From Access to Usability: Proxy-Supervised Fine-Tuning of Satlas for Land-Cover Mapping —A Case Study of Sinuiju, North Korea

Kihun Kim[1]*, **Yongseung Cho**[2], **Jinwon Noh**[3], **Jueun Hwang**[4], and **Beom Kim**[5]

[1] Institute for the North Korean Studies, Dongguk University, ROK
[2] Institute for Planetary Health, Yonsei University, Wonju, ROK
[3] Institute for Planetary Health; Division of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, Wonju, ROK
[4] Department of Urban Administration, University of Seoul, Seoul, ROK
[5] Department of Mathematics, Florida State University, Tallahassee, FL, USA

*Contact: Kihun Kim (bdfather@gmail.com)

ABSTRACT

Land-cover (LC) information is critical for environmental governance, yet its benefits remain unevenly distributed. Many access-restricted regions, such as North Korea, face a usability gap: although satellite imagery is openly available, the absence of reliable local ground truth prevents its conversion into actionable intelligence. This study introduces a reproducible workflow that pairs proxy supervision with foundation-model fine-tuning. We generate proxy LC labels from high-resolution Google Earth imagery and use them to fine-tune the Satlas Pretrain foundation model on Sentinel-2 (RGB+NIR, 10 m), building a region-tuned North Korea Satellite-based Segmentation Model (NKSSM).

On an independent test set, the fine-tuned model achieved mIoU = 0.7075 ± 0.0309 (1,000-bootstrap 95% CI) and a substantial absolute gain over a conservative baseline. Critically, North Pyongan Province (where Sinuiju is located) was absent from all development splits, so applying the model to Sinuiju (2019-2025) constitutes an out-of-distribution deployment; the resulting annual maps are temporally coherent and align with documented local changes (e.g., new transport corridor construction, temporary repurposing of a logistics depot, and post-flood redevelopment). We do not claim nationwide generalization; training tiles are concentrated in the southwestern lowlands (~60%). Area estimates are reported with an MAE-to-area propagation summarized as a conservative ±20% envelope.

This study demonstrates NKSSM not as a nationwide classifier but as a region-specific, reproducible workflow that turns open satellite data and proxy labels into usable LC information under severe data scarcity. These results demonstrate a practical, reproducible pathway for converting open satellite data into credible, usable LC intelligence in label-scarce settings, reframing data democratization from access to operational usability.

KEYWORDS

Land-cover mapping; Earth Observation; proxy supervision; foundation models; Satlas Pretrain; Sentinel-2; fine-tuning; out-of-distribution deployment; data democratization; North Korea (Sinuiju)

1. Introduction

Reliable, frequently updated land-cover (LC) information underpins climate adaptation, disaster risk reduction, agriculture and water resources management, and sustainable urban planning. Across the full pipeline of producing, periodically updating, and independently validating LC information, Earth Observation (EO)—particularly satellite-based sensing—serves as core infrastructure.

Although access to satellite imagery has improved substantially in recent years, the benefits are not evenly distributed. In many regions affected by political restrictions, economic marginalization, or conflict, the primary barrier is not a lack of imagery but a lack of operational capacity to convert open satellite data into usable LC intelligence. Put differently, openness does not automatically translate into use: shortfalls in skilled personnel, computing resources, and standardized processing and validation workflows impede the conversion of data into knowledge. This gap shows that “access for all” does not immediately become “use by all,” and it underscores the need to build the operational backbone required for practical use. Consequently, expanding access alone cannot deliver actionable knowledge; deficits in utilization capacity now constitute a key obstacle to data democratization.

Contemporary global LC products—European Space Agency (ESA) WorldCover, Google/World Resources Institute (WRI) Dynamic World, and Esri LULC—provide valuable 10 m coverage, but independent evaluations show substantial variation in accuracy across regions, biomes, and classes (Venter et al., 2022; Xu et al., 2024). For example, continent-level overall accuracy for WorldCover ranges from ~72.5% to 82.1% (ESA, 2022), and heterogeneous landscapes and several countries (e.g., Mozambique, Tanzania, Nigeria, Spain) exhibit lower accuracies across products (Xu et al., 2024). These patterns align with domain shift and limits in the representativeness of training/validation data and class definitions, rather than a single technical flaw (Venter et al., 2022; Xu et al., 2024). Where locally representative reference data are sparse, rigorous evaluation becomes difficult (Olofsson et al., 2014), so off-the-shelf global maps may generalize unevenly.

Motivated by these constraints, we shift the focus from data access to operational usability. This study introduces a reproducible workflow that integrates proxy ground truth generation from high-resolution (HR) imagery with foundation-model fine-tuning on Sentinel-2 data. Using high-resolution Google Earth imagery to create proxy labels and adapting the Satlas foundation model to local conditions, the approach aims to enable reliable LC mapping where official labels are absent.

We select North Korea as a stringent test case: it is label-scarce and operationally constrained. Demonstrating a workable pipeline here is informative for other underserved, hard-to-validate regions, turning data democratization from a principle into a practical method.
This study pursues three aims:

- to develop methods for generating dependable LC labels where verified ground truth (GT) is unavailable;
- to assess whether foundation model fine-tuning with proxy supervision can deliver robust classification in restricted domains; and
- to examine whether such a workflow supports longitudinal monitoring of LC change that enables meaningful geographic interpretation.

The remainder of this paper is organized as follows. Section 2 reviews the literature on the importance of land-cover data, data democratization and global LC products, satellite-based studies of North Korea, and recent deep-learning and foundation-model developments in Earth Observation. Section 3 describes the dataset, proxy-label workflow, and fine-tuning strategy. Section 4 presents quantitative evaluation, contextual comparison, and qualitative validation. Section 5 analyzes multi-year LC change in Sinuiju (2019-2025), using annual 10 m maps derived from proxy-label generation and regional fine-tuning. This case functions as a label-scarce stress test of operational usability. Section 6 discusses broader implications—data democratization, reproducibility, and methodological limitations—and concludes with directions for future research.

2. Literature Review

2.1 Importance of Land-Cover Data and National Practices

LC information provides essential baseline data for environmental management, climate adaptation, disaster response, agriculture, and urban planning. It underpins assessments of ecosystem change, supports greenhouse-gas mitigation planning, and helps identify climate-vulnerable areas.

Given this importance, South Korea has invested in systematic national mapping: the Ministry of Climate, Energy, and Environment produces standardized LC maps on a recurring cycle to support spatial-data infrastructure and decision-making. Such comprehensive programs, however, remain concentrated in countries with sufficient institutional and technical capacity, leaving many politically restricted or economically fragile regions without comparable systems.

Beyond environmental monitoring, LC data now enable interdisciplinary analysis linking environmental conditions to social outcomes. In public health, meta-analytic evidence connects mapped greenness and related LC indicators with lower all-cause mortality and improved mental and cardiometabolic outcomes (Twohig-Bennett & Jones, 2018). In urban and socioeconomic analysis, globally consistent built-up layers—e.g., the Global Human Settlement Layer and the World Settlement Footprint—support standardized measurement of urbanization patterns and their dynamics (Florczyk et al., 2019; Marconcini et al., 2020).

2.2 Concept and Limitations of Data Democratization

Data democratization in geospatial contexts concerns equitable access and the practical, auditable use of data to support transparent, participatory decision-making (Craglia & Shanley, 2015; Džanko et al., 2024). While open satellite missions such as Landsat and Sentinel have markedly expanded access, substantial regional differences persist in usability due to gaps in compute resources, skills, and reproducible workflows (Thapa et al., 2019; Džanko et al., 2024).

Consequently, genuine democratization requires not only open data but also operational frameworks and community capacity that enable equitable, repeatable, and verifiable use (Džanko et al., 2024); reproducible mapping architectures can help address this need (Saah et al., 2020).

2.3 Global Land-Cover Products Used as Baselines

We treat three 10 m global LC products as contextual baselines: ESA WorldCover 2021 v200 (Sentinel-1/-2 features with a Random Forest pipeline and expert-rule refinements), Google/

WRI Dynamic World (a semi-supervised FCNN that outputs per-pixel class probabilities with near-real-time latency), and Esri LULC (a deep-learning model trained on Sentinel-2). Their class taxonomies, temporal footprints, and QA procedures differ. Broader evidence on region- and biome-dependent accuracy variability and limits in the representativeness of global products is summarized in the Introduction and revisited in Section 4.6. Here we provide a brief description of the three 10 m global LC products used as contextual baselines for the North Korea Satellite-based Segmentation Model (NKSSM).

2.4 Satellite-Based Research on North Korea

North Korea exemplifies a data-scarce environment in which field surveys are restricted and satellite imagery remains the only feasible observation source.

Recent studies have therefore relied on remote-sensing pipelines tailored to these constraints. At the national scale, Piao et al. (2021) used a Random Forest classifier with time-series imagery to analyze land-use/land-cover (LULC) change across 1990-2020, explicitly noting the challenge of on-the-ground verification and the need for remote methods suited to inaccessible areas. Building at the local scale, Piao et al. (2023) constructed semi-permanent sample points from multiple LULC products and classified Landsat time-series with Random Forest (overall accuracy 97.66 ± 1.36%, Cohen's kappa = 0.95 ± 0.03). For Pyongyang (2000-2020), they report increases in built-up and forest area, decreases in cropland, and rising landscape fragmentation measured via FRAGSTATS—while emphasizing that North Korea's inaccessibility necessitates such product-based validation approaches.

Complementing these efforts, Kim et al. (2024) introduce a domain-adaptation method within a phenological classification framework to classify North Korea using South-Korea-trained models (overall accuracy 81.31%), explicitly positioning domain adaptation as a practical response to the absence of local labels.

Taken together, this literature shows a clear trajectory: where conventional GT is unavailable, researchers turn to multi-source sampling, time-series classification, and domain adaptation to build usable evidence. These strategies do not negate the limits of in-situ validation, but they demonstrate workable pathways for monitoring LC dynamics in North Korea.

2.5 Emergence of Deep Learning and Foundation Models

Recent advances in deep learning have markedly improved the accuracy and scalability of LC mapping, including in label-scarce settings. Encoder-decoder architectures (e.g., DeepLab) and Transformer backbones learn hierarchical spectral-spatial features directly from imagery, reducing dependence on hand-crafted indices and rules (Chen et al., 2018; Ma et al., 2019; Zhu

et al., 2017). Beyond architectures, transfer learning and domain adaptation have become practical strategies when locally verified labels are limited (Alem et al., 2022).
Building on this trajectory, foundation models—large models pre-trained on broad, heterogeneous corpora—serve as general-purpose backbones that can be adapted to specific regions and tasks (Bommasani et al., 2021). In EO, the Satlas Pretrain release provides an open, multi-sensor basis (e.g., Sentinel-2 and NAIP) trained across diverse geographies and tasks to support robust downstream fine-tuning (Bastani et al., 2023). This combination of scale, multi-task signals, and openness makes foundation-model adaptation a compelling option where curated GT is scarce.

Against this backdrop, our study pairs proxy supervision (HR imagery-derived labels) with targeted fine-tuning of a foundation model to build a reproducible, North-Korea-specific LC workflow. The aim is not to claim universality, but to demonstrate that—with transparent preprocessing, consistent alignment, and auditable inference—domain adaptation from a strong pretrain can yield usable, inspectable 10 m maps even where in-situ labels are unavailable. Accordingly, Section 3 details the dataset, proxy-label workflow, and fine-tuning setup.

3. Methodology

3.1 Research Design Overview

We frame the task as operational data democratization: turning widely accessible EO data into usable LC information in settings—such as North Korea— where in-situ labels and institutional capacity are limited. We present a low-cost, reproducible workflow that converts open imagery into proxy supervision and fine-tunes an open foundation (Satlas Pretrain) model to local conditions. The goal is to produce reliable 10 m LC maps under data scarcity—not by adding new sensors or infrastructure, but by making existing data practically usable.

The workflow consists of three stages:

- generation of proxy masks from HR imagery;
- spatial and temporal alignment with Sentinel-2 imagery;
- and training and validation of a region-specific model, the NKSSM.

All stages emphasize reproducibility and quantified uncertainty as core design principles. This three-stage workflow—proxy-label generation, Sentinel-2 alignment, and NKSSM training/validation—is summarized in Figure 1.

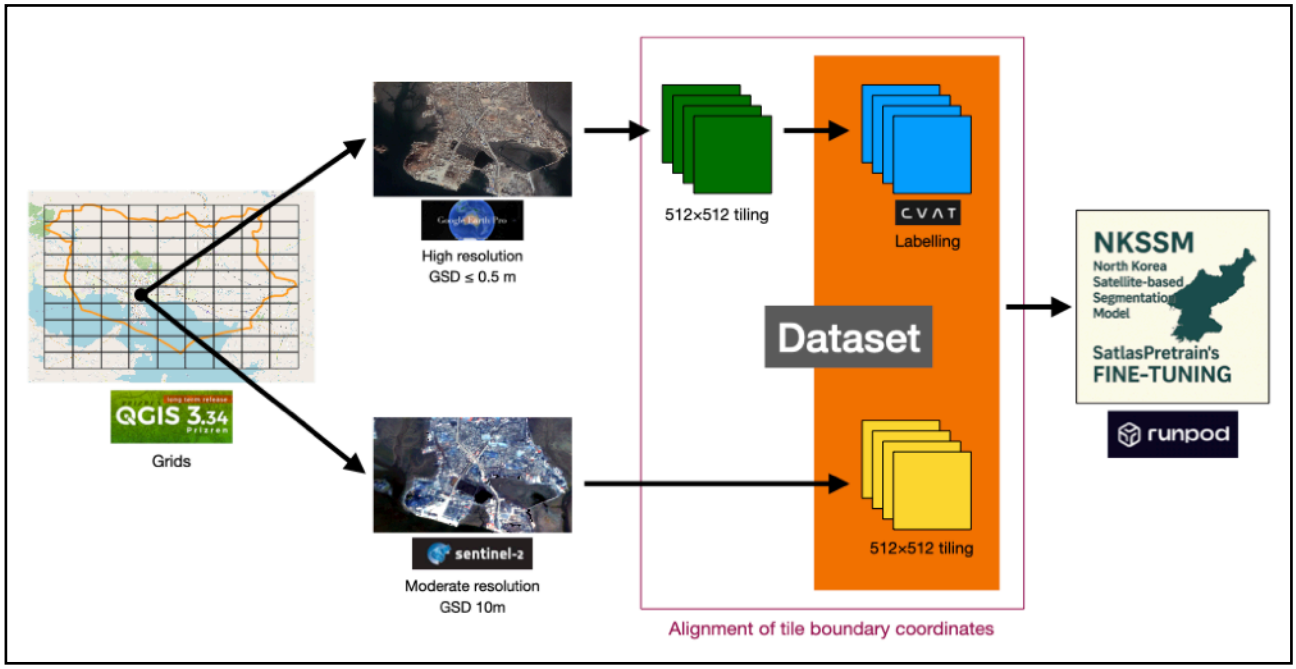


Figure 1. Workflow for Proxy Label Generation and Fine-Tuning of the NKSSM Model. The workflow integrates HR (≤0.5 m) Google Earth and moderate-resolution (10 m) Sentinel-2 imagery, aligned on QGIS grids (512×512 px). HR tiles are labeled in CVAT to produce proxy masks matched with Sentinel-2 tiles, forming a reproducible training dataset. This dataset fine-tunes the Satlas Pretrain foundation model to build the NKSSM, emphasizing coordinate alignment, reproducibility, and transparency under label scarcity.
Source: Google Earth; Sentinel-2 (Google Earth Engine)

3.2 Proxy Label Generation and Alignment

In the absence of in-situ validation data, HR Google Earth imagery (≤ 0.5 m) was used to create proxy labels. The imagery was divided into 512 × 512 tiles, manually annotated via visual interpretation and polygon editing, and spatially aligned with Sentinel-2 L2A imagery (10 m) of identical extent. Only Sentinel-2 scenes with ≤ 20% cloud cover were retained; within those scenes, tiles contaminated by clouds were excluded. Acquisitions were constrained to August-September to reduce seasonal/phenological variability in North Korea.

At this stage—after rice transplanting but before full vegetative growth—paddy fields retain water reflection, which can occasionally appear as Waterbody. Nonetheless, selecting this late-summer window increases spectral contrast between impervious Built-up surfaces and Cropland and stabilizes crop canopies, which we expect to mitigate confusion primarily between Built-up and Cropland. By contrast, Cropland-Woody Vegetation confusion does not necessarily decrease in this period and may persist, especially where tall crops (e.g., maize, sorghum) exhibit woody-like spectral signatures.

