Global-PCG-10: a 10-m global map of plastic-covered greenhouses derived from Sentinel-2 in 2020

Bowen Niu[1], Quanlong Feng[1*], Bingwen Qiu[2], Shuai Su[1], Xinmin Zhang[1], Rongji Cui[1], Xinhong Zhang[1], Fanli Sun[1], Wenhui Yan[1], Siyuan Zhao[1], Hanyu Shi[1], Cong Ou[3], Xiaolu Yan[4], Jianhua Gong[5], Gaofei Yin[6], Jianxi Huang[1,6], Jiantao Liu[7], Bingbo Gao[1], Xiaochuang Yao[1], Jianyu Yang[1], Dehai Zhu[1]

[1] College of Land Science and Technology, China Agricultural University, Beijing 100193, China
[2] Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education
[3] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[4] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, 6708 PB Wageningen, The Netherlands
[5] National Engineering Research Center for Geoinformatics, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[6] Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China
[7] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, Shandong, China

Correspondence to: Quanlong Feng (fengql@cau.edu.cn)

Abstract. Plastic-covered greenhouse (PCG) is widely used in agricultural production due to its temperature control, water conservation, and wind protection characteristics, significantly enhancing crop yields and economic benefits. However, its long-term and extensive use can lead to environmental issues, such as the accumulation of local toxic gases and the degradation of soil physicochemical properties. Therefore, obtaining a comprehensive distribution of PCGs is essential. To monitor PCGs on a large scale, this study developed a novel approach for producing the first global 10-meters PCGs dataset (Global-PCG-10) with high-quality. Firstly, the globe was divided into multiple 5-degree grids, and grids for classification were organized based on global cropland layer. Then, multi-temporal Sentinel-2 data and initial labels of PCGs were obtained through Google Earth Engine (GEE) to create a training set for deep learning. Next, initial labels were optimized with the active learning strategy combined with the deep learning model, APC-Net. Finally, the PCGs classification results were predicted, spatially analyzed, and compared with publicly released land use and land cover (LULC) datasets. Experimental results indicate that the proposed Global-PCG-10 dataset has a high overall accuracy of 92.08%. The global area of PCGs is 14,259.85 km², and 69.24% of PCGs are located in Asia, covering around 9,874.51 km². China has the largest PCGs area of 8,224.90 km², accounting for 57.67% of the globe and 83.29% of Asia. Comparisons with other LULC datasets revealed that PCGs, which should be classified as cropland, are often misclassified as bareland, impervious surfaces, ice/snow, etc.

1 Introduction

With the rapid development of modern plastic industry, agricultural plastic-covered greenhouses (PCGs) have been spreading widely around the globe. According to statistics, the total PCGs area of the world has reached to 1.3 million

35   hectares (Tong et al., 2024), accounting for 8 ‰ of the global cropland (global cropland data source from: *https://www.fao.org/faostat/en/#data/RL*). One important reason for the widespread of PCGs is the role in the increase of both crop yield and quality. Local climatic conditions could be greatly improved for crops with an increased accumulated temperature and a decreased water evapotranspiration, which is significant especially for regions with adverse climatic situations (Liu and Xin, 2023; Lu et al., 2018). In addition, the globe witnesses a speedy growth of PCGs recently. Countries

40   that have a large area of PCGs mainly include China, Spain, Italy, Vietnam, etc. (Feng et al., 2022a; Jiménez-Lao et al., 2020; Veettil et al., 2023; Wu et al., 2016).

Although PCGs play a key role in modern agriculture for the improvement of crop yield and quality, the demerits of PCGs could not be neglected. Firstly, PCGs increase both the cropping and land use intensity. Due to the improvement of hydrothermal conditions, the crops could now be harvested twice or three times within one year, leading to the

45   overexploitation of soil nutrients and underground water resources. Along with the cropping intensity, the usage of chemical fertilizers and pesticides has also been increased, which would lead to the widespread of soil contamination. Therefore, the existence of PCGs could be viewed as an important indicator of agricultural non-point source pollution. Secondly, PCGs have changed the pattern of water evapotranspiration, which hinders the circulation of water and may cause the microclimate anomaly. Finally, PCGs contribute more greenhouse gas emissions than other farmlands (Niu et al., 2023b; Wang et al.,

50   2022).

Therefore, it is of great significance to acquire the accurate spatial distribution of PCGs worldwide to understand where and how much PCGs are located and constructed globally. Due to the large-scale coverage and cost-effectiveness, Earth Observation technology especially satellite remote sensing has been widely used for PCGs classification. Commonly used satellites consist of Landsat-5/7/8, IKONOS, QuickBird, WorldView-1/2/3 and ESA's Sentinel-2, which all belong to

55   multispectral satellites (Hao et al., 2019; Ou and Wang, 2022). In addition to the aforementioned multispectral data, some researchers have used free RGB remote sensing imagery from Google Earth for PCGs classification (Niu et al., 2023a; Zhang et al., 2021b). There are three main kinds of classification methods used for PCG mapping: spectral based methods, machine learning based methods, and deep learning-based methods.

Spectral based methods tried to construct a spectral index which is sensitive to plastic greenhouses. The differences of

60   PCGs and the background could be enlarged by these spectral indexes, where a threshold is used to extract PCGs (Aguilar et al., 2022; Zhao et al., 2004). The merit of spectral index is the unnecessary for training samples, while the demerit is the uncertainty in obtaining the suitable threshold in large-scale regions. This is because the best threshold for PCGs extraction may be different in different regions, which is influenced by the spectral variations of both PCGs and the background (González-Yebra et al., 2018; Lu et al., 2014). Aguilar et al., (2016) and Yang et al., (2017) independently developed

65   greenhouse indices, the Moment Distance Index (MDI) and the Plastic Greenhouse Index (PGI), using Landsat satellite data. Similarly, Zhang et al., (2022a) derived the Advanced Plastic Greenhouse Index (APGI) from Sentinel-2 imagery through band calculations.

Earth System
Science
Data

Open Access
Discussions

In terms of machine learning based methods, decision tree, support vector machine (SVM) and random forest (RF) are commonly used supervised classifiers for PCGs mapping. Compared with spectral index, the merit of machine learning is its

70 robustness. However, its drawback is the reliance on labeled samples, where the quantity, quality and diversity of these samples significantly affect classification performance (Qiu et al., 2022, 2024; Zhang et al., 2024). Additionally, "salt-and-pepper effect" is unneglectable in machine learning classifications (Du et al., 2022). Recently, Google Earth Engine (GEE) provides a popular remote sensing cloud platform, which integrates the aforementioned machine learning methods and provides vast volume of multi-source, multi-temporal remote sensing data, along with powerful cloud computing service

75 (Feng et al., 2024; Li et al., 2023; Zhang et al., 2022c). This significantly enhances efficiency in large-scale mapping applications such as global LULC mapping, global wetland mapping, etc. Zhang et al., (2020, 2021a, 2022b, 2023, 2024b) used GEE to develop a comprehensive technical workflow for generating multiple global land cover data products, including a global impervious surface dataset, a global 30-meter LULC dataset, etc. With regard to PCGs classification, Ou et al., (2021) generated a 30-year PCGs distribution map in Shandong Province, China, using Landsat series satellite data and the

80 RF classifier on the GEE cloud platform. Similarly, by utilizing multi-year Landsat series satellite data and RF classifier on GEE, Gao et al., (2022) produced a 20-year greenhouse distribution map for the Guanzhong Plain area in Shaanxi Province, China. Besides, in our previous study, we utilized RF together with a partition modeling strategy to generate the first publicly released 30-m national PCGs map of China with an overall accuracy of 87% (Feng et al., 2021). Furthermore, we also tackled the long-neglected issue of the confusion between PCGs and plastic-mulched farmlands (PMFs) by introducing

85 multi-temporal observations (i.e., film-on and film-off) to exclude PMFs from PCGs (Feng et al., 2022b).

In recent years, deep learning has achieved remarkable success in the fields of computer vision (CV) and natural language processing (NLP). Unlike classical machine learning methods, which can only capture the shallow features of input data, deep learning has a deeper neural network structure and can effectively learn images' semantic features, leading to a better robustness and generalization ability (Chen et al., 2022; Niu et al., 2022; Zhang et al., 2023a). However, the

90 performance of deep learning models heavily depends on the quantity and quality of training samples, which calls for a huge workload for sample labeling. The deep learning model has also been applied to PCGs and PMFs classification. For instance. Zhou et al., (2024) developed a general framework for extracting PCGs, integrating prior knowledge with deep learning models. Li et al., (2022), Liu et al., (2023), and Chen et al., (2021) employed Google Earth imagery as their data source, selecting specific regions within Shandong of China or other smaller areas as study sites, and built deep learning models for

95 PCGs extraction. Additionally, Ma et al., (2021) and Chen et al., (2023) applied high-resolution remote sensing imagery (1-m resolution) and an object detection model to extract PCGs across China, however, these datasets have not been released publicly. In our previous study, we have proposed a dilated and non-local convolutional neural network (DNCNN) for the accurate delineation of PCGs in several key regions from China, Saudi Arabia, Turkey and Spain and achieved a high OA of about 90% (Feng et al., 2021; Niu et al., 2023a). In May 2024, the University of Copenhagen released the first publicly

100 accessible global dataset for large-scale PCGs mapping (Tong et al., 2024), which was derived from PlanetScope commercial satellite data and a deep learning model, UNet. While the dataset has good precision and coverage, acquiring

such high-resolution, world-covered commercial satellite data is costly for most researchers. Therefore, how to use open access satellite data such as Sentinel-2 and Landsat to generate a global PCGs map still remains a challenging task.

105 To address this issue, this study utilizes Sentinel-2 satellite image, one of the most influential open-source remote sensing datasets globally, for PCGs extraction. The temporal consistency and continuous, all-weather Earth observation capabilities of Sentinel-2 data effectively mitigate the temporal inconsistency found in commercial high-resolution datasets. Meanwhile, Sentinel-2 offers the highest resolution open-source remote sensing data, making it well-suited for PCGs classification tasks. With such high data quality that is widely preferred by researchers, many existing data products are derived from Sentinel-2, allowing our Global-PCG-10 dataset to integrate with these data products seamlessly. For data 110 organization, we have designed a global grid system to facilitate PCGs data indexing and accessibility for researchers. However, Sentinel-2 satellite image is not perfect. Unlike high-resolution data, the 10-meter resolution Sentinel-2 data contains a significant number of mixed pixels, which poses challenges for accurate extraction of PCGs. To address this issue, we utilize multi-temporal Sentinel-2 data to enhance the differentiation between PCGs and other confused land covers such as PMFs and bareland. Moreover, we have designed a framework that integrates active learning into deep learning model, 115 improving the robustness of the latter when dealing with large-scale PCGs mapping tasks.

Overall, we proposed a novel framework to generate the firstly publicly released 10-m global PCGs map in 2020 derived from Sentinel-2. We also analyzed the spatial pattern of PCGs around the globe together with the driving force behind. Furthermore, we validated the accuracy of Global-PCG-10 and compared with other studies to further show its merits and demerits.

