SiDroForest: A comprehensive forest inventory of Siberian boreal forest investigations including drone-based point ~~cloud products~~clouds, individually labelled trees, synthetically generated tree crowns and Sentinel-2 labelled image patches

5 Femke van Geffen[1,2], Birgit Heim[1], Frederic Brieger[1,3], Rongwei Geng[1,4,5], Iuliia A. Shevtsova[1,2], Luise Schulte[1,2], Simone M. Stuenzi[1,6], Nadine Bernhardt[1,7], Elena I. Troeva[8], Luidmila A. Pestryakova[9], Evgenij S. Zakharov[8,9], Bringfried Pflug[10], Ulrike Herzschuh[1,2,11], Stefan Kruse[1]

[1]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI), Research Unit Potsdam, Germany

[2]University of Potsdam, Institute of Biochemistry and Biology, Potsdam, Germany

10 [3]Carleton University, Department of Geography and Environmental Studies Ottawa, Canada

[4]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[5]University of Chinese Academy of Sciences, Beijing, China

[6]Humboldt-Universität zu Berlin, Geography Department, Unter den Linden, Berlin, Germany

15 [7]Julius Kühn-Institut Bundesforschungsinstitut für Kulturpflanzen, Quedlinburg, Germany

[8]Institute for Biological Problems of the Cryolithozone, Russian Academy of Sciences, Siberian Branch, Yakutsk, Russia

[9]North-Eastern Federal University of Yakutsk, Institute of Natural Sciences (NEFU), Yakutsk, Russia

[10]German Aerospace Center (DLR), Berlin, Germany

20 [11]University of Potsdam, Institute of Environmental Science and Geography, Potsdam, Germany

Correspondence to: Stefan Kruse (stefan.kruse@awi.de), Femke van Geffen (femke.van.geffen@awi.de)

Abstract

30 The SiDroForest data collection is an attempt to remedy the scarcity of ~~tree level~~ forest structure data in the circumboreal region~~, whilst in addition,~~ by providing ~~as part of the data collection,~~ adjusted and labelled tree level and vegetation plot level data for machine learning and upscaling ~~practices. Publicly available comprehensive datasets on tree level forest structure are rare, due to the involvement of governmental agencies, public sectors, and private actors that all influence the availability of these datasets.~~

purposes. We present ~~here~~ datasets of vegetation composition and tree and plot level forest structure ~~of~~for two important

35 vegetation transition zones in Siberia, Russia~~:~~; the summergreen–evergreen transition zone in Central Yakutia and the tundra–taiga transition zone in Chukotka (NE Siberia). ~~The SiDroForest collection contains a variety of data mainly based on unmanned aerial vehicle (UAV) and field data collected from 64 vegetation plots during fieldwork jointly performed by the Alfred Wegener Institute for Polar and Marine Research (AWI) and the North-Eastern Federal University of Yakutsk (NEFU) during the Chukotka 2018 expedition to Siberia.~~

40 The SiDroForest data collection consists of four ~~separate~~ datasets. ~~The fieldwork locations are the anchors~~ that ~~bind the~~contain different complementary data types that together ~~based~~support in-depth analyses from different~~ on the location of the vegetation plot~~ perspectives of Siberian Forest plot data for multi-purpose applications.

i) The first dataset ~~(Kruse et al., 2021, https://doi.pangaea.de/10.1594/PANGAEA.933263)~~ provides Unmanned Arial Vehicle (UAV)-borne data products covering the vegetation plots surveyed during fieldwork (Kruse et al., 2021,

45 https://doi.pangaea.de/10.1594/PANGAEA.933263). The dataset includes structure from motion (SfM) point clouds~~, a variety of~~ and Red Green Blue (RGB) and Red Green Near Infrared (RGN) orthomosaics. From the orthomosaics, point-cloud products were created such as the Digital Elevation Model (DEM), Canopy Height Model (CHM), Digital Surface Model (DSM and the Digital Terrain Model (DTM~~ and Red Green Blue (RGB) and Red Green Near Infrared (RGN) orthomosaics. Forest structure and vegetation composition data are crucial in the assessment of whether a forest is to act as a carbon sink under

50 changing climate conditions. Fieldwork and UAV~~). The point cloud products ~~can~~ provide ~~such data in depth~~information on the three-dimensional (3D) structure of the forest at each plot.

ii) The second dataset contains spatial data in the form of ~~points~~point and polygon shape files of 872 labelled individual trees and shrubs that were recorded during fieldwork at the same vegetation plots ~~with information on tree height, crown diameter, and species~~ (van Geffen et al., 2021c, https://doi.pangaea.de/10.1594/PANGAEA.932821). The dataset contains information

55 on tree height, crown diameter, and species type. These tree- and shrub-individual labelled point and polygon shape files were generated ~~and are located~~ on top of the UAV RGB orthoimages. The individual ~~number links to the~~tree information collected during the expedition such as tree height, crown diameter and vitality are provided in table format. This dataset can be used to link individual information on trees to the location of the specific tree in the SfM point clouds, providing for example, opportunity to validate the extracted tree height from the first dataset. The dataset provides unique insights into the ~~vegetation~~

60 ~~composition and also allows future monitoring of the~~ current state of individual trees and ~~the contents of the recorded vegetation plots at large.~~shrubs and allows for monitoring the effects of climate change on these individuals in the future.

iii) The third dataset contains a synthesis of 10 000 generated images and masks that have the tree crowns of two species of larch (*Larix gmelinii* and *Larix cajanderi*) automatically extracted from the RGB UAV images in the common objects in context (COCO) format (van Geffen et al., 2021a, https://doi.pangaea.de/10.1594/PANGAEA.932795). ~~The~~As machine learning algorithms need a large dataset to train on, the synthetic dataset was ~~created~~ specifically created to be used for machine learning algorithms to detect Siberian larch species.

iv) ~~If publicly available forest-structure datasets at tree level are rarely available for Siberia, even fewer ready-to-use tree and plot-level data are available for machine learning approaches, for example optimised data formats containing annotated vegetation categories.~~ The fourth ~~set~~dataset contains Sentinel-2 Level-2 bottom of atmosphere processed labelled image patches with seasonal information and annotated vegetation categories covering the vegetation plots (van Geffen et al., 2021b, https://doi.pangaea.de/10.1594/PANGAEA.933268). The dataset is created with the aim of providing a small ready-to use validation and training data set to be used in various vegetation-related machine-learning tasks. It enhances the data collection as it allows classification of a larger area with the provided vegetation classes.

The SiDroForest data collection serves a variety of user communities. ~~First of all, the UAV-derived top of canopy structure information, orthomosaics and~~ The detailed vegetation ~~information in the labelled data set provide detailed information on forest type, structure and composition for scientific communities with ecological and biological applications. The detailed Land Cover and Vegetation~~cover and structure information in the first two data sets are of use for ecological applications, on one hand for summergreen and evergreen needle-leaf forests and also for tundra–taiga ecotones. The first two data sets further support the generation and validation of land cover remote sensing products in radar and optical remote sensing. In addition to providing information on forest structure and vegetation composition of the vegetation plots, ~~parts of the SiDroForest dataset~~the third and fourth datasets are prepared ~~to be used~~ as training and validation data for machine learning purposes. For example, the synthetic tree crown dataset is generated from the raw UAV images and optimized to be used in neural networks. Furthermore, the fourth SiDroForest data set contains ~~standardized~~ Sentinel-2 labelled image patches processed to a high standard that provide training data on vegetation class categories for machine learning classification with JavaScript Object Notation (JSON) labels provided. The ~~SiDroForst~~SiDroForest ~~data collective serves as a basis to add future data collected during expeditions, creating a larger dataset in the upcoming years that can provide~~ data collection adds unique insights into remote hard to reach ~~boreal~~circumboreal forest regions ~~of Siberia~~.

1 Introduction

Circumpolar boreal forests represent close to 30% of all forested areas and are changing in response to climate, with potentially important feedback mechanisms to regional and global climate through altered carbon cycles and albedo dynamics (e.g., Loranty et al., 2018). These forests are located primarily in Alaska, Canada, and Russia. ~~Forests are three-dimensional systems~~

~~whose biophysical structure plays major roles in ecosystem function and diversity.~~ Forest structure is a crucial component in the assessment of whether a forest is likely to act as a carbon sink or source under changing climate (e.g., Schepaschenko et al., 2021). ~~On one hand, publicly~~Publicly available comprehensive datasets on forest structure are rare, due to the involvement of governmental agencies, public sectors, and private actors who all influence the availability of these datasets. ~~On the other hand, there are national and international collaborative initiatives that aim to distribute high quality forest data collections.~~ That said, the Arctic-Boreal Vulnerability Experiment (ABoVE) run by the NASA Terrestrial Ecology Program provides open-source data collections from boreal and arctic regions in Alaska and Canada (ABoVE Science Definition Team, 2014). ~~Also~~Globally, the Forest Observation System (FOS, http://forest-observation-system.net/) ~~that is supporting~~provides publicly available forest data for Earth Observation (validation and algorithm development ~~efforts~~) such as described in Chave et al. (2019) ~~provides a publicly available~~ and a global Above Ground Biomass (AGB) database (Schepaschenko et al~~.~~., 2019) containing a high number of plot level ~~data~~datasets from the boreal forest domain. Schepaschenko et al. (2017) used inventories from the old Soviet Forest Inventory and Planning System (FIP) and the new Russian National Forest Inventory (NFI) to compile and publish a highly comprehensive forest AGB data collection at plot level specifically for Eurasia. This data collection (Schepaschenko et al., 2017) and FOS (Schepaschenko et al. 2019) both distribute aggregated plot level~~. The data collection presented here provides open-source forest structure-related data at plot and also at tree level for boreal forests in central and North Eastern Siberia, Russia~~ information.

However, there is still a lack of usable data for satellite and UAV imagery classification tasks for the boreal zone as a whole. Also, there is a lack of usable training data for automatic needle-leaf tree crown detection. The central and eastern Siberian boreal zones with their forest types are especially underserved as there are no open-source UAV forest data available. ~~Still~~Also, for the circum-boreal, still, few data are publicly available at tree level or plot level that are ready-to use for machine learning applications in the field of remote sensing, for example optimised data containing annotated vegetation categories. ~~We therefore aim to provide tree level as well as plot level boreal forest datasets, with parts of the data collection that can be further used for remote sensing applications and machine learning analyses .~~

~~The data presented here includes a variety of data types contained in four datasets. The data collection is underpinned by fieldwork data from vegetation plots and unmanned aerial vehicle (UAV) acquisitions from a German-Russian expedition in 2018 by the Alfred Wegener Institute for Polar and Marine Research (AWI) together with the North-Eastern Federal University of Yakutsk (NEFU) to central and eastern Siberia. Likethis, the fieldwork locations are the anchors that bind the data types together based on the location of the vegetation plots.~~

The SiDroForest data collection provides ~~these~~open-source forest structure-related data at tree, plot and upscaling levels for boreal forests in central and North Eastern Siberia, Russia. At individual tree level, the data consists of conventional forest inventory data such as tree height, tree crown ~~diameters, tree~~ diameter and species ~~and tree stand density. The data acquired in the field span from forest inventories at the species level, tree height information and density for each vegetation plot derived from in situ assessments, to UAV camera data of the field plots that were processed to structure from motion (SfM) point clouds that give structural forest information, UAV orthoimages of the plots, and automatically extracted tree crowns in the~~

~~form of shape file products.~~labels. The individual tree-level data labelling per plot provides ~~already~~ opportunities for further machine learning applications~~.~~ in the form of validation data. At plot level, the data collection contains UAV structure from motion (SfM) point clouds, georeferenced orthoimages and products derived from point clouds providing structural forest information. On top of these state-of-the art forest inventory data and SiDroForest UAV products that are ~~already~~ enriched by labelling, we prepared two data sets that can be directly used for machine learning in remote sensing applications. One data set is a synthetically generated image data set on tree crowns in the common objects in context (COCO) format (Lin et al., 2013) that we constructed from selected UAV Red Green Blue (RGB) imagery from plot data. The other data set fit for machine learning contains labelled Sentinel-2 (S-2) image patches covering the vegetation plots related to the vegetation composition. These labelled S-2 image patches can, e.g., be used for machine learning training for a boreal forest land cover classification using S-2 satellite images.

~~Photogrammetric UAV-borne products (i.e., SfM point clouds, digital elevation products, RGB orthomosaics) that extend the ground-based inventories, have already a long application history in forestry and well-defined methodological standards (e.g., Jensen et al., 2016; Panagiotidis et al., 2017). Currently, the use of UAVs in environmental applications is undergoing a growing use in forestry and environmental science in general because of the landscape-level potential, the flexibility of the data generation and low costs (Fraser et al., 2016). Methods for the product generation of forest structure data from UAV imagery are advanced. For example, automatic tree crown detection in UAV Red-Green-Blue (RGB) imagery has become increasingly popular in the world of computer vision and machine learning. For the United States, the National Ecological Observatory Network (NEON; Weinstein2021) dataset containing 100 million tree crowns, derived from UAV LiDAR data that facilitated such an analysis, was created using machine learning tools such as DeepForest (Weinstein et al., 2019). Automatic tree crown detection can be performed with UAV data using a variety of deep learning algorithms. For example, the Mask R-CNN was used by Hao et al. (2021) to detect tree crown and canopy height of Chinese fir in a plantation in China. The UAV data was used to obtain high-resolution images of the study site. Tree crown width and tree height of Chinese fir was manually extracted from this UAV imagery using a combination of labelled ground-truth data and canopy height model (CHM) information and served as validation data. However, considerable effort is needed for the validation and training data generation and there is still a lack of usable data for image classification and instance segmentation tasks as a whole. Also, effort is needed for the extraction of tree crowns and there is a lack of usable RGB training data for automatic tree crown detection. The central and eastern Siberian boreal zones with their forest types are especially underserved as there are no open-source UAV forest data available. SiDroForest addresses this gap by providing on one hand extracted tree crowns of the forest plots of shape files, on the other hand we created from the UAV RGB photos a training data set that can be used for machine learning for this boreal forest type of Central and Eastern Siberia. For this data set, we manually extracted tree crowns to compose 10 000 synthetic RGB images that can be used to train machine-learning algorithms to identify larch tree crowns in UAV RGB images in the common objects in context (COCO) format (Lin et al., 2013).~~

In addition, as an enhanced data set ready-to use for remote sensing applications, we provide high quality labelled S-2 image patches with vegetation categories that contains seasonal information and can be used for upscaling purposes and machine-

160 learning classifications.

It is increasingly common in data science and environmental science to use multiple data types within one analysis. For example, S-2 images and metadata, topography data, CHM, as well as their combinations, were used to predict growing stock volume using deep neural networks in four forestry districts in central Finland (Astola et al., 2021). Another example of the use of multiple data types in non-machine learning remote sensing is the work by Wang et al. (2020) where above-ground

165 biomass estimation was performed using field plots, UAV-LiDAR strip data, and S-2 imagery. In this model, the partial-coverage UAV-LiDAR data were used to link ground measurements to S-2 data. These recent studies show the need for well-labelled publicly available data to link the data types together and for performance testing of remote sensing algorithms. In these studies, the testing data preparation was undertaken within the project: For example, Thanh et al (2018) compared the performance of three common machine learning algorithms; a support vector machine (SVM), a random forest (RF) and k

170 nearest neighbours (K-NN) on S-2 data from Vietnam. In order to validate the performance of these algorithms, the training data (training and testing samples) were collected based on the manual interpretation of the original S-2 data and high-resolution imagery obtained from Google Earth and 135 labelled land cover polygons were produced This study is a good example of manually labelled data creation for a specific task and specific research area to be able to use supervised classification tools. The work done by Abdi (2020) shows a similar study that assesses the performance of four machine

175 learning algorithms for land cover classification of boreal forests. Here too, the validation and training data is manually created to assess the performance of the algorithms.

The SiDroForest labelled S-2 image patches collection provides annotated image patches for around 60 sites. As each S-2 image patch consists of nine units (pixels) of 100 m^2 extent each, it amounts to around 550 annotated validation and training units. In its current stage, the SiDroForest S-2 data collection is not published with performance testing, and is by us not

180 considered as a benchmark data set for Remote Sensing image interpretation, (e.g. as defined in Long et al., 2020). The SiDroForest labelled S-2 image patches collection is available as a small training and validation data set providing so far underrepresented vegetation categories, that will save future users time when attempting to classify vegetation of Central Siberian and Eastern Siberian boreal forests. cover.

By making SiDroForest public, we aim to remedy public data scarcity on UAV data of boreal forest plots, on tree level forest

185 data, and specifically for annotated data for the boreal forests in Central and North-Eastern Siberia and encourage the use of the data presented here for further analyses and machine-learning tasks.

1.1. Study Region

The data collection we provide specifically contains tree level,plotlevel, plot level, and plotupscaling level forest-structure data from important boreal transition zones located in central and eastern Siberia that are specifically vulnerable to climate change,:

190 these are the tundra–taiga (in Chukotka) and the summergreen–evergreen (in Central Yakutia) transition zones (Fig. 1).