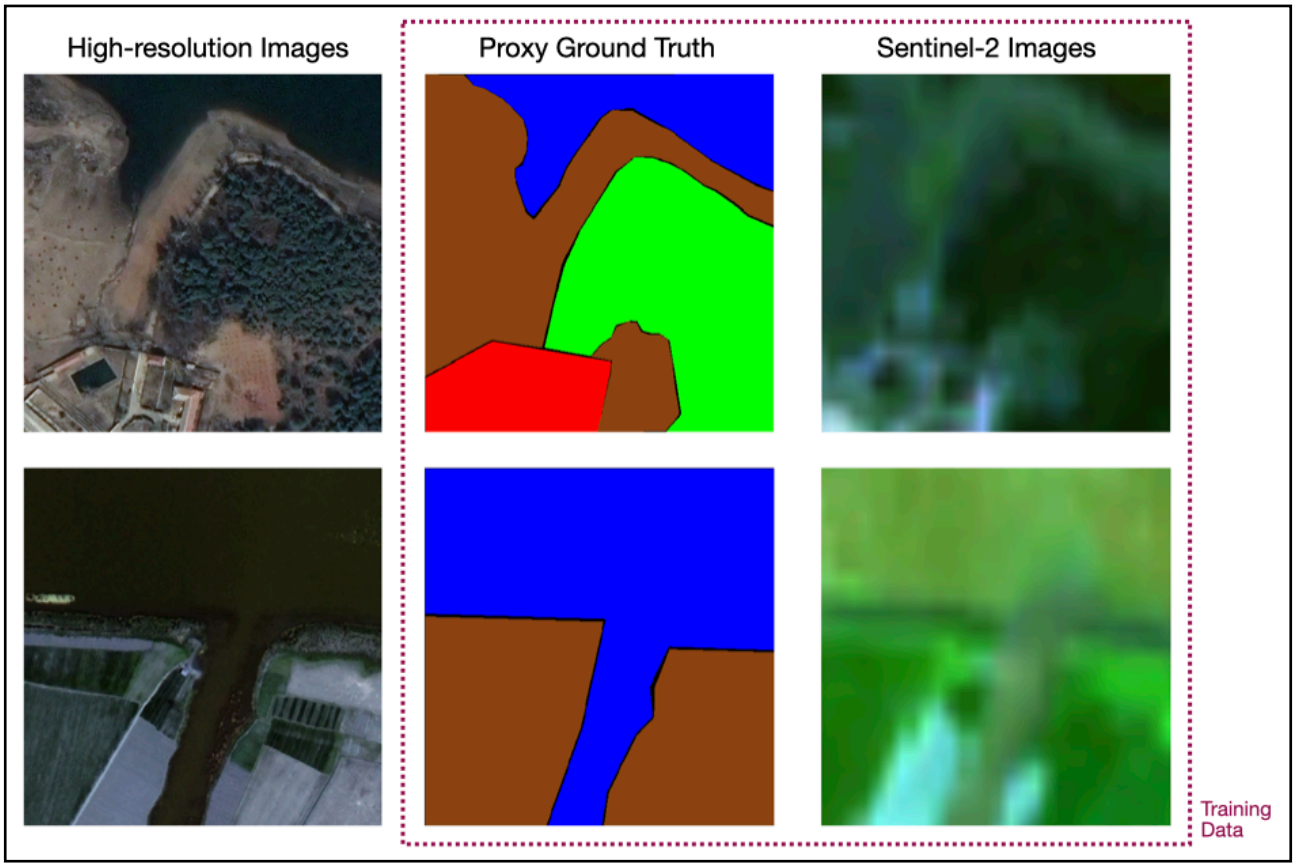


Figure 2. Image-Mask Pair Example. Example of how HR imagery (≤ 0.5 m) from Google Earth was visually interpreted and polygon-labeled to create proxy GT masks. These masks were then spatially aligned with Sentinel-2 L2A imagery (10 m) of identical spatial extent to form training pairs. Although minor geometric and temporal offsets remain due to sensor and acquisition differences, the aligned dataset provides a visual reference suitable for supervised learning under data-scarce conditions.
Source: Google Earth; Sentinel-2 (Google Earth Engine)

While the labels were carefully aligned to Sentinel-2 geometry, minor geometric and temporal discrepancies remain inevitable. The Google Earth basemap slightly differs from Sentinel-2 in viewing angle, projection system, and acquisition timing. These geometric differences may result in subtle parallax distortions or small boundary offsets, particularly near slopes and high-rise districts. Temporal differences, in turn, can produce date-to-date appearance changes (e.g., crop phenology, water extent, illumination/shadow), especially in floodplains and croplands. Spatial misalignment cannot be fully eliminated and is documented as a limitation. By contrast, temporal mismatch was partially mitigated by selecting the closest available Sentinel-2 acquisition dates for alignment with each Google Earth scene; nevertheless, perfect temporal synchronization was not always possible given Google Earth's irregular update cadence.

As a result, the annotated dataset should be understood as a proxy reference—a near-contemporary visual approximation of the Earth's surface—rather than an absolute GT.

3.3 Dataset Construction, Standardization, and Spatial Independence Validation

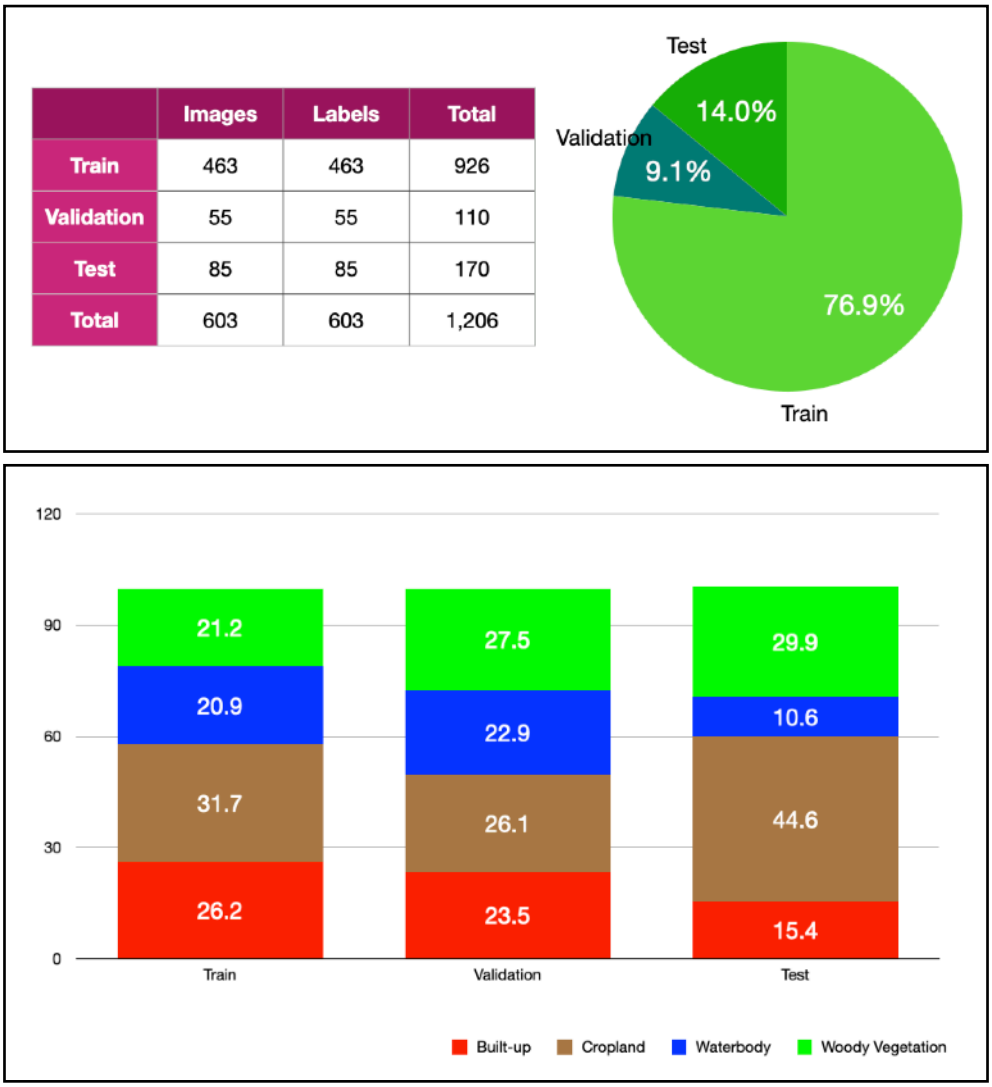
Four Sentinel-2 L2A bands (RGB+NIR, 10 m resolution) were used for analysis.
All imagery and Google Earth-derived masks were reprojected and resampled onto a common 10 m raster grid in EPSG:4326, ensuring pixel-wise alignment between Sentinel-2 inputs and proxy labels.

The dataset comprises a total of 603 image-mask pairs (463 training, 55 validation, 85 test). Proxy labels were generated from the nearest available Google Earth dates and aligned with August-September Sentinel-2 L2A composites. We quantified the pixel-wise class distribution (per-class pixel counts and proportions) in the proxy masks to characterize class prevalence in each split and to verify that the training and validation subsets were not severely imbalanced; the test set was left in its naturally occurring class proportions. Tile footprints (512×512 geographic extents) were held constant.

To evaluate the potential spatial leakage between data splits, we computed the shortest edge-to-edge distance between all tile pairs belonging to the training (463), validation (55), and test (85) subsets. The distance was defined not as the separation between tile centroids, but as the minimum Euclidean distance between the outer boundaries (edges) of two tiles. Specifically, pairs of tiles that overlapped (distance = 0) or whose boundaries were directly adjacent (distance = 0) were treated as contiguous, while distances were computed only when tiles were spatially separated.

Among the total of 69,495 inter-split pairs, 1,281 pairs (1.84%) were located within 1 km under an edge-to-edge proximity measure. At this threshold, short-distance proximity was most

frequent between the train-validation splits (2.45% of train-val pairs) and lower for combinations involving the test split (train-test 1.49%, val-test 1.50%), indicating that the test set remained largely spatially independent. These results indicate limited localized adjacency; however, because the test set remains largely spatially independent, spatial leakage is unlikely to materially affect the reported test metrics.



	Images	Labels	Total
Train	463	463	926
Validation	55	55	110
Test	85	85	170
Total	603	603	1,206

Figure 3. Dataset Composition, Split Ratios, and Class Distributions. The dataset consists of 603 image-mask pairs, divided into training (463, 76.9%), validation (55, 9.1%), and test (85, 14.0%) subsets. Each split maintains consistent geographic coverage and temporal alignment (August-September).The lower chart shows pixel-wise class distributions for the four LC classes in the training and validation splits; the classes are approximately balanced, with a slightly higher Cropland share in training. Class-balance checks were applied to training/validation to support stable learning; the test split was not constrained to match these pixel-level distributions so that it reflects the natural class mix in the area of interest.

The 1 km threshold used in this study is not an absolute criterion. Previous studies (Feng et al., 2023; Roberts et al., 2022) have empirically demonstrated that training and validation samples in close spatial proximity can lead to inflated validation accuracy and reduced generalization performance. Similarly, Schmitt et al. (2019) constructed the SEN12MS dataset using geographically separated block partitions to prevent such information leakage. However, these

studies emphasize the importance of sufficient spatial separation rather than prescribing any specific numerical threshold. Therefore, the 1 km threshold in this study serves as a conservative heuristic buffer distance, established as an operational criterion to assess the spatial independence of data splits.

As shown in Table 1, the dataset exhibits a clear geographical imbalance. The South and North Hwanghae Provinces account for approximately 362 tiles, or about 60% of the total 603 tiles, indicating that the data are concentrated in the southwestern lowland regions. In contrast, northern mountainous areas such as Jagang, Ryanggang, and North Hamgyong Provinces are underrepresented, limiting the dataset's ability to capture the full range of geomorphological and LC variations across North Korea.

Table 1. Distribution of Dataset Tiles by Province

Province	Train	Validation	Test
South Hwanghae	123	18	8
North Hwanghae	161	20	32
South Pyongan	64	5	15
Jagang	21	2	4
South Hamgyong	59	7	12
North Hamgyong	17	3	5
Ryanggang	6	-	4
Kangwon	10	-	6
Total	463	55	85

This regional bias may cause the model to overfit to lowland cropland patterns and perform less reliably in high-altitude or forest-dominant regions. Recognizing this limitation, we highlight the need for additional samples from mountainous provinces and more balanced data augmentation across regions to improve the model's spatial generalization. Expanding regional diversity is essential to achieve nationwide representativeness.

As North Pyongan is absent from Table 1, the Sinuiju application in Section 5 constitutes an out-of-distribution (OOD) evaluation with respect to all development splits (train/validation/test).

3.4 Class Taxonomy

For stable learning at 10 m resolution, the classification scheme was simplified to four core classes: Built-up, Cropland, Woody Vegetation, and Waterbody. Bare Land and Grassland were excluded at this stage because they are often small, transitional (e.g., fallow fields or

construction zones), and spectrally similar to Cropland or Built-up areas, which can destabilize training.
Future work will introduce a Bare Land class to improve interpretation of ecological transitions and forest recovery. We will employ digital elevation model (DEM)-based weak supervision and partial labeling, expand the optical input from 4 to 9 Sentinel-2 bands, and integrate Sentinel-1 synthetic aperture radar (SAR) to add structure- and moisture-sensitive features that enhance class separability (e.g., Built-up vs Bare, Cropland vs Woody Vegetation) and reduce sensitivity to illumination and shadow effects.

3.5 Model Architecture and Training Settings

NKSSM is built upon the Satlas Pretrain model (ResNet-50 FPN backbone) developed by the Allen Institute for AI. The input configuration uses four channels (RGB + NIR), and the output layer was adapted to four target classes. The loss function combines Lovász-Softmax (70%) and Focal Loss (30%), optimizing both region- and boundary-level accuracy.

Training was performed for up to 100 epochs with early stopping if no improvement was observed by epoch 70. After epoch 30, automatic mixed precision (AMP) and Stochastic Weight Averaging (SWA) were applied to enhance generalization.

Dynamic class weighting alleviated imbalance and boundary sensitivity, with representative weights: Built-up = 0.67, Cropland = 1.58, Waterbody = 2.00, Woody Vegetation = 2.00. All experiments were run in the customized Satlas-based multi-core training environment, tracking Pixel Accuracy, Precision, Recall, F1, mean Intersection over Union (mIoU), and Cohen's kappa metrics.

4. Results

4.1 Training Stability and Reproducibility

We assessed reproducibility over four independent runs (seeds 33, 42, 72, 333). For each run, we recorded the single epoch with the highest validation mIoU, and we summarize those four per-seed best values here. The mean of the per-seed best validation mIoUs was 0.5901 (sample standard deviation, SD 0.0049, variance 2.44×10^{-5}; range 0.5843–0.5962, i.e., 0.0119 or ≈2.0% of the mean), indicating low between-run variability. The epochs at which these best validation scores occurred ranged from 40–68, consistent with stable convergence across seeds; early stopping halted training near these peaks, suggesting appropriate stopping behavior under the monitored criterion. Overall, NKSSM demonstrated consistent convergence and statistical stability, independent of random initialization.

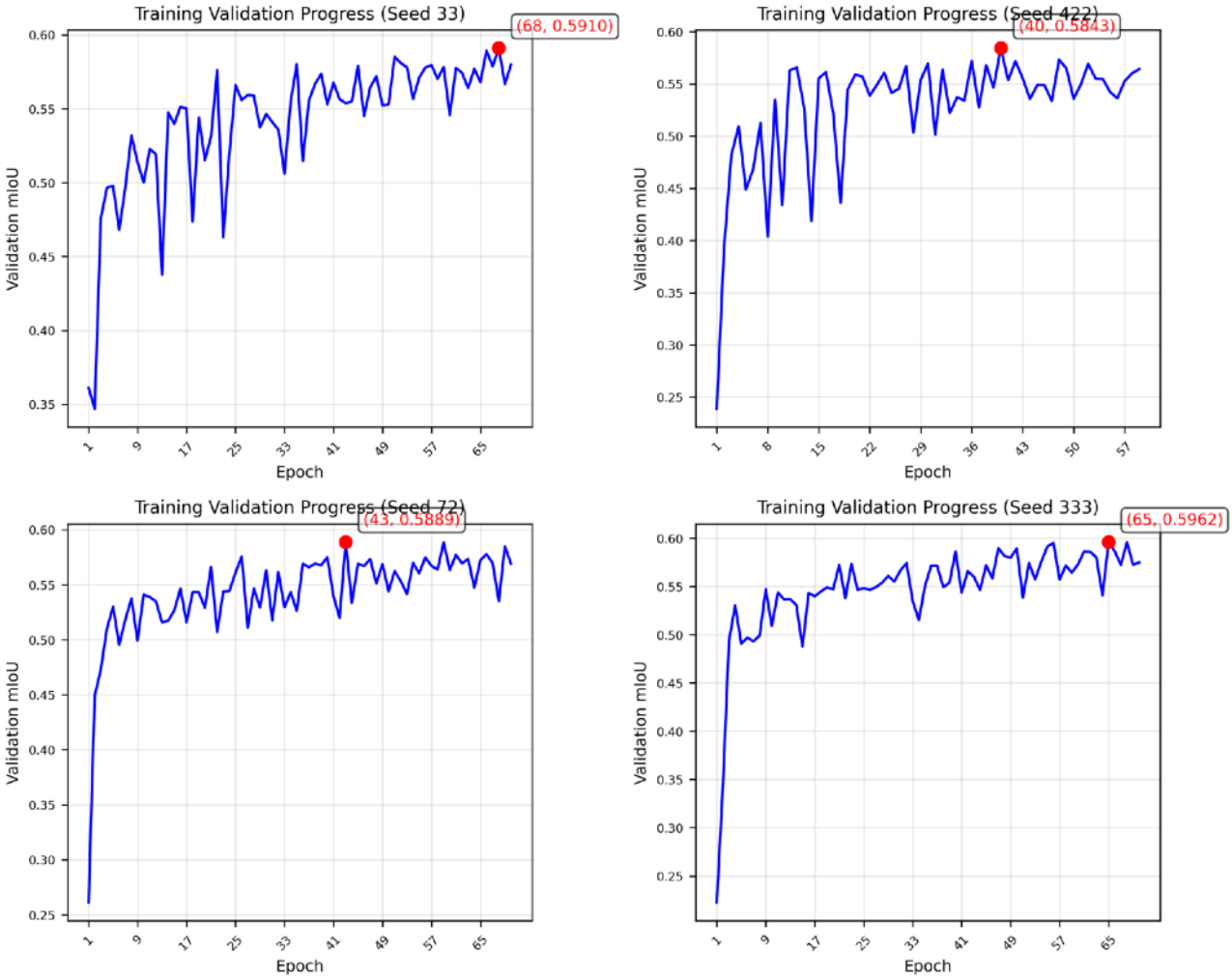


Figure 4. Validation mIoU Progress Across Four Random Seeds. Training progress of the NKSSM model under four independent random initializations (Seeds 33, 42, 72, 333). Each curve represents validation mIoU evolution over epochs, with the best-performing epoch marked in red. Across runs, the mean of the per-seed best validation mIoU reached 0.5901 ± 0.0049, showing minimal variability (range: 0.5843–0.5962) and stable convergence within 40–68 epochs. These results confirm the model's strong reproducibility and robustness against random initialization, with early stopping effectively preventing overfitting.

