Data Fusion Contest 2020 (DFC2020 [17])*: this dataset mirrors Sen12MS with the same land classes but with less noisy, refined annotations and realistic data imbalance (contrarily to Sen12MS, which is class balanced). The dataset spans seven geographic regions, of which we show the results on the Kippa Ring region in Tab. 1.

- *EuroSAT* [6], which contains multi-spectral Sentinel-2 images of 64 px by 64 px with 13 spectral bands. It features nine land use and land cover classes. The dataset is artificially balanced with 2500 to 3000 images per class.

- *NWPU-RESISC45* [4], which contains RGB images of 256 px by 256 px at different resolutions of 45 diverse classes. 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*.

- The *Floating Marine Objects* dataset [11], which contains Sentinel-2 images with hand-annotated labels of marine debris in 26 coastal regions across the globe. 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 31st, 2018. In Tab. 1, 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.

- *DENETHOR* [9] is a crop type mapping dataset that provides PlanetScope (4 bands, 3m resolution) and Sentinel-2 images from nine crop categories. In Tab. 1, we use one PlanetScope scene from May 8th, 2018. For each field parcel, we cropped a rectangular image of 128 px by 128 px enclosing it. We select only field parcels larger than 30 000 $m^2$ to maintain a certain homogeneity after rescaling. We selected three classes (*wheat*, *corn*, *meadow*) to obtain an annotated image dataset of 640 images.

- *AnthroProtect* [21] was gathered to measure the presence of human influence from Sentinel-2 imagery in Fennoscandia. The images depict areas designated as naturally protected areas and 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.

### 3.3. Comparison methods

We compare METEOR to self-supervised learning (SSL) approaches, pre-trained on either multi-spectral satellite (SSLTRANSRS [16] and SSL4EO [25]), RGB satellite data (SECO [10]), or on natural RGB images (SWAV [2] and DINO [3]). As further comparisons, we train a BASELINE to classify all ten classes present in the training areas of the Sen12MS dataset in a supervised way. We also add two ResNet-50 additional baselines, one initialized on ImageNet weights (IMAGENET) and another with random initialization (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 [20]. We also generate MOSAIKS [13] features dynamically for each downstream task from the training data and predict the test data with a random forest classifier.

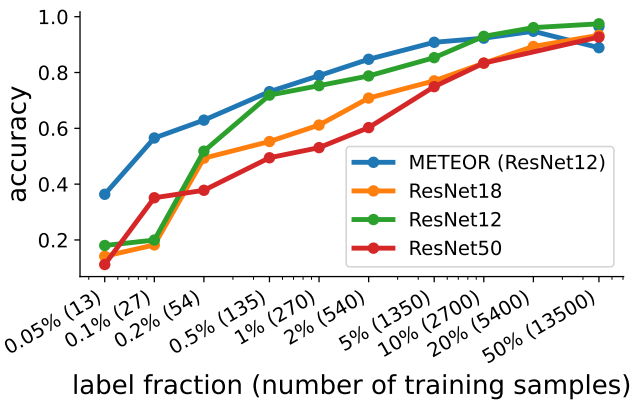



Figure 2. **Performance of METEOR vs supervised baselines as a function of the number of shots used in the EuroSAT downstream task**. In low shot scenarios, METEOR provides better representations, is easier to fine-tune to the specific task and better transfers from the task used for training the meta-model.

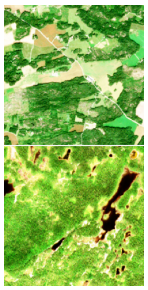
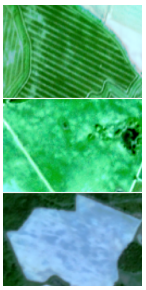
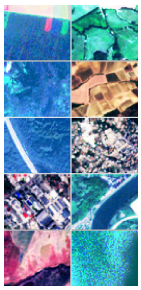
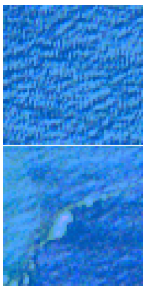
| | 5-shot problem | human influence | crop type mapping | land cover classification | | marine debris | urban scenes |
|---|---|---|---|---|---|---|---|
| | dataset | AnthPr. [21] | DENETHOR [9] | DFC2020-KR [19] | EuroSAT [6] | fl. obj. [11] | NWPU-Urban [4] |
| | spatial res. | 10m | 3m | 10m | 10m | 10m | $<$ 1m |
| | spectral res. | 10 bands | 4 bands | 13 bands | 13 bands | 12 bands | 3 bands |
| | # classes | 2 | 3 | 5 | 10 | 2 | 5 |
| | # 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 [2] | 4.2 | **96.7** | 69.8 | 54.2 | **67.7** | 65.4 | 70.4 |
| MOSAIKS [13] | 4.3 | 86.4 | **76.4** | 82.3 | 57.9 | 88.8 | 54.0 |
| DINO [3] | 5.0 | 91.2 | 66.2 | 56.6 | 61.3 | 65.1 | **70.6** |
| SECO [10] | 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 [25] | 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 [20] | 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 |
| |  |  |  |  |  |  | |