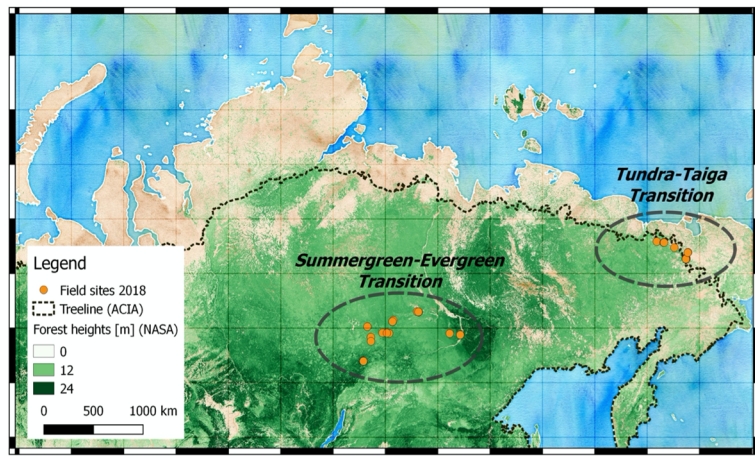



Figure 1: Overview of the Siberian transition zones ~~–:~~: the tundra–taiga transition in Chukotka and the summergreen–evergreen transition ~~–~~in Central Yakutia that were covered by the 2018 ~~Russian-German~~ Chukotka expedition (~~NEFU, RU and AWI, DE).~~orange points represent 2018 field sites with vegetation plots). The overview map (background OpenStreetMap ©) shows forest coverage by green colour-coded NASA forest height~~,~~ (Simard et al. (2011) and the Northern treeline (CAVM Team 2003, for Arctic Climate Impact Assessment (ACIA)).

The tundra–taiga transition zone occurs where boreal forests reach their maximum northwards position and form a treeline ecotone (MacDonald et al., 2007). Here, the transition from open forest stands ~~in the south and~~with decreasing stand densities towards treeless tundra in the north takes place. A warming climate drives the transition from tundra in the tundra–taiga transition zone to open taiga forests (Rees et al., 2020). During the snow-covered season, the taiga has a lower albedo than tundra due to the trees that emerge above the snow. A change from tundra to taiga albedo can result in a positive feedback loop of vegetation change which, in combination with the warming climate, may lead to dramatic environmental changes in the Arctic (Bonan, 2008). Remote-sensing data have been previously used to assess vegetation dynamics ~~and their changes~~ in Chukotka. Through vegetation monitoring using Landsat satellite data, Shevtsova et al. (2020) report that shrubification has expanded by 20% in area in the tundra–taiga zone and by 40% in the northern taiga as well as tree infilling occurring in the northern taiga. Extensive satellite remote-sensing work was done by Montesano et al. to assess the vegetation dynamics in Siberia using LiDAR and synthetic aperture radar data (2014) and Landsat satellite data (2016). To be able to expand on these satellite-derived remote-sensing findings, in-depth monitoring at a vegetation plot level in this region is important. Clear overviews of species distribution over the varying types of land cover are useful to study the impacts of climate change on the eastern Siberian treeline that is not yet well enough studied, in part due to sparse data being available for the region (Shevtsova et al., 2021). Our open-access data collection will ~~greatly~~considerably improve insights into the tundra–taiga transition zone.

~~Yakutia hosts the~~ The second relevant forest transition zone included in the SiDroForest data collection~~:~~ is the summergreen–evergreen transition zone~~.~~ in Central Yakutia. Summergreen needle-leaf tree species ~~on the plots~~ covered in the SiDroForest data collection consist of two species of larch trees: *Larix gmelinii* and *Larix cajanderi*. The evergreen species present are pine

and spruce: *Pinus sibirica* (Only at Lake Khamra plots)/, /*Pinus sylvestris* and *Picea obovata*. In forests, the light-demanding summergreen *Larix* trees are outcompeted by evergreen tree taxa (Troeva et al., 2010). Yet it is an open question as to how *Larix* forests, once established, hinder their replacement by evergreen forests and thus maintain a vegetation–climate equilibrium (Mamet et al., 2019). This self-stabilisation that takes place in the *Larix*-dominated forests in central and eastern Siberia most likely results from a combination of unique climate drivers for the region, such as vegetation, climate, fire, and permafrost (Herzschuh et al., 2020). Datasets such as the one presented here are a snapshot of the current state that can be used to monitor individual trees over time to gain a clearer idea aboutinsight into the vegetation dynamics of the region.

2 Data collectionSiDroForest data and methods

The SiDroForest data collection contains a variety of data types that were selected to create a comprehensive insight into the boreal forests in Central and North-Eastern Siberia. Each data type has been enhanced to best use the data for vegetation-related analyses and several of them for machine-learning tasks. The combined data types aim to provide a clear picture of the current state of the vegetation cover in central Yakutia and Chukotka.the most comprehensive insights into the boreal forest in Siberia.

The SiDroForest data collection is divided into four datasets (Fig. 2):

1. UAV based SfM point clouds, point-cloud products, and orthomosaics from UAV image data (orange hexagon symbols) *of expedition vegetation plots* (orange hexagon symbols*in Chukotka and Central Yakutia in summer 2018* (mint green rectangle).

2. Individual labelled trees surveyed during the fieldwork expeditions,, including information on height, tree crown, and species (green shield symbols).. These tree-individual labelled point and polygon shape files (light green symbols) were generated and are linked to the UAV RGB orthoimages *of the expedition vegetation plots*.

3. Synthetically created Siberian larch tree crown dataset of 10,000 instances in Microsoft's common objects in context (COCO) format (purple triangle symbols). The images and masks contain the tree crowns of two species of larch (*Larix gmelinii* and *Larix cajanderi*) manually extracted from selected RGB UAV images.

4. Sentinel-2 Level-2 bottom of atmosphere labelled image patches with seasonal information (red shape symbol) covering *the expedition vegetation plots* (red shape symbol)..

Each data type has been enhanced to best use the data for vegetation-related analyses. Dataset three and four have additionally been optimized and annotated for machine-learning tasks. Machine learning tasks often require validation data and also the annotated datasets one and two contain data for such an application. The combined data types aim to provide a multi-purpose application data set on the current state of the vegetation cover in Central Yakutia and Chukotka.

The SiDroForest products are in common software formats: there are point and polygonal shape files (shp), raster files are in the georeferenced tagged image format (tif), geotiff, and shapefile formats and JavaScript Object Notation (JSON) can be read and visualized in any open source and commercial GIS and Remote Sensing software tools and a wide range of libraries in R, python and other programming languages. The point clouds are provided in the standard LASer (LAS) binary file format that

250 can be handled in any software that supports this format such as CloudCompare (CloudCompare, 2021) or R (R, 2020) or Python libraries specifically developed for this datatype.

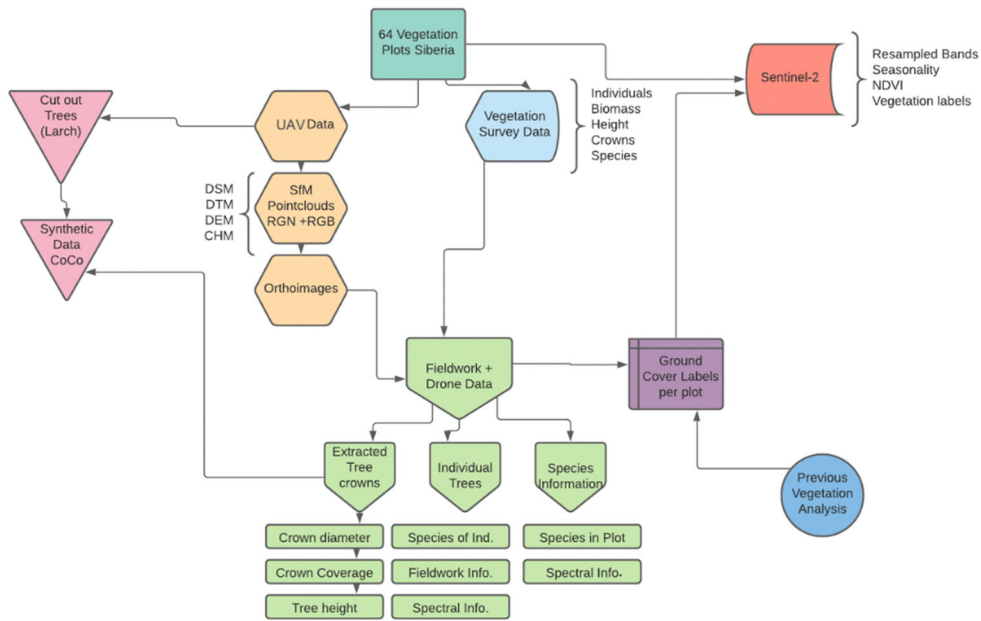



Figure 2: Overview of the four datasets all related to the 2018 expedition plots (UAV derived products, individual labelled shape files, synthetically created Siberian larch tree crown dataset, Sentinel-2 labelled image patches) **and their content and interconnections in SiDroForest**. (see text for details on the labels).

~~C~~2.1 SiDroForest field data

~~The fieldwork data include plant taxa and forest-structure inventories from a total of 64 vegetation plots, with 39 vegetation plots in the tundra–taiga transition zone in central Chukotka and 25 vegetation plots in the summergreen–evergreen transition zone in central Yakutia. All data types included in this dataset are linked to each other using a two-letter code signifying the subregion (Table 1) and the vegetation plot numbers.~~

~~The AWI and NEFU expeditions in summer 2018 covered both the tundra–taiga transition zone in eastern Siberia and the summergreen–evergreen transition zone in central Yakutia (Fig. 1). They~~Extensive expeditions from the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) from Germany in cooperation with the North-Eastern Federal University of Yakutsk (NEFU), Yakutia, in summer 2018 covered a bioclimatic gradient ranging from treeless tundra via extremely open larch forest with mean tree heights around 5 m close to Lake Ilirney in central Chukotka (tundra–taiga ~~ecotone~~transition zone) in north-eastern Siberia to dense mixed tree species stands near Lake Khamra in south-western Central Yakutia (~~boreal forest biome~~summergreen–evergreen transition zone) (Fig. 1). The larger regions were subdivided into 12 subregions that were named based on the nearest city or lake to the plots~~. In~~: in Chukotka, we defined three subregions encompassing 39 plots (Fig. 3a) and nine subregions encompassing 25 plots for Central Yakutia (Fig. 3b). The vegetation plots

have different ~~larch~~ tree cover: from treeless tundra to open larch forests on slopes and in lowlands, with tree density depending on slope and slope aspect. All data types included in this dataset are linked to each other using a two-letter code signifying the subregion (Table 1) and the vegetation plot numbers.

275 **Table 1: Overview of vegetation plots per transition zone, region, and subregion along with the subregion codes.**

Transition zone	Geographical region	Subregions	Subregion codes	Plot ~~name~~names
Taiga to tundra transition zone	Central Chukotka	Bilibino Lake Ilirney Lake Rauchuagytgyn	BI LI LR	EN18000; 18028-35 (n = ~~7~~9) EN18001-18027 (n = ~~25~~27) EN1805~~1~~0-18055 (n = 5)
Summergreen to evergreen transition zone	Central Yakutia	Yakutsk Magaras Vilnuyi Nyurba Suntar West Suntar Mirny Mirny-Lensk Lake Khamra	YA MA VI NY SW SU MI ML LK	EN18061 (n = 1) EN18062 (n = 1) EN18063-66 (n = 4) EN18067-70 (n = 4) EN18071 (n = 1) EN18072-74 (n = 3) EN18075-76 (n = 2) EN18077-78 (n = 2) EN18079-83 (n = 5)

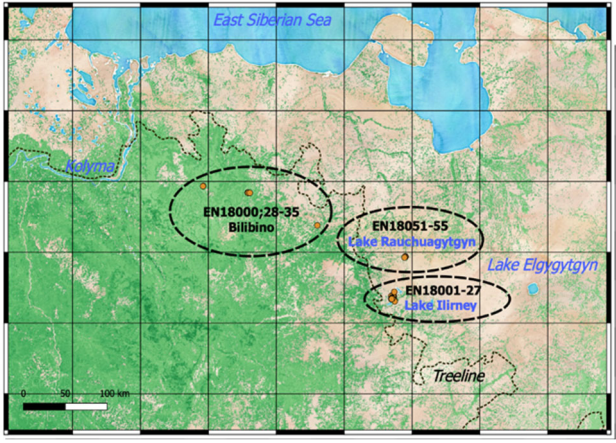



Figure 3a. Subregions and plots (red points) for Chukotka~~.~~: Bilibino (BI) (EN18000, 18028–35), Lake Ilirney (LI) (EN18001–27), and Rauchuagytgyn (RA) (EN18051–55). See also Table 1.
The overview map (background OpenStreetMap ©) shows forest coverage by green colour-coded NASA forest height (Simard et al. (2011) and the Northern treeline (CAVM Team, for Arctic Climate Impact Assessment (ACIA)).

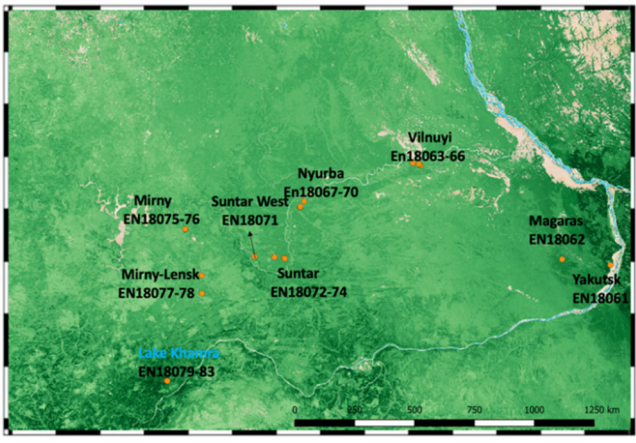



Figure 3b. Subregions and plots (red points) for Central Yakutia~~.~~: Yakutsk (YA) (EN18061), Magaras (MA) (EN18062), Vilnuyi (VI) (EN18063-66), Nyurba (NY) (EN18067–70), Suntar West (SW) (EN18071), Suntar (SU) (EN18072–74), Mirny (MI) (EN18075–76), Mirny-Lensk (ML) (EN18077–78) and Lake Khamra (LK) (EN18079–83). See also Table 1.
The overview map (background OpenStreetMap ©) shows forest coverage by green colour-coded NASA forest height (Simard et al. (2011).

A detailed vegetation inventory was conducted for each of the plots visited during fieldwork. Fifteen-metre radius circular plots for the ~~projective~~projected cover of trees and tall shrubs (~~see Appendix~~ Fig. A1) were set within 30 m x 30 m rectangular vegetation plots for ground projective cover of vegetation taxa. The plots and the field data collection are described in further detail in Shevtsova et al. (2019, 2020a,b,c, 2021).

~~Two~~In the field, two 30-m-long tape measures were laid out along the main cardinal directions, intersecting in the plot centre, marking the main axes of a circular area with a radius of 15 m ~~(Appendix Fig. A4).~~ A minimum of ten individuals of each tree and shrub species present were selected per plot. For each individual tree we measured the stem diameter at breast height and at the base. The tree crown diameter, tree height, and vitality were estimated ~~(~~as described in Brieger et al~~.~~ (2019). ~~For Plot~~ There were three deviations from the standard method of vegetation inventory. On plot EN1814 and EN1865, all trees were recorded~~. Plot~~, and plot EN18070 was recorded ~~differently to the other plots. Here, the area covered in the vegetation plot was not 15x 15 m but~~by a transect with three segments: edge, transition, and centre.

We post-fieldwork assigned 11 vegetation classes to the 64 plots (table A1). The class ~~to each plot that~~ assignment ~~–~~was ~~determined~~ based on the previous ~~work~~ classes determined by Shevtsova et al. (2020a) for Chukotka ~~and a~~. For plots in Central Yakutia~~.~~, we applied a similar method incorporating principal component analysis (PCA~~) and~~), tree density information from the UAV data ~~for the Yakutia plots. In addition, the~~, and recorded tree species information per plot ~~was taken into account when assigning the classes (Appendix Fig. A2 for Chukotka and Appendix Fig. A3 for Yakutia). A framework of 11 vegetation classes that can be applied to all the 64 plots and are further described in the results section was derived.~~ (Fig. A2, A3 show the field data information).

In addition to the ~~field vegetation and the specific~~fieldwork forest inventories that were obtained, 60 of the 64 vegetation plots were overflown with a consumer grade DJI Phantom-4 quadcopter carrying ~~a~~ MAPIR Survey-3W Red~~–~~ Green Blue RGB and Red Green Near-infrared (~~RGNIR) camera~~RGN) cameras to obtain spatially mapped detailed forest structure information in 2 and 3 dimensions (2D, 3D). The UAV imagery covered a minimum areal extent of 50 m x 50 m over the 15 m radius ~~and 30 m x 30 m vegetation~~ plots with a standardised flight plan following a double-grid in near-nadir position and a circular flight facing the plot centre at take-off elevation (~~Appendix~~ Fig. A4). Further details are described in Brieger et al. (2019).

2.2 SiDroForest dataset 1: Structure from motion (SfM) point ~~clouds, point~~-cloud products and orthomosaics

2.2.1 SfM point ~~clouds~~cloud -products of the plots

Due to the availability of multiple overlapping images from different camera viewpoints, point-cloud processing and the generation of 3D products and successive generation of orthoimages were possible. We manually rejected images that had been taken during take-off and landing, as well as under- or over-exposed images, from further processing (see also Brieger et al., 2019). The remaining images were used to generate the 3D SfM point clouds and related products directly from the point-cloud data.