Table 2. Seed-wise Validation Performance and Convergence Epochs

Seed	Best mIoU (Validation)	Each of Best Performance
33	0.5910	68
42	0.5843	40
72	0.5889	43
333	0.5962	65

4.2 Independent Test Evaluation and Model Selection

To assess the model's generalization capability, an independent test set of 85 tiles was evaluated using four random seeds (33, 42, 72, 333).

For each seed, we selected the single checkpoint (epoch) with the highest validation mIoU and then evaluated that checkpoint on the held-out test set. Model selection was based exclusively on validation performance; the test set remained untouched until final reporting. Test results (mean ± sample SD across seeds, n = 4) were: mIoU = 0.6976 ± 0.0094, Pixel Accuracy = 0.8313 ± 0.0061, Cohen's kappa = 0.7487 ± 0.0099, and mean absolute error (MAE) = 0.2808 ± 0.0109.

Table 3. Seed-wise Best Performance on the Independent Test Set

Seed	mIoU	Pixel Accuracy	Cohen's kappa	MAE	Epoch of Best Model
33	0.7112	0.8412	0.7626	0.2661	68
42	0.6929	0.8297	0.7457	0.2861	40
72	0.6972	0.8300	0.7460	0.2765	43
333	0.6889	0.8265	0.7385	0.2926	65
Mean±SD	0.6976 ± 0.0094	0.8313 ± 0.0061	0.7487 ± 0.0099	0.2808 ± 0.0109	

Among the four runs, the seed-33 checkpoint (epoch 68) achieved the best test performance (mIoU = 0.7112; Cohen's kappa = 0.7626; MAE = 0.2661) and was therefore used as the final model for the bootstrap analysis (Section 4.3) and qualitative validation (Section 4.5).

The validation set (55 tiles; mean of per-seed best checkpoints) achieved mIoU = 0.5901, whereas the test set (85 tiles), evaluated with the seed-33 checkpoint (epoch 68), reached mIoU = 0.7112. This difference does not result from model overfitting but rather from differences in sample-size effects and spatial heterogeneity inherent in the partitioning strategy. The validation set contained fewer samples, which increased statistical variance.

4.3 Bootstrap Validation of Statistical Robustness

The final model (seed 33, epoch 68) was subjected to 1,000 bootstrap resamplings to quantify statistical confidence and uncertainty. All metrics converged within narrow 95% confidence intervals, indicating strong statistical robustness and consistent performance.

The bootstrap mean mIoU reached 0.7075 ± 0.0309, Pixel Accuracy = 0.8410 ± 0.0174, and Cohen's kappa = 0.7608 ± 0.0278, confirming that all core indicators remained within stable and statistically tight confidence bounds. The overall MAE averaged 0.2693 ± 0.0355, further indicating that prediction errors were consistently small across samples.

Among the four LC classes, Woody Vegetation (0.7740) and Cropland (0.7261) showed the highest IoUs, while Built-up (0.6447) and Waterbody (0.6853) were slightly lower, reflecting the spectral heterogeneity and mixed-pixel effects typical of urban and river-edge regions. All class-wise SDs remained below 0.08, demonstrating stable class-level segmentation performance.

Table 4. Overall Performance Metrics Based on 1,000 Bootstrap Resampling

Performance Metric	Mean	95% CI (Lower-Upper)	SD
mIoU	0.7075	[0.6447, 0.7662]	0.0309
Pixel Accuracy	0.8410	[0.8049, 0.8738]	0.0174
Cohen's kappa	0.7608	[0.7037, 0.8117]	0.0278
MAE	0.2693	[0.2028, 0.3414]	0.0355

Table 5. Class-Wise IoU Statistics from Bootstrap Evaluation

Class	Mean	95% CI (Lower-Upper)	SD
Built-up	0.6447	[0.5429, 0.7321]	0.0483
Cropland	0.7261	[0.6661, 0.7788]	0.0284
Waterbody	0.6853	[0.5183, 0.8194]	0.0777
Woody Vegetation	0.7740	[0.6847, 0.8489]	0.0428

Table 6. Class-Wise MAE Statistics from Bootstrap Evaluation

Class	Mean	95% CI (Lower-Upper)	SD
Built-up	0.3428	[0.2399, 0.4655]	0.0572
Cropland	0.1946	[0.1257, 0.2773]	0.0392
Waterbody	0.2982	[0.1352, 0.5071]	0.0930
Woody Vegetation	0.2417	[0.1224, 0.3973]	0.0715

The class-wise MAE values are lowest for Cropland (0.1946) and Woody Vegetation (0.2417), and higher for Built-up (0.3428) and Waterbody (0.2982), reflecting patterns consistent with the boundary complexity observed in Section 4.4 and the known spectral heterogeneity of urban and river-edge environments. Across 1,000 bootstrap replicates, the 95% CI widths range ≈0.152–0.372 across classes (narrowest for Cropland, widest for Waterbody), reflecting class prevalence and boundary complexity.

Taken together, these results confirm the statistical reliability and reproducibility of the NKSSM framework. Unless noted, all overall metrics refer to seed-33 (epoch 68) bootstrap means.

4.4. Boundary-Sensitivity Analysis

To assess how precisely the model delineates LC boundaries, we conducted a boundary-sensitivity evaluation using the Boundary-F1 Score (BF1) and Trimap-IoU metrics. These indices quantify segmentation accuracy within narrow boundary regions (1–3 pixels) and are particularly sensitive to mixed-pixel effects in Sentinel-2 imagery (10 m GSD). For consistency, reference masks boundaries were extracted via a morphological gradient, and all overall aggregates are micro-averaged (boundary-weighted) across classes; classes with zero GT boundary pixels in a bootstrap replicate were excluded from that replicate's aggregation. Uncertainty was estimated with 1,000 bootstrap resamples.

Table 7. Boundary-F1 and Trimap-IoU Performance

Class	Boundary-F1	Trimap-IoU (1 px)	Trimap-IoU (2 px)	Trimap-IoU (3 px)
Overall (micro, boundary-weighted)	0.4255 [0.3870, 0.4680]	0.6348 [0.6006, 0.6668]	0.6196 [0.5850, 0.6532]	0.6048 [0.5704, 0.6400]
Built-up	0.4737 [0.3879, 0.5569]	0.3620 [0.2930, 0.4238]	0.3578 [0.2904, 0.4178]	0.3556 [0.2895, 0.4157]
Cropland	0.3210 [0.2694, 0.3752]	0.7100 [0.6451, 0.7629]	0.6887 [0.6256, 0.7394]	0.6618 [0.6049, 0.7115]
Waterbody	0.5561 [0.4683, 0.6524]	0.4618 [0.3567, 0.5737]	0.4567 [0.3514, 0.5652]	0.4488 [0.3488, 0.5560]
Woody Vegetation	0.4724 [0.3881, 0.5590]	0.5710 [0.4778, 0.6597]	0.5536 [0.4633, 0.6419]	0.5406 [0.4526, 0.6289]

The overall (micro-averaged) Boundary-F1 was 0.4255 with a 95% CI of [0.3870, 0.4680]. Class-wise differences mirrored the geometric and spectral complexity of each LC type: Waterbody and Built-up achieved the highest boundary precision (BF1 = 0.5561 and 0.4737, respectively), indicating the model's ability to capture linear and sharply defined features. Cropland showed lower boundary accuracy (BF1 = 0.3210), while Woody Vegetation was intermediate (BF1 = 0.4724). At the all-classes level, Trimap-IoU (micro) declined gently as the

margin widened, from 0.6348 at 1 px (95% CI [0.6006, 0.6668]) to 0.6196 at 2 px ([0.5850, 0.6532]) and 0.6048 at 3 px ([0.5704, 0.6400]). Class-wise values show the same pattern—e.g., Woody Vegetation: 0.5710 → 0.5406; Cropland: 0.7100 → 0.6618; Built-up: 0.3620 → 0.3556; Waterbody: 0.4618 → 0.4488—indicating consistent edge stability under relaxed boundary tolerances. The tight alignment among 1-3 px IoUs suggests robustness to small positional perturbations.

Overall, under 10 m Sentinel-2 resolution, the NKSSM exhibits geometrically coherent boundaries and reliably identifies narrow, fragmented, or elongated landscape features, supporting its utility for change detection and time-series mapping.

4.5 Performance Comparison: Before vs. After Fine-tuning

For fair comparison under the same 4-band constraint, the Before (Satlas Pretrain) baseline used the Satlas Sentinel-2 backbone (Sentinel2 ResNet50 SI MS, 9-channel input) with the encoder frozen and a learnable 1×1 adapter projecting 4 → 9 channels. Only the adapter and head were trained (linear-probe-plus setup), with the same dataset, loss, and hyperparameters as the fine-tuned NKSSM. This setup served as a conservative, reproducible baseline to isolate the direct effect of regional fine-tuning.

To evaluate the impact of fine-tuning, we compared the original Satlas Pretrain model with the fine-tuned NKSSM on identical 85 test tiles using the same Sentinel-2 imagery (August-September window), ensuring strict comparability. The After (Fine-tuned) figures represent the mean of 1,000 bootstrap replicates with 95% confidence intervals (CIs).

Fine-tuning with region-specific proxy labels produced large and consistent gains: Pixel Accuracy increased from 0.6908 [0.6437-0.7375] to 0.8410 [0.8049-0.8738] (+0.1502; +21.7%), mIoU from 0.5005 [0.4454-0.5587] to 0.7075 [0.6447-0.7662] (+0.2070; +41.4%), and Cohen's kappa from 0.5636 [0.4972-0.6274] to 0.7608 [0.7037-0.8117] (+0.1972; +35.0%). MAE decreased from 0.5298 [0.4605-0.6067] to 0.2693 [0.2028-0.3414] (–0.2605; –49.2%), indicating markedly fewer per-pixel errors.

Table 8. Performance Comparison before and after Fine-tuning with 95% Confidence Intervals

Metric	Before (Satlas Pretrain)	After (Fine-tuned NKSSM)	Δ (Absolute)	Δ (%)
mIoU	0.5005 [0.4454, 0.5587]	0.7075 [0.6447, 0.7662]	0.2070	41.4%
Pixel Accuracy	0.6908 [0.6437, 0.7375]	0.8410 [0.8049, 0.8738]	0.1502	21.7%
Cohen's kappa	0.5636 [0.4972, 0.6274]	0.7608 [0.7037, 0.8117]	0.1972	35.0%
MAE	0.5298 [0.4605, 0.6067]	0.2693 [0.2028, 0.3414]	-0.2605	-49.2%

These improvements demonstrate that fine-tuning on locally aligned proxy data effectively adapts the Satlas model's global representation to the distinct spatial and spectral conditions of North Korea, transforming a generic satellite foundation model into a domain-optimized and reproducible mapping system suitable for label-scarce and access-restricted regions such as North Korea.

4.6 Comparison with Global Land-Cover Products

On the held-out test dataset, the NKSSM achieved Pixel Accuracy = 0.8410 ± 0.0174 (bootstrap mean ± sample SD over 1,000 replicates), with accompanying summaries mIoU = 0.7075 ± 0.0309, Cohen's kappa = 0.7608 ± 0.0278, and MAE = 0.2693 ± 0.0355. In this study, pixel accuracy is equivalent to overall accuracy (OA), defined as the proportion of correctly classified pixels over all pixels in the confusion matrix.

For recent global 10 m LC products, provider-reported OAs are as follows: ESA WorldCover 2021 v200 reports a global OA of 76.7% ± 0.5%, with continent-level OAs roughly ~72.5-82.1% (ESA, 2022). Google/WRI Dynamic World reports 73.8% overall agreement against 409 expert-labeled validation tiles (Brown et al., 2022). Esri LULC is described as over 85% accuracy in the original IGARSS paper, and provider documentation for the annual composites cites OA ≥ 91% under strict-consensus labels and OA ≥ 76% under majority-consensus labels (Karra et al., 2021).

Independent cross-comparisons show substantial variation by region, biome, and class. Using a globally sampled reference, Venter et al. (2022) report OA of approximately Esri ≈ 75%, Dynamic World ≈ 72%, and WorldCover ≈ 65%, with strong class-wise differences and landscape effects. At global/continental/large-country scales, Xu et al. (2024) find global OA spanning ~73.4-83.8%, with notable continental and country-level dispersion, and recommend stronger regional validation and careful class-schema alignment.

Direct, like-for-like comparison requires caution because scope, class taxonomy, and validation protocols differ: WorldCover 2021 v200 uses 11 classes, Dynamic World 9, and Esri LULC 10, whereas NKSSM maps 4 classes (Built-up, Cropland, Waterbody, Woody Vegetation). Differences in class granularity, boundary conventions, temporal footprints, and reference-label construction (e.g., strict vs. majority consensus) can affect OA magnitudes and error patterns.

Given these differences noted above, the results nevertheless suggest that region-specific fine-tuning can achieve competitive or superior performance for target regions compared to globally generalized models at the same spatial resolution. This interpretation aligns with Venter et al. (2022)—who note reduced accuracy in specific regions due to limited local representativeness and sparse regional training data—and Xu et al. (2024), and NKSSM

addresses these limitations by incorporating regionally tailored proxy labels that capture local spatial and spectral characteristics.

Table 9. Comparative Performance of Global LC Products and the Regionally Fine-Tuned NKSSM

Model	Region	Resolution	# of Classes	OA / Pixel Accuracy	Remarks
ESA WorldCover 2021 v200	Global	10m	11	0.767 ± 0.005	Sentinel-1/2 fusion, Random Forest; provider-reported global OA = 76.7% ± 0.5% (ESA, 2022)
Google/WRI Dynamic World	Global	10m	9	0.738	Sentinel-2; semi-supervised FCNN; overall agreement = 73.8% vs 409 expert tiles (Brown et al., 2022)
Esri LULC	Global	10m	10	≥ 0.91 (strict); ≥ 0.76 (majority)	Sentinel-2; deep learning; annual composites: OA ≥ 91% (strict-consensus), ≥ 76% (majority-consensus) (Karra et al., 2021)
NKSSM	North Korea	10m	4	0.8410 ± 0.0174	Satlas Pretrain fine-tuned with proxy labels; held-out test dataset, 1,000-rep bootstrap; mIoU = 0.7075 ± 0.0309, Cohen's kappa = 0.7608 ± 0.0278

Note. For global products (WorldCover, Dynamic World, Esri LULC), overall accuracies are provider-reported values. NKSSM accuracy refers to this study's held-out test dataset.

4.7 Qualitative Evaluation Results

The 85 test tiles were visually inspected by grouping them according to their tile-level mIoU values—high (≥ 0.80), medium (0.60-0.79), and low (< 0.60)—to assess the spatial correspondence between NKSSM predictions and the proxy ground-truth masks. This qualitative evaluation focused on (i) whether large-scale landform structures are preserved, (ii) how accurately class boundaries are delineated, and (iii) the extent to which fine details are maintained. Figures 5-7 summarize these examples: high-mIoU tiles in Figure 5, medium in Figure 6, and low in Figure 7.

In each figure, every row shows, from left to right, the HR reference image used to derive the proxy labels, the resulting proxy mask, the August-September Sentinel-2 input tile, and the NKSSM prediction. Thus, the proxy masks are manually interpreted labels derived from HR imagery, whereas the prediction masks are produced from Sentinel-2 imagery using NKSSM; visual comparison therefore evaluates how closely the model output reproduces the proxy-labeled surfaces under a common spatial geometry.

Overall, the visual assessment closely tracked the quantitative grouping: on average, high-mIoU tiles showed strong preservation of structures, clear and stable boundaries, and relatively well-preserved details; mid-range tiles exhibited minor but localized degradations; and low-mIoU tiles displayed frequent boundary irregularities and loss of fine features. This alignment indicates that the tile-level mIoU scores provide a meaningful summary of the visual quality of NKSSM outputs.

4.7.1 High Group (≥ 0.80 mIoU)

The high group exhibited near-perfect spatial alignment with the proxy labels. All four classes—Built-up, Cropland, Waterbody, and Woody Vegetation—were clearly delineated with sharp, coherent boundaries and minimal class confusion. Large-scale landform structures and block-level patterns were faithfully reconstructed, and many fine details were preserved; however, very narrow linear features (e.g., small intra-urban roads) were sometimes absorbed into adjacent classes rather than explicitly resolved.

