A. Dataset details

In the following, we provide additional details on our dataset. For once, we summarize the metadata complementing our raw sensory data in Appendix A.1. In Appendix A.2 we describe the auxiliary Sentinel-2 [1, 17] images. Finally, we provide additional information on the different sampling densities of our dataset in Appendix A.3.

A.1. Planet metadata

The commercial Planet Fusion data constitutes the core part of the *DynamicEarthNet* dataset. In addition to the surface reflectance values (RGB+near-infrared) that we use in the main paper, Planet provides additional quality assurance (QA) information. The purpose of this is to denote which parts of the data are raw observations and which parts are gap-filled with temporally close observations. For every pixel, the QA product gives the distance and direction to the day of the observation. For example, a pixel value of -1 implies that the pixel has been filled from the previous day.

A.2. Sentinel 2 auxiliary images

Sentinel-2 (S2) images are publicly available through the open data policy of the European Space Agency's (ESA) Copernicus Program. The mission collects images of all landmasses every 5 days at a resolution of 10m per pixel [17]. While the temporal and spatial resolution of S2 time-series imagery is smaller than the Planet data, S2 collects 13 channels compared to 4 channels of Planet Fusion. In certain scenarios, the additional channels, particularly in the short-wave infrared spectrum, may provide useful auxiliary information about changes on the ground.

In order to encourage cross-research between Planet Fusion and S2 data, we accompany our dataset with monthly images of S2 data from the same locations. The Sentinel-2 images are composite images which means they have been created from multiple S2 images throughout the month. This allows for a direct comparison of the effectiveness of different sources of satellite imagery.

Our Sentinel-2 data is provided as a so-called Bottom-Of-Atmosphere product which includes the correction of distortions to the surface reflectance values caused by atmospheric interference. The S2 pre-processing quality is relatively low compared to the analysis-ready Planet Fusion product. For some areas of interest (AOIs), the collected S2 data suffer from occlusions through cloud coverage for all S2 images in a month. This naturally compromises the quality of the monthly composites. We have collected affected months for all AOIs manually in a designated S2 quality assessment spreadsheet that we provide, together with the dataset. 26% of monthly S2 composites suffer from minor quality issues and around 5% have major quality issues. When the community explores applications of S2 data with *DynamicEarthNet*, we advise to investigate whether considered cubes or months are potentially impacted.

A.3. Temporal densities

In our experiments in Sec. 5.2 and Sec. 5.3, we use three different temporal sampling densities for both the spatio-temporal and semi-supervised baselines:

- The monthly setting (fully supervised) shows the first day of each month, resulting in a one-to-one correspondence between input images and labels.
- For the weekly setting, we feed the architectures with samples from the 1st, 5th, 10th, 15th, 20th and 25th days of each month.
- The daily setting uses all the available images in a considered month, as well as the corresponding monthly label.

In Fig. 4, we show the images of 5 time-series with a weekly sampling density.

B. Evaluation protocol details

In the following, we motivate our design choices for the metric proposed in Sec. 4 and compare it to other existing metrics.

B.1. Semantic change

In contrast to semantic segmentation, semantic change segmentation focuses on the changed parts of a given semantic map. Similar to how boundary segmentation restricts evaluation to the boundary pixels, our proposed metric is restricted to changed pixels. We consider several options on how to restrict this subset. In the following, we refer to pixels that have changed their semantic class from one timestep to the next as changed pixels:

R1. We restrict the evaluation to the set of changed pixels, as predicted by the considered method.

R2. We restrict the evaluation to the set of changed pixels defined by the ground-truth semantic maps.

R3. We restrict the evaluation to the intersection of R1 and R2, which is the set of true positives.