The SfM point clouds were constructed with Agisoft PhotoScan Professional© (Agisoft, 2018) according to methods described in Brieger et al. (2019). Tracked Global Positioning System (GPS) information was automatically integrated into the images during this process. The parameters were tuned with the highest resolution settings to capture as much detail of the complex tree structures as possible. The depth filtering in the dense cloud generation was changed from the default to a mild filtering to preserve more detail especially in tree crowns (details in Brieger et al., 2019). The RGB point clouds have been further segmented into two separate point clouds with a cloth simulation filter (CSF) (Zhang et al., 2016) as described in Brieger et al. (2019) to produce two RGB point clouds. One of the point clouds contains the points of the ground and low vegetation (named here 'groundonly') and the other contains the points of the higher vegetation (named here 'treesonly') (Fig. 4).

~~The   dataset   contains~~

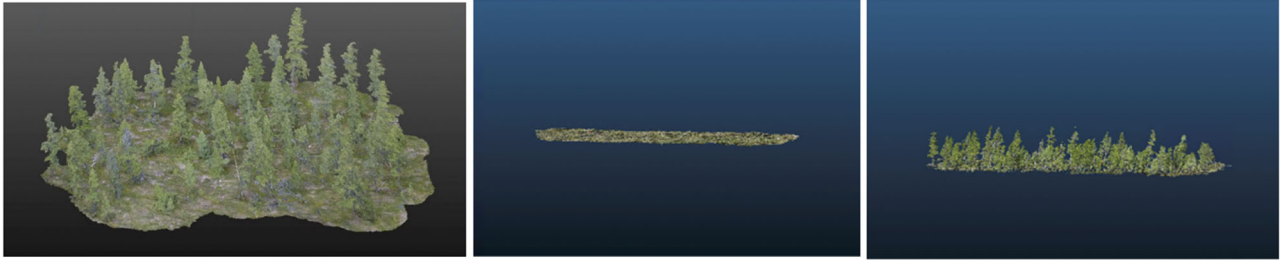



Figure 4: Left: example of a full Red Green Blue (RGB ~~and RGNIR point clouds. The RGB point clouds have been~~) structure from motion (SfM) point cloud for plot EN18074. Centre: segmented ~~into two separate point clouds with a cloth simulation filter (CSF) as described in Brieger et al. (2019). The result is two different~~ RGB point cloud containing only the points of the ground layer named *groundonly*. Right: segmented RGB point cloud ~~types: one containing~~with the above-ground ~~points and very low vegetation (~~named ~~here 'groundonly') and one containing the points of the higher vegetation (~~named here 'treesonly'~~) (Fig. 4)~~. treesonly.

We chose to ~~present~~segment the RGB point clouds ~~segmented~~ into ~~these two point clouds,~~ 'groundonly' and 'treesonly', because it reduces the size of the individual point clouds and ~~it is~~at the same time it remains easy for users to merge them together. It can also be interesting to have the two segmented when attempting to analyse the below-canopy vegetation or ground-cover classes. Plots with dense vegetation such as EN18077 and EN18063 could not be segmented into ~~groundonly~~'groundonly' and ~~treesonly~~'treesonly' due to the ground not being visible in the images. ~~The point clouds are provided in the standard LASer (LAS) binary file format. The~~The final SiDroForest dataset includes three point-cloud types per plot: ~~treesonly~~'treesonly' and ~~groundonly~~'groundonly' in RGB and the full point cloud in ~~RGNIR.~~

RGN. The ~~following section describes the data products that were~~ created ~~from the RGB~~ point ~~clouds~~cloud products include: a digital terrain model (DTM), a digital surface model (DSM), a canopy height model (CHM), and a digital elevation model (DEM). ~~All point cloud products were generated from the point clouds generated in Agisoft PhotoScan Professional©.~~ The point-cloud products were ~~then~~ produced in R (R Core Team, 2020) and exported as georeferenced geotiff raster files at 3 cm x 3 cm pixel resolution in the respective Universal Transverse Mercator (UTM) projection~~. During the creation~~ of the field site location. The DEM~~, the choice was made to crop it~~ products were cropped to a defined area in the form of a polygon (called here the *outer polygon*) due to the better quality of the points within this region. The *outer polygon* is the area covering the

camera positions plus a buffer of 5 m. In addition to the clipped product versions and the shapefiles of the *outer polygon*, the fully covered area that was not clipped to the *outer polygon* is also supplied for the orthomosaics and the point clouds to give the user a dynamic dataset to work with.

DTM: The definition of a DTM is that the surface represents the ground level with all natural and built features above the ground removed. ~~To create the~~ The DTM ~~for each plot,~~ is created from the ~~full point clouds were separated into two subsets: the higher vegetation (*treesonly*) and the ground~~ RGB ground cover and lower vegetation (*groundonly*) ~~using a Cloth Simulation Filter (Zhang et al., 2016) and the ground points rasterised and a minimum moving window filter applied to generate the DTM, as described in detail in Brieger et al. (2019). The~~ point cloud~~.~~, therefore, the SiDroForest DTM ~~in 3-cm grid resolution~~ represents the top of the canopy of the lowest vegetation canopy layer in case of low-structure vegetation.

DSM: The definition of a DSM is that the surface represents the highest-level elevation including natural and built features. The DSM is produced from the full point cloud, and interpolated between the highest points in each grid cell representing in case of a forest plot the top of the highest ~~vegetation~~tree canopy layer.

~~DSMs were created in R for all available plots with a 3-cm-resolution grid used to rasterise the full point cloud, where the elevation value of the highest point in each cell is assigned as the new raster value (Brieger et al., 2019).~~ CHM: The ~~3 cm resolution~~definition of a CHM is that it represents the difference between the DSM and the DTM (CHM = DSM – DTM), and thus normalises the DSM to the ground. ~~Each pixel represents the vegetation height above the ground (Brieger et al., 2019). The CHM represents the original point-cloud data.~~ Because the CHM is derived from a subtraction of the DSM and the DTM, it may contain no data values where the tree crown covers a large amount of ground and the ground data are missing due to this coverage. The SiDroForest CHM represents the vegetation height above the ground.

DEM: The DEM is a quantitative representation of the elevation of Earth's surface. The SiDroForest DEM provides the terrain relief referenced to the vertical datum of the World Geodetic System 1984 (WGS84) without the lowest canopy layer in contrast to the SiDroForest DTM that contains the lowest ground vegetation layer. ~~The 3-cm resolution DEM was cropped to the *outer polygon*.~~

2.2.~~3~~2 Orthomosaics of the plots

~~From each full RGNIR and RGB point cloud and overhead photo images orthomosaics were created.~~ UAV-derived orthomosaics are geometrically corrected images that are by standard georeferenced ~~including the distortions~~ by topography (the relief) and vegetation (the top-of-canopy elevation). The orthomosaics were constructed from the multiple RGB and ~~RGNIR~~RGN overhead photo images that were corrected for perspective and scale with Agisoft PhotoScan Professional© (Agisoft, 2018) using structure from motion/multi-view stereo (SfM-MVS) techniques ~~(~~as described in detail in Brieger et al~~.,~~. (2019). The ~~RGNIR~~RGN orthomosaics have been co-registered to the RGB point clouds using the ground control points (GCPs) distributed in the field to make the ~~RGNIR~~RGN and RGB point clouds align. The orthomosaics were exported ~~from Agisoft PhotoScan Professional (Agisoft, 2018)~~ as georeferenced geotiff ~~raster~~ files at 3 cm x 3 cm pixel resolution in the respective Universal Transverse Mercator (UTM) zone projections.

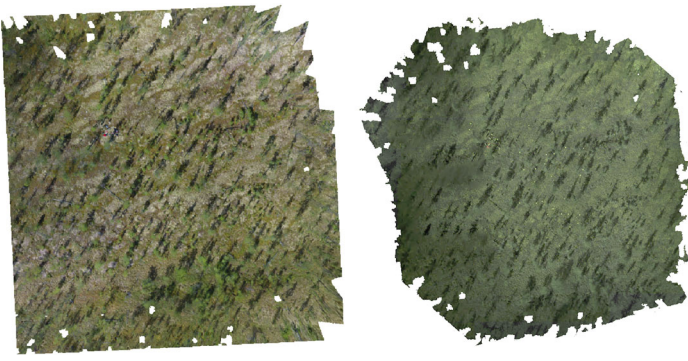



Figure 5: Example of Red Green Blue (RGB displayed with Red on Red, Green on Green and Blue on Blue as true colour) (left) and Red Green Near-infrared (RGN, displayed with Green on Red, NIR on Green, and Red on Blue) (right) orthomosaics for plot EN18000.

375

Not all RGB orthomosaics have the same high quality, as varying flight ~~plans~~ or weather conditions affected the construction of the final products. This resulted in 'blurry' parts in some of the orthomosaics. ~~The affected plots are:~~ (EN18030, EN1878, and EN1879~~. The~~). These blurry regions affect less than 20% of the image ~~and the rest is unaffected~~, therefore the orthomosaics of these plots are included in the data publication. Figure 5 shows an example of the high-quality type of an RGB and RGN

380 orthoimage product.

2.2.~~4~~3 Automated extracted tree crown polygons

~~The polygons of the tree crowns were created within an automated extraction workflow.~~ The SiDroForest data collection also contains 19 342 automatically detected tree-crown polygons (Kruse et al. 2021b). The tree crowns were captured in the CHM by watershed segmentation analysis using the R package ForestTools (Plowright, 2019) and successive automatic generation

385 of a polygon around them following Brieger et al. (2019). This automated tree-crown detection algorithm was run for all plots and the resulting shapefiles are provided for each plot that contained trees. Quality assurance was performed for each plot by carefully examining each plot based on expert knowledge and assigning a quality score of Q1 (good quality), Q2 (medium quality), or Q3 (poor quality) to the shapefile products.

2.3 SiDroForest dataset 2: Individual labelled trees

390 The individuals from within the 15-m-radius vegetation survey plots~~In order to make assumptions and predictions about the content of the vegetation plots it is important to link the labelled individual trees from the fieldwork to the processed orthoimages.~~ ~~This also allows future monitoring of the individual trees and the contents of the recorded vegetation plots at large. We produced an individual labelled trees dataset containing 872 trees and shrubs that were surveyed in Siberia during the two-month fieldwork expedition in 2018 (Kruse et al., 2019). The individual species were from within the 15-m-radius~~

395 ~~vegetation survey plots and measurements such as height and crown diameter were collected.~~

~~The individuals~~ that could be located in the orthoimages were marked in a *point* and *polygon* shapefile that outlines the tree crown of the individual tree, containing the individual number of the tree, the species, and its form (tree ~~-~~or shrub). The form attribute was added because in the Chukotka plots there are *Pinus* species that are ~~shrubs which can be misleading~~not the Pinus tree but the Pinus shrub form. The tree ID, exemplified in Figure 6, is the first letter of the genus of tree and the total number of individuals recorded (e.g., L259 is the 259th *Larix* specimen). The total number of *Larix* recorded is a cumulative number over all the plots recorded. The individual number was recorded during fieldwork and corresponds to information stored in the extensive database of Kruse et al. (2020) containing measurements concerning the individual tree, which are now also accessed via the SiDroForest dataset~~.~~ in form of attributed shape files.

The point shapefiles also include the geographical x and y coordinates of the point in decimal degrees. The individual number can be used to link the tree or shrub to the rest of the information collected during the expedition such as tree height, crown diameter, and vitality. This information is provided in form of a csv file in Kruse et al. (2021a).

In addition to the two shapefiles that are linked to the individually recorded trees, another shapefile is provided per plot with species-level information (Fig. 6). It contains a minimum of ten labelled polygon shapefiles that cover trees or large shrubs (>1.3 m height). These labelled polygons only cover the inside of the tree or shrub to minimize noise from the ground layer for classification purposes. For the species polygon, trees and shrubs that were seen in the rest of the orthoimages were also included, not only the individuals ~~included in~~from the fieldwork records.

●   EN18026_Individual_Point_Labelled
●   EN18026_positions_DINA4-cards

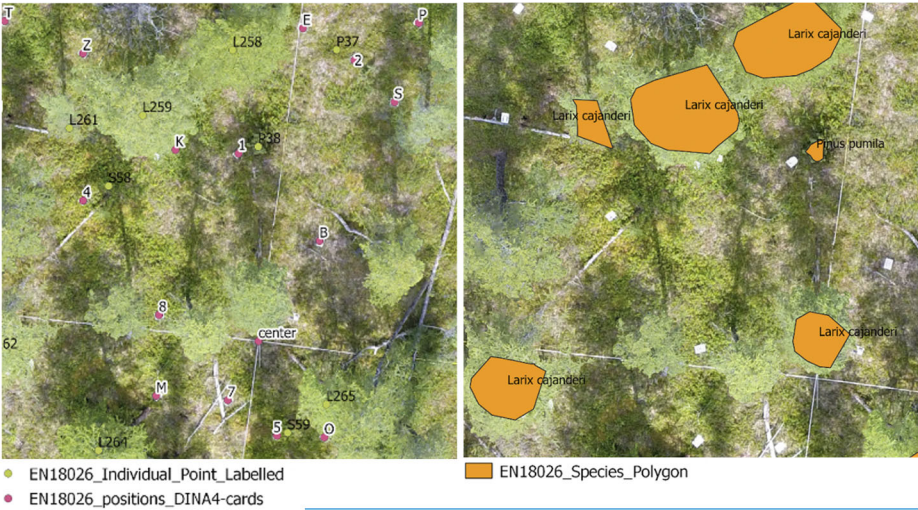
    ▮   EN18026_Species_Polygon

Legend
- ● Individual tree point
- ▢ Species polygon

Figure 6: ~~Left:~~ examples of the individual point labels and ~~right:~~ examples of species polygons. ~~The white bordered letters and numbers were used to mark~~Where possible the species polygon overlaps the individuals labelled in the field~~.~~, e.g., the larch at L34, L35 and L36. Additionally, the *Pinus pumila* were not recorded in the field but is added in the species shapefile. Both shapefiles are ~~visualised~~visualized on the Red ~~green blue~~ Green Blue (RGB) orthoimage of plot ~~EN18026~~EN18004.

2.4 SiDroForest dataset 3: Synthetic larch tree crowns

~~This synthetic Siberian larch tree crown dataset was created to enhance the data collective for upscaling and machine learning purposes.~~ The synthetic dataset contains larch (*Larix gmelinii* (Rupr.) ~~Rupr.~~ and *Larix cajanderi* (Mayr). tree crowns extracted from the onboard camera RGB images of five selected vegetation plots from fieldwork, placed on top of fully-resized images from the same UAV flights.

To create the dataset, backgrounds and foregrounds were needed. The RGB images included for the backgrounds were from the field plots: EN18062 (62.17° N 127.81° E), EN18068 (63.07° N 117.98° E), EN18074 (62.22° N 117.02° E), EN18078 (61.57° N 114.29° E), and EN18083 (59.97° N 113° E), located in Central Yakutia, Siberia (Fig. 7).

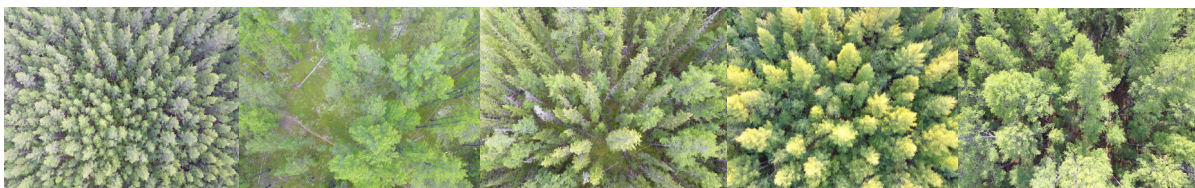



Figure 7: Examples of Red ~~green blue~~ Green Blue (RGB) images of plots from the selected unmanned aerial vehicle (UAV) flights in the following order: EN18063, EN18068, EN18074, EN18078 and EN18083.

430

Figure 8: Example of a Red ~~green blue~~ Green Blue (RGB) image that was excluded from the 35 images for plot EN18068.

The plots were selected based on their vegetation content and their spectral differences, as well as UAV flight angles and the clarity of the UAV RGB images. For each plot, 35 images were selected in order of acquisition, starting at the fifteenth image in the flight to establish the backgrounds for the dataset. The first fifteen images were excluded because they often contain a

435 visual representation of the research team ~~(Fig. 8).~~ for example, Fig. 8). Excluding these images reduces noise in the dataset as we aimed to include only forest and natural terrain in the images.

The raw UAV RGB images were cropped to 640 by 480 pixels at a resolution of 72 dots per inch (dpi). These are later rescaled to 448 by 448 pixels in the process of the dataset creation. In total there are 175 cropped backgrounds.

The foregrounds used in the dataset consist of 117 tree crowns and were manually cut out using Gimp V2.10 software (Gimp,

440 2019) to ensure that they were all *Larix* trees (see Fig. 9). Of the tree crowns, 15% from the margins of the image were included to make sure that the algorithm does not rely on a full tree crown in order to detect a tree.

The COCO format for the SiDroForest synthetic dataset is stored in a JavaScript Object Notation (JSON) file that contains the mask and image name, the colour category that was used to create the mask ~~(Appendix Fig. A5),~~ the category the image falls under, which in this case is 'larch' and the super category which is 'tree'~~.~~ (an example is shown in Table A2). This way the

445 created masks are connected to the created images.

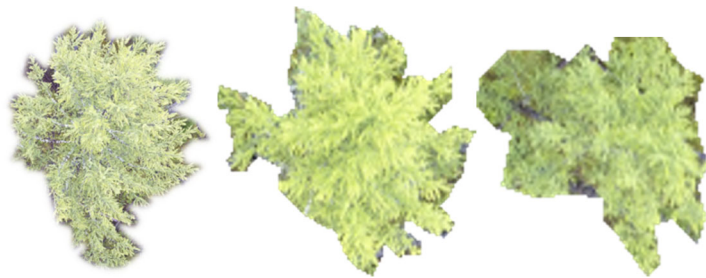



Figure 9: Examples of cut out tree crowns.