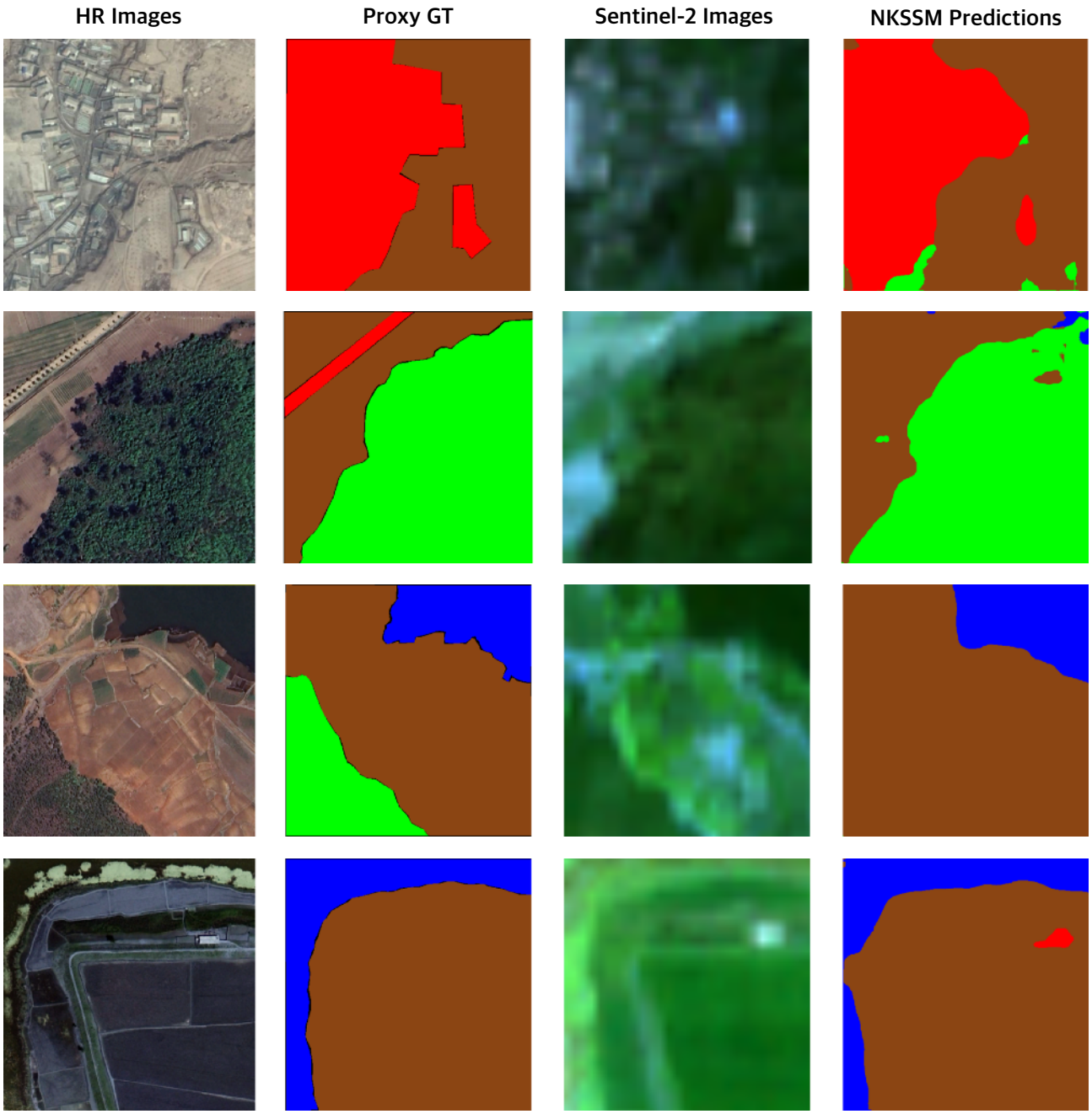


Figure 5. Representative Tiles from the High-Performance Group (mIoU ≥ 0.80). Examples of NKSSM outputs showing close agreement with proxy labels. All four classes are clearly delineated, with relatively well-preserved boundaries and minimal confusion.
Source: Google Earth; Sentinel-2 (Google Earth Engine)

4.7.2 Mid Group (0.60-0.79 mIoU)

The mid group retained correct large-scale landform structures, but boundary and detail quality was noticeably degraded compared with the high group. Across classes, mapped patches were generally in the right locations and exhibited broadly correct shapes, yet their edges were more irregular and small annexes were sometimes omitted or merged into neighboring areas. As a result, the maps remain interpretable at regional scale but are less reliable for parcel-level or edge-focused analysis.

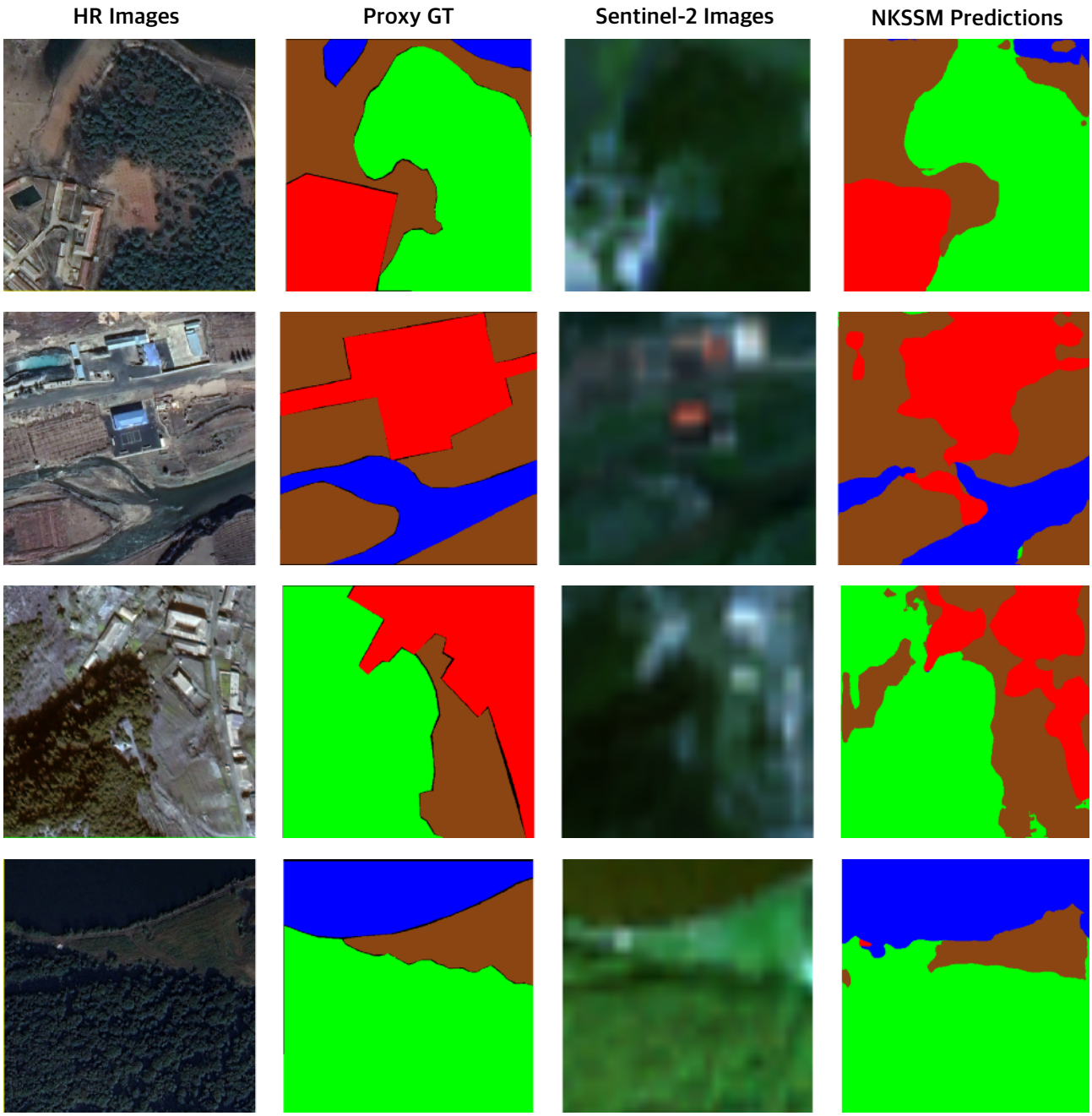


Figure 6. Representative Tiles from the Mid-Performance Group (0.60-0.79 mIoU). Tiles retain large-scale structures and broadly correct class shapes, but boundaries are more irregular with occasional class swaps. The maps support regional interpretation yet are less reliable for parcel-level or edge-focused analysis.
Source: Google Earth; Sentinel-2 (Google Earth Engine)

4.7.3 Low Group (< 0.60 mIoU)

The low group demonstrated frequent spatial mismatches and irregular class boundaries. Built-up areas were sometimes over- or under-estimated, and transitions between Cropland and Woody Vegetation were wide and unstable. Waterbody shapes occasionally differed in width or connectivity, and small features were often smoothed or merged.

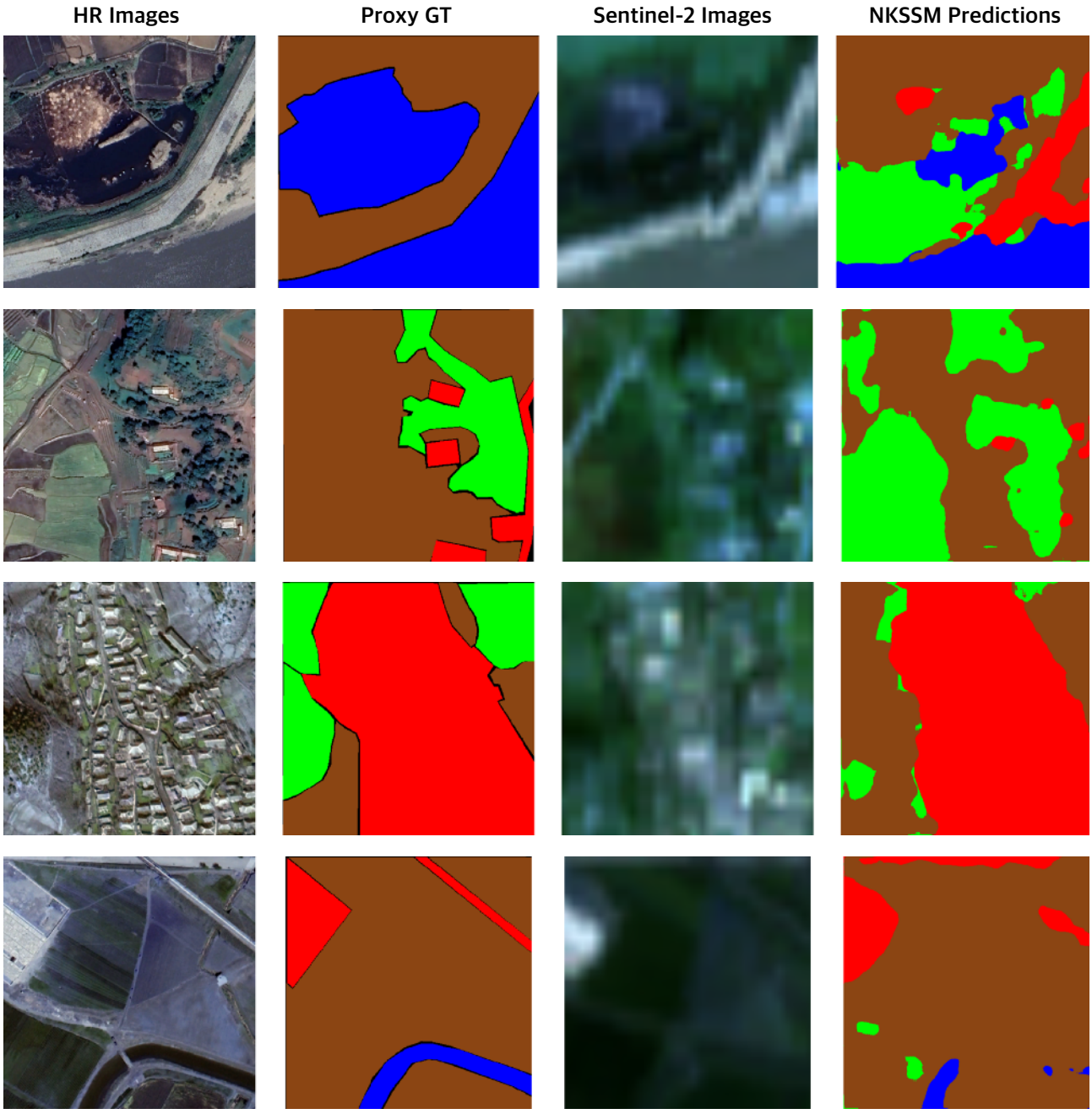


Figure 7. Representative Tiles from the Low-Performance Group (mIoU < 0.60). Examples showing NKSSM predictions with frequent spatial mismatches and irregular class boundaries. Built-up areas are occasionally over- or under-estimated, while Cropland-Woody Vegetation transitions appear unstable. Waterbody shapes differ in extent or connectivity, and fine details are often smoothed or merged.
Source: Google Earth; Sentinel-2 (Google Earth Engine)

5. Regional Application: Sinuiju, North Korea Case Study

5.1. Study Area and Context

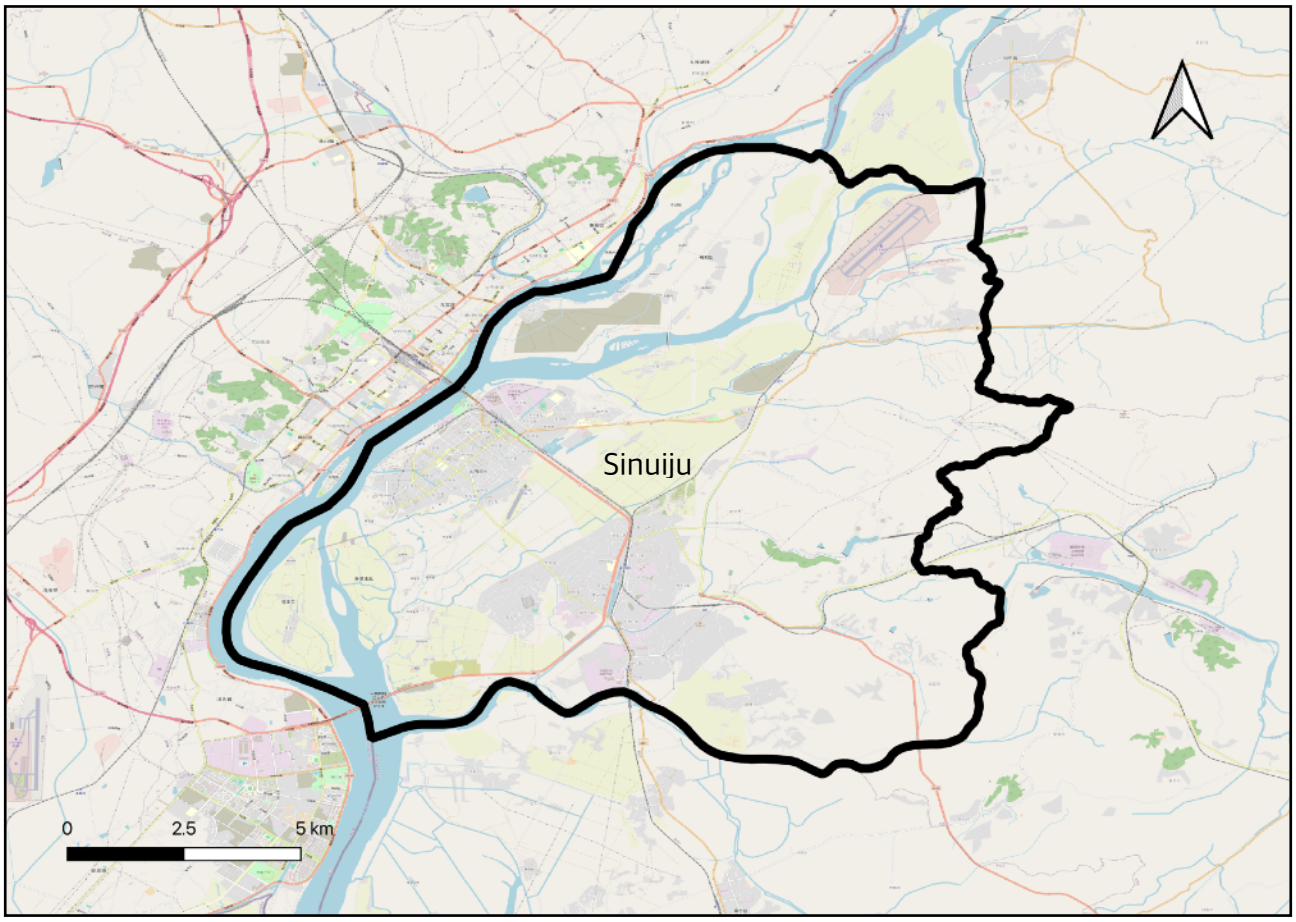


Figure 8. Map of Sinuiju, North Korea. Sinuiju presents a compact blend of urban, agricultural, forest, and river-island landscapes, offering an ideal setting for assessing NKSSM's 10 m segmentation fidelity and temporal stability.
Source: OpenStreetMap.

Sinuiju is a border city and the capital of North Pyongan Province in North Korea, located at 40°06′N, 124°25′E, across the Yalu River from Dandong, China. According to OpenStreetMap (OSM), its total area—including parts of the Yalu River—is approximately 190 km²; while our spatial datasets are handled in WGS 84 (EPSG:4326), this area estimate was calculated after projecting the OSM boundary to WGS 84 / UTM Zone 52N (EPSG:32652). Sinuiju is situated on a broad alluvial plain with gentle low hills.