120 **2 Dataset**

2.1 Cropland Layer

It should be noted that almost all of the PCGs lie in the cropland, where other land covers such as forest, water bodies and grassland witness no PCGs. Therefore, we resort to the global cropland layer to eliminate the classification errors (i.e., mainly false positives) in the regions that have a very low probability of PCGs. Nonetheless, there still might be PCGs that 125 lie outside of the cropland layer. To tackle this issue, firstly, we divided the globe into a total of 2,592 grids with a size of 5° ×5° in the WGS84 projection, while retaining those grids that contained cropland cover. These retained grids extend, which is larger than the initial cropland, were designated as the first-level classification unit for data organization (i.e., blue grids in Figure 1). Each 5°×5° grid was further divided into 25 grids of 1°× 1°, which served as the second-level classification unit. Ultimately, we retained the grids that contained PCGs predictions in the second-level classification units (i.e., orange grids in 130 Figure 1). In specific, we compared a series of open-access global LULC maps and selected GLC_FCS30D (Zhang et al., 2023b) as cropland layer due to its good performance.
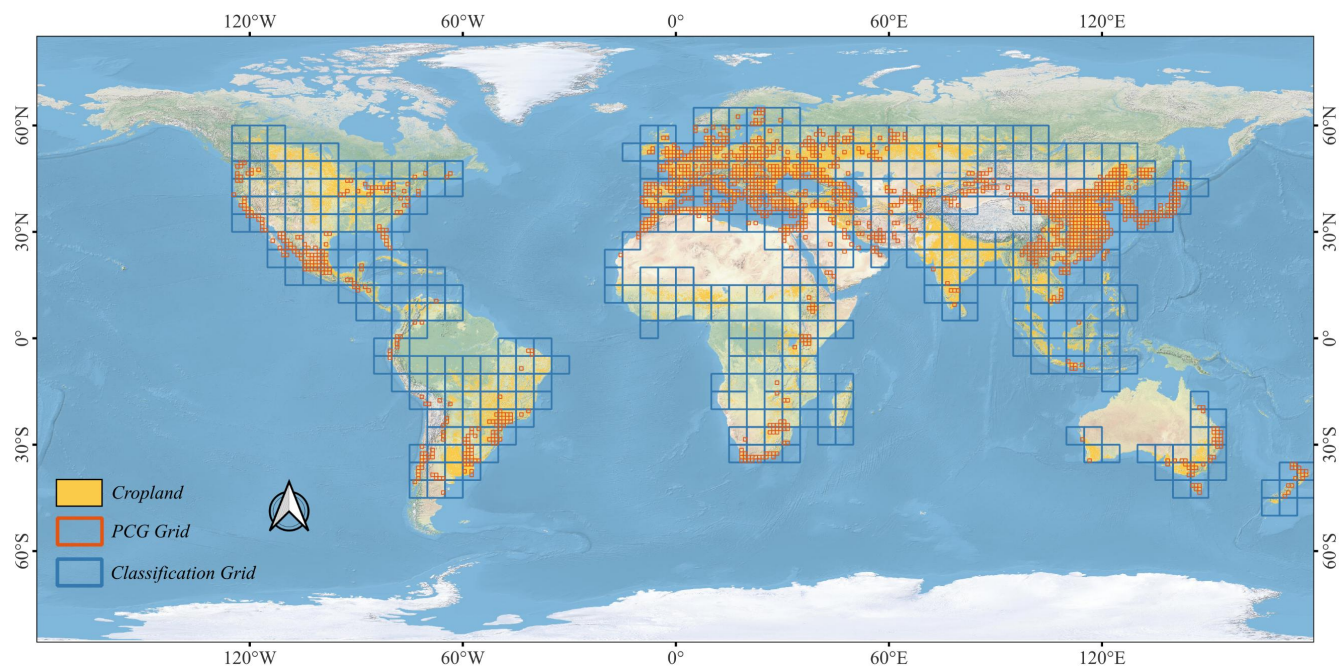
Figure 1. Spatial distribution of Cropland, PCGs grid and classification grid. (Note*: The yellow pixels indicate the cropland layer, which sources from the GLC_FCS30D cropland category (Zhang et al., 2023b), the orange grids stand for 1-degree grids that contain PCGs classification results, and the blue grids represent the original 5-degree grids used for PCGs classification.)

2.2 Satellite datasets

Sentinel-2 multispectral images were used in this study. As the important part of ESA's Copernicus Programme, Sentinel-2 aims to provide global Earth Observation data at a fine scale with 10 meters captured by MultiSpectral Instrument (MSI) with a total of 13 bands and a swath width of 290 km. Actually, Sentinel-2 is a constellation consisting of two satellites, i.e., Sentinel-2A and Sentinel-2B, which are in the same sun-synchronous orbit while phased at 180° to each other. Several reg-edge bands that are very sensitive to vegetation have been designed in Sentinel-2, which could capture a more detailed conditions of vegetated regions than other satellites such as Landsat and MODIS.

In addition, Sentinel-2 has two major merits over Landsat for PCGs mapping around the globe. Firstly, Sentinel-2 has a finer spatial resolution of 10-m. When compared with Landsat data at 30-m resolution, PCGs on Sentinel-2 images show a rather neat and tidy boundaries. Besides, Sentinel-2 witness much less mixed pixels than Landsat due to the increase of spatial resolution. Secondly, the revisit time of Sentinel-2 is 5 days at the equator while 2-3 days at mid-latitudes, which is much shorter than the 16 days of Landsat. The frequent revisit of Sentinel-2 is very important for large-scale PCGs mapping performance, since it increases the possibility to composite could-free images, especially for those cloudy and rainy regions around the world.

3 Method

150

Figure 2 depicts the overall workflow of this study, which consists of three stages. 1) Stage-1: generating initial PCGs labels via random forest classifier and the GEE cloud platform, which aims to release human labor in PCG label annotation; 2) Stage-2: producing accurate PCGs classification results through a deep learning model combined with an active learning strategy, which adopts a coarse-to-fine procedure to generate high-quality PCGs maps; 3) Stage-3: finalizing global PCGs
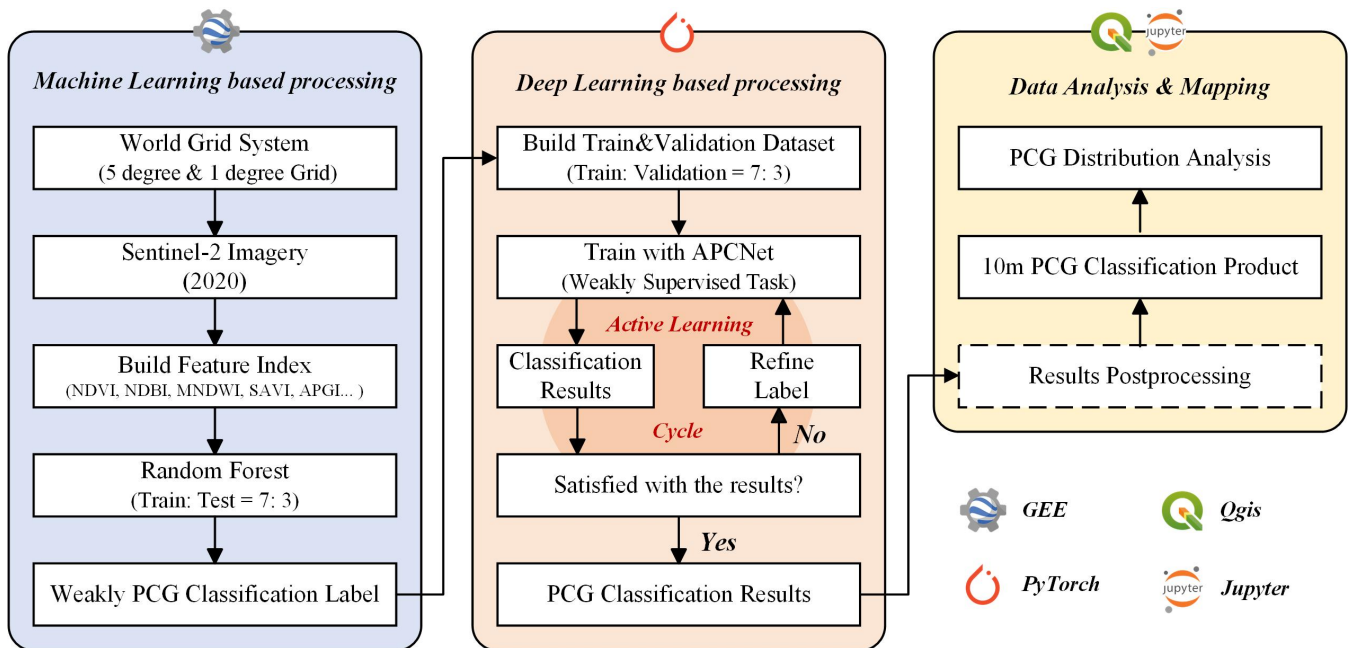
155 mapping and conducting spatial analysis.

Figure 2. Schematic flowchart to produce the Global-PCG-10 dataset.

Specifically, firstly, both PCG and non-PCG samples were labeled on GEE platform in the format of point. The locations of samples were acquired from both visual inspection on very high resolution images on Google Earth and our

160 previous field survey records. Multi-temporal and cloudless Sentinel-2 images in 2020 were composited by GEE, from which a multi-dimensional feature space for PCGs classification was constructed, including spectral indices (i.e., NDVI, MNDWI, etc.), texture features and recently published plastic greenhouse indexes (i.e., APGI, PGHI, etc.). The RF classifier was then used to generate initial labels. Afterwards, the label dataset was split into training and validation sets in 8:2, and the APC-Net model was built using the PyTorch framework. The initial and weak labels were refined through the active learning

165 strategy with the APC-Net model, continuously improving the PCGs classification performance. Finally, post-processing was applied to the PCGs classification results to eliminate isolated noises and then followed by spatial analysis and mapping.

3.1 Stage-1: PCGs weak label generation

In the field of large-scale remote sensing classification, the quantity and quality of labels are very important. However, if using only human annotation, it would be time-consuming to acquire enough samples for global PCGs classification. To

170   tackle this issue, we first employed GEE and random forest to generation the initial PCGs classification maps, from which samples (i.e., denoted as weak samples) are refined to train a deep semantic segmentation model.

Figure 3. Reference samples and their position in the world. (Note*. Blue circle represents the "salt and pepper effect" in the reference samples. The size of the reference samples is 512×512, and the background map of the *'Position'* column is from
175   **©Google Earth imagery.)**

The weakly labeled samples used in this study were generated through the RF classifier on GEE platform and sourced from regions with large amount of PCGs. The size of each sample is 512×512. Figure 3 shows several examples from six typical regions, including Weifang, China (Asia); Almeria, Spain (Europe); Uruapan, Mexico (North America); Campinas, Argentina (South America); Agadir, Morocco (Africa); and Coffs, Australia (Oceania). Figure 3 depicts that the "salt-and-
180 pepper effect" exists in RF classification results. This is also the reason why we introduce the second stage (i.e., deep learning & active learning) to refine PCGs maps. Notably, misclassifications are in areas with highly reflective surfaces, such as factory rooftops, beaches, deserts, and bareland.

3.1.1 Multi-temporal Sentinel-2 imagery collection

The Sentinel-2 images were loaded through *ee.ImageCollection()* function on GEE, and generated cloud-free images
185 for selected time periods by image property *"CLOUDY_PIXEL_PERCENTAGE"*, which could minimize the impact of cloud cover. Meanwhile, a total of seven bands (B1, B2, B3, B4, B8, B11, B12) from Sentinel-2 data are loaded for feature extraction and RF classifier.