The extracted tree crowns were rotated, rescaled, and repositioned across the images using the cocosynth algorithm developed by Kelley (2009) resulting in a diverse synthetic dataset that contains 10,000 images for training purposes and ~~2000~~22,000 images for validation purposes for complex machine-learning neural networks. In addition, the data are saved in the Microsoft COCO format (Lin et al., ~~2013~~2014) and can be easily loaded as a dataset for networks such as the Mask R-CNN, U-Nets, or the Faster R-NN. These are neural networks for instance-segmentation tasks that have become more frequently used over the years for forest monitoring purposes. The Synthetic dataset contains images and labels in the COCO format and can be loaded into most programming languages such as R (R Development Team, 2020) and Python.

2.5 SiDroForest dataset 4: Sentinel-2 satellite image patches

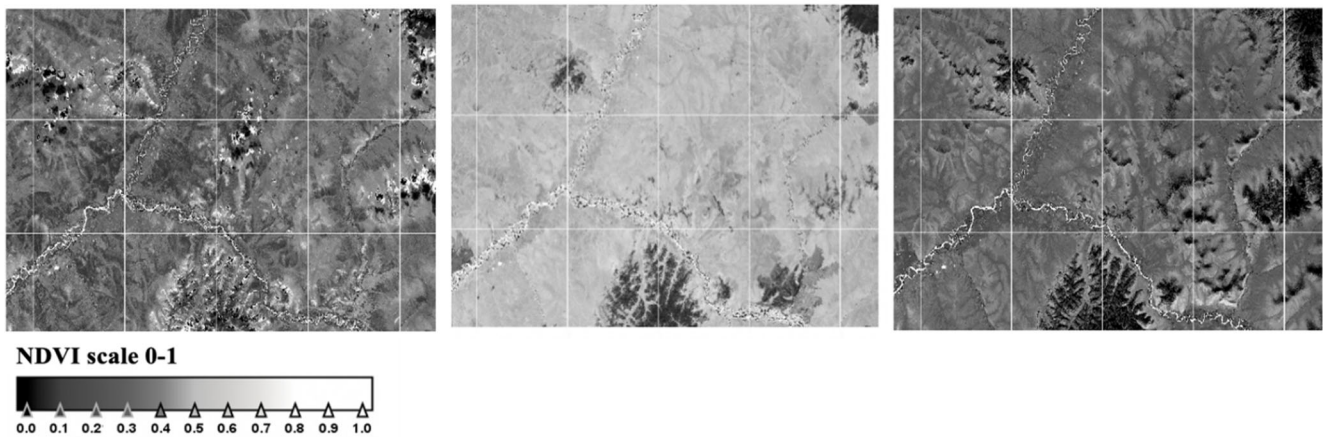
Sentinel-2 (S-2) is an ESA optical satellite mission providing satellite imagery globally and ~~is~~ freely available, which facilitates low-cost broad-scale analyses of circumpolar boreal forests. The S-2 mission is composed of two identical satellites that were launched in 2015 and 2017 (ESA, 2015). The S-2 imagery has 13 multispectral bands, where in the native spatial resolution four bands have the highest (i.e., 10 m) spatial pixel resolution covering the visible wavelength region with three spectral bands (blue, green, red), and one spectral band in the near infrared (NIR). An overview of the S-2 spectral bands can be seen in Appendix Table ~~A1~~A3.

The best possible acquisitions of S-2 data, that is, cloudless and without smoke from forest fires, were retrieved from the ESA archive from the years 2016 to 2020 for three distinct time stamps: early summer (May to June, depending on latitude), peak summer (mid-July to early August), and late summer (late August to September). The ~~ESA~~ S-2 ~~L1C~~Level-1C (top of atmosphere) image data were processed to Level-2A (bottom of atmosphere) surface reflectance using the newest version of the atmospheric correction processor Sen2Cor (ESA Sen2Cor, 2020). Atmospheric correction processing was performed mainly with the default configuration which uses a rural aerosol model with a start visibility parameter of 40 km corresponding to aerosol optical thickness of 0.20 at 550 nm. Actual aerosol optical thickness is determined during the atmospheric correction processing. The ~~only~~ two non-default settings were further enhancements such as the use of the Copernicus DEM for terrain correction (Copernicus, 2021) and the use of vertical column ozone content from L1C-metadata instead of a fixed value of 331 Dobson units.

The data provided in SiDroForest are optimised for vegetation-related analyses, such as resampling ~~the~~all bands to 10 m spatial resolution to make them comparable at the same resolution and removing the 60 m bands that support atmospheric correction but are not optimal for land surface classification. ~~The dataset presented here contains subsets of S-2 acquisitions that cover all the 64 locations where fieldwork was performed in Siberia in 2018. The spectral Bands 5/6/7/8A/11/12 were resampled from 20 m to 10 m.~~

The NDVI was calculated using (B8 − B4) / (B8 + B4) and masked for surface waters using the water-mask provided with the L2A-product. Areas of snow and lake and river ice in early season acquisition NIR bands were masked using an adaptable optimised threshold. ~~For each site,~~The dataset presented here contains 12 subregions (sites) of S-2 acquisitions that cover all the 64 locations where fieldwork was performed in Siberia in 2018 (Table A1) with the three seasonal time-stamps ~~were~~

included and an the water-masked NDVI band was added (Fig. 10 shows an example of the Bilibino subregion NDVI product in Early, Peak and Late Summer).

NDVI scale 0-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

485 **Figure 10: Sentinel-2 NDVI in greyscale forof the three time stamps periods for the Bilibino, subregion in Chukotka. Left: early summer, centre: peak summer, and right: late summer.**

TheIn a further step, the pre-processed S-2 imagery with the spectral bands 2,3,4 (visible), 5,6,7,8A (NIR), and 11,12 (SWIR; short-wave infrared) at 10 m resampled spatial resolution and the additional water-masked NDVI band are cropped to 30 m x 30 m image patches around the centre coordinate of the vegetation plot using UTM oriented shapefiles. These shapefiles and

490 the JSON-annotated image patches receive the annotation of one of the 11 vegetation classes derived from fieldwork and analysis of the UAV data, described in section 2. 2. 1, as attributions (overview on vegetation categories, e.g., Table 2A1A1). The labels are also stored in athe JSON file for each plot in accordance with the patch labelling in BigEarthNet-S2 (Sumbul et al.,. 2019). JSON is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays. It is often used in machine learning as the standard for

495 stored labels.

3. Results

3.1 Dataset 1: SfM point clouds and point-cloud products

The For most of the plots, especially for the Chukotka plots, the total number of RGB and RGNIRRGN point-cloud points were compared per plot. In order to get the full number of points for the RGB point clouds, the *treesonly* (with 'treesonly' and

500 *groundonly* 'groundonly' segmented points were added together (Fig. 11). For most of the plots, the total number of point-cloud points) were of a similar magnitude, especially for the Chukotka plots. (Fig. 11). With higher vegetation structure, the NIR reflectance enables more data points in RGNIRRGN than the RGB point clouds over the high and dense Central Yakutian forest plots. Not all plots have all data available due to technical problems during the data collection (plots EN18035 EN18055, EN18069, EN18071, and EN18076 have no UAV data).

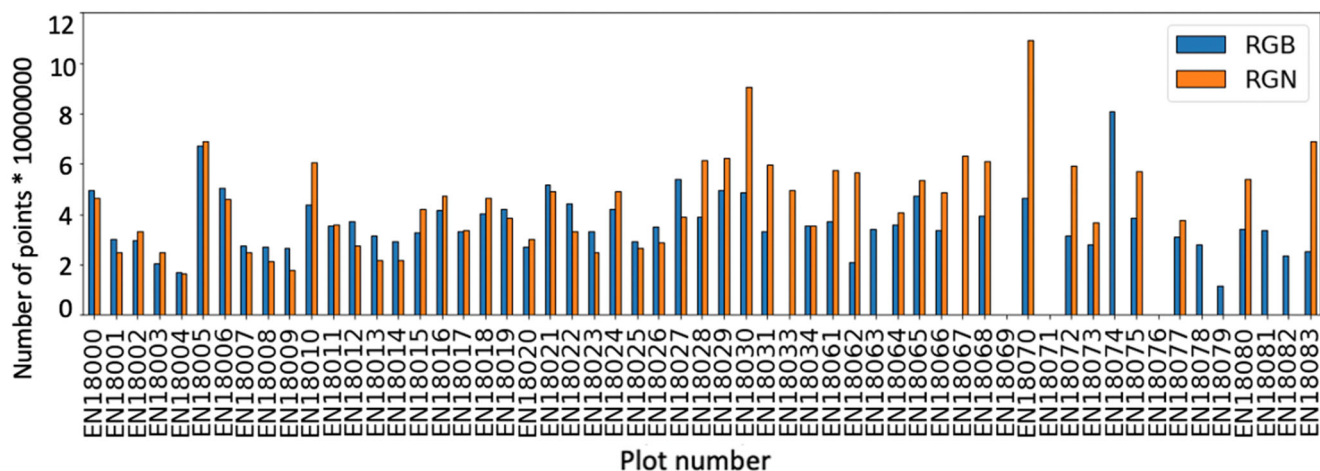
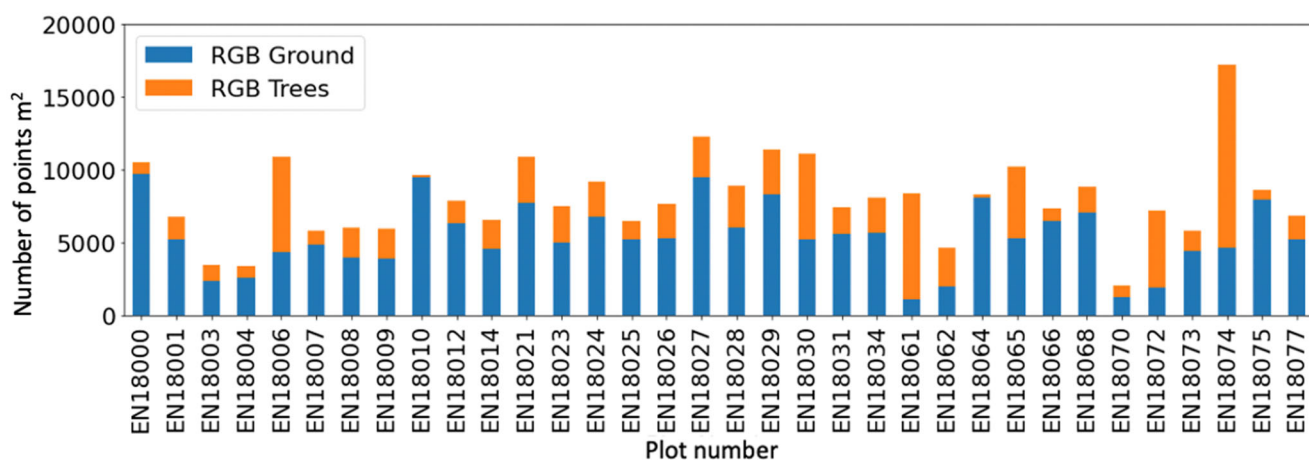



Figure 11: Comparison of the number of points in Red ~~green near~~ Green Near infrared (RGN; orange bars) and the combination of the two Red ~~green blue~~Red Green Blue (RGB) *~~groundonly~~'groundonly'* and *~~treesonly~~'treesonly'* point clouds (RGB; blue bars).

510   **Figure 12: Ground and above-ground points per segmented point cloud per m².**

For the segmented RGB point clouds, the ground to above-ground ratio ~~was inspected (Fig.~~ confirms that ~~12).~~ the plots that have substantially more points in the above ground (*treesonly*) part, i.e~~.~~. a large proportion the point cloud is concentrated the forest canopy if the plots also have more vegetation cover in the higher vegetation layer (Fig. 12)~~.~~.

~~Point cloud products~~

515   ~~For each plot, the generated~~The SiDroForest point ~~ ~~cloud products provide ~~insights into~~high-quality 2D and 3D data on the forest stand structure, the tree height and density and the ground surface elevation ~~and density~~ of the plots (see example for EN18077 in Fig. 13).

520

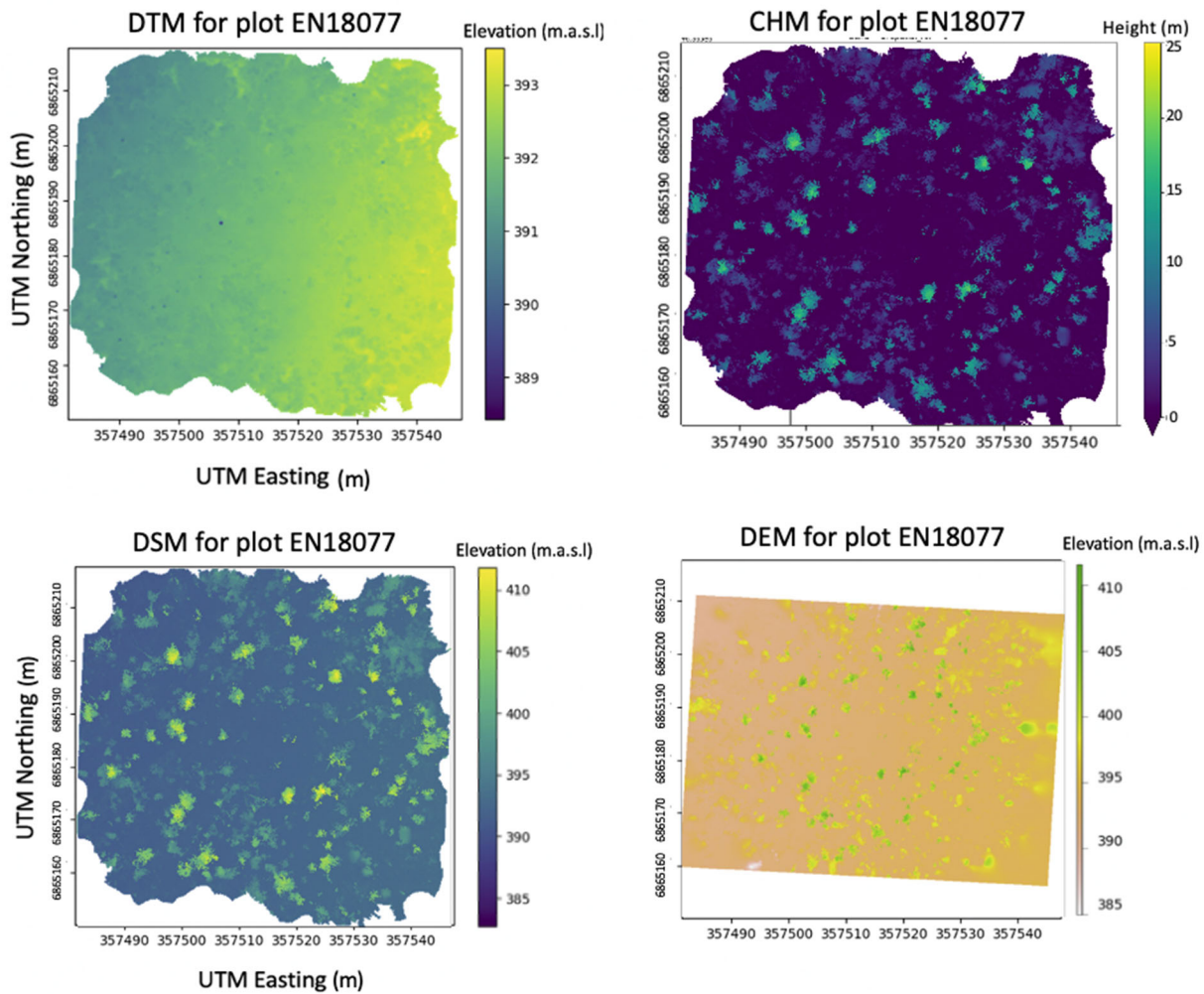



Figure 13: Digital terrain model (DTM), canopy height model (CHM), digital surface model (DSM), and digital elevation model (DEM) for plot EN18077. The DEM products are cropped, the DTM, CHM and DSM are not cropped.

~~Automated tree~~The SiDroForest data collection contains 19,342 automatically detected tree-crown ~~polygons per plot~~ ~~The~~ polygon. In contrast to the high quality 2D, 3D point cloud products, the automatic tree-crown detection algorithm ~~detected~~

525 ~~tree crowns for each plot. The tree crown detection~~ was not equally successful for each plot. For this reason, ~~a~~the quality control label (Q1, Q2, Q3) ~~was~~ included with every shapefile in the name is already a useful indicator for the possible applications of this product. Fig. ~~. Q1 represents the best quality and with the increasing number the~~1414 shows an example of the different quality ~~decreases (Fig. 14).~~scores. Each generated tree crown also has an attribute table assigned that contains information on tree height ~~etc. (Table A2, Appendix).~~ vitality, and crown diameter among others. (Table A4) providing useful

530 metadata information.