Table 1. **Quantitative comparison of METEOR with several state-of-the-art methods (rows) across different heterogeneous Earth observation datasets (columns)**. Each evaluated task is characterized by a different number of spectral bands, number of classes, and spatial resolution. METEOR achieves the best average rank of 3.6, 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 ($*$: $p < 0.05$; $**$: $p < 0.01$) from the BASELINE, PROTO, IMAGENET, SCRATCH models. This comparison shows that METEOR achieves state-of-the-art few-shot performance across various remote sensing problems. Example images of each class for each downstream task are shown in the bottom row.

## 4. Experiments

In Fig. 2, we compare the METEOR model, fine-tuned on the EUROSAT downstream task in several few- and many-shot settings: METEOR shows competitive performances in all settings and is particularly accurate in few-shot scenarios, showing that METEOR can leverage the knowledge from the meta-learning task while adapting efficiently to the new task better than fully supervised models.

In Tab. 1, we compare the performance of several SSL and few-shot methods detailed in Sec. 3.3. Across all methods, METEOR shows the best average rank (3.6 out of 11), closely followed by the large-scale contrastive pre-training methods SWAW (4.2) and DINO (5.0). This is surprising since those methods are trained only on RGB data. On the contrary, existing SSL methods trained on multispectral data perform worse on average (e.g., SSLTRANSRS (5.3), SSL4EO (5.5)) but still show some significant improvements over the baseline (6.8) or basic pre-training (e.g., on IMAGENET (8.8)).

## 5. Discussion and Conclusion

In this short paper, we presented the results of a few-shot model for satellite remote sensing data named METEOR. Learned from global land cover class examples, METEOR improves on the model-agnostic meta-learning strategy to learn to learn different tasks and therefore is naturally performant in fine-tuning new classification problems involving a variable number of classes, different sensors (the meta-model is trained on both multispectral and radar data) and only a very limited number of labeled samples.

Our experiments confirmed that utilizing transfer- and meta-learning for different-but-related tasks is important for addressing meaningful problems with limited training data. This is particularly evident in naturally unbalanced problems.

Future directions to further improve the model comprise, for example, a more diversified meta-training encompassing a larger variability of thematic classification tasks that can be encountered in remote sensing.

The source code and pre-trained weights of METEOR are available under `https://github.com/marcCoru/meteor`

## References


[1] Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein (Editors). *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing*. Wiley & Sons, 2021. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020. 3, 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 9650–9660, 2021. 3, 4

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 3, 4

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 1, 2

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 12(7):2217–2226, 2019. 3, 4

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015. 2

[8] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 1

[9] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andrés Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longépé, Timothy Davis, Giovanni Marchisio, 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. 3, 4

[10] Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 9414–9423, 2021. 3, 4

[11] Jamila Mifdal, Nicolas Longépé, and Marc Rußwurm. Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:285–293, 2021. 3, 4

[12] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 1

[13] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):1–11, 2021. 3, 4

[14] M. Russwurm, S. Wang, B. Kellenberger, R. Roscher, and D. Tuia. Instance norm improves meta-learning in class-imbalanced land cover classification. In *NeurIPS workshop DistShift*, 2022. 2

[15] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 200–201, 2020. 3

[16] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022. 3, 4

[17] Michael Schmitt, Lloyd Hughes, Pedram Ghamisi, Naoto Yokoya, and Ronny Hänsch. IEEE GRSS Data Fusion Contest. *IEEE Dataport*, 2020. 3

[18] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. 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*, volume IV-2/W7, pages 153–160, 2019. 2, 3

[19] Michael Schmitt and Yu-Lun Wu. Remote sensing image classification with the sen12ms dataset. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-2-2021, pages 101–106, 2021. 3, 4

[20] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3, 4

[21] Timo T Stomberg, Taylor Stone, Johannes Leonhardt, and Ribana Roscher. Exploring wilderness using explainable machine learning in satellite imagery. *arXiv preprint arXiv:2203.00379*, 2022. 3, 4

[22] Devis Tuia, Ribana Roscher, Jan Dirk Wegner, Nathan Jacobs, Xiaoxiang Zhu, and Gustau Camps-Valls. Toward a collective agenda on ai for earth science data analysis. *IEEE Geoscience and Remote Sensing Magazine*, 9(2):88–104, 2021. 1

[23] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[24] Johannes Von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, and João Sacramento. Learning where to learn: Gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 2

[25] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for

self-supervised learning in earth observation. *arXiv preprint arXiv:2211.07044*, 2022. 3, 4

[26] Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jingming Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10):875–883, 2013. 1