The region is rich in water resources: in addition to the Yalu River, the Sapgyocheon stream flows along the southern border with Ryongchon County. Numerous river islands have formed over time due to sediment accumulation in the Yalu River. Among them, Wihwa Island (12.2 km²), Taji Island (13.4 km²), Ryucho Island (5.3 km²), and Im Island (6.2 km²) belong to Sinuiju. One-third of the city's farmland is located on these islands and the alluvial plains along the

Sapgyocheon. Forests are more concentrated in the eastern part of the city, primarily composed of pine and oak trees.

Sinuiju is a predominantly urban area, with rural surroundings. The urban center formed around Sinuiju Chongnyon Station, located near the border, in the area sometimes referred to as North Sinuiju. About 4 km to the south lies Nam Sinuiju Station, around which another city area, South Sinuiju, has developed; the two districts are connected by both the Pyongui railway line and several road connections. Both North Sinuiju and South Sinuiju are descriptive labels rather than official administrative names.

The compact juxtaposition of urban, agricultural, forest, and river-island landscapes makes Sinuiju an effective test area for evaluating the NKSSM's 10 m segmentation fidelity, boundary performance, and temporal consistency.

5.2 Data and Land-Cover Map Construction

A LC map assigns every pixel to a surface class. In this study, we generate a high-fidelity 10 m annual LC series for Sinuiju, North Korea, covering 2019–2025. The dataset is designed to be (i) spatially consistent across years, (ii) statistically analyzable in terms of class areas, ratios, and transitions, and (iii) reproducible without post-hoc tuning.

We use four classes—Built-up, Cropland, Woody Vegetation, Waterbody—balancing semantic clarity with 10 m stability.

Annual inputs are Sentinel-2 L2A (10 m) constrained to August–September to reduce clouds/ seasonal drift and maintain comparable illumination geometry. All rasters are EPSG:4326 and share an identical tiling layout; cloud-heavy tiles are excluded by a ≤ 20% selection rule.

Annual maps are generated by NKSSM (RGB+NIR 4-band) with direct softmax→argmax assignment. No post-processing (no smoothing/morphology/threshold-tuning) is applied. This guarantees exact reproducibility (same input ⇒ same output), avoids researcher subjectivity across years, and keeps downstream uncertainty modeling simple.

Importantly, no North Pyongan tiles were included in the training/validation/test sets (Table 1), so all results in Section 5 constitute an OOD deployment.

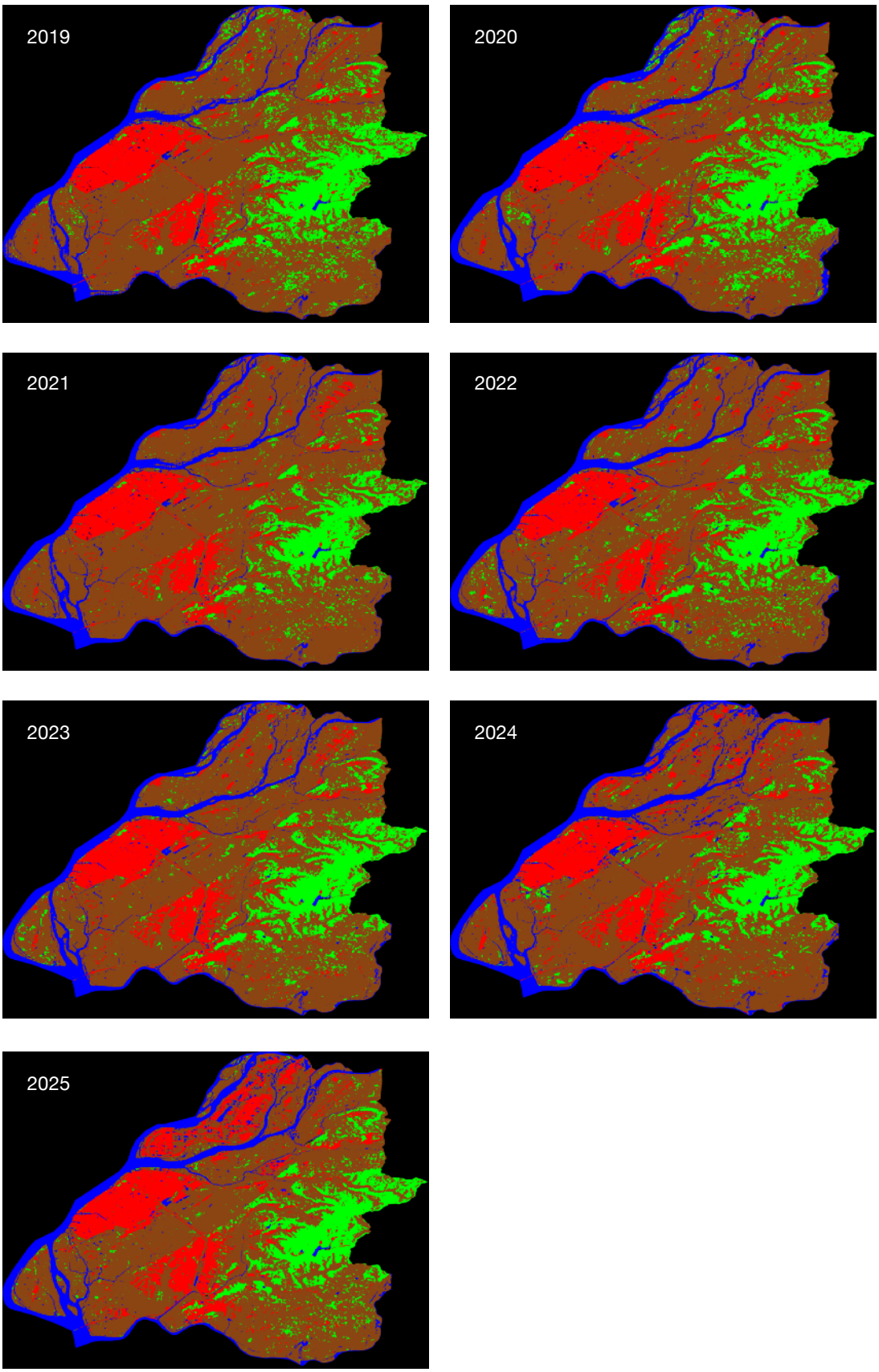


Figure 9. Annual Land-Cover Maps of Sinuiju (2019–2025, Aug–Sep). Time series of NKSSM-derived land-cover maps showing Built-up (red), Cropland (brown), Woody Vegetation (green), and Waterbody (blue) from 2019 to 2025. All maps were produced directly from Sentinel-2 L2A imagery (10 m) restricted to August–September to ensure consistent illumination and phenological conditions, with no manual or rule-based post-processing applied.

5.3 Annual Land-Cover Composition

In terms of annual land-cover composition, Cropland remains the dominant LC type throughout the observation period, occupying roughly two-thirds of Sinuiju's total area. Built-up areas gradually expand from ~12% in 2019 to ~16% in 2025, consistent with the positive slope (+1.10 km^2/yr) identified in the Theil-Sen trend analysis (Section 5.4). Waterbody proportions show a modest increase from ~8% to ~11%, indicating stable yet slightly expanding surface-water representation at 10 m resolution. Woody Vegetation varies modestly between ~10% and 15%, reflecting localized dynamics rather than systematic change. Across all four classes, interannual variability remains small, with SDs of only ~1-3%, confirming that year-to-year proportions are highly stable.

Table 10. Annual Land-Cover Composition (%)

Class	2019	2020	2021	2022	2023	2024	2025	Mean ± SD
Built-up	11.81	13.85	11.69	11.99	12.74	14.19	16.25	13.22 ± 1.60
Cropland	65.52	62.01	69.45	66.20	65.10	64.80	60.19	64.75 ± 2.98
Waterbody	7.72	9.53	8.62	9.31	9.56	10.47	10.88	9.44 ± 1.00
Woody Vegetation	14.84	15.02	10.16	12.50	12.59	10.50	12.66	12.90 ± 1.86
No Data	0.12	0.09	0.09	0.01	0.01	0.04	0.02	0.05 ± 0.04

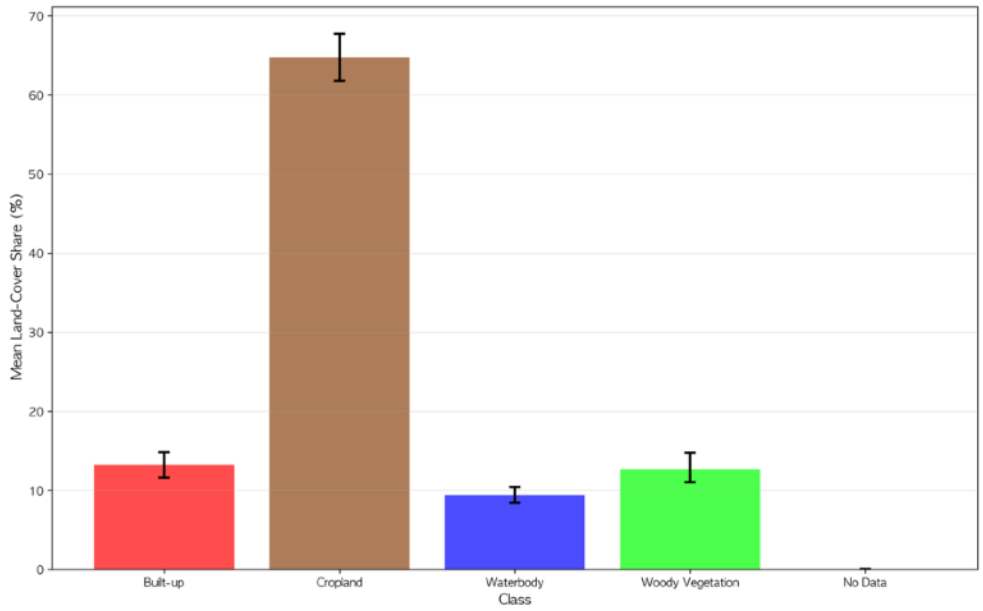


Figure 10. Mean Land-Cover Composition in Sinuiju (2019-2025). Mean proportional share of each LC class, averaged across the 2019-2025 period. Cropland dominates the landscape (~65%), followed by Built-up (~14%), Woody Vegetation (~12%), and Waterbody (~9%). Error bars denote ±1 SD across annual estimates, indicating stable class proportions with minimal interannual fluctuation.

All interannual variations fall within the ±20% area-level uncertainty adopted in Section 5.4, confirming that NKSSM's year-to-year LC estimates are statistically robust despite inherent classification noise.

5.4 Class-Specific Area Uncertainty and Robust Trends

Section 4 estimates the test-set performance of the final NKSSM model (seed 33) using 1,000 bootstraps, yielding mIoU 0.7075 ± 0.0309, Pixel Accuracy 0.8410 ± 0.0174, Cohen's kappa 0.7608 ± 0.0278, and mean MAE 0.2693 ± 0.0355. These are pixel-level metrics. Because directly propagating pixel errors to city-scale areas (km²) would overstate uncertainty due to spatial clustering of misclassifications, we employ class-specific area bands scaled by pixel MAE.

Table 11. Annual Land-Cover Areas by Class in Sinuiju, 2019-2025 (km²)

Class	2019	2020	2021	2022	2023	2024	2025
Built-up	22.515	25.458	22.287	22.86	24.304	27.069	30.991
Cropland	124.938	118.242	132.429	126.242	124.138	123.573	114.773
Waterbody	14.719	18.172	16.432	17.745	18.224	19.958	20.744
Woody Vegetation	28.301	28.651	19.373	23.832	24.006	20.023	24.137
No Data	0.219	0.169	0.17	0.014	0.021	0.07	0.047
Total	190.692	190.692	190.691	190.693	190.693	190.693	190.692

Note. Values are point estimates of annual class areas (km²). Trend analysis in Section 5.4 applies class-specific uncertainty bands of ±25.5% (Built-up), ±14.5% (Cropland), ±22.1% (Waterbody), and ±18.0% (Woody Vegetation), renormalized each year to a total area of 190.7 km².

The scaling rule is:

$$r_c = 0.20 \times \frac{\mathrm{MAE}_c}{\mathrm{MAE}_{\mathrm{overall}}} \quad (0.10 \le r_c \le 0.35)$$

$\mathrm{CI}_t^{(c)} = \left[(1 - r_c)A_t^{(c)}, (1 + r_c)A_t^{(c)}\right]$. and for year t with mapped area $A_t^{(c)}$,

Lower and upper bounds are then linearly renormalized per year (separately) so that class sums equal 190.7 km² (OSM boundary). Under this rule the bands are: Built-up ±25.5%, Cropland ±14.5%, Waterbody ±22.1%, Woody Vegetation ±18.0%.

For 2019-2025, we regress annual class areas using the Theil-Sen estimator and form 95% confidence intervals for slopes via 1,000 bootstrap perturbations within the class-specific bands. A result is labeled robust when the slope CI excludes zero. We also report Mann-Kendall (MK) results using a two-sided α = 0.05 as the primary threshold and α = 0.10 as a pre-

specified secondary criterion—appropriate for an exploratory analysis with a short time series (n = 7), where statistical power is limited; findings at the 10% level are described as marginal.

The four classes follow distinct trajectories. Built-up increases at +1.11 km^2 yr^{-1} (95% CI [+0.01, +2.39], robust), with MK tau = 0.62, p = 0.072 (marginal at the 10% level). Waterbody increases at +1.00 km^2 yr^{-1} (95% CI [+0.01, +1.14], robust), with MK tau = 0.81, p = 0.016 (significant at 5%); this indicates a gradual expansion of visible surface water at 10 m. By contrast, Cropland shows –1.33 km^2 yr^{-1} (95% CI [–4.07, +0.36], not robust; MK tau = –0.43, p = 0.230) and Woody Vegetation –0.90 km^2 yr^{-1} (95% CI [–1.88, +0.26], not robust; MK tau = –0.14, p = 0.764); both display declines that are not statistically confirmed.

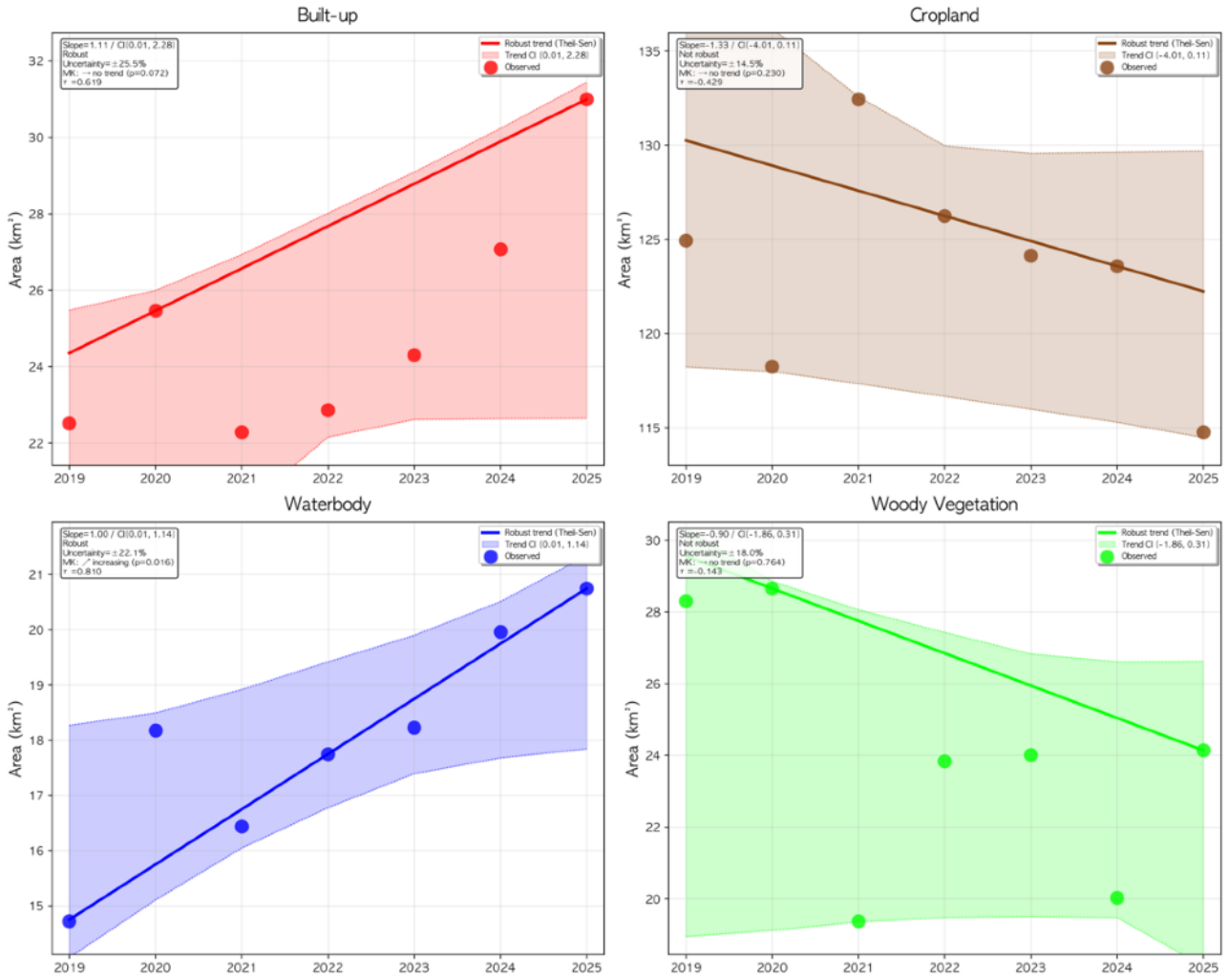


Figure 11. Class-wise Area Trends in Sinuiju (2019-2025). Annual LC area trajectories for Built-up, Cropland, Waterbody, and Woody Vegetation classes. Shaded regions denote 95% confidence intervals estimated via 1,000 bootstraps within class-specific uncertainty bands. Built-up and Waterbody exhibit robust positive trends under the Theil-Sen estimator, while Cropland and Woody Vegetation show modest, non-significant declines. Mann-Kendall statistics support monotonic increases for Built-up (MK tau = 0.62, p = 0.072) and Waterbody (MK tau = 0.81, p = 0.016).