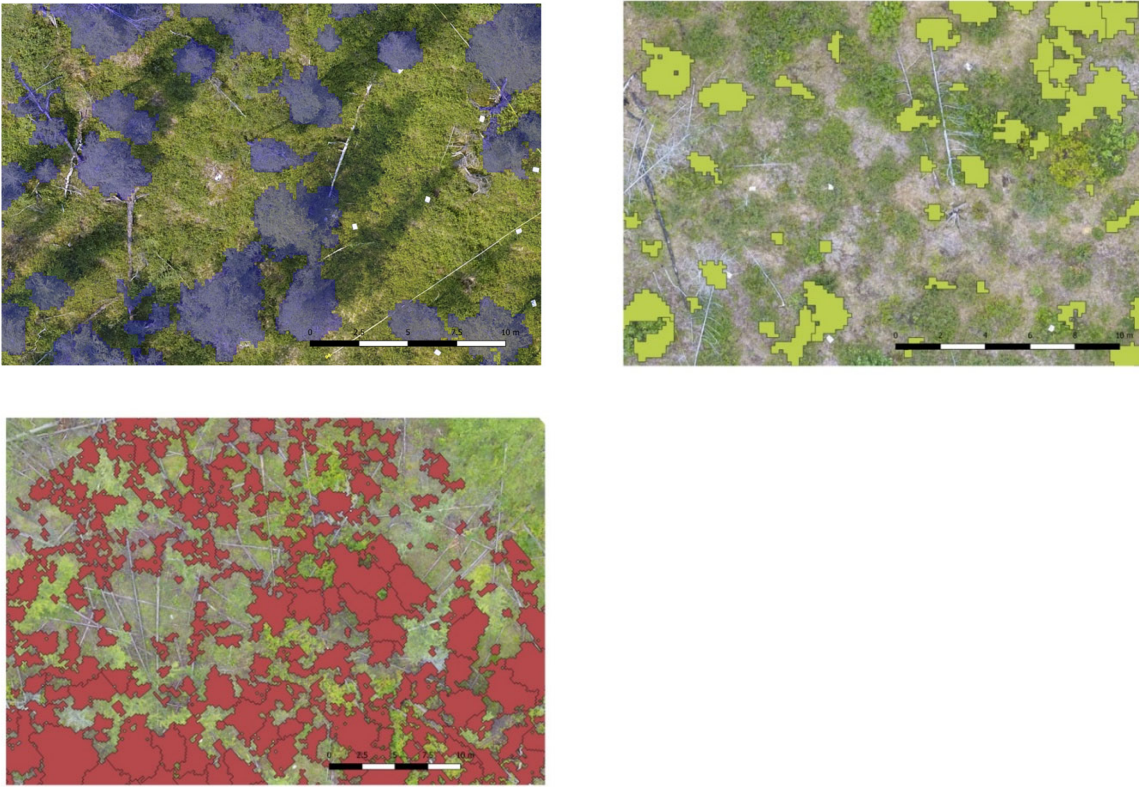



Figure 14: Top left: Crown polygons for plot EN18014 with Q1 quality score. Top right: Crown polygons for plot EN18014 with Q2 quality score. Bottom: Crown polygons for plot EN18014 with Q3 quality score. The scale bar represents 10 meters.

Each plot has a different number of automatic tree crowns detected, ~~this number depends~~depending on the density and the quality of the detected crowns in the plot. The percentage of crowns covering each plot was calculated to show the coverage of trees per plot (Fig. 15). Low tree crown cover staying below 50% coverage characterize the vegetation plots in the tundra-taiga transition zone in Chukotka, whereas reach crown coverage of higher then 50% up to 80 and 90% in some of the plots in the summergreen-evergreeen transition zone in Central Yakutia. However, also in the Central Yakutian boreal zone a tree crown coverage between 30 to 60% only characterise most of the field forest plots.

535

540

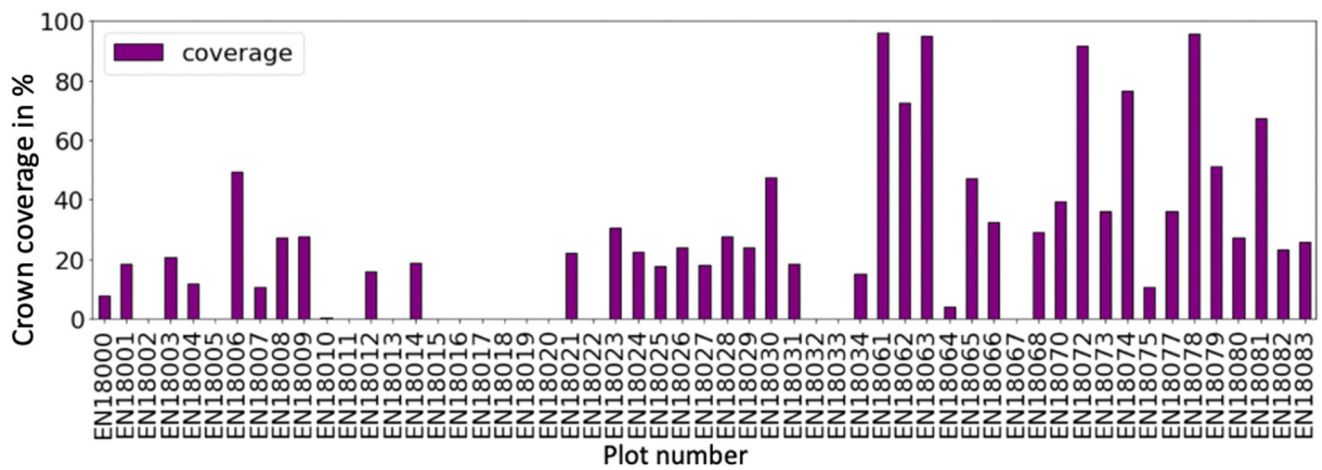



Figure 15: The percentage of the crown coverage in the orthomosaics per plot. A high percentage reflects a denser forest.

3.2 Dataset 2: Individual labelled trees

In order to make assumptions and predictions about the content of the vegetation plots it is important to link the labelled individual trees from the fieldwork to the processed orthoimages. ~~We located 872 trees and large shrubs that were recorded~~

545 ~~during fieldwork in the orthoimages. A minimum of ten individual trees were recorded per plot with the exception of plots EN18014 and EN18028, where all trees present were recorded. The 872 trees and large shrubs recorded in the field were selected as representative individuals for their respective plots. Not all individual trees that were recorded in the field were also located in the orthomosaics (Fig. 16). The mean tree height and mean crown diameter per plot can be seen in the Appendix (Fig. A6 and Fig. A7).~~

550

We located 872 trees and large shrubs in the orthoimages that were surveyed in Siberia during the two-month fieldwork expedition in 2018 (Kruse et al., 2019) (Fig. 16).

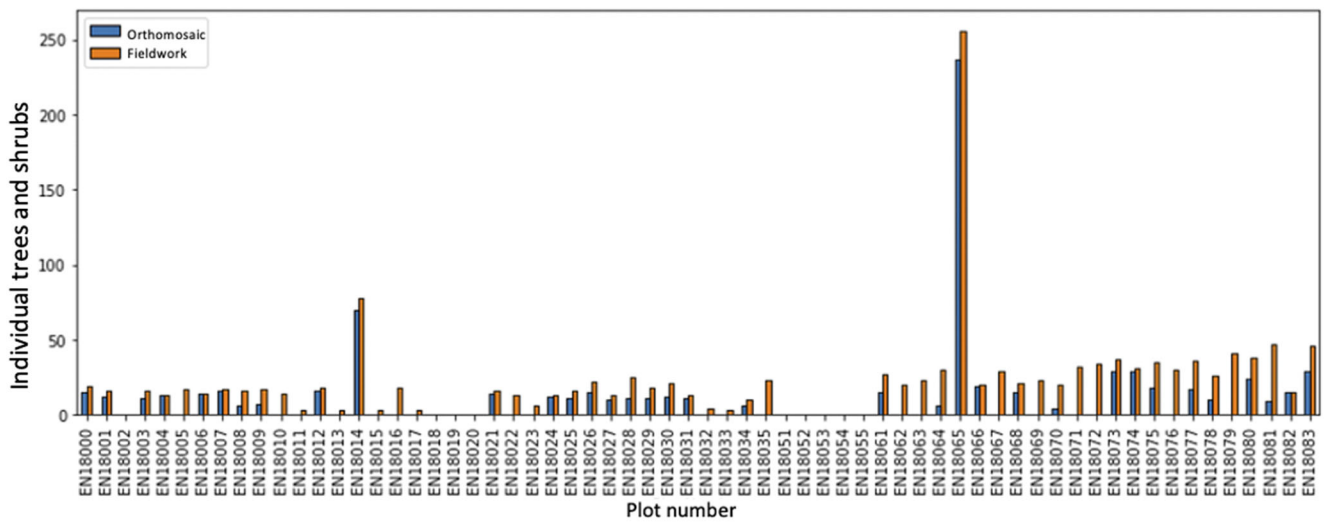



Figure 16: Number of individual trees recorded in the field (orange colour) and visually identified and relocated ~~in the red green~~ ((blue colour) in the Red Green Blue (RGB) orthomosaics per plot. For plots EN18014 and EN18065 all trees were recorded that were present on the plot.

For each tree or shrub from fieldwork visually identified in the orthoimages, the created point and polygon shapefiles contain information about the tree or shrub species visible in the orthoimages. The field data on species distribution (trees and tall shrubs) and on mean tree height and mean crown diameter per plot can be seen in the Appendix (Fig. A2, A3, A6, ~~and~~ A7.).). For each located individual, the three shapefiles pinpoint the location, ~~add~~provide a unique identifier, and record the species information and ~~are~~cancan be overlain by users on the RGB or ~~RGNIR~~RGN orthoimages of the plots as a useful visualisation (example in Fig. 17). The individual number links to the information collected during the expedition such as tree height, crown diameter and vitality. This dataset can be used to link individual trees in the SfM point clouds, providing unique insights into the vegetation composition and also allows future monitoring of the individual trees and the contents of the recorded vegetation plots.

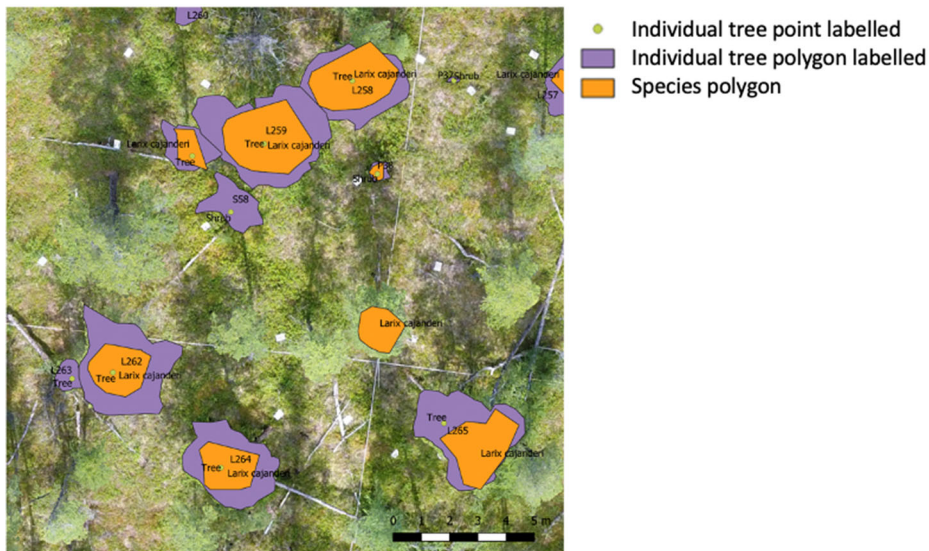



Figure 17: Overview of the three types of shapefiles included in the individual labelled trees dataset ~~visualised~~visualized on top of a red–green–blue (RGB) orthoimage.

3.3 Dataset 3: Synthetic dataset results

570   This synthetic Siberian larch tree crown dataset was created to enhance the data collective for upscaling and machine-learning purposes. The synthetic larch tree crown RGB image database has many different larch-dominated forest structures and contains 10,000 synthetically produced images. This creates a large diversity of spatial and spectral features for machine-learning tasks. Examples of the results for the synthetic larch tree crowns include the RGB images that were generated and the accompanying masks that are used for the instance segmentation and object detection tasks as shown in Figure 18.

575

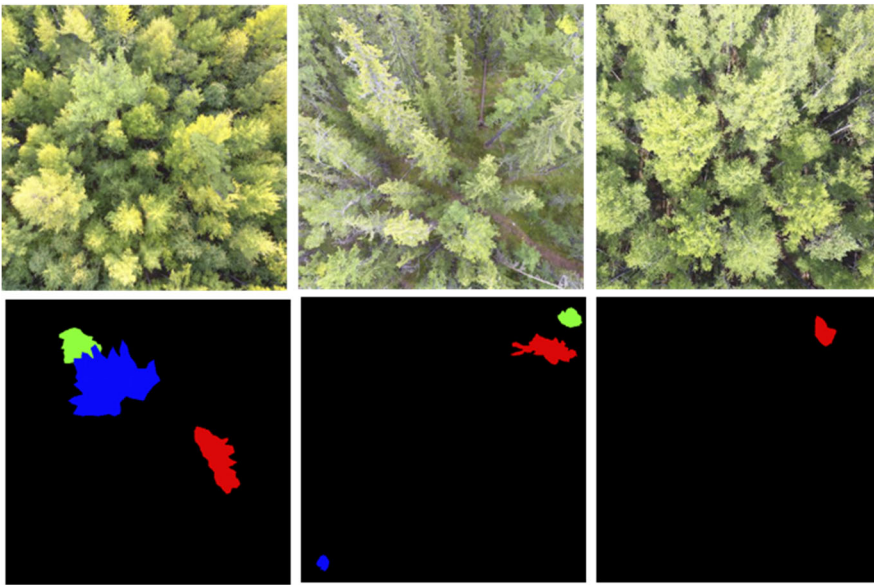



Figure 18: Examples of synthetic images and corresponding masks generated. The images show three drone flight images with a cut out larch tree overlay. The masks below show the location of the placed trees in the form of masks. Each mask is assigned a different colour to distinguish the masks.

580 ~~Examples of the results for the synthetic larch tree crowns include the RGB images that were generated and the accompanying masks that are used for the instance segmentation and object detection tasks as shown in Figure 18.~~ The synthetic larch tree crown RGB image database has many different larch dominated forest structures. This creates a large diversity of spatial and spectral features for machine learning tasks.

3.2 Dataset 4: Sentinel-2 labelled image patches

585 The labelled S-2 image patch dataset comprises 30 m × 30 m labelled multi-band (10 multispectral bands + NDVI) image patches ~~per plot,~~ with ~~11 bands and~~ vegetation labels (table A1) assigned ~~to each plot. The information per patch is stored in a JSON file. Figure 19 provides a schematic overview. It also includes~~ and three seasonal ~~time-stamps:~~ representations (early summer, peak summer, and late summer.~~ A~~) for 63 plots and 12 subregions (sites) (table A1) with the same multi-band format. As each 30 m x 30 m S-2 image patch consists of nine units (pixels) of 100 m² extent each, it amounts to around 550 annotated

590 validation and training units. Figure 19 provides a schematic overview of the contents of the dataset, a visual representation of the 11 vegetation classes can be seen in Figure 20. ~~In total there are 11 vegetation classes for Chukotka and Yakutia. The~~ For easy re-use and machine learning purposes the vegetation classes ~~were added~~are in the file name for each patch as well as in the JSON file. The classes and their ~~distribution~~representation in the labelled S-2 image patches are shown in ~~Figure 21. A~~ table ~~with all the class labels assigned per plot can be seen in Appendix Table A3~~table 2.

4 VIS bands,
4 NIR bands,
2 SWIR bands,
1 NDVI masked

Vegetation Classes
1. Graminiod Tundra
2. Forest Tundra and Shrub Tundra
3. Prostrate Herb Tundra
4. Open Canopy Pine with Lichen
5. Open Canopy Pine
6. Closed Canopy Pine
7. Open Canopy Mixed Forest
8. Closed Canopy Mixed Forest
9. Open Canopy Larch
10. Closed Canopy Larch
11. Closed Canopy Spruce

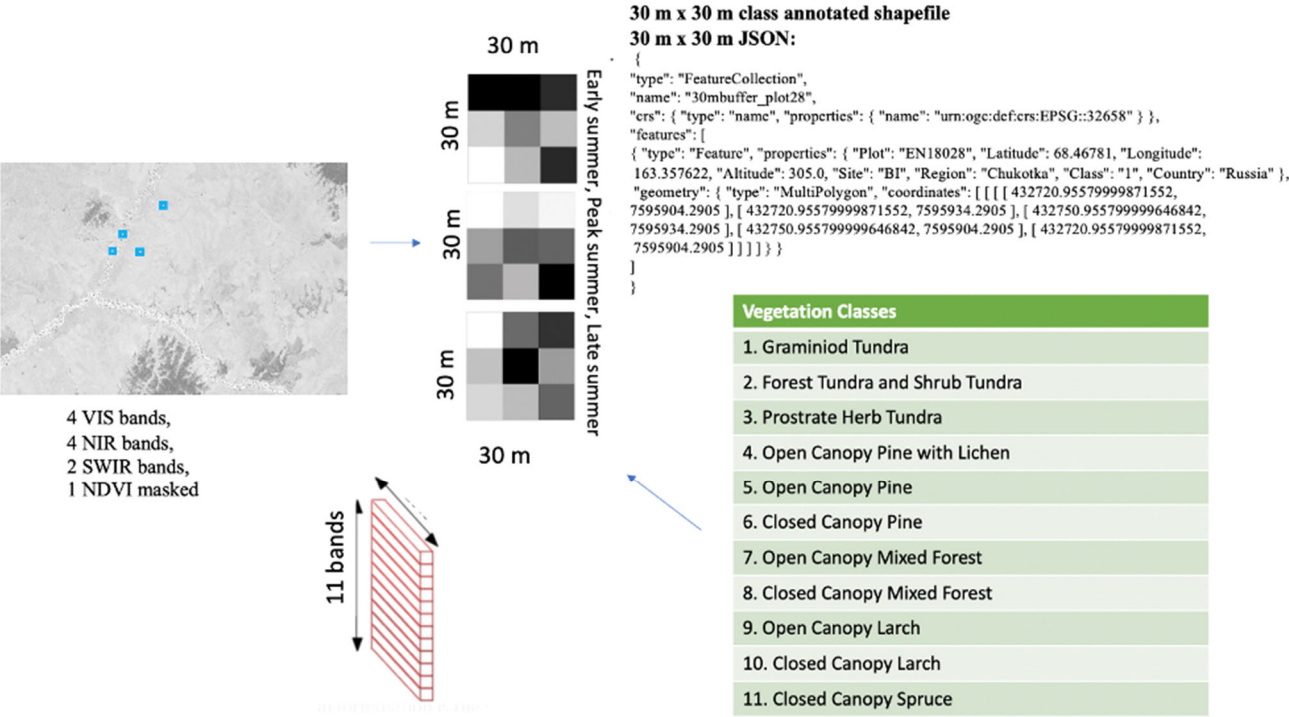



Figure 19: Overview of the products in the Sentinel-2 labelled image data set, exemplified for plot EN18028.