Based on our previous research (Feng et al., 2022a), using Sentinel-2 satellite data for PCGs extraction often encounters confusion with PMFs and bareland. This is mainly due to the spectral similarities among PCGs, PMFs and bareland. To
190 address this issue, we introduced multi-temporal observations to enhance inter-class separability. Here, as an example, we selected 100 sample points for PCGs, PMFs and bareland in part of Gansu province, Northwest China (Figure 4). From the NDVI time-series spectral curves in 2020 (Figure 4c), it is observed that both spring (highlighted in grey) and summer (highlighted in green) witness the differences between PCGs, PMFs and bareland. As a result, we selected multi-spectral Sentinel-2 images in spring (April-June) and summer (July-September) periods as input data in this study area. By
195 incorporating multi-temporal data, we can mitigate the effect of feature confusion in single time-phase images, and get more precise PCGs classification results.
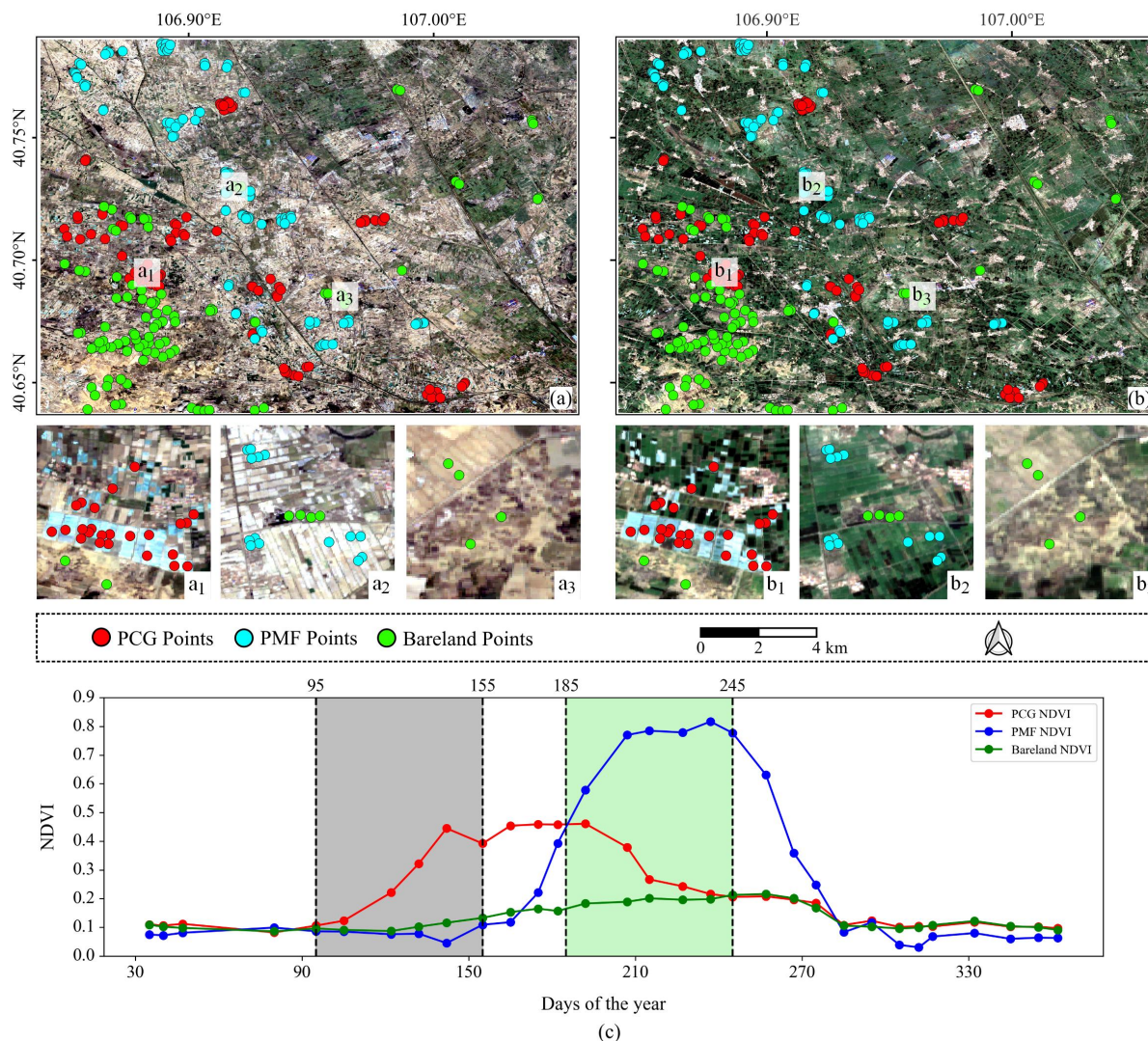
Figure 4. Multiple-temporal NDVI profile of bareland, PCGs and PMFs in a representative sub-region of Gansu Province, China. (a) Sentinel-2's spring true color image. (b) Sentinel-2's summer true color image. (c) Time-series NDVI value trend of bareland, PCGs and PMFs.

3.1.2 Feature extraction

The role of feature extraction is to transform remote sensing data from the original pixel space to the feature space, in which the difference and separability of PCGs and non-PCGs would be further enlarged. Specifically, a multi-dimensional and robust feature space is constructed considering the integration of spectral features and texture features.

1> Spectral features

According to our previous study (Feng et al., 2022a) and other relevant researches (Ou et al., 2021; Zhang et al., 2022a), we mainly consider the following spectral indices, including Normalized Difference Vegetation Index (NDVI) (Huang et al.,

2021), Soil Adjusted Vegetation Index (SAVI) (Huete, 1988), Normalized Difference Built-up Index (NDBI) (Zha et al., 2003) and Modified Normalized Difference Water Index (MNDWI) (Xu, 2006). Moreover, several recently published PCG

210   indexes are also included, consisting of PMLI (Lu et al., 2014), APGI (Zhang et al., 2022a), PGHI (Ji et al., 2020), PGI (Yang et al., 2017), and RPGI (Yang et al., 2017).

2> Textual features

It should be noted that PCGs have very distinct geometrical and textual characteristics, manifesting a rather regular rectangular appearance. Therefore, the inclusion of textual features could assist in the separation of PCGs from non-PCGs.

215   Specifically, we consider the six widely used textural features that derived from grey-level co-occurrence matrix (GLCM), including mean (MEA), standard deviation (STD), homogeneity (HOM), dissimilarity (DIS), entropy (ENT) and angular second moment (ASM).

3.1.3 Random Forest

Random Forest is utilized as the PCGs classification model to generate the initial samples. RF belongs to an ensemble

220   learning method while the base classifier is decision tree. The final output of RF is determined by the majority vote from all the decision trees involved. RF modeling involves two random selection steps. Firstly, the training samples of each base decision tree is randomly selected through bootstrapping. Secondly, the features used to split each node of the decision tree is also randomly selected. These two random processes effectively increase the robustness of RF on multi-dimensional data, and enable RF to cope with collinearity where the latter is an unavoidable issue in remote sensing data. Due to its simplicity

225   and robustness, RF has been widely adopted in remote sensing applications such as urban vegetation mapping, water extraction, crop classification and achieves promising performance (Mei et al., 2024; Sui et al., 2022; Zhang et al., 2024a). In this study, the RF classifier was configured using the *ee.Classifier.smileRandomForest()* function in GEE, the parameters set as follows: *numberOfTrees* was set to 150, and *variablesPerSplit* was set to 4.

3.2 Stage-2: Coarse-to-fine PCGs classification via deep learning

230   ### 3.2.1 APC-Net model

This study utilized our previously proposed deep semantic segmentation model, APC-Net, as the core model to derive final PCGs maps in a coarse-to-fine manner (Niu et al., 2023a). This network effectively combines local with global information to improve the model's capability in complex landscapes worldwide through multi-scale feature learning.
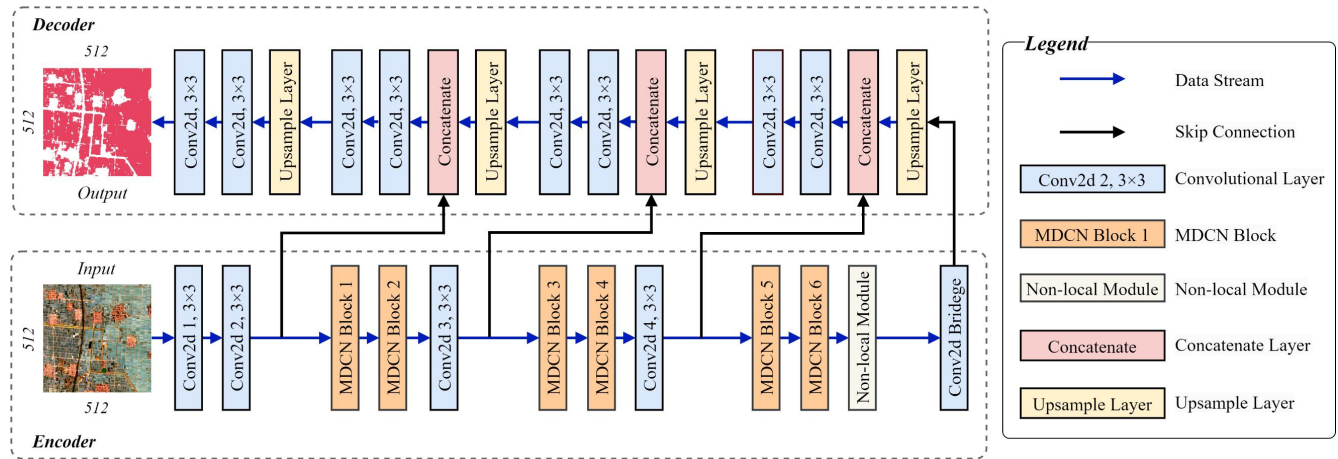
235 **Figure 5. Overview of proposed APC-Net model**

Specifically, APC-Net is divided into two major parts: the encoder and the decoder (Figure 5). The encoder, which serve as the core component of APC-Net, receives remote sensing images with a size of 512×512 and extracts highly representative features using its multi-level network structure, which improves the internal consistency within the same land covers while enhancing the distinction between different land cover categories. The encoder consists of convolutional layers,

240 MDCN (i.e., multi-scale dilated convolutional network) modules and a non-local module. Herein, the MDCN modules effectively learn local features at multiple scales by integrating multi-scale dilation convolution, addressing the issue of scale variation in PCGs classification. The non-local module focuses on capturing global contextual information and enhances the model's understanding of the whole image scene. The decoder's role is to recover information from the downsampled feature maps generated by the encoder and produce the final segmentation map. The decoder employs bilinear interpolation

245 for upsampling and skip connections from encoder to further fuse and refine the features. Finally, the output is PCGs classification results with the same size (512×512) as the input images.

3.2.2 Active Learning strategy

In this study, the active learning strategy is employed to optimize initial labels by refining and reorganizing via human intervention. It aims to reduce the false-positive rate hence to improve the classification accuracy. It works as follows. First,

250 the APC-Net model is trained on the initial weak labels (Input Dataset in Figure 6), which are generated from the RF classifier, and saving the best model weights. Then, these weights are applied to predict the results of the input dataset, producing a set of updated labels. Subsequently, the classification performance is evaluated according to both accuracy evaluation and visual inspection. If the results do not meet the expected standard, initial labels with significant updates are selected to form a new input training dataset. The process is repeated until satisfactory results are achieved, or until

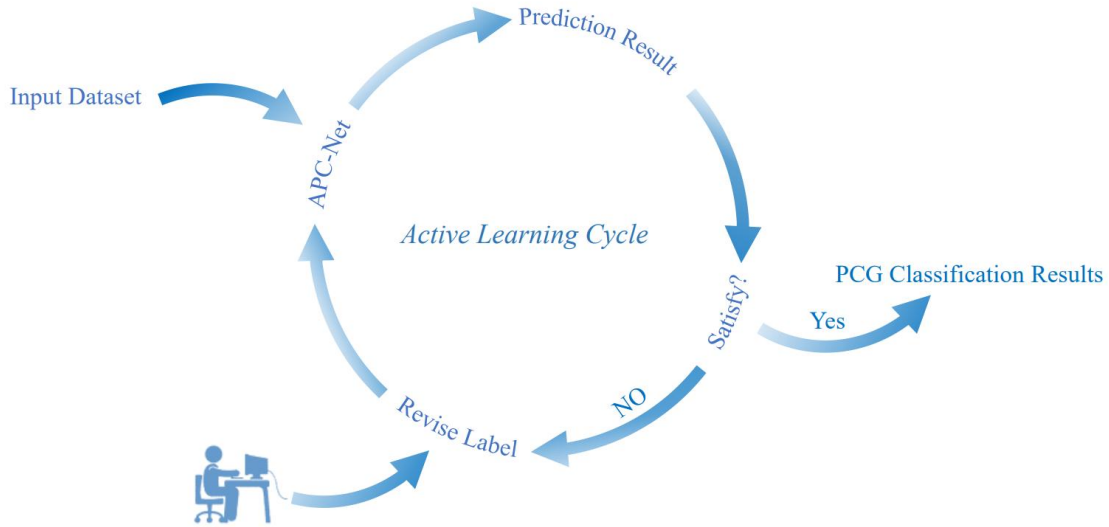255 performance stabilizes with no further improvements.

Figure 6. Active learning strategy utilized.

3.2.3 Training Details

In this section, we provide a detailed description of the model training process, including the design of the hybrid loss
260   function, the choice of optimizer and the configuration of model hyperparameters. Specifically, this study combines Cross
Entropy Loss (CE Loss) and Dice Loss to create a hybrid loss function for the PCGs semantic segmentation task. In this
hybrid loss function, CE loss primarily measures the discrepancy between labels and model predictions, while Dice loss
mitigate the issue of class imbalance, ensuring robust model performance even dealing with underrepresented categories.

$$CE\ Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{1}$$

$$Dice\ Loss = 1 - \frac{2\sum_{i=1}^{N} y_i\,p_i + \varepsilon}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} p_i^2 + \varepsilon} \tag{2}$$

where N denotes the number of samples, y_i is the true label of the i th sample, p_i is the corresponding predicted
probability, and ε is a very small constant in case of a division by zero error.

A hybrid loss function is designed by combining the merits of CE loss and Dice loss for semantic segmentation. The
formula as follows, where α represents the weight ratio between the two loss terms, equals to 0.2.
270   $$Hybrid\ Loss\ =\ \alpha \times CE\ Loss\ +\ (1 - \alpha) \times Dice\ Loss \tag{3}$$

Regarding the optimizer, the widely-used Adam optimizer was applied for training APC-Net, with an initial learning
rate of 1e-4. APC-Net was constructed using the PyTorch 1.12.1 framework. The dataset was divided into training and
validation sets using an 8:2 ratio. The training set comprises 14,825 samples, while the validation set includes 3,707 samples.