Using the set of R1 or R3 has the disadvantage that it couples the semantic change performance with the binary change performance. Only the pixels that are predicted as changed are potentially also evaluated for the semantic change score. Hence, errors in the binary change influence the semantic change score, which potentially opens the metric to misconduct. One can easily imagine a method that reduces the set of predicted change artificially to a single pixel for which the semantic class is predicted with very high confidence. Then the SC score would be perfect (1.0), while the BC score would be close to 0. The resulting overall SCS score would be around 0.5, which is much higher than the scores reported in Tab. 5. Such behavior is completely undesired and leads to a metric that is not aligned

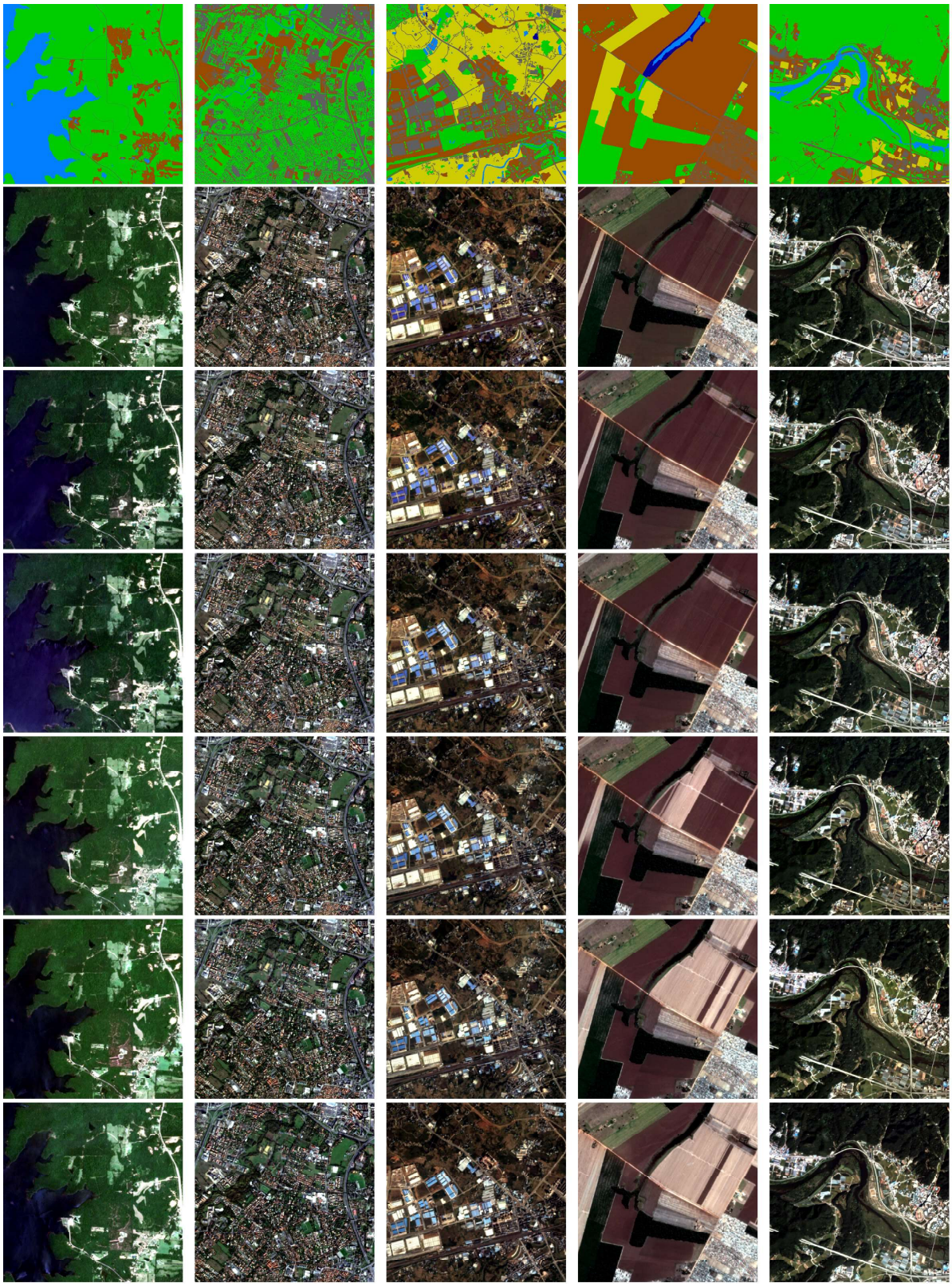

Figure 4. **Training set samples.** We visualize 5 sample time series (one per column) from the training set of the presented *DynamicEarthNet* dataset. Each sequence illustrates weekly samples (row 2-7) and the corresponding annotated monthly labels (1st row).

with human intuition, with results that are hard to interpret. Thus, we use the second option R2 to compute the SC metric. This makes the errors decoupled and the scores easy and intuitive to interpret.

