D9 – Model training data package v3

Deliverable Overview

Lead(s)	AWI
Contributor(s)	GFZ, DLR, FZJ
Work Package	WP 2.1 (Data collation, pre-processing, curation: Amazon & Polar)
Stage	Foundation Stage (I)
Duration	PM 11 (03/2025) – PM 22 (02/2026; planned), PM 24 (04/2026; revised)
Dependencies	None
Outcome (Type)	Data, Report
Link	
Status	☐ To be done ☐ In progress ☒ **Completed**

Executive Summary

This document describes datasets prepared for pretraining and fine-tuning the 3D-ABC FM over the Arctic–boreal permafrost region. All datasets are temporally co-located where possible and reprojected into a unified UTM tiling system based on the Military Grid Reference System. The dataset spans 4,994 tiles in the circumpolar Northern Hemisphere.

Results

For each dataset, the report describes the criteria for data selection, data preparation workflow and progress, and the delivered data. The primary data sources are:

- HLS: 30 m optical imagery, composited with cloud-minimization and patch-based tiling.
- TanDEM-X: Interferometric data at 20 m resolution, including coherence, vertical wavenumber, and incidence angle, ~10,000 acquisitions processed and expanding.
- ICESat-2: Spaceborne photon-counting lidar providing terrain elevation and canopy height percentiles, with >100 million high-quality samples.
- Copernicus 30 m GLO-30 DEM, providing terrain representation derived largely from TDX data.
- ERA5-Land: Climate Reanalysis data at 9 km resolution, including air temperature, surface temperature, pressure, solar radiation, and precipitation.

An additional fine-tuning dataset group includes:

- Terrain derivatives (slope, aspect, TPI, TWI) derived from DEM at 30 m resolution
- Permafrost data (extent, active layer thickness, soil temperature) at 1 km resolution
- Soil carbon reference data (~50,000 observations) from WoSIS Snapshot
- Airborne laser scanning data for forest structure and canopy height modeling