Spatially, Built-up growth coincides with localized Cropland contraction, while Woody Vegetation exhibits oscillatory variability. This pattern appears to be driven by seasonal

phenology and spectral ambiguity in late-August/September imagery, in which tall summer crops can temporarily mimic woody signatures.

Table 12. Trend Summary under MAE-Scaled Area Bands

Class	Slope (㎢/yr)	95% CI	Robust	MK tau	MK p-value	MK Significance
Built-up	1.11	[0.01, 2.39]	robust	0.62	0.072	marginal (α=0.10), not significant at 0.05
Cropland	-1.33	[-4.07, 0.36]	not robust	-0.43	0.230	not significant
Waterbody	1.00	[0.01, 1.14]	robust	0.81	0.016	significant (α=0.05)
Woody Vegetation	-0.90	[-1.88, 0.26]	not robust	-0.14	0.764	not significant

In sum, Built-up and Waterbody show robust monotonic increases, whereas Cropland and Woody Vegetation exhibit interannual variability without statistically significant monotonic trends.

5.5 Land-Cover Transition Dynamics in Sinuiju: Persistence and Exchanges

Over the 2019-2025 period, the LC transitions in Sinuiju reveal a mixed pattern of stable categories and dynamic boundary zones. Ratio-normalized transition matrices were computed for each consecutive year pair (2019→2020, ⋯, 2024→2025) and aggregated to evaluate cumulative changes across the seven-year span. These transition patterns are consistent with the year-to-year net change structure in Figure 12, where Built-up and Waterbody exhibit small but persistent positive increments, while Cropland and Woody Vegetation show alternating gains and losses driven by localized boundary adjustments and phenological shifts.

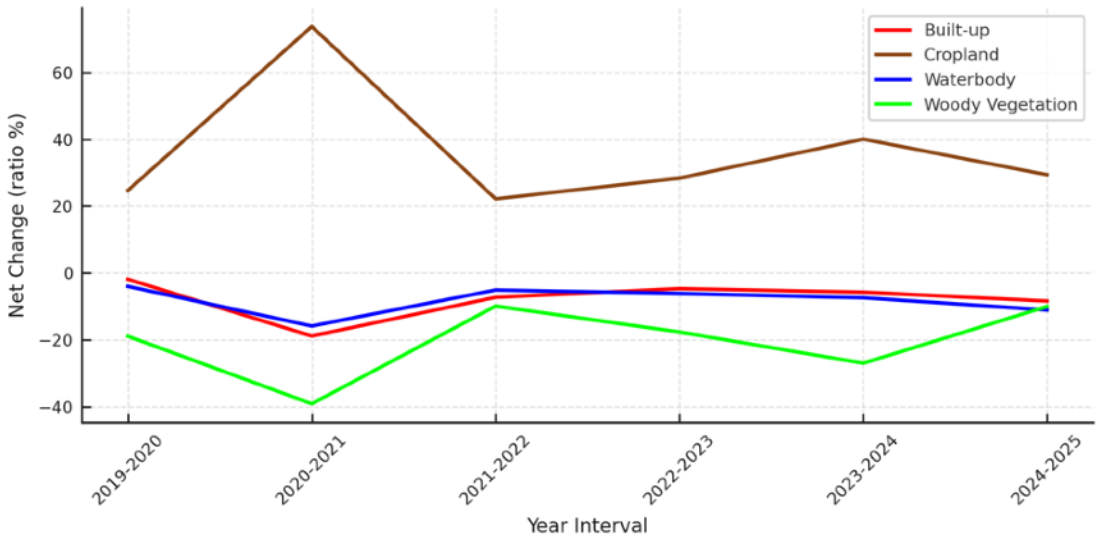


Figure 12. Annual Net Change by Class (2019-2025). Normalized annual percentage changes in LC area. Built-up and Waterbody show consistent increases, whereas Cropland and Woody Vegetation fluctuate with short-term gains and losses.

Over the 2019-2025 period, cumulative transitions show that Built-up and Waterbody classes exhibit high self-retention (≈80%), indicating that urban and surface-water areas remained spatially consistent at the 10 m scale throughout the period. In contrast, Cropland and Woody Vegetation display bidirectional exchanges in the range of 18-27%. Approximately 5-7% of cropland consistently transitioned into Built-up. Woody Vegetation shows both inflows from cropland and localized edge loss, while Waterbody remained generally stable apart from minor seasonal fluctuations in the Yalu River floodplain.

Table 13. Characteristic Transition Structure, 2019-2025 (ratio %, representative ranges)

From \ To	Built-up	Cropland	Waterbody	Woody Vegetation
Built-up	high (~80)	low	low	low
Cropland	modest	~60-65	low-mid	~18-27
Waterbody	very low	low	high (~80)	low
Woody Vegetation	low	~15-20	low	~70-75

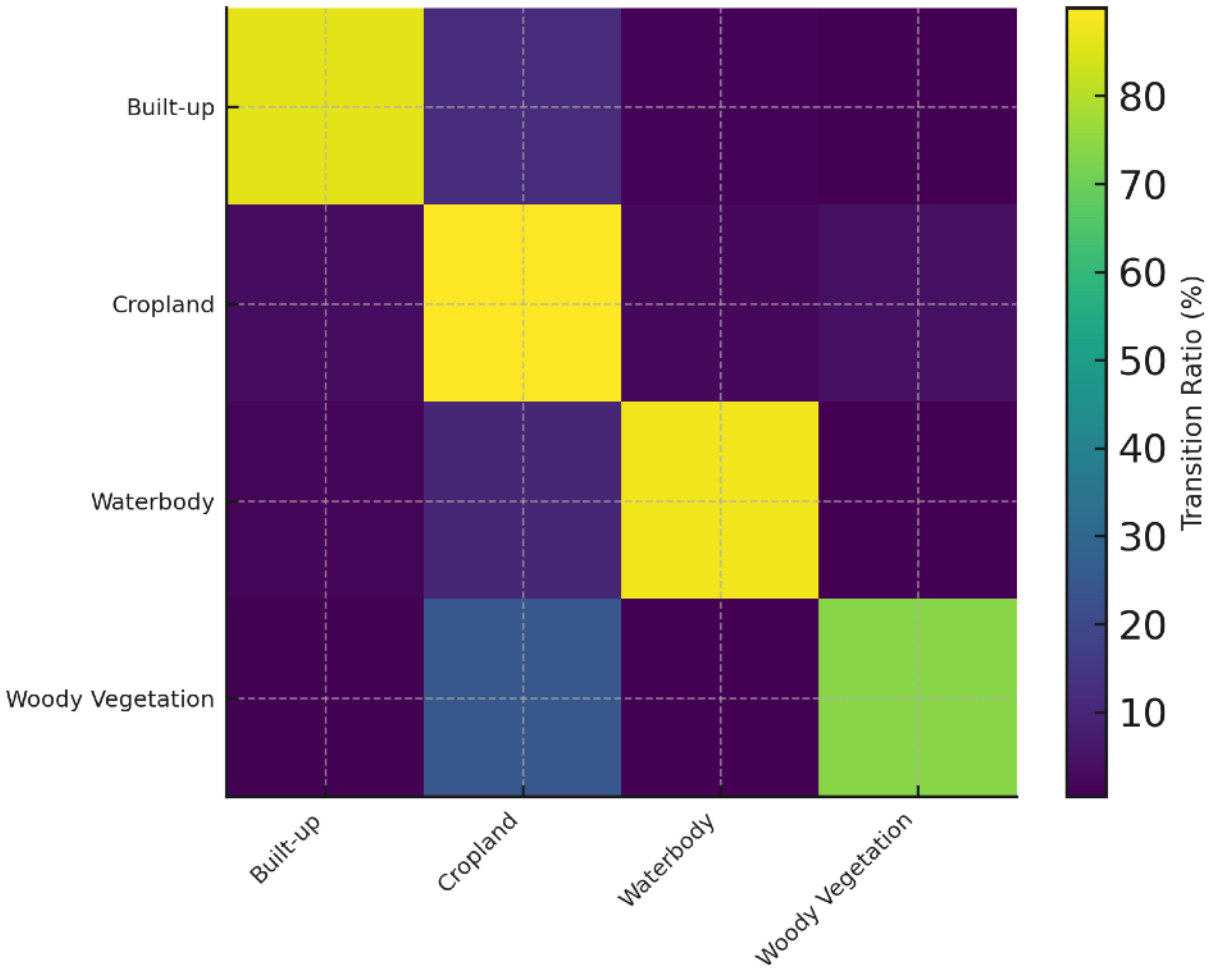


Figure 13. Average Transition Matrix, 2019-2025 (Ratio %). Average ratio-normalized land-cover transitions across all year pairs (2019-2025). Built-up and Waterbody show strong self-retention (≈80%), while Cropland and Woody Vegetation exhibit bidirectional exchanges (≈18-27%) and modest Cropland→Built-up flows (≈5-7%). These patterns reflect stable urban and water areas alongside dynamic agricultural-vegetation boundaries.

These patterns, observed cumulatively between 2019 and 2025, align with the overall direction of change captured by the temporal trend analysis—an expansion of Built-up and Waterbody areas accompanied by weak declines in Cropland and Woody Vegetation.

5.6 Qualitative Validation of Spatial and Temporal Consistency

5.6.1 Spatial Correspondence between NKSSM Predictions and Sentinel-2 Imagery (2025)

To verify spatial consistency, the 2025 LC map generated by the NKSSM—composed of 4,359 predicted tiles (512×512 each)—was compared with Sentinel-2 composite imagery (August-September 2025) within a common geospatial reference (EPSG:4326).

Both maps exhibit strong spatial correspondence across urban and peri-urban zones. The dual-core configuration of Sinuiju—comprising North and South Sinuiju—is distinctly represented, linked by continuous railway and roadway corridors. The downstream Yalu River islands (Wihwa, Taji, Ryucho, and Im) retain their distinct morphology in both representations. The riverine meanders—including those of the Sapkyocheon—are also precisely aligned. Key landmarks—including the eastern forest zone, the Uiju Airfield runway, and the New Yalu River Bridge—are distinctly preserved. Taken together, these correspondences demonstrate the spatial plausibility and structural fidelity of the NKSSM predictions for 2025.

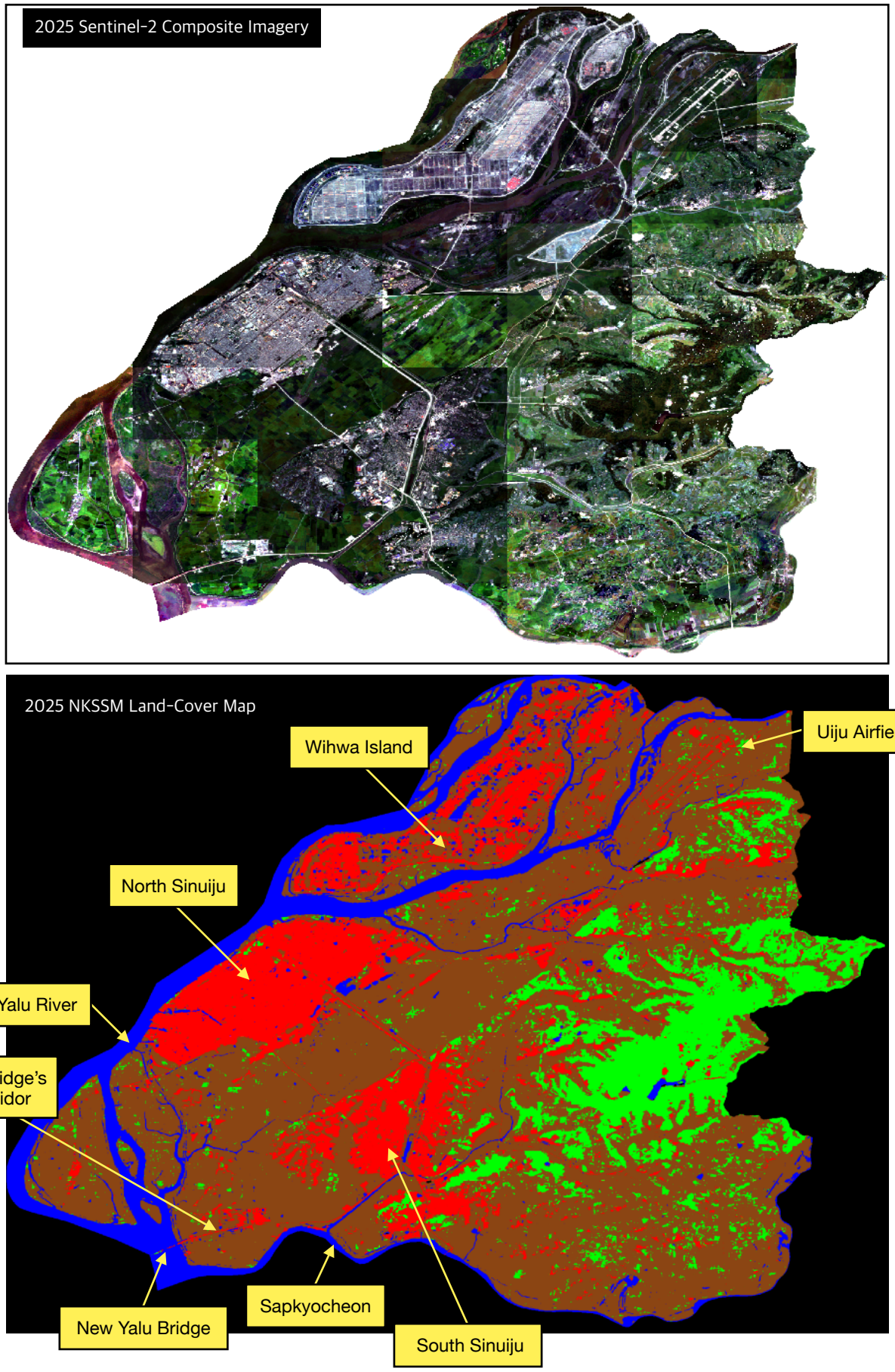


Figure 14. Sentinel-2 Composite Imagery and NKSSM Land-Cover Map (2025). The 2025 NKSSM LC map aligns closely with Sentinel-2 composite imagery. The strong correspondence indicates high spatial plausibility and structural fidelity of the predictions.
Source: Sentinel-2 L2A (Google Earth Engine).

5.6.2 Qualitative Verification through Representative Transitions (2019-2025)

When comparing annual LC maps of Sinuiju (Figure 9; location referenced in Figure 13), three representative transitions illustrate the recent LC evolution of Sinuiju.

- Formation of the New Yalu Bridge and its road corridor: Following the bridge's completion in 2014, a previously absent road corridor becomes clearly identifiable from 2022 onward. In Figure 9, no connector road is visible between the North Korea abutment of the New Yalu Bridge and South Sinuiju in 2019-2020. Beginning in 2021, a continuous linear feature consistent with a paved carriageway appears and remains traceable thereafter. This timeline aligns with reporting that the connector-road project began in late 2019 (OhmyNews, May 4, 2020) and with on-site imagery of the construction zone captured on April 28, 2020 (Figure 14). Taken together, the cartographic and documentary evidence indicate that the bridge-to-South Sinuiju segment was under construction during 2019-2020 and had reached a level of completion by 2021, thereby establishing the new transport axis visible from 2022.

Figure 15. Construction site of the New Yalu Bridge connector road. Photographed on April 18, 2020. The works shown form the connector road between the bridge and South Sinuiju, which was under construction in 2019-2020 and established as a continuous transport corridor by 2021-2022.
Source: Jongchol Park.

- Transformation and restoration of Uiju Airfield: The runway of Uiju Airfield, located northeast of Sinuiju, is clearly visible in 2019-2020. From 2021 to 2024, however, the same area appears in red on the LC maps—an indication of Built-up. This red zone gradually contracts over subsequent years and, by 2025, the runway once again becomes distinctly recognizable. These changes correspond closely to reports that the airfield's runway was temporarily repurposed as a logistics depot during the COVID-19 period in 2021 and was subsequently cleared and restored by 2025 for renewed aviation use.

- Flood damage and redevelopment of Wihwa Island: Severe flooding in 2024 reshaped the island's land surface; by 2025, new residential zones and greenhouse farms had emerged. According to Figure 9, Wihwa Island was largely composed of Cropland up to 2024, represented in brown on the LC maps. In 2025, however, the same areas appear in red, indicating Built-up zones. This visual evidence supports the observation that the flood-affected agricultural land was subsequently redeveloped for residential and greenhouse use, reflecting a rapid transformation in land utilization.