B.2. Comparison

Even though there exists no unified evaluation protocol for semantic change segmentation, there are a few metrics that focus on certain aspects of the task. In the following, we discuss the different options and compare their efficacy for the task of semantic change segmentation.

Pixel accuracy. Pixel accuracy, also referred to as overall accuracy, is one of the simplest measures for (binary) segmentation problems. It is defined as the ratio of correctly classified pixels to all pixels. In settings like ours, in which there are 2 classes for binary change and a high imbalance between them, the pixel accuracy is not able to report meaningful insights. In our setting, 95% of all pixels do not change. Thus, a score of 95% can be obtained by predicting no change all the time. Therefore, we refrain from using pixel accuracy as a metric.

mIoU. The standard mean intersection-over-union addresses the immediate shortcomings of the vanilla pixel accuracy metric. It is possible to use it for both, binary change and semantic change. However, using the mIoU metric for binary change directly, *i.e.* computing the mean IoU of the 2 classes, suffers also from the imbalance issues discussed for the pixel accuracy. Thus, the proposed BC metric computes the IoU of only the change class, rather than both the change and no-change class. For the semantic change, we however apply mIoU, *i.e.* computing the mean over all semantic classes. As explained in the previous subsection, an insightful change metric should focus on the changed regions. We, therefore, refrain from using mIoU on the whole image but compute the scores solely on the changed pixels.

Cohen's kappa. Previous works [25, 28, 36] have used Cohen's kappa to measure the performance in similar settings. Cohen's kappa is a statistical measure of the agreement between the predictions and ground-truth. It is more robust compared to pixel accuracy as it takes the agreement occurring by pure chance into account. However, this measure is not as informative as mIoU. It does not offer insights into the performance of individual classes. Moreover, since scores are not aggregated per class, the performance of classes with high appearance rates will dominate the score and therefore lead to an overall higher score. For more details about the dataset imbalance, we refer to Tab. 2. We thus choose to adapt the well-established IoU measure for our needs.

		SCS (↑)	BC(↑)	SC(↑)
bi-temp	CAC [21]	17.8	10.1	25.4
	U-TAE [34]	**19.1**	9.5	**28.7**
	U-ConvLSTM [26]	19.0	**10.2**	27.8
	3D-Unet [26]	17.6	10.2	25.0
multi-temp	CAC [21]	**27.7**	23.6	**31.8**
	U-TAE [34]	27.6	23.4	31.8
	U-ConvLSTM [26]	27.5	**24.2**	30.7
	3D-Unet [26]	25.3	21.2	29.4

Table 6. **Quantitative results of our metric variant on our test set.** The first row shows the bi-temporal, and the second row shows the multi-temporal results on weekly data. The first row results are identical to the weekly results in Tab. 5.

B.3. Correcting wrong predictions

Our proposed metric requires a separate binary change map $\hat{\mathbf{b}}$ and semantic map $\hat{\mathbf{y}}$. It is therefore not limited to the special case of computing the binary change $\hat{\mathbf{b}}$ directly from the predicted semantic maps for two consecutive timesteps $\hat{\mathbf{y}}_{t-1}$ and $\hat{\mathbf{y}}_t$. This provides additional flexibility, as it is often preferable to decouple the semantic maps from the change predictions [25, 28]. Moreover, it allows for the correction of previous mistakes in online methods that obtain predictions frame-by-frame for an input time-series. As an example, suppose that a semantic class for a certain pixel is predicted wrong at a given timestep. If that pixel does not change in the next timestep, its prediction would either need to keep the wrong semantic class or predict a different semantic class. However, predicting a different semantic class would automatically be recognized as a predicted change, resulting in an error in the binary change. Thus, there is no way to correct previous mistakes without introducing another one. This also holds for other types of errors. By requiring each method to pass explicitly a binary change map $\hat{\mathbf{b}}$ and semantic map $\hat{\mathbf{y}}$, this issue can be avoided. In our setting and the above example, the semantic class can be corrected without predicting a binary change for this pixel. This is especially important for methods that are used for both semantic segmentation as well as semantic change segmentation.