Each sample has a resolution of 512 × 512 pixels. Additionally, the model was trained with a GPU of NVIDIA GeForce
275 RTX 3090 with 24 GB memory, the Intel Core i7-12700KF CPU@5.00 GHz, and the Ubuntu 20.04 operating system.

3.3 Accuracy assessment

We adopt both qualitative and quantitative accuracy assessment to justify the classification performance of Global-
PCG-10. The former is to compare Global-PCG-10 with remote sensing images to check the obvious classification errors,
while the latter is to calculate a series of accuracy metrics including overall accuracy (OA), recall, precision, F1-score, etc.
280 Actually, OA is derived from the confusion matrix that are calculated in the test dataset, whose formulas are as follows.

$$OA = \frac{1}{N} \sum_{i=1}^{r} x_{ii} \tag{4}$$

where N denotes the total number of samples, r represents the number of classes, and x_{ii} refers to the diagonal elements
of the confusion matrix.

Considering that PCGs mapping belongs to a binary classification problem, therefore, the widely used metrics in binary
285 classification such a recall, precision and F1-score would also be a good choice to justify the performance of Global-PCG-10.
These metrics have also been used by Fu et al. (2021), where the three metrics are used for the accuracy evaluation of
China's marine aquaculture mapping results.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

290

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

where TP represents true positives, i.e., the number of correctly classified PCGs pixels, FN denotes false negatives, i.e.,
the number of PCGs pixels misclassified into non-PCGs, while FP stands for false negatives, i.e., the number of non-PCGs
pixels misclassified into PCGs. In general, recall and precision are contradictory to each other. A high recall also brings in a
high FP, which would lead to low prevision. On the other hand, F1-score is an integrated index that takes into consideration
295 of both recall and precision. F1-score has a value between 0 and 1, where a higher F1-score means a better classification
performance.

4 Results and discussion

4.1 Spatial pattern of Global-PCG-10

Figure 7 illustrates the global distribution of PCGs in the mapping unit of 0.1° grid instead of using per-pixel PCGs
300 classification results. This is because the predicted PCGs cover only a small fraction of the entire globe. If we put the per-
pixel results on map, PCGs would be overwhelmed by background regions. To tackle this issue, we change the mapping unit

from per-pixel (i.e., 10 m) to 0.1° grid through zonal analysis, which could enhance the visual effect of PCGs spatial distribution globally.

As depicted in Figure 7, the global PCGs mainly locate in East Asia and Mediterranean regions. Specifically, in East
305 Asia, China has the largest area of PCGs, where most PCGs are clustered in North China Plain (Eastern China), Liaohe Plain (Northeastern China), Sichuan Basin and intermountain basin of Yunnan Province (Southwestern China). In Mediterranean region, PCGs are mainly distributed along the coasts in Iberian Peninsula, Apennine Peninsula, Balkan Peninsula and Nile Delta. The widespread presence of PCGs in these regions, where are characterized by both a well-developed and a long history of farming, can be attributed to two key factors. First, the use of PCGs allows for the expansion of both acreage and
310 production of high-quality vegetables, fruits, flowers, and other cash crops. This is particularly beneficial in China, where it could effectively increase income of local farmers. Second, most of these PCGs are located in plains or basins, close to urban areas, and have relatively abundant water resources. These geographic advantages provide favorable conditions for irrigation and product marketing, which could help to ensure the efficiency and output of facility-based agriculture.

Meanwhile, Figure 7 also witnesses several regions with nearly no PCGs, including North America, Northern Eurasia,
315 Sub-Saharan Africa and Oceania. Two reasons may account for this. On one hand, in North America, the agricultural mode is large farms facilitated with advanced agricultural machinery and less workers. Considering that PCGs are rather labor intensive and not easy for machinery to work, therefore, they are not widespread in both United States and Canada. On the other hand, in areas like South America, sub-Saharan Africa, Northern Eurasia and Oceania, the lower level of agricultural development and limited infrastructure hinder the adoption and growth of PCGs. Additionally, in China and along the
320 Mediterranean coast region, profit-driven small holders are the majority. Under this circumstance, together with policy incentive, farmers choose to build PCGs to produce cash crops, leading to the prevalence of PCGs in these regions.

In addition, we calculate the PCG area along both longitude and latitude in an interval of 1° and depict the area histogram in Figure 7. It indicates that the global PCGs mainly locate in the Northern Hemisphere, especially between 30°N and 40°N with a peak at about 36°N, which accounts for 65.84% of the total PCGs area. Meanwhile, these regions just
325 correspond to North China Plain and Mediterranean region. From the perspective of longitude, most PCGs are clustered in the Eastern Hemisphere, while the Western Hemisphere only witnesses a high PCG density on the west side of the Mexican Plateau, the west side of the Chilean Cordillera, and the La Plata Plain (river inlets) of Argentina.
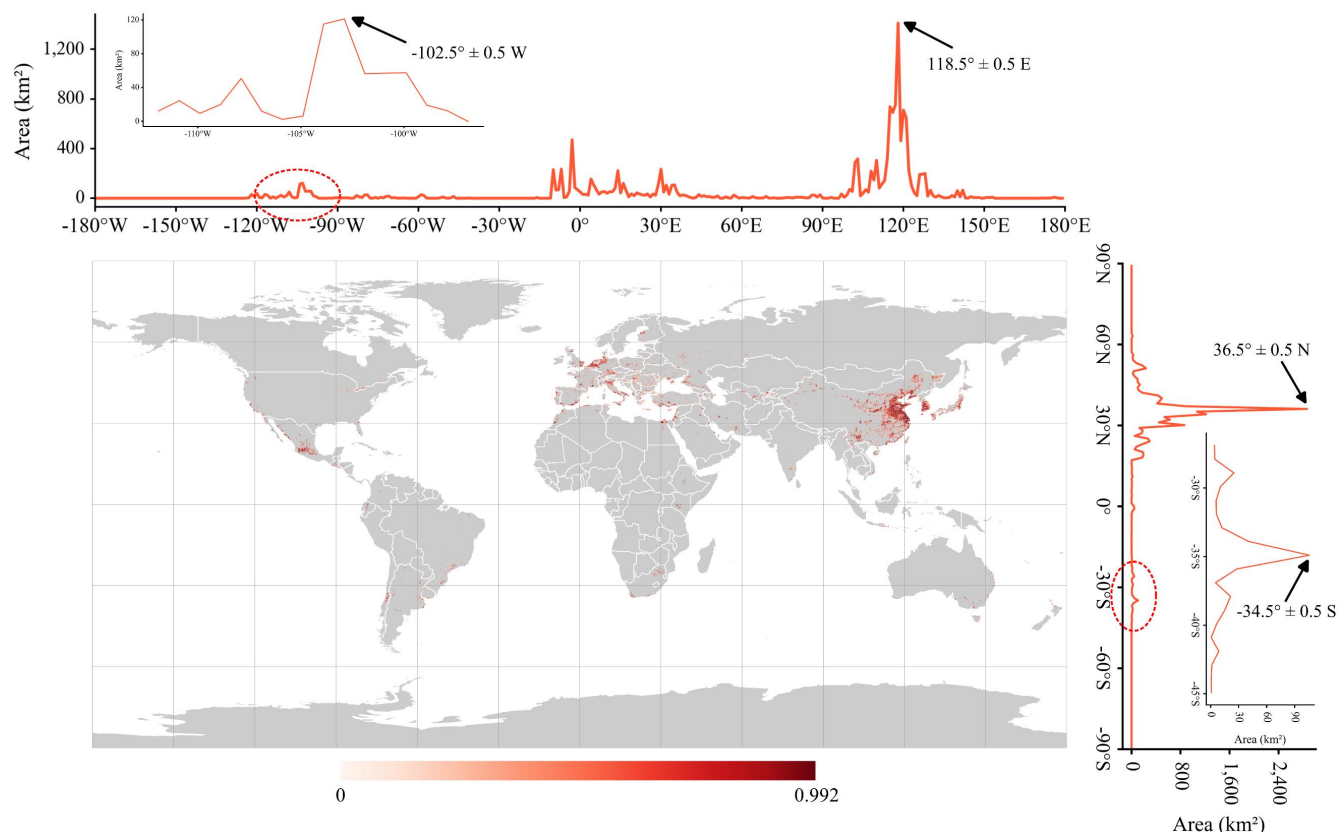
Figure 7. Global PCGs spatial distribution in 2020. The spatial resolution of the map is 0.1°.

330   Moreover, we calculate both the PCGs area and the ratio of PCGs to cropland area in each country. The former reflects the production scale of PCGs while the latter stands for the proportion or importance of PCGs in local agricultural activities. The total area of global PCGs reached to 14,259.85 km² in 2020, while Asia has the largest PCGs area of 9874.51 km², accounting for 69.24% of the total global PCGs. Europe ranks the second with a PCGs area of 2530.56 km² and a PCGs ratio of 17.75%. North America, Africa, South America and Oceania witness a decent PCG area of 819.12 km², 668.82 km²,

335   213.92 km² and 152.91 km², respectively. From the perspective of country, China ranks the first with a PCG area of 8224.90 km². Meanwhile, China accounts for 83.29% of PCGs in Asia and 57.67% in the globe. Spain ranks the second in the world and the first in Europe with a PCG area of 803.26 km². Other countries with a PCGs area over 500 km² include Mexico, Italy and South Korea. On the contrary, countries in Sub-Saharan Africa, Central Asia and other countries like Mongolia, Russia, United States and Canada, have very few PCGs.
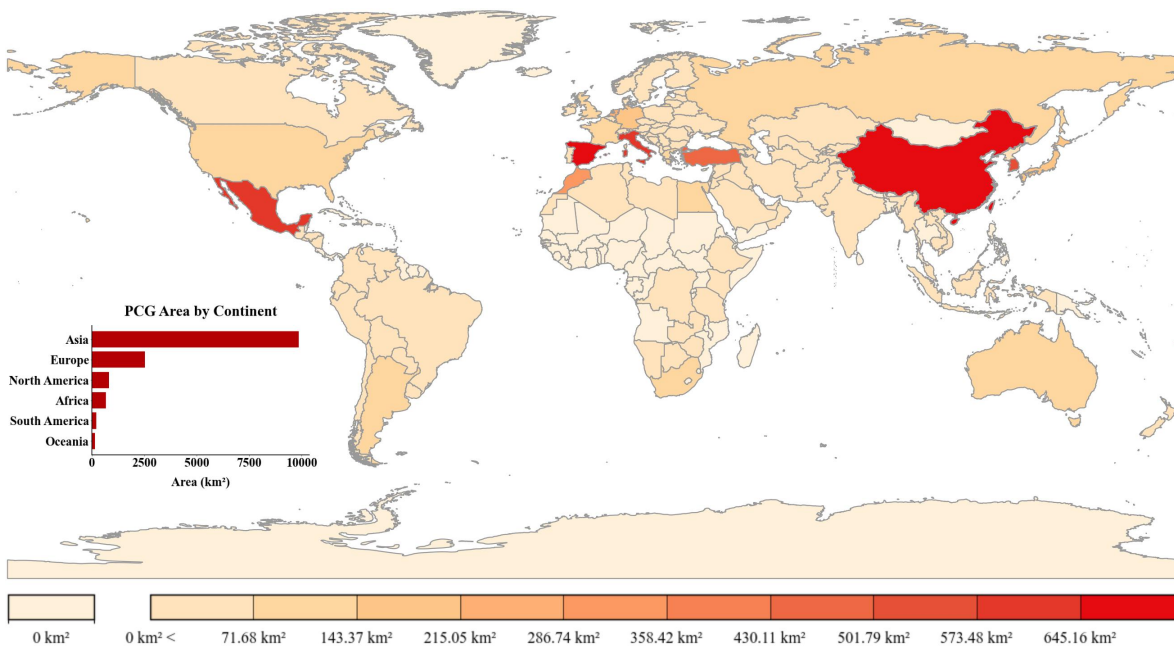
(a)
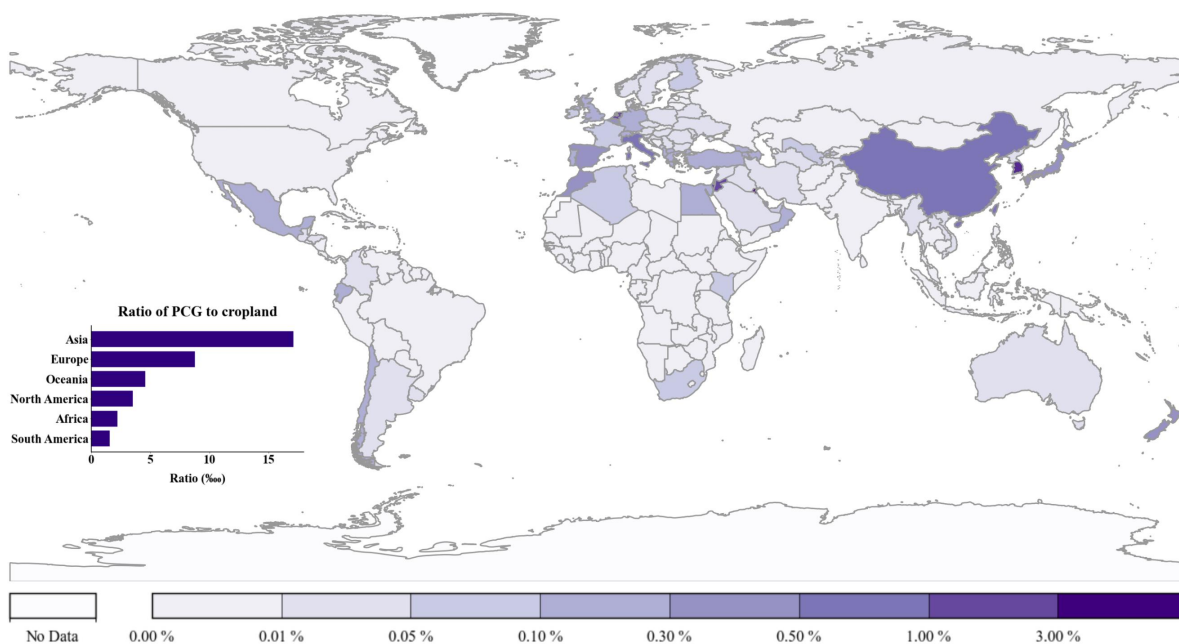
(b)