Figure 16. Newly Constructed Residential Houses and Greenhouse Farm Complex on Wihwa Island in 2025. These facilities occupy areas that shifted from Cropland to Built-up following severe 2024 flooding, reflecting rapid post-flood redevelopment. Source: Korea Central News Agency.

Together, these cases demonstrate the temporal coherence, spatial plausibility, and interpretive reliability of the NKSSM-derived predictions.

Overall, the NKSSM predictions from 2019 to 2025 demonstrate that the model consistently reproduces real-world surface changes without introducing random distortions or structural noise.

5.6.3 Reproducible Observation of the Unseen: Lessons from the Sinuiju Case

The Sinuiju case offers a modest but practical answer to the question raised in the introduction: how can open satellite data be transformed into usable and credible information where direct GT is unavailable?

By applying a transparent and reproducible workflow, this study shows that meaningful LC analysis is possible even in data-restricted regions.

The observed increase in built-up areas and waterbodies, together with the gradual decline of cropland, illustrates how Sinuiju has been quietly reshaped by both development and environmental change—visible in the construction of the New Yalu Bridge, the redevelopment of Uiju Airfield, and the post-flood reorganization of Wihwa Island.

While these findings are limited by temporal and geometric inconsistencies in proxy labels, they nonetheless suggest that reproducible methods can help narrow the usability gap that separates open data from actionable knowledge. In this sense, open satellite data and foundation-model AI, when carefully combined through reproducible workflows, may help make unseen places more visible.

6. Discussion

6.1 Data Democratization as Operational Usability

This study presents a reproducible, proxy-based workflow that fine-tunes a foundation model (Satlas Pretrain) using open satellite imagery for producing interpretable LC information for label-scarce regions. The application to Sinuiju, North Korea suggests that NKSSM functions not merely as a classifier but as a practical instrument for data democratization—transforming open-data accessibility into operational usability. Instead of relying on institutional ground surveys, the model establishes a form of proxy GT using freely available HR imagery, providing a practical basis for consistent and interpretable LC information over time.

In settings where in-situ validation is constrained, the proxy-label + fine-tuning pipeline provides a pragmatic form of usable GT: data that are not perfect, but sufficiently consistent, transparent, and reproducible—allowing meaningful longitudinal analysis (2019-2025) and transition accounting.

Such an approach may also contribute to the broader agenda of equitable EO, where open data are not only available but also usable.

In this sense, the study reframes the concept of data democratization from a question of access ("Is the data open?") to a question of use ("Can open data be systematically transformed into reliable, policy-relevant evidence?"). This shift emphasizes not technological capability but a procedural transparency—the ability for others to reproduce, verify, and extend the results using the same open resources.

We also note that the workflow can be viewed through the lens of appropriate technology—context-sensitive, affordable, and socially responsive uses of technology (often associated with Schumacher, 1973, and, later, Hazeltine & Bull, 1999). Rather than relying on capital-intensive infrastructures, the approach seeks what is feasible under real constraints. In this light, combining open satellite data with a foundation model through a transparent, reproducible pipeline offers one modest way to repurpose advanced computation for regions that are otherwise underserved, helping to turn open data into usable, credible, and more inclusive environmental knowledge.

6.2 Reproducibility and Transparent Assumptions

The proposed workflow emphasizes transparent preprocessing, consistent spatial alignment, and clearly defined evaluation procedures, which together support both statistical and procedural reproducibility. As detailed in Section 4, the final NKSSM model (seed 33, epoch 68)

achieved mean mIoU = 0.7075 ± 0.0309, pixel accuracy = 0.8410 ± 0.0174, Cohen's kappa = 0.7608 ± 0.0278, and mean MAE = 0.2693 ± 0.0355 under 1,000 bootstrap iterations. These metrics provide a statistically explicit summary of model performance on held-out tiles.

To avoid overestimating uncertainty when scaling from pixel- to area-level indicators, Section 5 applied a conservative ±20% perturbation envelope to map-integrated class areas (km^2). This approach makes the underlying assumptions fully explicit: no post-processing is used, all maps are restricted to August-September Sentinel-2 L2A imagery, a single geodetic reference (EPSG:4326) is maintained, and a consistent tiling scheme is applied. Given these design choices and openly documented parameters, independent researchers can, in principle, reproduce the full pipeline, replicate the reported statistics, and audit the sensitivity of results to key assumptions. In data-scarce settings, such transparent and reproducible design is essential for building confidence in LC estimates and for enabling others to verify, update, or extend the results.

6.3 Methodological Limitations

Several limitations qualify the scope of the NKSSM results. First, the proxy labels are derived from ~0.5 m Google Earth imagery and spatially aligned to Sentinel-2 L2A composites (see Section 3.2). Differences in viewing geometry, projection, and acquisition timing introduce residual geometric and temporal misalignment—especially near slopes, high-rise structures, and seasonally dynamic surfaces. These labels therefore function as near-contemporary proxy references rather than absolute GT.

Second, the August-September observation window and four-band Sentinel-2 input lead to known spectral ambiguities between tall summer crops and woody vegetation. As discussed in Section 3, this confusion reflects inherent data and phenology constraints and could be reduced in future work through multi-season compositing, additional spectral bands, or SAR and DEM integration.

Third, the simplified four-class scheme (Built-up, Cropland, Woody Vegetation, Waterbody) stabilizes training but reduces thematic granularity, limiting representation of transitional types such as Bare Land and Grassland.

Fourth, tile-wise inference without full overlap and blending likely yields conservative estimates of boundary-focused metrics (e.g., BF1, Trimap-IoU), while the decision to forgo morphological smoothing or post-hoc filtering preserves strict reproducibility at the cost of minor local irregularities along roads, riverbanks, and urban edges.

Finally, area-level uncertainty was propagated from pixel-level error using a conservative ±20% envelope based on the MAE-derived procedure in Section 5.4; this ensures that temporal and transition analyses remain commensurate with the 10 m spatial resolution but also highlights the need for cautious interpretation of fine-scale area differences.

Beyond these methodological constraints, the geographic representativeness of NKSSM remains limited. Approximately 60% of training tiles originate from the southern lowlands of the Hwanghae provinces, whereas mountainous regions (e.g., Jagang, Ryanggang, Kangwon) are underrepresented (Section 3.3). The current model should therefore be viewed not as a nationwide classifier for North Korea but as a proof-of-concept demonstrating a reproducible workflow for regional fine-tuning under data-scarce conditions. Future extensions should prioritize additional samples from northern and eastern provinces to improve spatial generalization and better capture the diversity of terrain, vegetation, and settlement patterns across the country. In this sense, Sinuiju functions as an OOD test case relative to all development splits, illustrating how the framework behaves when deployed beyond its primary training footprint.

7. Conclusion

This study demonstrates that credible LC information can be produced even in the absence of conventional GT by pairing proxy supervision with foundation-model fine-tuning. The NKSSM workflow—built on transparent preprocessing, consistent spatial alignment, and verifiable inference—was designed to convert open satellite accessibility into operational usability. Applied to Sinuiju for the 2019-2025 period, it produced reproducible 10 m annual LC maps that can be independently inspected, reconstructed, and audited using the same open resources.

Quantitative evaluation indicates that the final model (seed 33, epoch 68) performs robustly on held-out tiles: mean mIoU = 0.7075 ± 0.0309, Cohen's kappa = 0.7608 ± 0.0278, pixel accuracy = 0.8410 ± 0.0174, and mean MAE = 0.2693 ± 0.0355 across 1,000 bootstrap iterations (Section 4). For map-integrated area indicators, we adopted a conservative ±20% envelope based on the MAE-derived procedure of Section 5.4, avoiding the over-extension of pixel-level noise to city-scale quantities. Within this uncertainty bound, several directional tendencies are consistent: Built-up and Waterbody expanded, Cropland declined, and Woody Vegetation fluctuated without a persistent trend. Visual comparisons with Sentinel-2 composites and HR scenes corroborate the spatial plausibility of these patterns.

Several limitations qualify the interpretation of these results. Proxy labels derived from ~0.5 m Google Earth imagery exhibit residual geometric and temporal misalignment with Sentinel-2 composites, functioning as near-contemporary references rather than absolute GT. Seasonal and four-band spectral constraints in August-September imagery lead to known ambiguities between tall summer crops and woody vegetation. A simplified four-class taxonomy reduces thematic granularity, while tile-wise inference without full overlap and the absence of post-processing introduce minor boundary irregularities. Area-level indicators rely on ranges rather than point estimates, reflecting the ±(14.5-25.5)% uncertainty envelope used in Section 5.4.

Beyond methodological constraints, geographic representativeness remains limited: approximately 60% of training tiles originate from the southern lowlands of the Hwanghae provinces, whereas mountainous regions are underrepresented. NKSSM should therefore be understood not as a nationwide classifier for North Korea but as a proof-of-concept demonstrating how proxy supervision and foundation-model fine-tuning can support reproducible LC mapping under severe data scarcity.

The design choice to use single-season Sentinel-2 optical imagery (August-September) prioritized temporal consistency and methodological clarity for reproducibility testing. Although multi-season and multi-sensor data (e.g., SAR) were available, incorporating them would have introduced additional alignment and normalization challenges beyond the scope of this initial

framework. The moderate bidirectional transitions between Cropland and Woody Vegetation (18–27%) largely reflect expected spectral overlap in late-summer conditions.

Future work will extend this framework by integrating multi-season compositing, multi-sensor fusion (SAR-optical-DEM), and expanded spectral inputs up to nine bands. Establishing standardized proxy-labeling protocols—combining algorithmic reproducibility with contextual expertise—may further support cross-regional applications in diverse data-scarce environments. These developments are not proposed as definitive solutions but as practical steps toward transparent, reproducible, and socially usable EO workflows that convert open data into credible environmental knowledge. In access-restricted regions such as North Korea, these steps offer a concrete blueprint for turning satellite openness into usable land-cover intelligence that can be routinely updated, scrutinized, and improved.

Acknowledgements

The authors used AI tools—Claude (Anthropic) for coding assistance and ChatGPT (OpenAI) primarily for English translation and writing support. All analyses, interpretations, and conclusions were reviewed and finalized independently by the authors.

References

Alem, A., & Kumar, S. (2022). Transfer learning models for land cover and land use classification in remote sensing image. Applied Artificial Intelligence, 36(1), 2014192. https://doi.org/10.1080/08839514.2021.2014192

Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A., & Ranjan, R. (2023). SatlasPretrain: A large-scale dataset for remote sensing image understanding. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 16772-16782.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ⋯, Liang, P. (2021). On the opportunities and risks of foundation models. arXiv Preprint, arXiv:2108.07258. https://arxiv.org/abs/2108.07258

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., Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. Scientific Data, 9, 251. https://doi.org/10.1038/s41597-022-01307-4

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 833-851). https://doi.org/10.1007/978-3-030-01234-2_49

Craglia, M., & Shanley, L. (2015). Data democracy: Increased supply of geospatial information and expanded participatory processes in the production of data. International Journal of Digital Earth, 8(9), 679-693. https://doi.org/10.1080/17538947.2015.1008214

Džanko, E., Kozina, K., Cero, L., Marijić, A., & Horvat, M. (2024). Rethinking data democratization: Holistic approaches versus universal frameworks. Electronics, 13(21), 4170. https://doi.org/10.3390/electronics13214170

ESA. (2022). WorldCover 2021 Product Validation Report - Version 2.0. European Space Agency, ESA WorldCover Project. Retrieved from https://worldcover2021.esa.int/documentation

Feng, H., Wang, Y., Li, Z., Zhang, N., Zhang, Y., & Gao, Y. (2023). Information leakage in deep learning-based hyperspectral image classification: A survey. Remote Sensing, 15(15), 3793. https://doi.org/10.3390/rs15153793

Florczyk, A. J., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., Melchiorri, M., Pesaresi, M., Politis, P., Schiavina, M., Sabo, F., Zanchetta, L. (2019). GHSL Data Package 2019 (JRC117104). Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/290498

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850-853. https://doi.org/10.1126/science.1244693

Hazeltine, B., & Bull, C. (1999). Appropriate technology: Tools, choices, and implications. Academic Press.

Kim, K., Nam, W., Park, S., Sunwoo, J., Yoo, K., Jung, S., & Hwang, J. (2025). Regional geography of D.P.R. Korea for the era of exchange and cooperation (Vol. 1: Sinuiju City, Junggang County, Samjiyon City, Chongjin City, Kim Chaek City, Sinpho City, and Hamhung City). Seoul: Stream & Forest Publishing Co.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., & Brumby, S. P. (2021). Global land use/land cover with Sentinel-2 and deep learning. In Proceedings of IGARSS 2021—IEEE International Geoscience and Remote Sensing Symposium (pp. 4704-4707). IEEE. https://doi.org/10.1109/IGARSS47720.2021.9553499

Kim, J., Jo, H.-W., Kim, W., Jeong, Y., Park, E., Lee, S., Kim, M., & Lee, W.-K. (2024). Application of the domain adaptation method using a phenological classification framework for the land-cover classification of North Korea. Ecological Informatics, 81, 102616. https://doi.org/10.1016/j.ecoinf.2024.102616

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166-177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E. (2020). Outlining where humans live: The World Settlement Footprint 2015. Scientific Data, 7, 242. https://doi.org/10.1038/s41597-020-00580-5

OhmyNews. (2020, May 4). Construction of the New Yalu Bridge connector road on the North Korean side resumes; opening expected around October - Professor Park Jong Chol of Gyeongsang National University releases photos of the construction site: "Four-lane expressway." OhmyNews. https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0002638179

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148, 42-57. https://doi.org/10.1016/j.rse.2014.02.015

Piao, Y., Jeong, S., Park, S., & Lee, D. (2021). Analysis of land use and land cover change using time-series data and Random Forest in North Korea. Remote Sensing, 13(17), 3501. https://doi.org/10.3390/rs13173501

Piao, Y., Xiao, Y., Ma, F., Park, S., Lee, D., Mo, Y., & Kim, Y. (2023). Monitoring land use/land cover and landscape pattern changes at a local scale: A case study of Pyongyang, North Korea. Remote Sensing, 15(6), 1592. https://doi.org/10.3390/rs15061592

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881

Saah, D., Johnson, G. W., Ashmall, B., Tondapu, G., Tenneson, K., Patterson, M. S., Poortinga, A., Markert, K., Hanh, N., Aung, K. S., Schlichting, L., Matin, M., Uddin, K., Aryal, R. R., Dilger, J., Ellenburg, W. L., Flores-Anderson, A., Wiell, D., Lindquist, E., ⋯, Chishtie, F. A. (2019). Collect Earth Online: An online tool for systematic reference data collection in land cover and use applications. Environmental Modelling & Software, 118, 166-171. https://doi.org/10.1016/j.envsoft.2019.05.004

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). SEN12MS-A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-1, 141-146. https://doi.org/10.5194/isprs-annals-IV-1-141-2019

Schumacher, E. F. (1973). Small is beautiful: Economics as if people mattered. Blond & Briggs.

Thapa, R. B., Matin, M., & Bajracharya, B. (2019). Capacity building approach and application: Utilization of Earth observation data and geospatial information technology in the Hindu Kush Himalaya. Frontiers in Environmental Science, 7, 165. https://doi.org/10.3389/fenvs.2019.00165

Twohig-Bennett, C., & Jones, A. (2018). The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. Environmental Research, 166, 628-637. https://doi.org/10.1016/j.envres.2018.06.030

Tyukavina, A., Stehman, S. V., Foody, G. M., Bontemps, S., See, L., Olofsson, P., Tsendbazar, N.-E., Radoux, J., Komarova, A., Serre, B. M., Song, X.-P., d'Andrimont, R., Koren, G., Potapov, P., Bullock, E. L., Campbell, P., de Bruin, S., Defourny, P., Friedl, M. A., ⋯, Xiao, X. (2025). Land Cover and Change Map Accuracy Assessment and Area Estimation Good Practices Protocol. Version 0.1. CEOS Working Group on Calibration and Validation Land Product Validation Subgroup. https://pure.iiasa.ac.at/id/eprint/20873/

Venter, Z. S., Gaughan, A. E., Barton, M., Creutzig, F., Mahtta, R., Seto, K. C., & Schiavina, M. (2022). Global 10 m land use land cover datasets: A comparison of Dynamic World, WorldCover and Esri Land Cover. Remote Sensing, 14(16), 4101. https://doi.org/10.3390/rs14164101