B.4. Discussion on bi-temporal change

In Sec. 4.1, we define the problem as a bi-temporal semantic change segmentation that measures the SCS, SC, and BC scores for a given ground truth $\mathbf{y}_t$ and $\mathbf{y}_{t+1}$. Given that our dataset contains consistent multi-temporal land use and land cover ground-truth information, it allows us to extend the bi-temporal metric definition and calculate the scores on time intervals of varying lengths. Specifically, we investigate a variant of our bi-temporal semantic change segmentation metrics by measuring the change between all viable pairs of months (t to $t+1$, $t+2$, $t+3$, ...) at each

area of interest (24 x 23 = 552 pairs in total).

We report the resulting accuracies in Tab. 6. For the most part, the modified metric yields slightly higher values than our bi-temporal metric. We attribute this to the fact that the modified metric has a certain smoothing effect, *i.e.* less emphasis is placed on pinpointing the exact frame where change occurs. Throughout our dataset, we notice that different types of changes happen over different time periods (daily, weekly, monthly quarterly, or even seasonally/yearly). On the other hand, the smoothing effect of longer time intervals potentially under-penalizes prediction errors on small time intervals, which goes against one of the main motivations of having daily time-series observations. In our work, we ultimately prefer the bi-temporal setting and leave the detailed multi-temporal discussion as future work.

C. Implementation details

All the experiments are implemented in PyTorch. Our dataset contains 4 spectral bands (RGB + near-infrared). The theoretical valid range for all 4 channels is 1-32,767; however, in practice, the maximum value for the type of data contained in our dataset is 10,000. For data normalization, we calculate the mean and standard deviation per band, averaged over the whole dataset. The exact obtained values are

$$\begin{aligned} \text{mean} &= [1042.59, 915.62, 671.26, 2605.21] \quad \text{and} \\ \text{std} &= [957.96, 715.55, 596.94, 1059.90], \end{aligned}$$

respectively. For data augmentation, we randomly resize the images with a ratio between $[0.5, 2]$ and crop them to half the input resolution $(512, 512)$. Additionally, we apply random horizontal flips. As we specified in Sec. 3.3, due to the scarcity of the snow & ice class, we do not include them in the test and validation set. For the spatio-temporal architectures, we use the Adam optimizer with a learning rate of $1e-4$. The batch size is set to 4. We generally train our networks for up to 100 epochs. For the spatio-temporal experiment with daily samples, we use 200 epochs to ensure convergence. The reported results are taken from the epoch that achieves the highest validation accuracy. For the semi-supervised architecture, analogous to [21], we use the SGD optimizer with the poly learning rate decay policy. For both the supervised and unsupervised samples, we use a batch size of 8.

D. Additional qualitative results

Additional visualizations. We present additional qualitative visualizations corresponding to the results in Sec. 5.2. In Fig. 5, we depict a comparison of the different spatio-temporal baselines described in Sec. 5.1. Furthermore, we compare the effect of different temporal densities for the semi-supervised baseline CAC [21] in Fig. 6. The weekly training achieves the best results on the validation set, as indicated by the results in Tab. 4. This is mostly due to the fact that monthly and daily settings struggle to predict uncommon classes like wetlands (first example in Fig. 6) and agriculture (second example in Fig. 6). Note that these observations are consistent with the confusion matrices shown in Fig. 7.