Figure 8. PCGs area statistic mapping in 2020. (a) Area of PCGs. (b) Ratio of PCGs to cropland.

Figure 8b illustrates the ratio of PCGs area to cropland. It indicates that although China has the largest PCGs region, its PCGs ratio (0.64%) is relatively lower. The country with the highest PCGs ratio is Kuwait (5.23%). Other Mediterranean

16

countries such as Italy and Turkey also have a relatively higher PCG ratio. In Eastern Asia, South Korea has a high PCGs ratio although they have small PCGs regions, which manifests the important role of PCGs in these countries. The main reason is that these countries are mostly situated in mountainous, hilly or Gobi terrain conditions, leading to a limited

350 amount of usable cropland. In such conditions, PCGs, as a form of intensive facility-based agriculture, can overcome the limitations of the local natural climate. It also effectively optimizes the structure of the local agriculture industry, increases the diversity of agricultural products, and reduces reliance on imported fruits, vegetables, and other cash crops.

4.2 Reliability of Global-PCG-10

This section would analyze and justify the reliability of Global-PCG-10 from both visual inspection and accuracy

355 assessment results. Here, we present both the remote sensing images and the corresponding Global-PCG-10 maps from several typical regions across the world, including Shouguang of China, Lam Dong of Vietnam in Asia; Almeria of Spain, Ragasu of Italy in Europe; Outlet of Morocco, Damous of Algeria in Africa; Samis of the U.S., Zapotitan of Mexico in North America; Lisandro of Argentina, Bom Repouso of Brazil in South America; Robinvale of Australia, Manngatepere of New Zealand in Oceania. As shown in the Figure 9, our classification results are excellent in regions with a high density of

360 greenhouses, with virtually no noticeable omission of PCGs. Furthermore, no significant false-positive errors were observed, even in areas with a sparse distribution of PCGs (e.g., the United States, Brazil, etc.). Additionally, Global-PCG-10 demonstrates reliable recognition of PCGs in various global climate zones, including humid climates, Mediterranean climates, and others.

Details of Global-PCG-10 illustrate that we have achieved a very good PCGs map with accurate and neat boundaries

365 under the spatial resolution of 10 m. The confusion between PCGs and non- PCGs is not obvious and the speckle noises in the background have been greatly suppressed. Two reasons may account for this. First, the utilization of multi-temporal Sentinel-2 satellite image could reduce the misclassification and "salt and pepper effect" caused by PMFs, bareland, and other land cover classes. Second, the PCGs classification framework, which integrates active learning strategy into the deep learning model, enables a coarse-to-fine classification process.
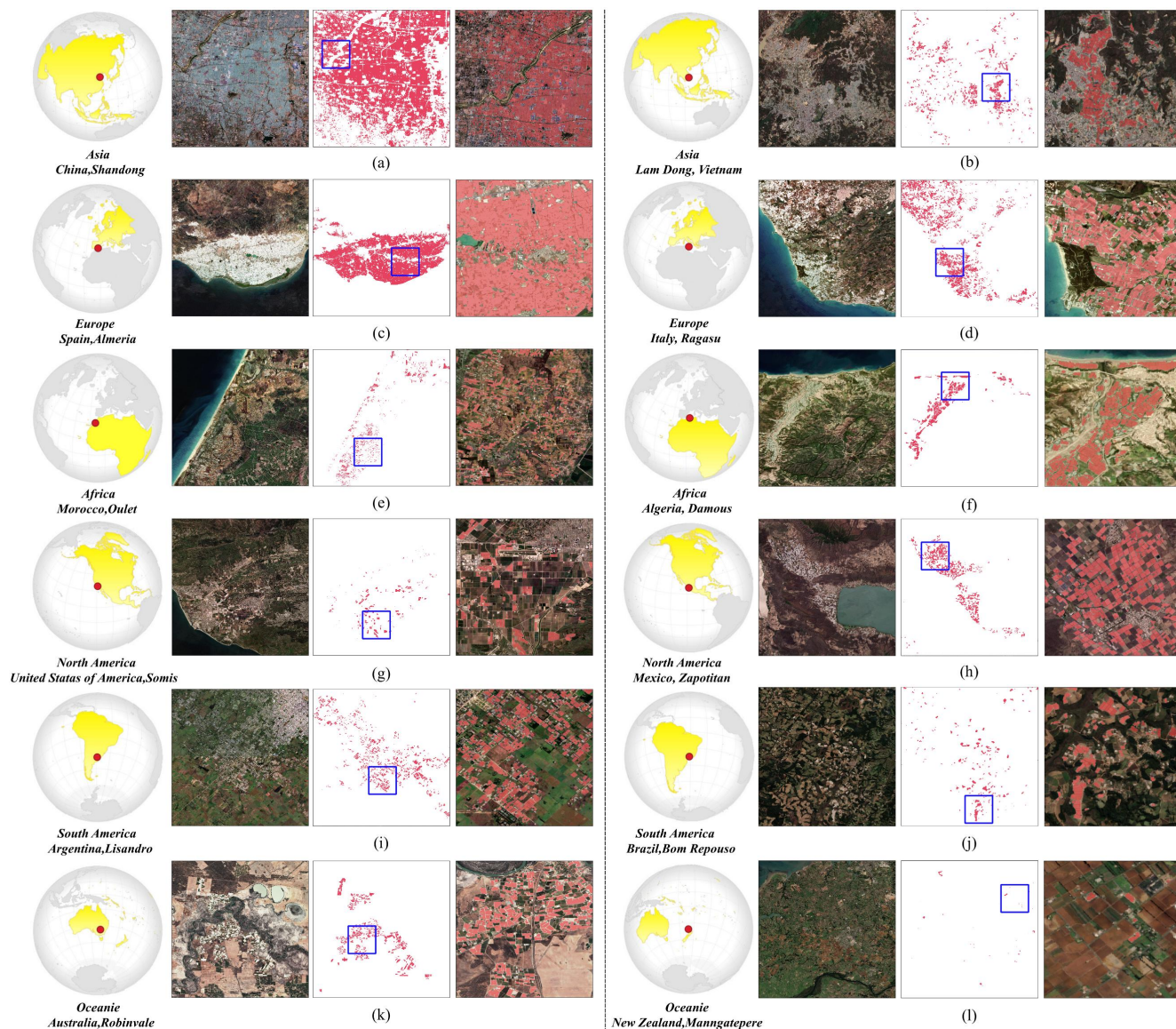
Figure 9. Details of the Global-PCG-10 map. (Note*. Taking (a) as an example, from left to right is the location in the world, RGB Sentinel-2 image, Global-PCG-10 map and detailed fused mapping result of the blue rectangle region.)

To further justify the reliability of Global-PCG-10 from a quantitative perspective, we have selected a global test dataset for accuracy assessment based on the sampling grid (Figure 10). The test dataset includes both PCGs and non- PCGs samples, with a number of 20,500 for each category. It should be noted that all the test samples and training samples are spatially independent from each other with no overlap to maintain the reasonability.
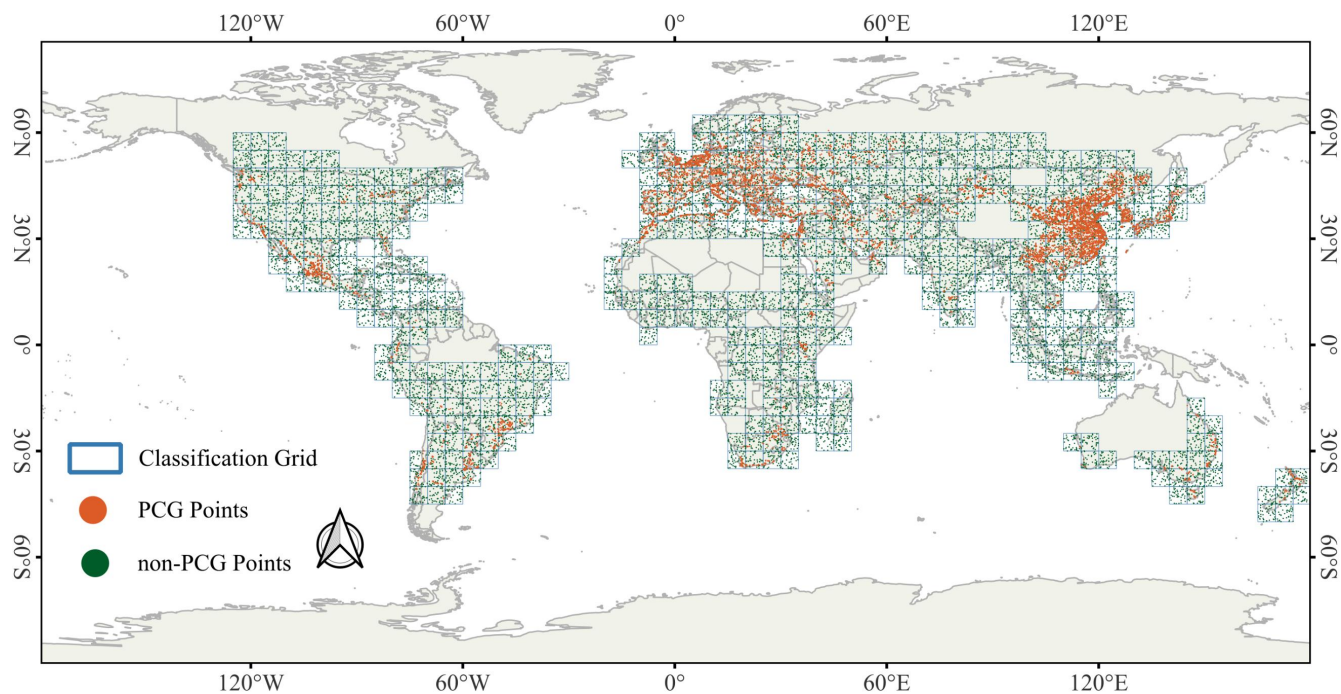
18

Figure 10. Spatial distribution of global test samples.

Table 1 indicates that Global-PCG-10 yields a high accuracy with an OA of 92.08%. Since the test dataset covers the
380   entire globe with a sufficient amount, it could verify the high quality of Global-PCG-10.

Table 1. Confusion matrix.