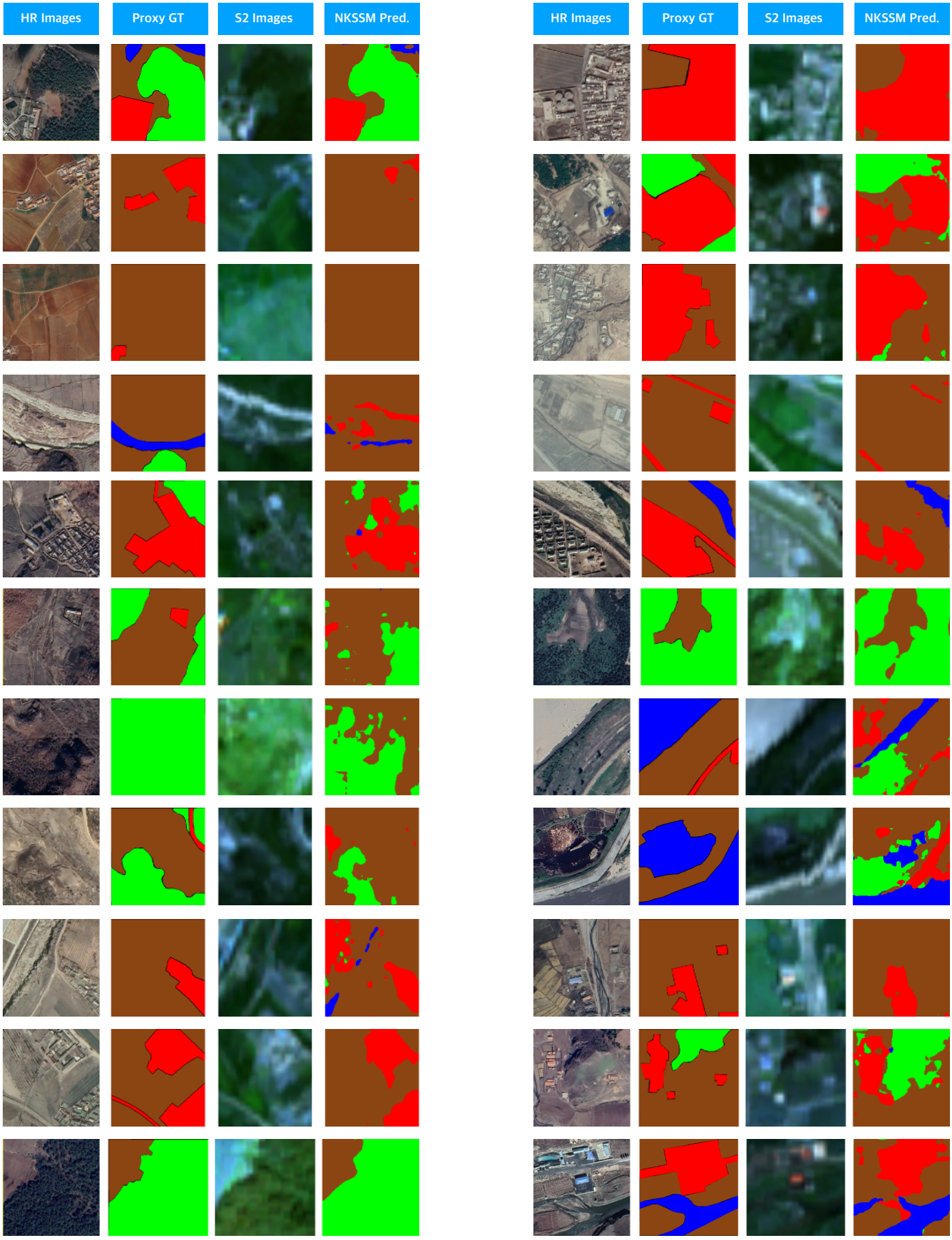
Xu, P., Tsendbazar, N.-E., Herold, M., de Bruin, S., Koopmans, M., Birch, T., Carter, S., Fritz, S., Lesiv, M., Mazur, E., Pickens, A., Potapov, P., Stolle, F., Tyukavina, A., Van De Kerchove, R., Zanaga, D. (2024). Comparative validation of recent 10 m-resolution global land cover maps. Remote Sensing of Environment, 311, 114316. https://doi.org/10.1016/j.rse.2024.114316

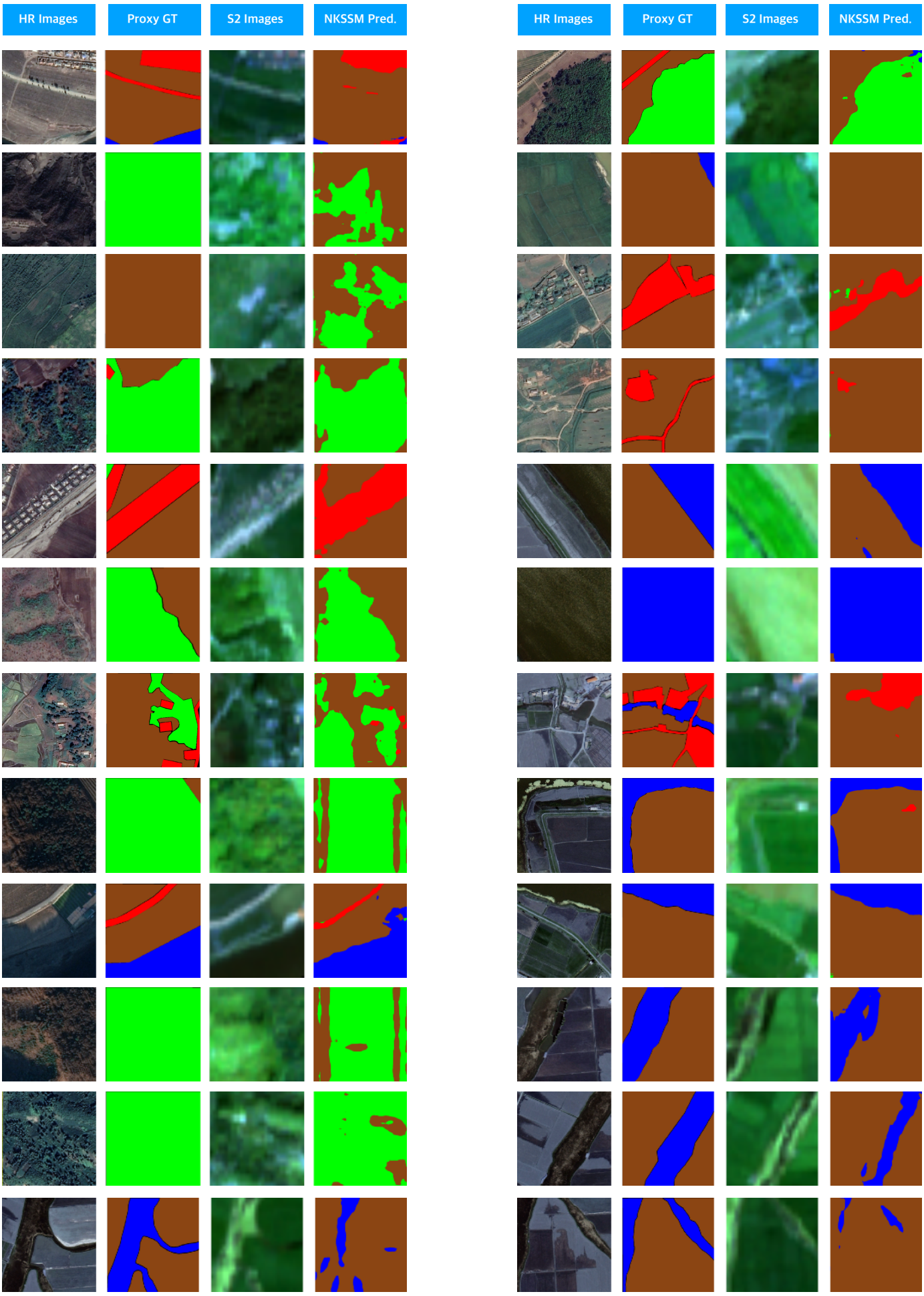
Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4), 8-36. https://doi.org/10.1109/MGRS.2017.2762307

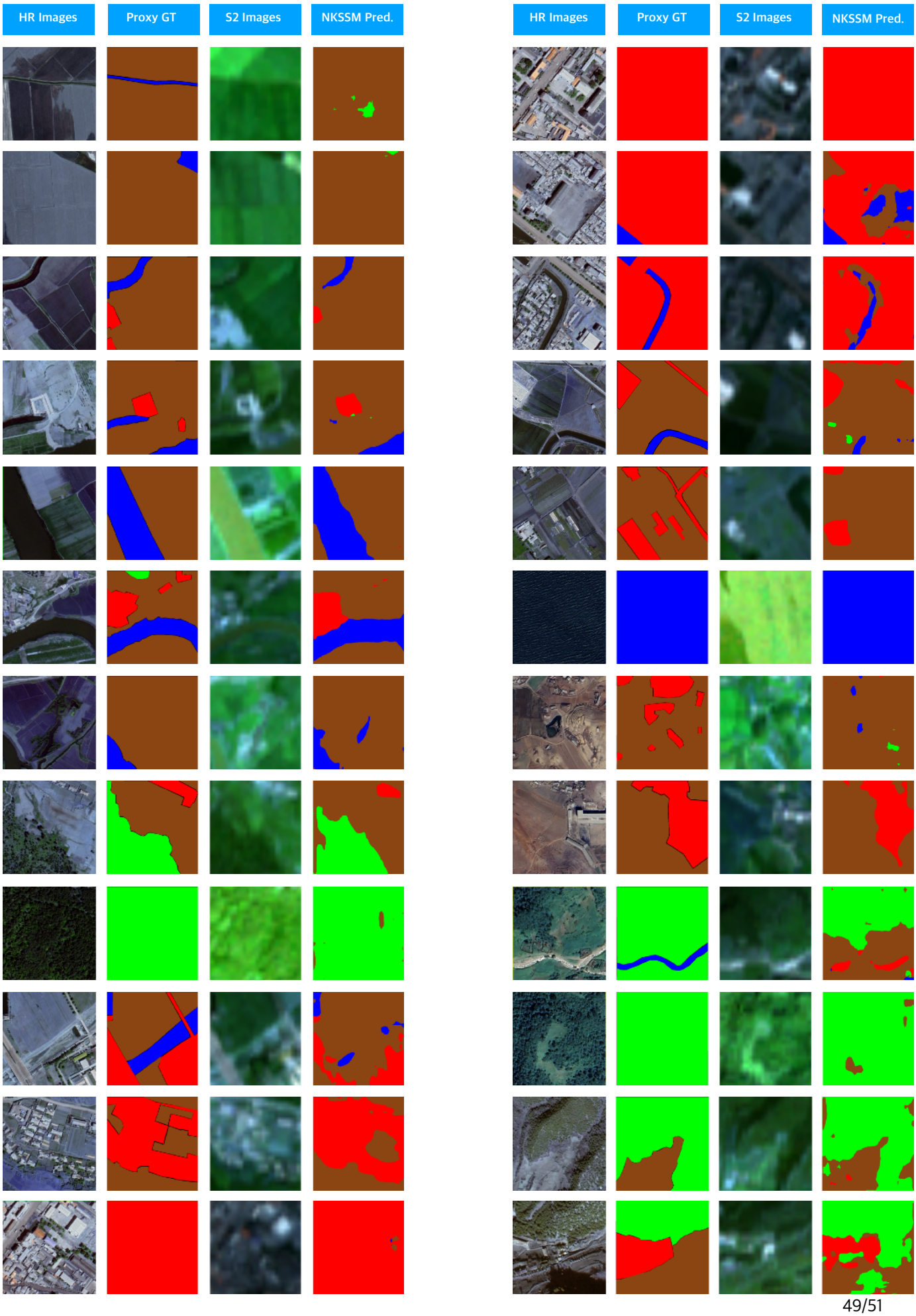
Appendix A. NKSSM Model Training Configuration and Environment

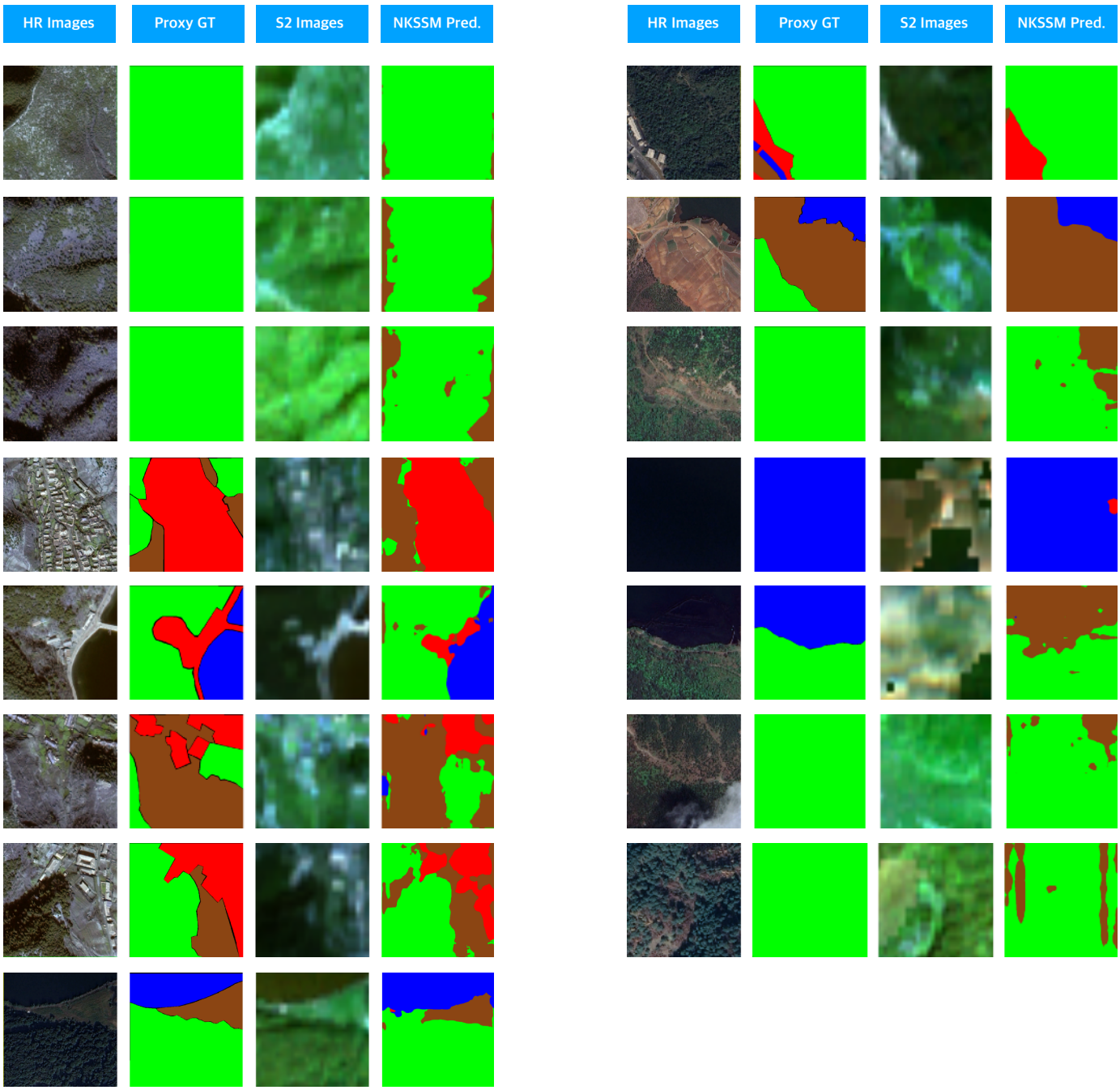
Item	Details
Model Architecture	Satlas Pretrain (ResNet-50 FPN encoder-decoder)
Input Bands	Sentinel-2 L2A RGB + NIR (4-band, 10 m GSD)
Output Classes	Built-up (0), Cropland (1), Woody Vegetation (2), Waterbody (3)
Loss Function	0.7 × Lovász-Softmax + 0.3 × Focal Loss
Training Epochs	Up to 100 epochs (early stopping at 70 epochs if no improvement)
Optimizer / Scheduler	AdamW (learning rate = 1×10^{-4}) + cosine annealing scheduler
Batch Size	8
Class Weights	Built-up 0.67 / Cropland 1.58 / Woody 2.00 / Waterbody 2.00
Training Techniques	Automatic Mixed Precision (AMP); Stochastic Weight Averaging (SWA after epoch 30)
Evaluation Metrics	Precision, Recall, F1, mIoU, Cohen's kappa, Boundary-F1, Trimap-IoU
Bootstrap Evaluation	1,000 resampling iterations with 95% confidence interval (CI) estimation
Random Seeds	33, 42, 72, 333 (four independent runs; seed 33 selected as final model)
Execution Environment	PyTorch 2.x + CUDA 11.x on NVIDIA L4 GPU (16 GB VRAM)
Result Characteristics	mIoU variance across seeds ± 0.01 → high reproducibility confirmed

Appendix B. Test-Set Qualitative Examples: HR Imagery, Proxy GT, Sentinel-2 Inputs, and NKSSM Predictions




HR Images
Proxy GT
S2 Images
NKSSM Pred.
HR Images
Proxy GT
S2 Images
NKSSM Pred.


HR Images
Proxy GT
S2 Images
NKSSM Pred.
HR Images
Proxy GT
S2 Images
NKSSM Pred.



Source: Google Earth; Sentinel-2 L2A (Google Earth Engine)

Appendix C. Bootstrap Evaluation Summary (n = 1,000 replicates)

Metric	Mean ± 95% CI	Across-Seed SD (mIoU unless noted)	P-value vs. baseline	Notes
mIoU	0.7075 ± 0.0309	± 0.0094	< 0.01	Baseline = random/ stratified uniform labeling
Cohen's kappa	0.7608 ± 0.0278	–	< 0.01	Chance-corrected agreement
MAE	0.2693 ± 0.0355	–	< 0.01	Pixel-wise error; mapped to ±20% area uncertainty in text
Boundary-F1	0.4255 ± 0.0796	–	< 0.01	Boundary sensitivity at narrow trimaps
Trimap-IoU (1-3 px)	0.6197 ± 0.0340	–	< 0.01	Mean over 1-3 px bands

- 33, 42, 72, 333 (four independent trainings; seed 33 selected as final model).
- Bootstrap unit: tile-level resampling (with replacement).
- CI method: percentile 95% CI from bootstrap distribution.
- p-value: proportion of bootstrap replicates where the model's metric ≤ baseline metric (two-sided where applicable).
- Interpretation: Small across-seed variance (±0.01 mIoU) and uniformly low p-values indicate high reproducibility and performance significantly above chance.