Confusion matrices. We provide confusion matrices to complement our results on LULC segmentation in Sec. 5.2. The main idea is to allow for a more fine-grained analysis in terms of the 6 semantic classes, see Fig. 7. We show results for both spatio-temporal methods and semi-supervised learning. Each confusion matrix depicts which classes typically get misclassified as certain other classes. For example, the overall uncommon class wetland frequently gets mislabeled as soil, see *e.g.* the first example in Fig. 6. Beyond that, one can also directly read the relative segmentation accuracy of each class in the diagonal entries. As can be expected, the predictions are overall more stable for the more common classes like forest & other vegetation and soil, see Tab. 2 for reference. Among the spatio-temporal methods, the 3D-Unet [26] setting yields the best results for the challenging impervious surface class, see *e.g.* the first example in Fig. 5. All in all, these results indicate that future approaches might benefit from reweighting the individual class labels for a more balanced training that can account for rare LULC classes.

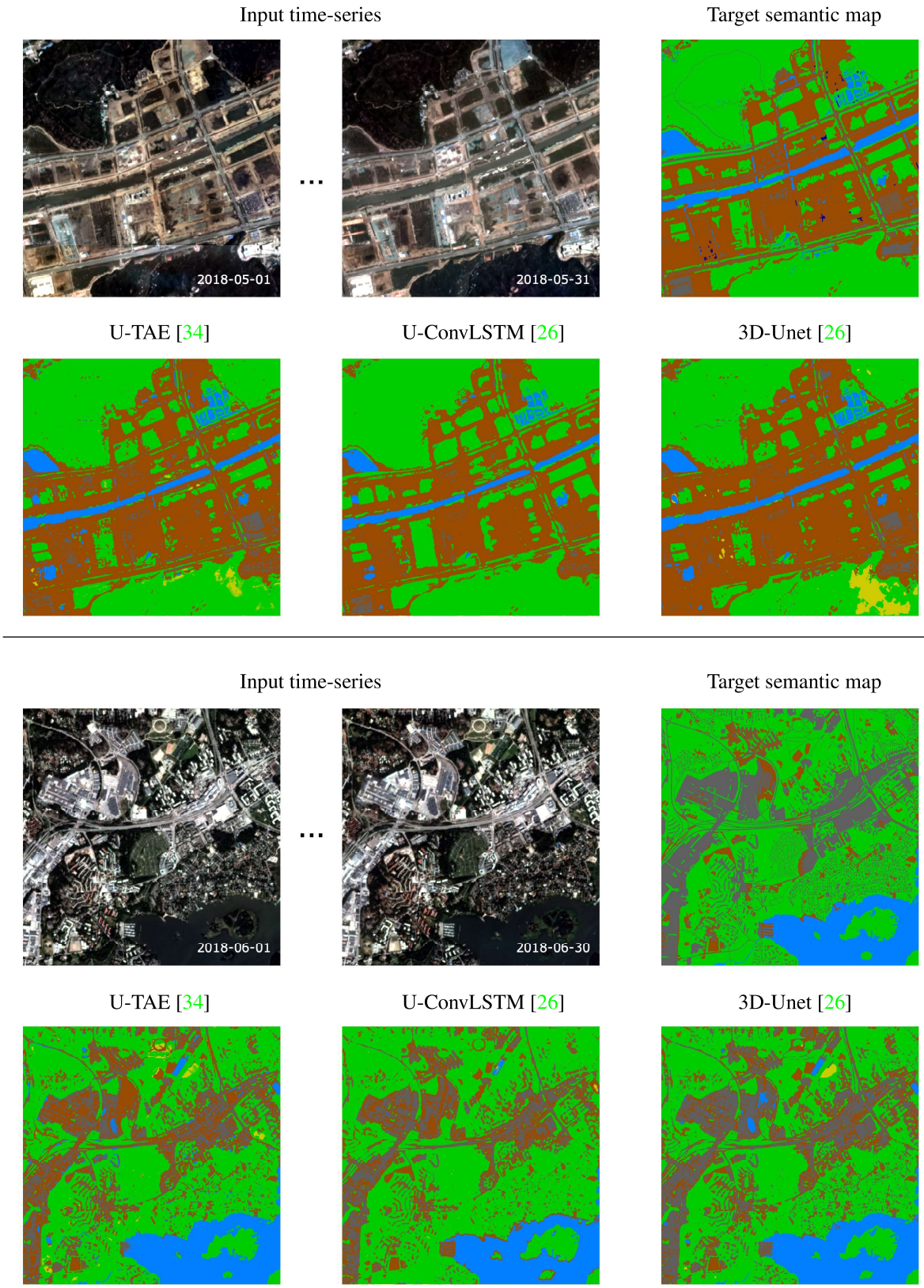

Figure 5. **Spatio-temporal predictions.** We show two qualitative comparisons of the spatio-temporal methods discussed in Sec. 5.1. Both examples are taken from our validation set. The methods take a sequence of 31 and 30 daily samples as inputs (top left) and predict a single semantic map for the whole month (bottom row). We furthermore show the ground-truth annotated map for comparison (top right).