Classification results	Validation Data		
	Non-PCG	PCG	UA (%)
Non-PCG	20498	2	99.99
PCG	3243	17257	84.18
PA (%)	86.30	99.99	
OA (%)	92.08		

Note*: PA, Producer's Accuracy; UA, User's Accuracy; OA, Overall Accuracy.

We also calculated other accuracy metrics and the results are in Table 2. It indicates that Global-PCG-10 has both a
high recall, precision and F1-score of 84.18%, 99.99% and 91.41%, respectively, which further justifies the performance of
385   the published dataset.

Table 2. Accuracy indicators.

Dataset	Recall (%)	Precision (%)	F1-score (%)	OA (%)
Global-PCG-10	84.18	99.99	91.41	92.08

4.3 Comparison with other studies

4.3.1 Comparison with global LULC dataset

As a distinct land cover category, PCGs should belong to cropland. However, to the best of our knowledge, PCG has
390 not been fully considered in previously released global LULC datasets. Therefore, in this section, we compare our Global-
PCG-10 with other released global LULC datasets. To ensure temporal consistency, we only selected LULC datasets
containing 2020 data products comparison, including Dynamic World (Brown et al., 2022), ESA World Cover (Zanaga et al.,
2021), ESRI Land Cover (Karra et al., 2021), FROM-GLC30 (Yu et al., 2022), and GLC_FCS30D (Zhang et al., 2023b).
Among these, Dynamic World, ESA World Cover, and ESRI Land Cover have a spatial resolution of 10 m, while FROM-
395 GLC30 and GLC_FCS30D have a spatial resolution of 30 m. Since the classification system of each dataset is different, this
study is based on the land cover types of the selected typical regions and uses GLC_FCS30 as a reference. And the land
cover types are unified into the following nine categories: Cropland, Shrubland, Grassland, Bareland, Wetland, Impervious
surface, Forest, Water Body and Ice/snow.

As shown in Figure 11, the left part presents the Sentinel-2 true color images of typical regions along with their
400 corresponding Global-PCG-10 maps, while the right part displays the mapping results from various LULC datasets. It is
quite clear that these LULC datasets have low classification performance in PCGs regions. Specifically, Dynamic World
erroneously classifies PCGs as impervious surfaces (Figure 11a, b, f, g), Ice/snow (Figure 11c, d, f, h), and bareland (Figure
11i). ESA World Cover misclassifies PCGs as impervious surfaces (Figure 11d, h) and bareland (Figure 11a, e, i). ESRI
Land Cover misclassifies PCGs as impervious surfaces (Figure 11b, f, i), bareland (Figure 11d), and grassland (Figure 11h).
405 FROM-GLC30 misclassifies them as impervious surfaces (Figure 11a, f, i, j) and grassland (Figure 11a, b, i, j).
GLC_FCS30D similarly misclassifies PCGs as impervious surfaces (Figure 11d, f, h).

Above all, PCGs are commonly misclassified into four categories: impervious surfaces, bareland, grassland, and
Ice/snow. Specifically, impervious surfaces and bareland, such as white-roofed factories, villages, and photovoltaic panels,
may share similar spectral or texture features with PCGs, leading to obvious misclassification. Additionally, the phenology
410 of crops grown within greenhouses can also affect the spectral features of PCGs, making them resemble grassland at certain
times and leading to misclassification. Besides, the reflectance of PCGs in some regions is similar to that of clouds and snow,
which might explain why PCGs are sometimes misclassified into these categories.

Meanwhile, Figure 11 also indicates some LULC datasets exhibit good performance in classifying PCGs into cropland.
For instance, Figure 11b, c of ESA World Cover, Figure 11a, c, e of ESRI Land Cover, Figure 11d, e, h of FROM-GLC30,
415 and Figure 11a, b, c, e, g, i of GLC_FCS30D. The classification results successfully identify PCGs as cropland with greater
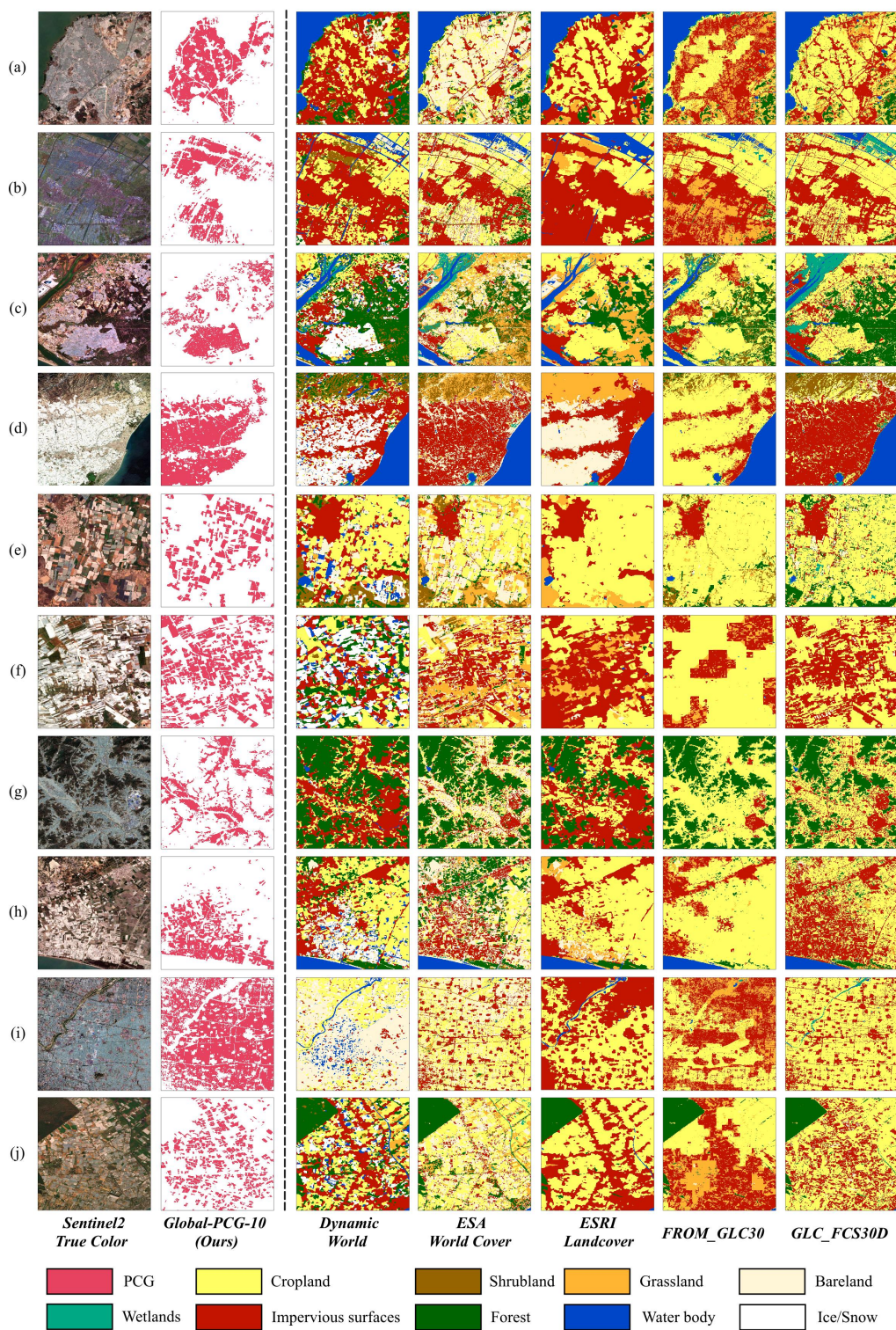precision in these cases.

Figure 11. Comparison of Global-PCG-10 with other global LULC products.

4.3.2 Comparison with other PCG dataset

420   Research on large-scale extraction of PCGs includes several excellent efforts. Our previous study released the first 30-m national-scale PCGs dataset of China in 2019 (Feng et al., 2022a), and in May 2024, the University of Copenhagen published a global 3-m PCGs dataset also in 2019, both of which are open access. Additionally, Wuhan University conducted PCGs mapping in 2016 using high-resolution satellite data and deep learning techniques (Chen et al., 2023; Ma et al., 2021), and the Chinese Academy of Sciences conducted a nearly 20-year extraction and spatial analysis of PCGs in
425   China using Landsat 5/8 data on GEE cloud platform (Liu and Xin, 2023; Ou et al., 2019). However, these two studies' datasets have not been publicly released. Therefore, this study only compares the 3-m PCGs dataset from the University of Copenhagen.
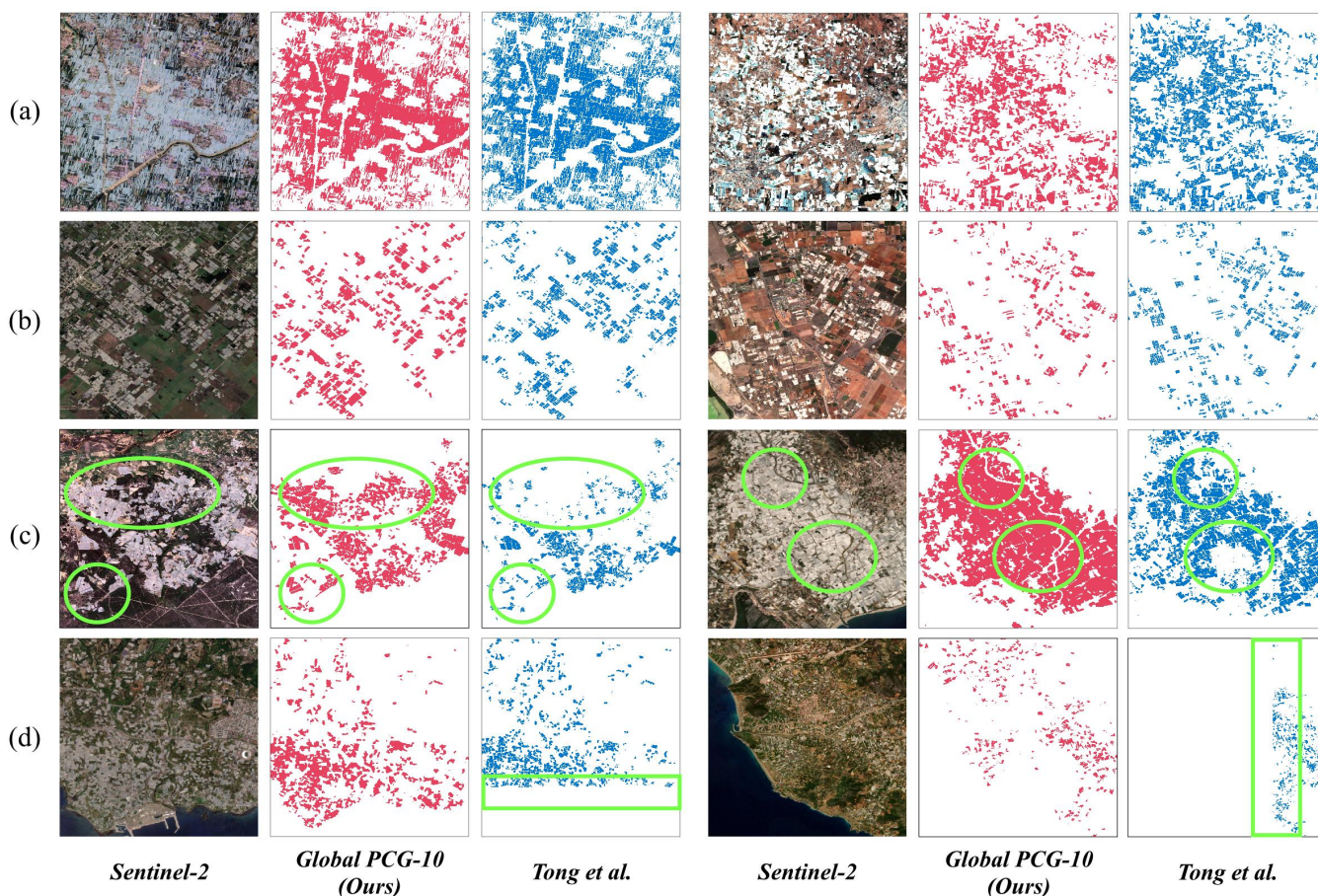
Figure 12. Comparison with other PCGs dataset. (Note*. Light red depicts our Global PCG-10, light blue represents Tong et al.,
430   **and green circles/rectangles mark the differences.)**

Figure 12a and b illustrate the mapping results in regions with different PCGs densities. In Figure 12a, it shows both datasets exhibit very similar spatial patterns in densely distributed PCGs regions. Figure 12b demonstrates that Global-PCG-10 can still accurately capture the greenhouse layout even in areas with sparse PCGs distribution, performing on par with

Tong et al., (2024) acquired from 3-m commercial satellites. Figure 12c highlights the case where Global-PCG-10 accurately detected greenhouses while Tong et al., (2024) missed. Figure 12d also shows a case of missed detection by Tong's dataset, but unlike Figure 12c, the missed detection in Figure 12d is more likely due to missing of satellite image. The rectangle indicates that the data is split into two parts by an "invisible" line. Such situation is commonly caused by improper data organization or gaps of source satellite data. On the contrary, we used Sentinel-2 satellite data, with its dual constellation providing a 5-day revisit cycle, which ensures minimal gaps in remote sensing imagery, thereby delivering high-quality data support for PCGs mapping globally.

