Data-Driven Mapping in the Summergreen–Evergreen Boreal Transition Zone in Eastern Siberia Using Seasonal Sentinel-2 Satellite Data and Machine Learning

Femke van Geffen

Summary

Boreal forests constitute the largest terrestrial biome on Earth and play a central role in the global climate system through their influence on carbon storage, surface energy balance, and land–atmosphere interactions (Bonan, 2008; Gauthier et al. 2015; Loranty et al., 2018). Within this biome, Eastern Siberia is characterized by a distinctive forest composition dominated by summergreen needleleaf forests, primarily *Larix* spp., which coexist and compete with evergreen needleleaf forests across an extensive transition zone shaped by permafrost conditions, fire regimes, and site-specific environmental constraints (Herzschuh et al., 2016, 2020; Kharuk et al., 2013a,b; Kruse et al., 2016, 2022; Loranty et al., 2021, Stuenzi et al., 2021). Changes in the distribution of summergreen and evergreen forest types in this region have implications for carbon cycling, surface albedo, and ecosystem functioning, underscoring the need for reliable, spatially explicit information on boreal forest composition.

Despite their ecological importance, labelled datasets on boreal forests in Eastern Siberia remain poorly represented and insufficiently validated in existing global land cover and forest monitoring products. Global datasets differ substantially in spatial resolution and thematic detail, and their training data are often biased toward more accessible regions (Buchhorn et al. 2020a, Strahler et al. 2008), resulting in high uncertainty in forest type distributions across Siberia (Enguehard et al., 2024). In particular, functionally distinct forest types such as summergreen and evergreen needleleaf forests are frequently aggregated into a single forest class or misrepresented in forest transition zones (van Geffen et al., 2022; van Geffen et al., 2025; van Geffen et al. under review). At the same time, the scarcity of publicly available, machine-learning-ready reference datasets has limited the application of data-driven classification approaches in this region (Schepaschenko et al., 2017).

This thesis addresses these limitations by developing and evaluating data-driven approaches for boreal forest type mapping in Eastern Siberia, with a specific focus on the summergreen–evergreen forest transition zone. The overarching objective is to improve forest type mapping in a data-scarce region by integrating multi-scale reference data, operational satellite observations, and machine-learning methods. The work is structured around four research questions that address the creation of reference datasets, the performance of satellite-based classification, the comparison with global land cover products, and the identification of key limitations and uncertainties.

To address the first research question, this thesis introduces the SiDroForest (Siberian Drone-mapped Forest Inventory) dataset as an open-access, multi-scale reference data collection combining field-based forest inventories, UAV-derived forest structure products, synthetically generated tree-crown imagery, and labelled Sentinel-2 image patches (Kruse et al., 2020; Kruse et al. 2021; van Geffen et al., 2021a,b,c; van Geffen et al., 2022). The dataset provides one of the first machine-learning-ready reference frameworks for boreal forests in Eastern Siberia and enables explicit linkage between tree-level observations, plot-scale structure, and satellite-based representations. The analyses demonstrate that while UAV and field data provide high ecological fidelity, they also introduce spatial uncertainty related

to canopy overlap, plot representativeness, and the logistical constraints of high-latitude field campaigns (Brieger et al., 2019; Schepaschenko et al., 2017).

Building on this reference framework, the second research question examines the potential of operational multispectral satellite data for boreal forest type classification. A benchmark dataset for Sentinel-2–based forest type classification is developed and evaluated across different seasonal configurations and feature sets (van Geffen et al., 2025). The results show that Sentinel-2 imagery can reliably distinguish summergreen and evergreen needleleaf forests when phenological timing is optimized, with late-summer imagery providing the highest classification performance. Early- and peak-summer imagery yield lower separability, and multi-seasonal feature combinations do not improve performance in this region, contrasting with findings from temperate forest studies (Grabska et al., 2019; Immitzer et al., 2016; Li et al., 2003). Feature importance analyses indicate that shortwave infrared and near-infrared bands are particularly informative, reflecting differences in canopy moisture and structure between forest types (van Geffen et al., 2025).

The third research question assesses how regionally trained forest type classifications compare with existing global land cover products. A systematic comparison with the Copernicus Global Land Cover 100 m product (Buchhorn et al. 2020) reveals substantial discrepancies in forest type representation, including systematic underrepresentation of evergreen needleleaf forests and low agreement at the forest subclass level, especially in transition zones and near the treeline (van Geffen et al., 2025; van Geffen et al., under review). These findings demonstrate that higher spatial resolution alone does not guarantee improved thematic accuracy when training data and class definitions are not tailored to regional ecological conditions (Di Gregorio, 2005; Herold & Di Gregorio, 2016, White et al. 2021).

Finally, the synthesis of results addresses the fourth research question by identifying key limitations and sources of uncertainty in data-driven forest type mapping for Eastern Siberia. These include limited spatial representativeness of reference data, seasonal constraints on satellite data availability, mixed-pixel effects at medium spatial resolution, and restricted model transferability beyond well-sampled regions (Enguehard et al., 2024; van Geffen et al., 2025). At the same time, the results demonstrate that integrating multi-scale reference data with operational satellite observations provides a robust foundation for improving forest monitoring in data-scarce boreal regions.

Overall, this thesis contributes new reference datasets, benchmark classification approaches, and empirical insights into boreal forest type mapping in the Siberian summergreen–evergreen forest transition zone. By linking local field and UAV observations with regional satellite-based mapping and critically evaluating global land cover products, the work advances data-driven forest monitoring approaches for one of the most climatically sensitive forest regions on Earth.