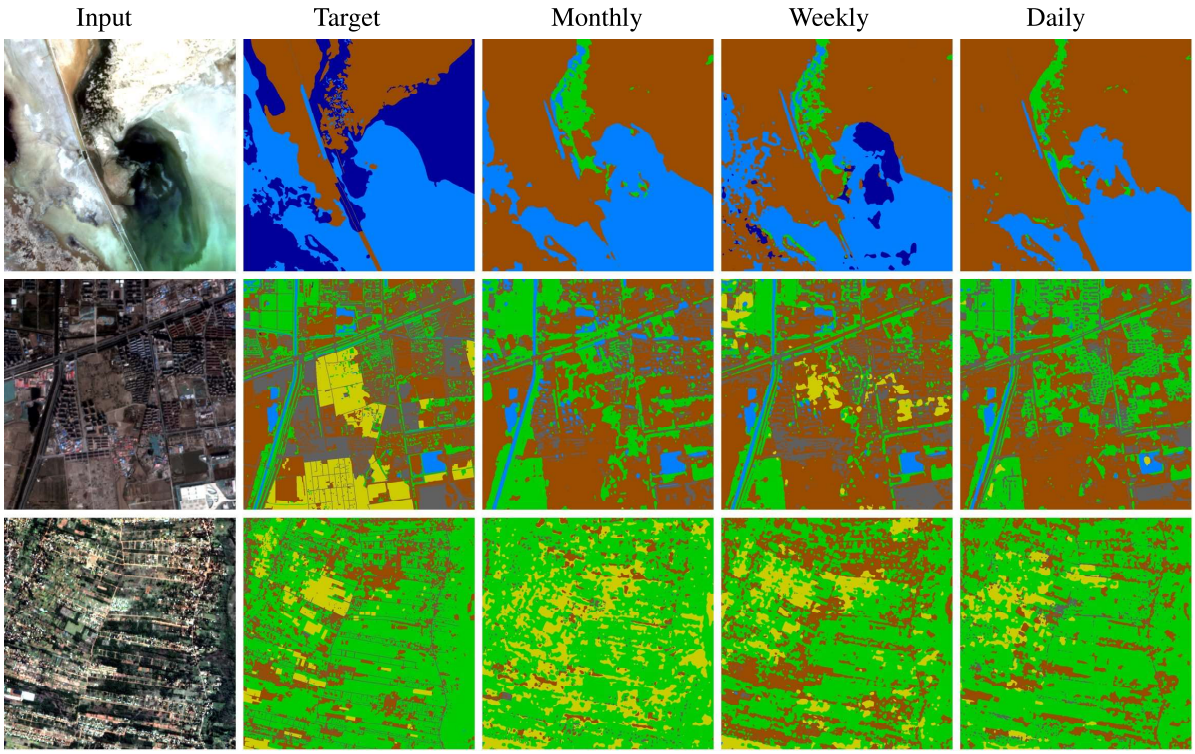

Figure 6. **CAC [21] predictions.** We show sample predictions by the semi-supervised baseline CAC [21] for three different examples from our validation set. For each example, we depict the input sample (1st column), the ground-truth semantic map (2nd column), as well as the predictions of [21] for the monthly, weekly, and daily training setup (3rd-5th column) respectively.