4.4 Advantages and limitations of Global-PCG-10

As previously mentioned, Global-PCG-10 is a globally comprehensive, high-quality dataset built on open-source Sentinel-2 data, demonstrating high classification accuracy. The Global-PCG-10 utilizes free and open-source Sentinel-2 data, significantly reducing dataset production costs. Additionally, it features a more organized data structure. In the future, this dataset is expected to become an important data source in the fields of agriculture, environment, ecology, and land use and cover change (LUCC) research. To enhance mapping precision, this study introduces multi-temporal remote sensing imagery data, which effectively mitigates the confusion of PMFs and bareland with PCGs. Additionally, we adopted an approach that combines the active learning strategy and the deep learning model. This method allows the model to learn more robust PCG features by optimizing weak labels, which could significantly reduce false positives and improve the classification performance.

However, there are still some limitations of Global-PCG-10. Firstly, small sized or isolated PCGs are not fully extracted due to the limitation of 10 m spatial resolution of Sentinel-2 data. Future study would introduce remote sensing data with an even higher spatial resolution than 10 m to tackle this issue. Secondly, we now only provide the global PCGs map for the year of 2020. In the future, we would insist in producing time-series global PCGs map for further understanding of PCGs and its environmental effect. Thirdly, the technical approach proposed in our research, which integrates deep learning with active learning strategies, is still a semi-automatic weak label updating process and has not yet achieved fully end-to-end task training. In future research, we will explore end-to-end weak label updating methods to achieve a more efficient and automated data processing workflow.

5 Code and data availability

The code for generating the initial labels of PCGs is publicly available via the following link on Google Earth Engine: *https://github.com/MrSuperNiu/Greenhouse_Classification_GEE*. It consists of feature extraction, RF classification, etc. Additionally, the code of APC-Net is accessible through the following link: *https://github.com/MrSuperNiu/APCNet*.

The Global-PCG-10 dataset is stored on figshare, and can be downloaded here: *https://doi.org/10.6084/m9.figshare.27731148.v2* (Niu et al., 2024). The dataset contains 245 5°×5° grids, and each of them is named using the grid's *'id'* attribute. Within each 5°×5° grid file, there are 1°×1° TIF files, named in the format

gridID_subgridID_PCG_Result.tif. Here, *gridID* represents the *'id'* of the 5°×5° grid containing the 1°×1° subgrid, and *subgridID* represents the *'id'* of the 1°×1° grid. The 5°×5° grids and 1°×1° grids, corresponding to the *Classification Grid* and *PCG Grid* in Figure 1, are saved in SHP format within the Global-PCG-10 dataset to facilitate users in indexing the corresponding PCG classification **.tif* files. Additionally, the cropland and PCG area statistic data are contained in the Excel

470   file, *SATA_Cropland&PCG.xlsx*. Please refer to the attached supplementary document *Supplementary File.docx* for more detailed data organization information.

6 Conclusions

As an important representative of facility agriculture, PCGs play a crucial role in enhancing crop yields and increasing local agricultural income through their ability to retain soil moisture and temperature. This study constructed the first global

475   PCGs dataset with a high spatial resolution of 10 meters based on deep learning and active learning. Specifically, we first divided the globe into 2,592 grids with a size of 5°×5° and retained those containing cropland as classification units. Then, we obtained pre-processed multi-temporal Sentinel-2 data through GEE and used random forest to generate initial labels to build the training dataset. Next, we developed a classification workflow that integrates the active-learning and deep learning to optimize weak labels, enhance model robustness, and reduce false positives. Subsequently, we used the trained deep

480   learning model to predict the global distribution of PCGs, generating the Global-PCG-10 dataset. Finally, we analyzed the spatial distribution patterns and driving forces of global PCGs, compared the proposed dataset with other open-source datasets.

Experimental results show that the global PCGs area is approximately 14,259.85 km² in 2020. PCGs are mainly distributed between 30°N and 40°N, accounting for about 65.84% of the total area. Asia holds the most extensive area of

485   PCGs, covering approximately 9874.5 km², accounting for 69.24% of the global total. China, not only has the largest area of PCGs in Asia but also ranks first worldwide, with a PCGs area of 8,224.90 km², making up 57.67% of the global and 83.29% of the Asia. We validated the Global-PCG-10 dataset using 40,500 randomly sampled points, which indicates that the overall accuracy is satisfactory of 92.08%.

Additionally, we compared the Global-PCG-10 dataset with several open-source global LULC datasets. The findings

490   reveal that PCGs, which should be classified as cropland, are often misclassified as bareland, impervious surfaces, Ice/snow, and grassland in those LULC datasets, which could negatively impact the estimation of global cropland areas. Compared to other publicly available PCG datasets, Global-PCG-10 demonstrates excellent accuracy in the distribution of PCGs. Besides, it offers a better data organization for relevant researchers. In future research, we will continue generating time-series global PCGs maps from multi-source remote sensing data.

495   **Author contributions**

Bowen Niu: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Sampling, Data Curation, Writing, Visualization. **Quanlong Feng**: Conceptualization, Methodology, Formal analysis, Investigation, Writing,

Supervision, Project administration. **Bingwen Qiu:** Methodology, Formal analysis, Writing, Supervision, **Shuai Su, Xinmin Zhang and Rongji Cui:** Investigation, Resources, Sampling. **Xinhong Zhang, Fanli Sun, Wenhui Yan, Siyuan Zhao,**

500 **Hanyu Shi and Xiaolu Yan:** Resources, Sampling. **Cong Ou, Jianhua Gong, Gaofei Yin, Jianxi Huang, Jiantao Liu, Bingbo Gao, Xiaochuang Yao, Jianyu Yang and Dehai Zhu:** Data Curation.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

505   Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

510 ## Financial support

References

Aguilar, M., Nemmaoui, A., Novelli, A., Aguilar, F., and García Lorca, A.: Object-Based Greenhouse Mapping Using Very High Resolution Satellite Data and Landsat 8 Time Series, Remote Sens., 8, 513, https://doi.org/10.3390/rs8060513, 2016.

515 Aguilar, M. A., Jiménez-Lao, R., Ladisa, C., Aguilar, F. J., and Tarantino, E.: Comparison of spectral indices extracted from Sentinel-2 images to map plastic covered greenhouses through an object-based approach, GIScience Remote Sens., 59, 822–842, https://doi.org/10.1080/15481603.2022.2071057, 2022.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., and Tait, A. M.:

520 Dynamic World, Near real-time global 10 m land use land cover mapping, Sci. Data, 9, 251, https://doi.org/10.1038/s41597-022-01307-4, 2022.

Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., Gong, J., and Liu, J.: Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network, Int. J. Appl. Earth Obs. Geoinformation, 109, 102794, https://doi.org/10.1016/j.jag.2022.102794, 2022.

525   Chen, D., Ma, A., Zheng, Z., and Zhong, Y.: Large-scale agricultural greenhouse extraction for remote sensing imagery based on layout attention network: A case study of China, ISPRS J. Photogramm. Remote Sens., 200, 73–88, https://doi.org/10.1016/j.isprsjprs.2023.04.020, 2023.

Chen, W., Xu, Y., Zhang, Z., Yang, L., Pan, X., and Jia, Z.: Mapping agricultural plastic greenhouses using Google Earth images and deep learning, Comput. Electron. Agric., 191, 106552, https://doi.org/10.1016/j.compag.2021.106552, 2021.

530   Du, Z., Yang, J., Ou, C., and Zhang, T.: Agricultural Land Abandonment and Retirement Mapping in the Northern China Crop-Pasture Band Using Temporal Consistency Check and Trajectory-Based Change Detection Approach, IEEE Trans. Geosci. Remote Sens., 60, 1–12, https://doi.org/10.1109/TGRS.2021.3121816, 2022.

Feng, Q., Niu, B., Chen, B., Ren, Y., Zhu, D., Yang, J., Liu, J., Ou, C., and Li, B.: Mapping of plastic greenhouses and mulching films from very high resolution remote sensing imagery based on a dilated and non-local convolutional neural 535   network, Int. J. Appl. Earth Obs. Geoinformation, 102, 102441, https://doi.org/10.1016/j.jag.2021.102441, 2021.

Feng, Q., Niu, B., Zhu, D., Yao, X., Liu, Y., Ou, C., Chen, B., Yang, J., Guo, H., and Liu, J.: A dataset of remote sensing-based classification for agricultural plastic greenhouses in China in 2019, China Sci. Data, 6, 21.86101.1/noda.2021.0009.zh, https://doi.org/10.11922/noda.2021.0009.zh, 2022a.

Feng, Q., Niu, B., Zhu, D., Liu, Y., Ou, C., and Liu, J.: Classification of Agricultural Plastic Cover Based on Multi-kernel 540   Active Learning and Multi-source Data Fusion, Trans. Chin. Soc. Agric. Mach., 53, 177–185, https://doi.org 10.6041/j.issn.1000-1298.2022.02.018, 2022b.

Feng, Q., Niu, B., Ren, Y., Su, S., Wang, J., Shi, H., Yang, J., and Han, M.: A 10-m national-scale map of Data Descriptor ground-mounted photovoltaic power stations in China of 2020, Sci. Data, https://doi.org/10.1038/s41597-024-02994-x, 2024.

Gao, C., Wu, Q., Dyck, M., Lv, J., and He, H.: Greenhouse area detection in Guanzhong Plain, Shaanxi, China: spatio-545   temporal change and suitability classification, Int. J. Digit. Earth, 15, 226–248, https://doi.org/10.1080/17538947.2021.2023667, 2022.

González-Yebra, Ó., Aguilar, M. A., Nemmaoui, A., and Aguilar, F. J.: Methodological proposal to assess plastic greenhouses land cover change from the combination of archival aerial orthoimages and Landsat data, Biosyst. Eng., 175, 36–51, https://doi.org/10.1016/j.biosystemseng.2018.08.009, 2018.

550   Hao, P., Chen, Z., Tang, H., Li, D., and Li, H.: New Workflow of Plastic-Mulched Farmland Mapping using Multi-Temporal Sentinel-2 data, Remote Sens., 11, 1353, https://doi.org/10.3390/rs11111353, 2019.

Huang, S., Tang, L., Hupy, J. P., Wang, Y., and Shao, G.: A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing, J. For. Res., 32, 1–6, https://doi.org/10.1007/s11676-020-01155-1, 2021.

555   Huete, A. R.: A soil-adjusted vegetation index (SAVI), Remote Sens. Environ., 25, 295–309, https://doi.org/10.1016/0034-4257(88)90106-X, 1988.

Earth System
Open Access Science
Data Discussions

Ji, L., Zhang, L., Shen, Y., Li, X., Liu, W., Chai, Q., Zhang, R., and Chen, D.: Object-Based Mapping of Plastic Greenhouses with Scattered Distribution in Complex Land Cover Using Landsat 8 OLI Images: A Case Study in Xuzhou, China, J. Indian Soc. Remote Sens., 48, 287–303, https://doi.org/10.1007/s12524-019-01081-8, 2020.

560 Jiménez-Lao, R., Aguilar, F. J., Nemmaoui, A., and Aguilar, M. A.: Remote Sensing of Agricultural Greenhouses and Plastic-Mulched Farmland: An Analysis of Worldwide Research, Remote Sens., 12, 2649, https://doi.org/10.3390/rs12162649, 2020.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P.: Global land use / land cover with Sentinel 2 and deep learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS,

565 IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 4704–4707, https://doi.org/10.1109/IGARSS47720.2021.9553499, 2021.