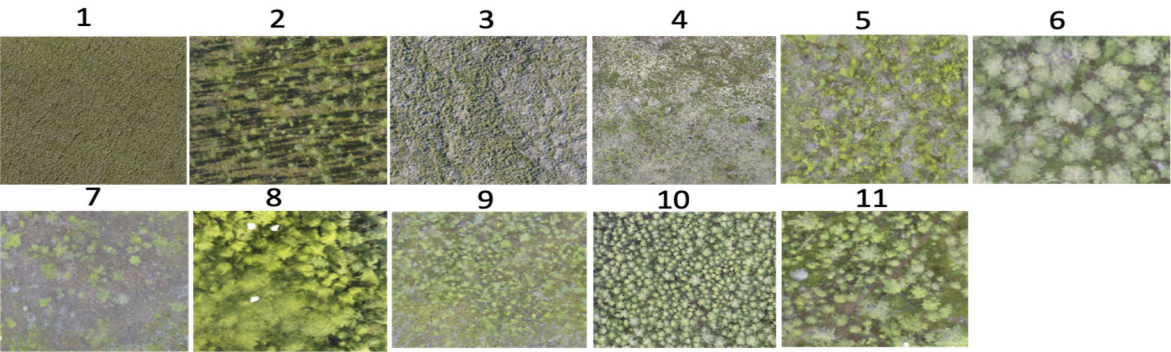



Figure 20: Visual representation of the 11 vegetation classes from the orthomosaics (see also Appendix Table A3A1). The varying quality of the orthomosaics is described in 2.2.322.

Table 2: Vegetation class labels per plot and percentage of plots for each classification.

Vegetation label	~~Percentage~~Fraction of plots with this label
Graminoid tundra	39%
Forest tundra and shrub tundra	4%
Prostrate herb tundra	21%
Open canopy pine with lichen	2%
Open canopy pine	2%
Closed canopy pine	2%
Open canopy mixed forest	10%
Closed canopy mixed forest	4%
Open canopy larch	4%
Closed canopy larch	10%
Closed canopy spruce	2%

4. Discussion

4.1 ~~A comprehensive~~Uniqueness of the SiDroForest comprehensive data collection on Siberian boreal forests

605 To date, the most relevant open-source datasets available on boreal and arctic vegetation data are from the long-term ABoVE NASA Terrestrial Ecology Program, focusing on boreal and arctic regions in Alaska and Canada. The ABoVE data collections contain field-based, airborne, and satellite sensor-derived data, providing a foundation for improving the analysis and modelling capabilities needed to understand and predict climate change in the arctic and boreal regions. ~~There are~~In 2021, there were fifty vegetation-related datasets published so far in the ABoVE Science Cloud (ASC): eleven thematic maps, mostly

610 derived from remote sensing and focused on Alaska, nine vegetation-variable related mapped remote-sensing products, mostly covering large regions, one time series product extracted for the footprint of a flux tower, and six ground-based vegetation related data collections, including data from ten terrestrial LiDAR vegetation plots (Maguire et al~~., 2016., 2016~~2020) and 24 vegetation plot surveys. The ~~circum arctic~~circumarctic vegetation map north of the treeline (CAVM Team, 2003) is one circumarctic product, the other forty-nine datasets are all located in Alaska.

615 SiDroForest provides a new comprehensive data collection ~~of~~with a variety of data types that were selected to create the most useful insights into ~~the boreal forests,~~ specifically the larch-dominated forests representative of ~~central and~~eastern Siberia. The focus of the SiDroForest data collection is, at this stage, not to provide thematic maps or upscaled remote-sensing products but to provide a rich, open data source on ground-based and UAV-derived information and labelled data types enhanced to best use the data for vegetation-related analyses and machine-learning tasks. ~~The data types combined aim to provide a clear

620 picture of the current state of the vegetation in the tundra taiga and summergreen evergreen transition zones.
In this newly assembled SiDroForest data collection we provide i) labelled field data that are new data from the field studies in 2018 in~~For For Eastern Siberia ~~for the Chukotka and Yakutia regions and ii) link SiDroForest data in the PANGAEA data repository to the~~, we had already published ~~datasets via their metadata. Publications by~~ in Shevtsova et al. (2019, 2020b) ~~include the ~~2016 and 2018 vegetation inventories on the projective vegetation cover, ~~while the total ~~and ~~partitioned~~2018

biomass data ~~of the 2018~~(Shevtsova et al. 2020c) of vegetation plots for the tundra–taiga transition zone in Chukotka ~~are found in Shevtsova et al. (2020c). Kruse et al. (2020, 2021) published two.~~. Tree level forest ~~inventories for the 2018 vegetation~~inventory data from Eastern Siberian forest plots~~, one for the tundra–taiga in Chukotka and one for the summergreen–evergreen transition zone in central Yakutia.~~ were published in Kruse et al. (2020a) and Miesner et al. (2022). As well, we published in Brieger et al. (2019a,b) ~~supply~~a first version of ten ultra-high resolution photogrammetric point clouds from the UAV overflights in 2018 over forest vegetation plots~~.~~ in Central Yakutia. For these ten plots, the construction of RGB SfM point clouds was evaluated and optimised and was then used to process all RGB and ~~RGNIR~~RGN SiDroForest point clouds. ~~We~~In the SiDroForest data collection, we provide ~~here~~ the complete and comprehensive dataset of the full range of standardized SfM-derived products of the 2018 UAV acquisitions in Central Yakutia and Chukotka (Kruse et al. 2021b~~) including~~). In the SiDroForest data collection in addition to all ~~the RGNIR~~RGN and RGB point clouds from all ~~60~~63 overflown vegetation plots ~~with~~, we provide enhanced field data information such as the individually labelled tree dataset (van Geffen et al., 2021b). These existing field inventories (Shevtsova et al. 2019, 2020b,c; Kruse et al. 2020a, Miesner et al., 2022) are data publications optimized for ecological applications, and not for machine learning, and upscaling applications. In the PANGAEA data repository, the individual data sets for ecological applications and the SiDroForest data sets can all be linked to each other by the vegetation plot codes. With these interlinked data types, multi-purpose applications, and a more in-depth understanding of the Siberian boreal forests can be fostered.

4.2 High spatial resolution UAV-derived domain in forest data collections

The SiDroForest data collection is based on a large part on photogrammetric UAV-borne products (i.e., SfM point clouds, digital elevation products, RGB orthomosaics) following a long application history in forestry and well-defined methodological standards (e.g., Jensen et al., 2016; Panagiotidis et al., 2017). ~~The SiDroForest data collection also contains 19 342 automatically detected tree-crown polygons (Kruse et al. 2021b). Bieger et al. (2019) report the comparison of field measured and detected crown diameters (mean R² = 0.46, mean RMSE = 0.673 m, mean RMSE% = 24.9%). The weak correlation between observed and detected crown diameters is explained by Brieger et al. as being due to the quality of the available field data, which are estimations instead of direct measurements and therefore could have decreasing precision with increasing heights.~~

~~Automatic tree-crown detection~~Currently, the use of UAVs in environmental applications is undergoing an ever faster growing use in forestry and environmental science due to the landscape-level potential, the flexibility of the data generation and low costs (Fraser et al., ~~2016~~2017). The SiDroForest data collection is now extending our standard ground-based inventories. In addition to the photogrammetric UAV products, we undertook an automated tree-crown detection that has become more frequent due to the availability of state-of-the-art instance segmentation algorithms from the world of computer vision (Neuville, 2021). An example of previous work using a neural network tree-crown detection is Braga et al. (2020), where the Mask R-CNN (He, 2017) was used to perform the tree-crown detection and delineation. In another example, the Mask R-CNN was used by Hao et al. (2021) to detect tree crown and canopy height of Chinese fir in a plantation in China. Tree crown width

and tree height of Chinese fir were manually extracted from this UAV imagery using a combination of labelled ground-truth data and canopy height model (CHM) information and served as validation data. This exemplifies how the synthetic dataset in SiDroForest (van Geffen et al. ~~2021e~~2021a) could be used for analysis as the Mask R-CNN is trained with a COCO format dataset.

~~The format of how to provide extracted tree crowns can differ: Weinstein et al. (2021) provide~~For the United States, the National Ecological Observatory Network (NEON) provides a 100-million individual tree crowns dataset covering a large area and standardised LiDAR remote-sensing data ~~from the National Ecological Observatory Network (NEON) which is very impressive.~~(Weinstein et al., 2021) created using machine learning tools such as DeepForest (Weinstein et al., 2019). Here a CHM was used to filter out all canopy tops over 3 metres in height from 37 different NEON sites. The individual tree crowns in Weinstein et al. (2021) are represented by a bounding box shapefile that approximates the crown area and links to the tree attributes. The SiDroForest tree-crown dataset ~~(van Geffen et al., 2021e) provides the tree crowns in the form of a crown-delineating polygonal shape enriched with attributes. The SiDroForest tree-crown data~~ cannot cover a comparably large area as the NEON airborne LiDAR data collection extending over 1 km x 1 km tiles, ~~but it does offer~~and used RGB point cloud products and not lidar-derived CHMs. However, the SiDroForest tree-crown dataset provides 19,342 automatically detected tree-crown polygons in the form of a crown-delineating polygonal shape enriched with attributes offering plot-size coverage of tree crowns with useful data for machine learning and computer vision applications. ~~The SiDroForest tree-crown data (van Geffen et al., 2021e)~~The tree crown extraction with 19,342 tree crowns is not complete, what we addressed assigning quality scores to the products. Brieger et al. (2019) also report a weak correlation between observed and detected crown diameters (mean R^2 = 0.46, mean RMSE = 0.673 m, mean RMSE% = 24.9%). We assume that is due to the reduced quality of the available field data, which are subjective estimations instead of absolute measurements and therefore could have decreasing precision with increasing tree heights. The SiDroForest tree-crown data are specifically made to detect Siberian larches in different mixtures of mixed summergreen needle-leaf and evergreen needle-leaf forest. ~~Additionally, the synthetic SiDroForest dataset (van Geffen et al. 2021a) provides 10 000 images for training and 2000 validation images to train neural networks.~~

4.3 Upscaling using SiDroForest data types

It is increasingly common in data science and environmental science to use multiple data types within one analysis. For example, S-2 images and metadata, topography data, CHM, as well as their combinations, were used to predict growing stock volume using deep neural networks in four forestry districts in central Finland (Astola et al., 2021). Another example of the use of multiple data types in non-machine learning remote sensing is the work by Wang et al. (2020) where above-ground biomass estimation was performed using field plots, UAV-LiDAR strip data, and S-2 imagery. In Wang et al. (2020), the partial-coverage UAV-LiDAR data were used to link ground measurements to S-2 data. These recent studies show the need for well-labelled publicly available data to link the data types together and for performance testing of remote sensing algorithms. In these studies, the testing data preparation was undertaken within the project: For example, Thanh ~~et al.~~Noi and Kappas (2018) compared the performance of three common machine learning algorithms; a support vector machine (SVM), a

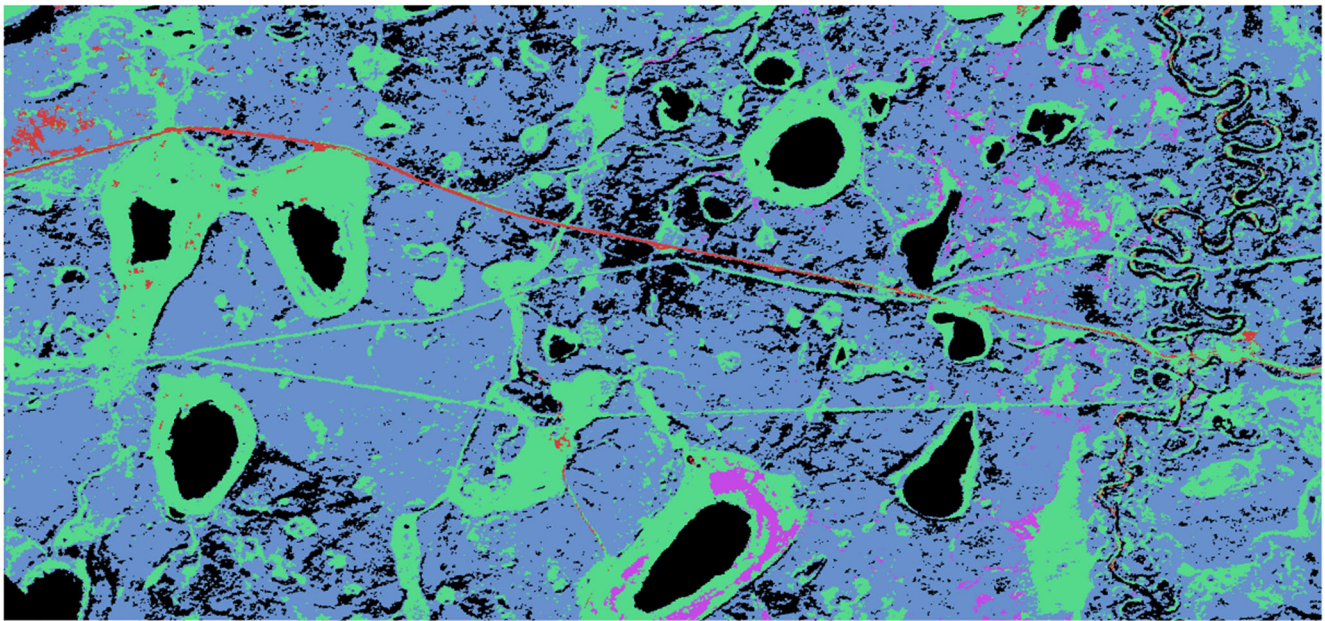
random forest (RF) and k nearest neighbours (K-NN) on S-2 data from Vietnam. In order to validate the performance of these algorithms, the training data (training and testing samples) were collected based on the manual interpretation of the original S-2 data and high-resolution imagery obtained from Google Earth and 135 labelled land cover polygons were produced. Thanh et al.Noi and Kappas (2018) is a good example of manually labelled data creation for a specific task and specific research area to be able to use supervised classification tools. The work done by Abdi (2020) shows a similar study that assesses the performance of four machine learning algorithms for land cover classification of boreal forests. Here too, the validation and training data is manually created to assess the performance of the algorithms.

However, despite the increased availability of satellite missions and open-source remote-sensing data and products, challenges remain that are particular to terrestrial high-latitude ecosystems. Seasonal challenges such as the combination of snow cover over a long time of the year, a short and rapidly progressing growing season, high cloud frequency, and low sun angles pose a problem for comprehensive remote-sensing collectionsapplications in the sub-Arctic regionhigh latitude regions (Beamish et al., 2020). SiDroForest aims to remedy this scarcity by providing this multi-source data set, for example the high-quality dataset of S-2 data linked to published field inventories (van Geffen et al., 2021b). The final labels for the S-2 labelled image patches provide ground truthare assigned from the in situ information neededof multiple data sets the first two data sets i) and ii) information that can now be upscaled to classify larger areas. by satellite image classification. By this, we assigned the labels with expert knowledge from the field data, still keeping all transparent, so that future users of these data sets can adapt the labelling to their applications, based for example on the detailed information in the tree level and plot level labelled data sets i) and ii) that we provide in this data collection together with the S-2 labelled image patches for training.

The Yakutia field data collection covered diverse plots as seen in the vegetation classes assigned (Table 2) which may pose a problem for classification as the classes are unevenly distributed. When the fieldwork was undertaken, multiple plot sites covering different classes were preferentially recorded in close proximity to each other for time-related reasons. The time spent in fieldwork is limited and expensive and a variety of different data can be collected close to each other. The diversity of the collected fieldwork data has advantages and disadvantages for machine learning. On the one hand it is good to have many different vegetation types covered in the field plots to log the diversity of the vegetation cover for the region. On the other hand, more ground-truth data plots in the same category will greatly improve classification of satellite data and too much diversity in the classes hinders a balanced classification. For example, label 4: Open Canopy Pine with Lichen, only occurs in one plot. Spectrally, this plot is different from the others due to the presence of the almost white coloured lichen. It was therefore important to label this plot differently from the others, even if this creates uneven and unbalanced labels.