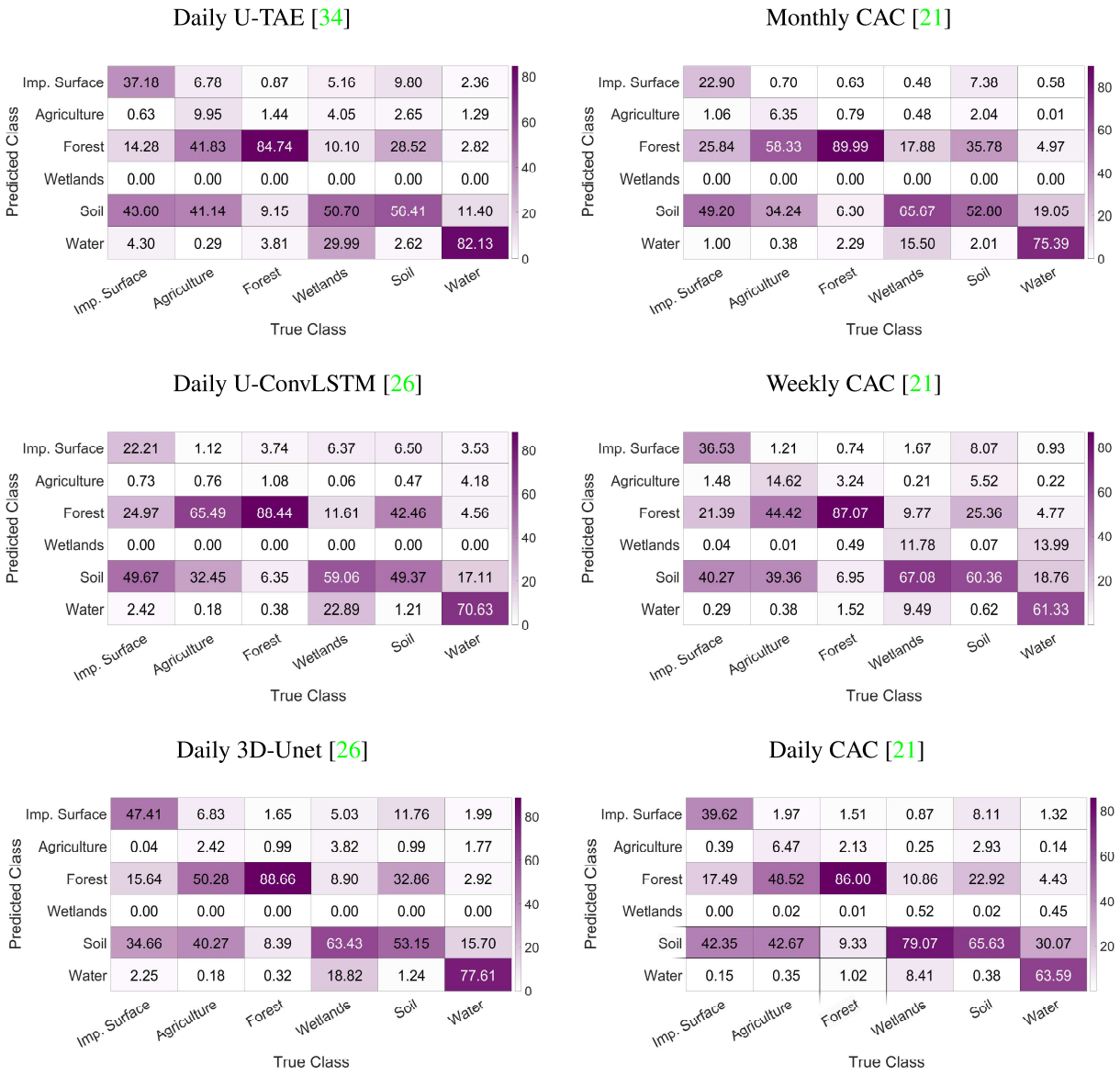

Figure 7. **Confusion matrices.** We show confusion matrices corresponding to the LULC segmentation results in Sec. 5.2 on the validation set. The goal is to provide a fine-grained analysis of which classes frequently get misclassified as certain other classes. Each column of an individual confusion matrix is normalized, meaning that it shows the relative distribution of predictions (in percent) for a given, true class. Results are shown for both spatio-temporal (left column) and semi-supervised baselines (right column) with three different settings each.