Li, H., Gan, Y., Wu, Y., and Guo, L.: EAGNet: A method for automatic extraction of agricultural greenhouses from high spatial resolution remote sensing images based on hybrid multi-attention, Comput. Electron. Agric., 202, 107431, https://doi.org/10.1016/j.compag.2022.107431, 2022.

570 Li, J., Wang, H., Wang, J., Zhang, J., Lan, Y., and Deng, Y.: Combining Multi-Source Data and Feature Optimization for Plastic-Covered Greenhouse Extraction and Mapping Using the Google Earth Engine: A Case in Central Yunnan Province, China, Remote Sens., 15, 3287, https://doi.org/10.3390/rs15133287, 2023.

Liu, X. and Xin, L.: Spatial and temporal evolution and greenhouse gas emissions of China's agricultural plastic greenhouses, Sci. Total Environ., 863, 160810, https://doi.org/10.1016/j.scitotenv.2022.160810, 2023.

575 Liu, X., He, W., and Zhang, H.: Cross-region plastic greenhouse segmentation and counting using the style transfer and dual-task networks, Comput. Electron. Agric., 207, 107766, https://doi.org/10.1016/j.compag.2023.107766, 2023.

Lu, L., Di, L., and Ye, Y.: A Decision-Tree Classifier for Extracting Transparent Plastic-Mulched Landcover from Landsat-5 TM Images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 7, 4548–4558, https://doi.org/10.1109/JSTARS.2014.2327226, 2014.

580 Lu, L., Tao, Y., and Di, L.: Object-Based Plastic-Mulched Landcover Extraction Using Integrated Sentinel-1 and Sentinel-2 Data, Remote Sens., 10, 1820, https://doi.org/10.3390/rs10111820, 2018.

Ma, A., Chen, D., Zhong, Y., Zheng, Z., and Zhang, L.: National-scale greenhouse mapping for high spatial resolution remote sensing imagery using a dense object dual-task deep learning framework: A case study of China, ISPRS J. Photogramm. Remote Sens., 181, 279–294, https://doi.org/10.1016/j.isprsjprs.2021.08.024, 2021.

585 Mei, Q., Zhang, Z., Han, J., Song, J., Dong, J., Wu, H., Xu, J., and Tao, F.: ChinaSoyArea10m: a dataset of soybean-planting areas with a spatial resolution of 10 m across China from 2017 to 2021, Earth Syst. Sci. Data, 16, 3213–3231, https://doi.org/10.5194/essd-16-3213-2024, 2024.

Niu, B., Feng, Q., Chen, B., Ou, C., Liu, Y., and Yang, J.: HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery, Comput. Electron. Agric., 201, 107297,

590 https://doi.org/10.1016/j.compag.2022.107297, 2022.

Niu, B., Feng, Q., Su, S., Yang, Z., Zhang, S., Liu, S., Wang, J., Yang, J., and Gong, J.: Semantic segmentation for plastic-covered greenhouses and plastic-mulched farmlands from VHR imagery, Int. J. Digit. Earth, 16, 4553–4572, https://doi.org/10.1080/17538947.2023.2275657, 2023a.

Niu, B., Feng, Q., Yang, J., Chen, B., Gao, B., Liu, J., Li, Y., and Gong, J.: Solid waste mapping based on very high
595 resolution remote sensing imagery and a novel deep learning approach, Geocarto Int., 38, 2164361, https://doi.org/10.1080/10106049.2022.2164361, 2023b.

Niu, B., Feng, Q., Qiu, B., Su, S., Zhang, X., Cui, R., Zhang, X., Sun, F., Yan, W., Zhao, S., Shi, H., Ou, C., Yan, X., Gong, J., Yin, G., Huang, J., Liu, J., Gao, B., Yao, X., Yang, J., and Zhu, D.: Global-PCG-10: a 10-m global map of plastic-covered greenhouses derived from Sentinel-2 in 2020, https://doi.org/10.6084/m9.figshare.27731148.v2, 2024.

600 Ou, C. and Wang, Y.: Tracking Spatio-Temporal Dynamics of Greenhouse-Led Cultivated Land and its Drivers in Shandong Province, China, Front. Environ. Sci., 10, 944422, https://doi.org/10.3389/fenvs.2022.944422, 2022.

Ou, C., Yang, J., Du, Z., Liu, Y., Feng, Q., and Zhu, D.: Long-Term Mapping of a Greenhouse in a Typical Protected Agricultural Region Using Landsat Imagery and the Google Earth Engine, Remote Sens., 12, 55, https://doi.org/10.3390/rs12010055, 2019.

605 Ou, C., Yang, J., Du, Z., Zhang, T., Niu, B., Feng, Q., Liu, Y., and Zhu, D.: Landsat-Derived Annual Maps of Agricultural Greenhouse in Shandong Province, China from 1989 to 2018, Remote Sens., 13, 4830, https://doi.org/10.3390/rs13234830, 2021.

Qiu, B., Hu, X., Chen, C., Yang, P., Zhu, X., Yan, C., and Jian, Z.: Maps of cropping patterns in China during 2015–2021., Sci. Data, https://doi.org/10.1038/s41597-022-01589-8, 2022.

610 Qiu, B., Chen, J., Wang, L., Zhang, C., and Wu, F.: National-scale 10-m maps of Data Descriptor cropland use intensity in China during 2018–2023, Sci. Data, https://doi.org/10.1038/s41597-024-03456-0, 2024.

Sui, Y., Feng, M., Wang, C., and Li, X.: A high-resolution inland surface water body dataset for the tundra and boreal forests of North America, Earth Syst. Sci. Data, 14, 3349–3363, https://doi.org/10.5194/essd-14-3349-2022, 2022.

Tong, X., Zhang, X., Fensholt, R., Jensen, P. R. D., Li, S., Larsen, M. N., Reiner, F., Tian, F., and Brandt, M.: Global area
615 boom for greenhouse cultivation revealed by satellite mapping, Nat. Food, 5, 513–523, https://doi.org/10.1038/s43016-024-00985-0, 2024.

Veettil, B. K., Van, D. D., Quang, N. X., and Hoai, P. N.: Remote sensing of plastic-covered greenhouses and plastic-mulched farmlands: Current trends and future perspectives, Land Degrad. Dev., 34, 591–609, https://doi.org/10.1002/ldr.4497, 2023.

620 Wang, K., Chen, W., Tian, J., Niu, F., Xing, Y., Wu, Y., Zhang, R., Zheng, J., and Xu, L.: Accumulation of microplastics in greenhouse soil after long-term plastic film mulching in Beijing, China, Sci. Total Environ., 828, 154544, https://doi.org/10.1016/j.scitotenv.2022.154544, 2022.

Wu, C., Deng, J. S., Wang, K., Ma, L., and Tahmassebi, A. R. S.: Object-based classification approach for greenhouse mapping using Landsat-8 imagery, Int J Agric Biol Eng, 9, 79–88, https://doi.org/10.3965/j.ijabe.20160901.1414, 2016.

625 Xu, H.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery, Int. J. Remote Sens., 27, 3025–3033, https://doi.org/10.1080/01431160600589179, 2006.

Yang, D., Chen, J., Zhou, Y., Chen, X., Chen, X., and Cao, X.: Mapping plastic greenhouse with medium spatial resolution satellite data: Development of a new spectral index, ISPRS J. Photogramm. Remote Sens., 128, 47–60, https://doi.org/10.1016/j.isprsjprs.2017.03.002, 2017.

630 Yu, L., Du, Z., Dong, R., Zheng, J., Tu, Y., Chen, X., Hao, P., Zhong, B., Peng, D., Zhao, J., Li, X., Yang, J., Fu, H., Yang, G., and Gong, P.: FROM-GLC Plus: toward near real-time and multi-resolution land cover mapping, GIScience Remote Sens., 59, 1026–1047, https://doi.org/10.1080/15481603.2022.2096184, 2022.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., 635 Tsendbazar, N.-E., Ramoino, F., and Arino, O.: ESA WorldCover 10 m 2020 v100 (v100), https://doi.org/10.5281/zenodo.5571936, 2021.

Zha, Y., Gao, J., and Ni, S.: Use of normalized difference built-up index in automatically mapping urban areas from TM imagery, Int. J. Remote Sens., 24, 583–594, https://doi.org/10.1080/01431160304987, 2003.

Zhang, L., Xie, Y., Zhu, X., Ma, Q., and Brocca, L.: CIrrMap250: Annual maps of China's irrigated cropland from 2000 to 640 2020 developed through multisource data integration, https://doi.org/10.5194/essd-2024-2, 29 January 2024a.

Zhang, P., Du, P., Guo, S., Zhang, W., Tang, P., Chen, J., and Zheng, H.: A novel index for robust and large-scale mapping of plastic greenhouse from Sentinel-2 images, Remote Sens. Environ., 276, 113042, https://doi.org/10.1016/j.rse.2022.113042, 2022a.

Zhang, P., Lin, C., Guo, S., Zhang, W., Fang, H., and Du, P.: A labor-free index-guided semantic segmentation approach for 645 urban vegetation mapping from high-resolution true color imagery, Int. J. Digit. Earth, 16, 1640–1660, https://doi.org/10.1080/17538947.2023.2207839, 2023a.

Zhang, T., Yang, J., Zhou, H., Dai, A., Gao, B., Tang, K., and Tan, D.: Abandoned Cropland Mapping With Phenology-Enhanced Change Vector Analysis and Semi-Supervised Learning in Different Cropping Intensity Areas, IEEE Trans. Geosci. Remote Sens., 62, 1–15, https://doi.org/10.1109/TGRS.2024.3374451, 2024b.

650 Zhang, X., Liu, L., Wu, C., Chen, X., Gao, Y., Xie, S., and Zhang, B.: Development of a global 30 m impervious surface map using multisource and multitemporal remote sensing datasets with the Google Earth Engine platform, Earth Syst. Sci. Data, 12, 1625–1648, https://doi.org/10.5194/essd-12-1625-2020, 2020.

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., and Mi, J.: GLC_FCS30: global land-cover product with fine classification system at 30 m using time-series Landsat imagery, Earth Syst. Sci. Data, 13, 2753–2776, https://doi.org/10.5194/essd-13-655 2753-2021, 2021a.

Zhang, X., Cheng, B., Chen, J., and Liang, C.: High-Resolution Boundary Refined Convolutional Neural Network for Automatic Agricultural Greenhouses Extraction from GaoFen-2 Satellite Imageries, Remote Sens., 13, 4237, https://doi.org/10.3390/rs13214237, 2021b.

Zhang, X., Liu, L., Zhao, T., Gao, Y., Chen, X., and Mi, J.: GISD30: global 30 m impervious-surface dynamic dataset from
660   1985 to 2020 using time-series Landsat imagery on the Google Earth Engine platform, Earth Syst. Sci. Data, 14, 1831–1856,
https://doi.org/10.5194/essd-14-1831-2022, 2022b.

Zhang, X., Xu, M., Wang, S., Huang, Y., and Xie, Z.: Mapping photovoltaic power plants in China using Landsat, random
forest, and Google Earth Engine, Earth Syst. Sci. Data, 14, 3743–3755, https://doi.org/10.5194/essd-14-3743-2022, 2022c.

Zhang, X., Zhao, T., Xu, H., Liu, W., Wang, J., Chen, X., and Liu, L.: GLC_FCS30D: The first global 30-m land-cover
665   dynamic monitoring product with a fine classification system from 1985 to 2022 using dense time-series Landsat imagery
and continuous change-detection method, https://doi.org/10.5194/essd-2023-320, 31 August 2023b.

Zhang, X., Liu, L., and Chen, X.: Global annual wetland dataset Data Descriptor at 30 m with a fine classification system
from 2000 to 2022, Sci. Data, https://doi.org/10.1038/s41597-024-03143-0, 2024c.

Zhao, G.-X., Lin, G., and Warner, T.: Utilizing Landsat TM Imagery to map greenhouses in Qingzhou, Shandong Province,
670   China, Pedosphere, 14, 1–7, https://doi.org/10.1007/BF02873091, 2004.

Zhou, C., Huang, J., Xiao, Y., Du, M., and Li, S.: A novel approach: Coupling prior knowledge and deep learning methods
for large-scale plastic greenhouse extraction using Sentinel-1/2 data, Int. J. Appl. Earth Obs. Geoinformation, 132, 104073,
https://doi.org/10.1016/j.jag.2024.104073, 2024.

675