An idea to remedy this in the future is to add in more fieldwork plot locations from upcoming expeditions and to supplement the plots with S-2 data that are similar in their spectral signal from the same region. If the user wants to classify S-2 based on the labelled patches, it is important to keep the unbalanced classes in mind.

The classes assigned to the S-2 image patches were tested with simple machine-learning algorithms. The patches were extracted for both Yakutia and Chukotka and used together to classify all sites. A Gaussian Naive Bayes performed best with

725 82% overall average accuracy per class for the Yakutia sites. The preliminary results for one of the Yakutia sites are shown in Figure ~~23~~21, chosen due to the diverse vegetation at the site, to show the classification potential.

Legend

Vilnuyi
- ⬛ Water
- 🟥 Barren area
- 🟦 Open Canopy mixed forest
- 🟪 Closed Canopy mixed forest
- 🟩 Closed canopy Larch

Figure ~~23~~21: Classification of the ~~Vilnuyi~~ Sentinel-2 ~~image patch~~ Vilnuyi subregion based on the vegetation labels in SiDroForest. This is an initial classification using a Naïve Bayes algorithm with additional classes water and barren areas.

730 **4.4 SiDroForest Labelling and Data Quality**

Labelling accurately is one of the most important aspects for a usable dataset for machine-learning purposes. If the labels are inconsistent or very uneven the classification tools will have trouble correctly identifying the classes. The SiDroForest data collection contains a variety of labels per dataset.

The labels for the *Individual Labelled Tree* dataset (van Geffen et al. 2021c) contain information on species and location of

735 the individual tree or shrub. These data have been verified and checked, yet in some instances two trees are located very close to each other or the location was not recorded correctly in the field and an individual tree or shrub could not be found in that case. The difference between the number of trees recorded in the field and located in the orthomosaics can be seen in Figure 15. The UAV images were inspected based on expert knowledge to locate the trees as accurately as possible. However, dense forest plots in Yakutia posed a problem for locating all the individuals correctly and not all individuals recorded in the field

740 could be located in the orthoimages for those plots. Figure 22 shows an example of dense forest plots.

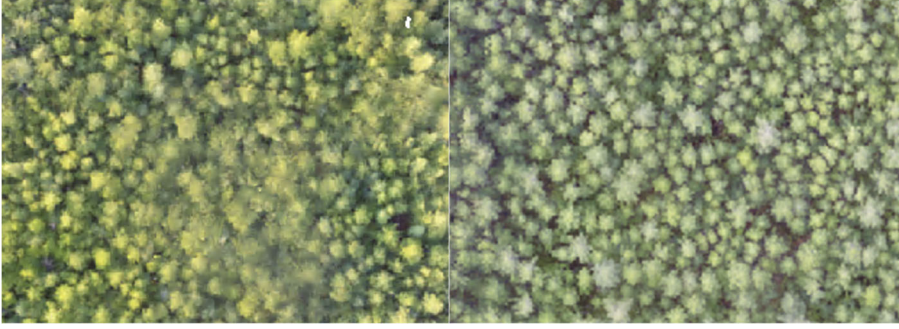



Figure 22: Dense forest Red Green Blue (RGB) orthomosaics for plots EN18077 and EN19063.

The SiDroForest synthetic dataset (van Geffen et al., 2021a) has written labels in the JSON format (table A2) that contain the higher category, or 'super category', 'Tree' and subcategory 'Larch'. The two categories exist in case there are more species added under the higher-level label 'Tree'. The current set identifies all larch trees, regardless of which species, since the sites covered contain two larch species: *Larix cajanderi* and *Larix gmelinii*. The two species of larch here have only the one label larch because the aim was to identify all larch trees in both Chukotka (solely *Larix cajanderi*) and Yakutia (predominantly *Larix gemelinii*). It would be an enhancement of the dataset in the future to distinguish between the two species of *Larix* in the labels as well. The dataset can be further enhanced by adding the other dominant tree species for the region: spruce and pine. The backgrounds were carefully selected for the synthetic dataset to create diverse scenes and forest information for the algorithm to learn from. This can help the algorithm detect larch trees on multiple backgrounds. However, it may also introduce noise into the dataset. As investigated by Xiao (2020), on one hand, there is evidence that models succeed by using background correlations but on the other hand, advances in classifiers have given rise to models that use foregrounds more effectively and are more robust to changes in the background. These findings suggest that the performance of the algorithm is more important than the consistency of the backgrounds in a dataset. However, it is still important to be aware of such interference, and extensive benchmarking is needed to evaluate the performance of an instance segmentation or object detection algorithm for the dataset, which we are planning to undertake.

The dataset also contains generated RGB images that should contain natural looking scenes. In practice, not all the RGB images look as natural as others (for example, parts of images in Fig. 23). The unnatural image construction is mostly due to variation in size compared to the images placed on them. Since there are 10,000 images in the dataset these unnatural images do not strongly undermine the natural ones and make up less than 10% of the total images.

Figure 23. Examples of unnatural looking generated images in the synthetic image dataset, the red arrows show the cut-out larch trees that were placed over the UAV images.

The SiDroForest data collection also provides labelled S-2 satellite image patches per vegetation plot (van Geffen et al. 2021b) that can be used as ground-truth data for machine-learning classifications. Though freely available and operationally downloadable, S-2 data are not ready-to-use. Despite a frequent acquisition rate at higher latitudes, S-2 data often contain clouds and finding a cloud- and haze-free acquisition can take time, even with cloud filtering. It is common practice that users produce labelled patches of satellite data that function as parameterisation for classification and upscaling purposes. For example, BigEarthNet (Sumbul, et al., 2019) is a large-scale open-source dataset that provides labelled S-2 image patches (now called BigEarthNet-S2, previously *BigEarthNet*) acquired between June 2017 and May 2018 over ten countries. Each patch includes a JSON file with the ground cover labels for the patch. In accordance with the structure of BigEarthNet-S2, the SiDroForest image patches are also accompanied by a JSON file that contains the class labels per image patch. BigEarthNet-S2 provides patches of larger area coverage to represent 'landscapes' such as estuaries. The purpose of the SiDroForest S-2 image patches and labels lies in the true representation of vegetation classes and evergreen needle-leaf mixed forest and the seasonal time stamps of early summer, peak summer, and late summer.

In its current stage, the SiDroForest S-2 data collection is not published with performance testing, and is by us not considered as a benchmark data set for Remote Sensing image interpretation (e.g., Long et al., 2020). The SiDroForest labelled S-2 image patches collection is available as a small training and validation data set providing so far underrepresented vegetation categories, that will save future users time when attempting to classify vegetation of Central Siberian and Eastern Siberian boreal forests.

5. Conclusions

The circum-boreal forests are covering large areas on the globe. Every new forest data set collected, processed further and published in a ready-to be used format for a wide range of biological and ecological applications is therefore quite rare and an important addition for scientific studies that aim to better understand global forest dynamics.

The datasets presented here provide a comprehensive overview of the vegetation structure of boreal forest using a variety of data types. The fieldwork locations are the anchors that bind all the data types in this data collection together. The datasets include fieldwork information from vegetation plots and UAV acquisitions from ~~64 vegetation plots during fieldwork jointly performed by~~extensive field expeditions in summer 2018 covering the ~~Alfred Wegener Institute for Polar~~tundra–taiga and ~~Marine Research (AWI)~~summergreen–evergreen forest transition zones in Chukotka and ~~the North~~Central Yakutia in Eastern ~~Federal University of Yakutsk (NEFU) during the Chukotka 2018 expedition to~~ Siberia.field expeditions. The data collection spans from forest inventories at the species level, tree height information and density for each vegetation plot, UAV-derived SfM point clouds that provide structural forest information, RGB and ~~RGNIR~~RGN orthoimages from the plots, to S-2 image patches of seasonal information annotated with vegetation categories that can be used for upscaling purposes to a larger region. Combining the data types within SiDroForest can lead to a better understanding of forest structures and vegetation ~~dynamics~~composition. The future states of boreal forest are still largely unpredictable: labelled field data and remote-sensing data provide the tools for machine-learning based applications to help forecast likely scenarios.

The increased use of machine-learning techniques in the field of remote sensing and forest analyses ~~call~~calls for more and better labelled data. If ~~tree level~~ forest structure data are rarely available for the tundra–taiga and summergreen–evergreen transition zones, even less ~~data~~ is available that can be used for machine learning, such as optimised data containing labelled vegetation. In addition, due to the remote nature of the dataset locations, obtaining ground-truth data is difficult and expensive. The current data ~~collective~~collection provides ~~a base~~rare data ~~collection. ed.~~on the Central Yakutian and North Eastern Siberian land cover, optimized on larch forest across the evergreen-summergreen transition zone and the northern tree line. Adding in future similar datasets derived from the Northern American boreal domain will consistently ~~to the SiDroForest dataset by~~enlarge~~enlarge~~ and will encompass more tree species and forest types in the upcoming ~~joint AWI-NEFU expeditions in the future, the usability and relevance of this data collective will grow with time~~years.~~years.~~ By making this data collection open source, we aim to remedy data scarcity on tree level forest data for the region and we encourage the use of the labelled tree level and ~~lot~~plot~~plot~~ level forest data sets presented here for further analyses and machine-learning tasks.

6. Data availability

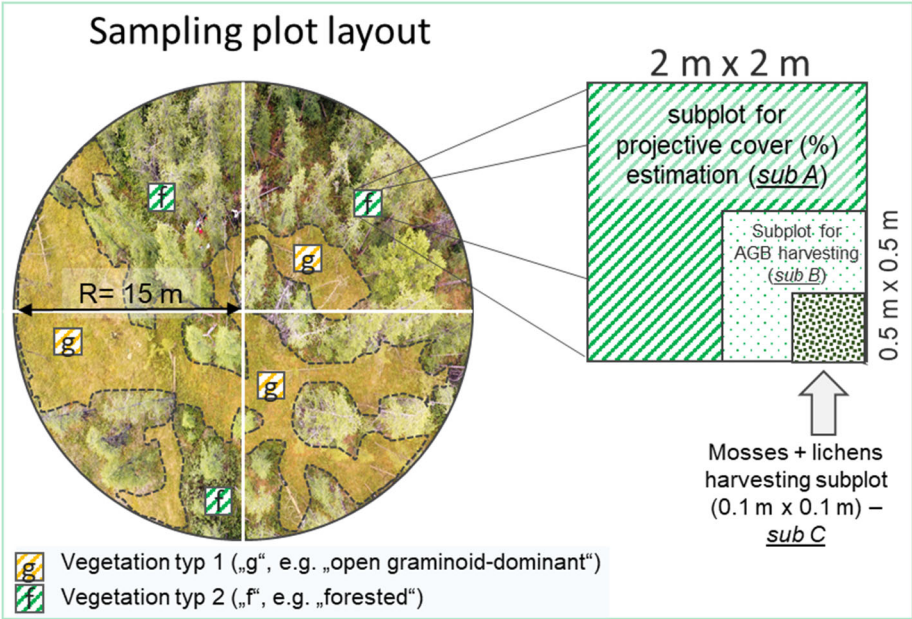
All four data sets of the SiDroForest Data collection are published in the PANGAEA data repository ~~together with read-me files with further technical details~~ and are available for download:

i) UAV-SfM point clouds, point-cloud products, and orthoimages: https://doi.pangaea.de/10.1594/PANGAEA.933263,

ii) Individual labelled trees: https://doi.pangaea.de/10.1594/PANGAEA.932821,

iii) Synthetically created tree crowns dataset: https://doi.pangaea.de/10.1594/PANGAEA.932795

iv) Sentinel-2 labelled image patches: https://doi.pangaea.de/10.1594/PANGAEA.933268

7. Appendix

820 **Figure A1: Sampling scheme of the 2018 expedition vegetation survey. Projective cover of tall shrubs and trees was estimated on a circular sample plot with a radius of 15 m (after Shevtsova et al. 2020).**

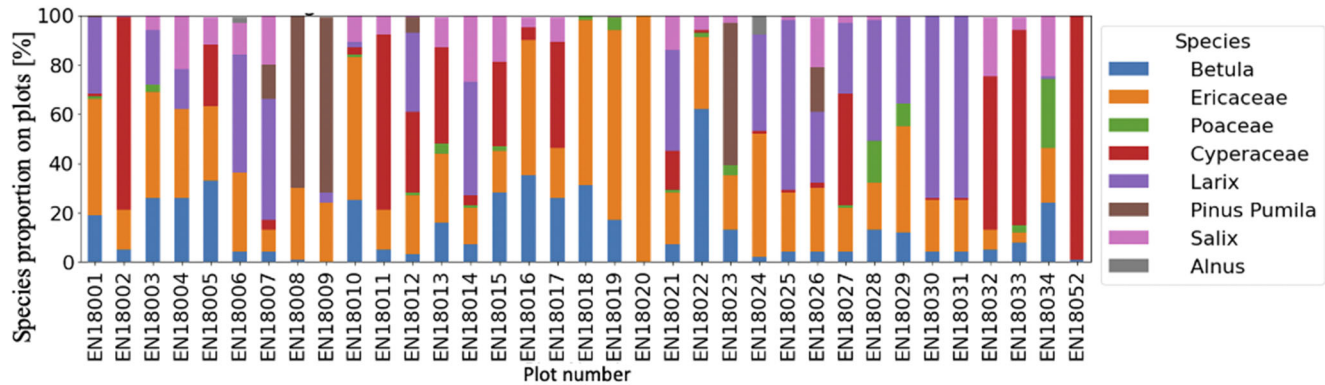



Figure A2: Percentage vegetation cover per plot in Chukotka for all recorded vegetation in the plots.

825

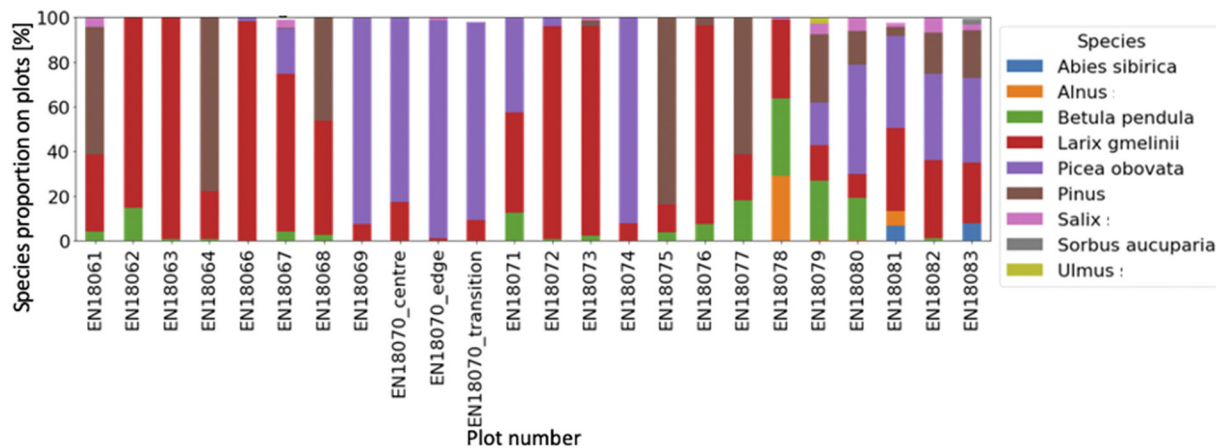



Figure A3: Percentage vegetation cover per plot in Yakutia for only large shrubs and trees (>1.3m).

830 **Figure A4: SiDroForest unmanned aerial vehicle (UAV) data acquisition and flight pattern consisting of a double grid (blue) and a circular mission (orange). The two 15 m long grid lines (red) divide the plot area into four quadrants of similar size (yellow). From Brieger et al. (2019).**

Figure A5: ~~:~~ **Mean tree height (m) per plot from fieldwork measurements**

840   **Figure ~~A7~~A6: Mean tree crown diameter (m) per plot from fieldwork measurements.**

845 **Figure ~~A8~~A7: Mean heights for trees and shrubs below 1.3 m for unmanned aerial vehicle (UAV)-derived heights (blue) and fieldwork-derived heights (orange).**

Table A1: An overview of the plots, the latitude and longitude of the central coordinates, the site name, the region (Chukotka or Yakutia), and the vegetation class labels for the 30 × 30 m S2-patches.

Plot Code	Latitude	Longitude	Site	Region	Vegetation Class
EN18001	67.39273	168.34662	Lake Ilirney	Chukotka	2
EN18002	67.38677	168.33673	Lake Ilirney	Chukotka	1
EN18003	67.39273	168.34702	Lake Ilirney	Chukotka	2
EN18004	67.39748	168.35122	Lake Ilirney	Chukotka	2
EN18005	67.41965	168.3875	Lake Ilirney	Chukotka	1
EN18007	67.40327	168.37196	Lake Ilirney	Chukotka	1
EN18008	67.40213	168.37528	Lake Ilirney	Chukotka	2
EN18009	67.40072	168.37968	Lake Ilirney	Chukotka	2
EN18011	67.40404	168.36425	Lake Ilirney	Chukotka	1
EN18012	67.40214	168.37807	Lake Ilirney	Chukotka	2
EN18013	67.40517	168.35530	Lake Ilirney	Chukotka	1
EN18014	67.39530	168.34910	Lake Ilirney	Chukotka	2
EN18015	67.42037	168.33061	Lake Ilirney	Chukotka	1
EN18016	67.42672	168.39004	Lake Ilirney	Chukotka	1
EN18017	67.43229	168.38337	Lake Ilirney	Chukotka	3
EN18018	67.456295	168.405961	Lake Ilirney	Chukotka	2
EN18019	67.457073	168.408963	Lake Ilirney	Chukotka	1
EN18020	67.459159	168.411934	Lake Ilirney	Chukotka	2
EN18021	67.392129	168.32881	Lake Ilirney	Chukotka	1
EN18022	67.401024	168.34800	Lake Ilirney	Chukotka	2
EN18023	67.399236	168.35128	Lake Ilirney	Chukotka	1
EN18024	67.370964	168.42636	Lake Ilirney	Chukotka	2
EN18025	67.367027	168.42381	Lake Ilirney	Chukotka	2
EN18026	67.396089	168.35429	Lake Ilirney	Chukotka	2
EN18027	67.393408	168.35905	Lake Ilirney	Chukotka	2

850

Plot Code	Latitude	Longitude	Site	Region	Vegetation Class
EN18028	68.467811	163.35762	Bilibino	Chukotka	1
EN18029	68.465606	163.35226	Bilibino	Chukotka	1
EN18030	68.405539	164.53273	Bilibino	Chukotka	2
EN18031	68.404918	164.54535	Bilibino	Chukotka	1
EN18032	68.404868	164.55118	Bilibino	Chukotka	2
EN18033	68.403212	164.55180	Bilibino	Chukotka	2
EN18034	68.403486	164.54804	Bilibino	Chukotka	1
EN18035	68.403166	164.59093	Bilibino	Chukotka	2
EN18051	67.802610	168.70471	Lake Rauchuagytgyn	Chukotka	1
EN18052	67.799410	168.7083	Lake Rauchuagytgyn	Chukotka	1
EN18053	67.797290	168.7107	Lake Rauchuagytgyn	Chukotka	1
EN18054	67.797660	168.6904	Lake Rauchuagytgyn	Chukotka	1
EN18055	67.791030	168.682500	Lake Rauchuagytgyn	Chukotka	3

Plot Code	Latitude	Longitude	Site	Region	Vegetation Class
EN18061	62.076376	129.618586	Yakutsk	Central Yakutia	6
EN18062	62.179065	127.805796	Magaras	Central Yakutia	10
EN18063	63.776636	122.501003	Vilnuyi	Central Yakutia	10
EN18064	63.814594	122.209683	Vilnuyi	Central Yakutia	4
EN18065	63.795223	122.443715	Vilnuyi	Central Yakutia	9
EN18066	63.7971189	122.438071	Vilnuyi	Central Yakutia	9
EN18067	63.076368	117.975342	Nyurba	Central Yakutia	8
EN18068	63.074232	117.98207	Nyurba	Central Yakutia	7
EN18069	63.173288	118.132507	Nyurba	Central Yakutia	11
EN18070	63.082914	117.984905	Nyurba	Central Yakutia	11
EN18071	62.225093	116.275603	Suntar West	Central Yakutia	8
EN18072	62.199571	117.379125	Suntar	Central Yakutia	10
EN18073	62.188712	117.409917	Suntar	Central Yakutia	9
EN18074	62.215192	117.021599	Suntar	Central Yakutia	11
EN18075	62.696991	113.676535	Mirny	Central Yakutia	7
EN18076	62.70089	113.67341	Mirny	Central Yakutia	10
EN18077	61.892568	114.288623	Mirny-Lensk	Central Yakutia	5
EN18078	61.575058	114.29995	Mirny-Lensk	Central Yakutia	10
EN18079	59.974919	112.958985	Lake Khamra	Central Yakutia	8
EN18080	59.977106	112.961379	Lake Khamra	Central Yakutia	7
EN18081	59.970583	112.987096	Lake Khamra	Central Yakutia	8
EN18082	59.97764	112.98218	Lake Khamra	Central Yakutia	7
EN18083	59.974714	113.002874	Lake Khamra	Central Yakutia	7

855 1 = Graminoid tundra; 2= Forest tundra and shrub tundra; 3= Prostrate herb tundra; 4= Open canopy pine with lichen; 5= Open canopy pine;6= Closed canopy pine; 7= Open canopy mixed forest; 8= Closed canopy mixed forest; 9 = Open canopy Larch; 10= Closed canopy Larch; 11= Closed canopy spruce

860

Table A2: Example of common objects in context (COCO) style annotation labels for the masks (1) and images (2).

1: "masks": {"images/00000000.jpg": {"mask": "masks/00000000.png", "color_categories": {"(255, 0, 0)": {"category": "larch", "super_category": "tree"}

2: {"info": {"description": "SiDroForest: Synthetic Tree Crowns", "url": "http://immersivelimit.com/datasets/test", "version": "1", "year": 2021, "contributor": "Femke van Geffen", "date_created": "12/04/2021"}"00000000.jpg", "width": 448, "height": 448, "id": 0}

Table A3: Overview of Sentinel-2 spectral bands, spatial resolution, and the central wavelength.

Sentinel-2 Bands	Central Wavelength (nm)	Pixel Length (m)
Band 1- Coastal aerosol	443	60
Band 2- Blue	490	10
Band 3- Green	560	10
Band 4- Red	665	10
Band 5- Vegetation Red Edge	705	20
Band 6- Vegetation Red Edge	740	20
Band 7- Vegetation Red Edge	783	20
Band 8- NIR	842	10
Band 8A- Vegetation Red Edge	865	20
Band 9- Water vapour	945	60
Band 10- SWIR-Cirrus	1,375	60
Band 11- SWIR-1	1,610	20
Band 12- SWIR-2	2,190	20

Table A4: Screenshot of the *crowns_polygon* shapefile attribute table for plot EN18077 as an example. Height: tree height in metres as identified with the tree top finding algorithm, crownAr: area of the tree crown in square metres, CrwnDmt: simplification of the crown diameter in metres assuming a circular crown, orgHght: maximum height value in metres recorded in the canopy height model (CHM) under the total crown polygon.

	layer	height	winRads	crownAr	crwnDmt	orgHght
1	1.000000000000...	3.9404476708941...	0.97732145190239	2.170800000041...	1.6625126777775...	4.649345874786...
2	1.000000000000...	0.476775185929...	0.821454883366...	0.388799999697...	0.703587616866...	0.78668737411499
3	1.000000000000...	0.615317266848...	0.827689277008...	0.194400000005...	0.497511575246...	1.480337023735...
4	1.000000000000...	4.726099067264...	1.012674458026...	2.624400000362...	1.827974250820...	5.248609066009...
5	1.000000000000...	1.785895599259...	0.880365301966...	0.323999999974...	0.642284681790...	3.317377567291...
6	1.000000000000...	2.465017358462...	0.910925781130...	0.939600000241...	1.093771400494...	3.295063734054...

8. Author contributions

Femke van Geffen FvG is the leading author of this manuscript and of most of the related data publications in the PANGAEA data repository. FvG wrote the manuscript together with Stefan Kruse SK, Birgit Heim BH, and Ulrike Herzschuh UH. Luidmila A Pestryakova LP and Evgenij S Zakharov EZ ~~organised~~ organized and facilitated the data collection for the expedition in Siberia and took part in the ~~data collection.~~field work. The majority of vegetation related ground fieldwork was performed by Iuliia A. Shevtsova IS, Luise Schulte LS, Simone Stünzi SS, Elena I. Troeva ET, Nadine Bernhardt NB, UH, Frederik Brieger FB and SK. SK and FB undertook the data processing and together with assistants constructed the products for the orthomosaics dataset, including the point-cloud products. Rongwei Geng RG and FvG supplemented the orthomosaics dataset and assigned vegetation labels to the plots based on vegetation classes by IS. BH and Bringfried Pflug BP processed the Sentinel-2 dataset. FvG created the synthetics dataset and identified the individuals in the individual labelled trees dataset. FvG cleaned, compiled, and constructed all four final datasets under supervision of SK as lead scientist on this project.

9. Competing Interests

The authors declare that they have no conflict of interest.

10. Acknowledgements

The SiDroForest data collection was created as part of a PhD project within the context of the HEIBRiDS graduate school. This research and particularly the field work has received funding from the ERC consolidator grant Glacial Legacy of Ulrike Herzschuh (grant no. 772852).

We thank our Russian and German colleagues from the joint Russian–German expedition 2018 for support in the field. Special thanks to the staff of the BIOM-laboratory in Yakutsk for their great overall support and scientific contributions. We thank Guido Grosse and Thomas Laepple (AWI) who provided us with computational resources for the point-cloud reconstruction from UAV-based data.

We greatly thank the three anonymous reviewers, everyone at ESSD who supported us and our topical editor Yuyu Zhou for helping us improve the manuscript.

11. References

905 Abdi, A.M.: Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GIScience & Remote Sensing, 57(1), 1-20, doi.org/10.1080/15481603.2019.1650447, 2020.

ABoVE Science Definition Team: A Concise Experiment Plan for the Arctic-Boreal Vulnerability Experiment. ORNL DAAC, Oak Ridge, Tennessee, USA, doi.org/10.3334/ORNLDAAC/1617, 2014.

910

Argolo dos Santos, R., Mantovani, E., Filgueiras, R., Fernandes Filho, I. E., Barbosa da Silva, A. C., and Venancio, L. P.: Actual Evapotranspiration and Biomass of Maize from a Red-Green-Near-Infrared (RGNIR) Sensor on Board an Unmanned Aerial Vehicle (UAV), Water, 12(9), 2359, doi.org/10.3390/w12092359, 2020.

915 Astola, H., Seitsonen, L., Halme, E., Molinier, M. and Lönnqvist, A.: Deep Neural Networks with Transfer Learning for Forest Variable Estimation Using Sentinel-2 Imagery in Boreal Forest, Remote Sensing, 13(12), 2392, doi.org/10.3390/rs13122392, 2021.

Beamish, A., Raynolds, M. K., Epstein, H., Frost, G. V., Macander, M. J., Bergstedt, H., Bartsch, A., Kruse, S., Miles, V., 920 Tanis, C. M, Heim, B., Fuchs, M., Chabrillat, S., Shevtsova, I., Verdonen, M., and Wagner, J.: Recent trends and remaining challenges for optical remote sensing of Arctic tundra vegetation: A review and outlook, Remote Sensing of Environment, 246, 111872, doi.org/10.1016/j.rse.2020.111872, 2020.

Bonan, G. B.: Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests, Science, 925 320(5882):1444-9., doi.org/10.1126/science.1155121, 2008.

Braga, J. R. G., Peripato, V., Dalagnol, R. P., Ferreira, M., Tarabalka, Y. O. C., Aragão L. E. F., de Campos Velho, H., Shiguemori, E. H., and Wagner, F. H.: Tree Crown Delineation Algorithm Based on a Convolutional Neural Network, Remote Sensing, 12(8),1288, doi.org/10.3390/rs12081288, 2020.

930

Brieger, F., Herzschuh, U., Pestryakova, L. A., Bookhagen, B., Zakharov, E. S., and Kruse, S.: Advances in the derivation of Northeast Siberian forest metrics using high-resolution UAV-based photogrammetric point cloud, Remote Sensing, 11(12), 1447, doi.org/10.3390/rs11121447, 2019.

935 CAVM Team: Circumpolar arctic vegetation map (1:7,500,000 scale). Conservation of Arctic Flora and Fauna (CAFF) Map No.1. Anchorage, AK: US Fish and Wildlife Service, 2003.

Chave, J., Davies, S. J., Phillips, O.L., Lewis, S.L., Sist, P., Schepaschenko, D., Armston, J., Baker, T. R., Coomes, D., Disney, M., Duncanson, L., Hérault, B., Labrière, N., Meyer, V., Réjou-Méchain, M., Scipal, K. and Saatchi, S.: Ground Data are Essential for Biomass Remote Sensing Missions, Surv Geophys, 40, 863–880, doi.org/10.1007/s10712-019-09528-w, 2019.

CloudCompare (version 2.10) [GPL software]. (2022). Retrieved from http://www.cloudcompare.org/

Conard, S. G., Sukhinin, A. I., Stocks, B. J., Cahoon, D. R., Davidenko, E. P., and Ivanova, G. A.: Determining effects of area burned and fire severity on carbon cycling and emissions in Siberia, Climatic Change, 55(1-2), 197-211, doi.org/10.1023/A:1020207710195, 2002.

Copernicus: Copernicus Digital Elevation Model Product Handbook: https://spacedata. copernicus. eu/web/cscda/datasetdetails?articleId=394198 accessed on 21/01/2021.

European Space Agency (ESA), Sentinel-2 S2MPC, Sen2Cor Software Release Note, S2-PDGS-MPC-L2A-SRN-V2.9.0, November 30, 2020, Sen2Cor v2.9 – STEP (esa. int), accessed on 06/05/2021.

European Space Agency (ESA), Sentinel-2 S2 MPC Level-2A Algorithm Theoretical Basis Document (ATBD), S2-PDGS-MPC-ATBD-L2A, Issue 2.9, 2021 https://sentinel.esa.int/web/sentinel/userguides/sentinel-2-msi/document-library accessed on 06/05/2021.

European Space Agency (ESA), Sentinel-2 User Handbook, Issue 1.2, 64 pp, 2015.

Fraser, R. H., Olthof, I., Lantz, T. C., and Schmitt, C.: UAV photogrammetry for mapping vegetation in the low-Arctic.,Arctic Science, 2(3), 79–102, doi.org/10.1139/as-2016-0008, 2016. He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), 2980–2988, Venice, Italy, Oct 22 –29, doi:10.1109/ICCV.2017.322, 2017.

Hao, Z., Lin, L., Post, C. J., Mikhailova, E. A., Li, M., Chen, Y., and Liu, J.: Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (Mask R-CNN), ISPRS Journ Photogrammetry and Remote Sensing, 178, 112-123, doi.org/10.1016/j.isprsjprs.2021.06.003, 2021.

Herzschuh, U.: Legacy of the Last Glacial on the present-day distribution of deciduous versus evergreen boreal forest. Global Ecology and Biogeography, 29(2), 198–206, doi.org/10.1111/geb.13018, 2020.

The GIMP Development Team: GIMP. Retrieved from https://www. gimp. org, 2019.

975 Gülei, S.: The determination of some stand parameters using SfM-based spatial 3D point cloud in forestry studies: an analysis of data production in pure coniferous young forest stand. Environmental Monitoring and Assessment, 191, 495, doi.org/10.1007/s10661-019-7628-4, 2019.

Jensen, J. L. R., and Mathews, A. J.: Assessment of image-based point cloud products to generate a bare earth surface and estimate canopy heights in a woodland ecosystem, Remote Sensing, 8, 50, doi.org/10.3390/rs8010050, 2016.

980

Kelley, A.: Complete Guide to Creating COCO Datasets, GitHub repository: https://github. com/akTwelve/cocosynth, 2019.

Kruse, S., Bolshiyanov, D., Grigoriev, M. N., Morgenstern, A., Pestryakova, L., Tsibizov, L. and Udke, A.: Russian-German Cooperation: Expeditions to Siberia in 2018, Reports on Polar and Marine Research. Alfred Wegener Institute for Polar and 985 Marine Research, 734, 257 p., doi:10.2312/BzPM_0734_2019, 2019a.

Kruse, S., Herzschuh, U., Stünzi, S., Vyse, S., and Zakharov, E: Sampling mixed sspecies boreal forests affected by disturbances and mountain lake mountain lake and alas lake coring in Central Yakutia. In Kruse, S., Bolshiyanov, D., Grigoriev, M. N., Morgenstern, A., Pestryakova, L., Tsibizov, L. and Udke, A. (Eds.), Russian-German Cooperation: 990 Expeditions to Siberia in 2018, Reports on polar and marine research (148–153). Bremerhaven: Alfred Wegener Institute for Polar and Marine Research. doi:10.2312/BzPM_0734_2019, 2019b.

Kruse, S., Herzschuh, U., Schulte, L., Stuenzi, S. M., Brieger, F., Zakharov, E. S. and Pestryakova, L. A.: Forest inventories on circular plots on the expedition Chukotka 2018, NE Russia, PANGAEA, doi.org/10.1594/PANGAEA.923638, 2020a.

995

Kruse, S., Kolmogorov, A. I., Pestryakova, L. A., and Herzschuh, U.: Long-lived larch clones may conserve adaptations that could restrict treeline migration in northern Siberia. Ecology and Evolution, 10(18), 10017–10030, doi.org/10.1002/ece3.6660, 2020b.

1000 Kruse, S., Farkas, L., Brieger, F., Geng, R., Heim, B., Pestryakova, L.A., Herzschuh, U. and van Geffen, F.: SiDroForest: Orthomosaics, SfM point clouds and products from aerial image data of expedition vegetation plots in 2018 in Central Yakutia and Chukotka, Siberia, PANGAEA, doi.pangaea.de/10.1594/PANGAEA.933263, 2021.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., and Zitnick, C. L.: Microsoft COCO: Common objects in context. In European Conference on Computer Vision, 740-755. Springer, Cham, September 6–12, 2014.

Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S. and Wilson, B. T.: Use of Remote Sensing Data to Improve the Efficiency of National Forest Inventories: A Case Study from the United States National Forest Inventory. Forests, 11(12), 1364, doi.org/10.3390/f11121364, 2020.

Long, Y., Xia, G.S.,Li, S., Yang, W., Yang, M., Y., Zhu X.X., Zhang, L., and Li D., DiRS: On Creating Benchmark Datasets for Remote Sensing Image Interpretation, CoRR,abs/2006.12485, https://arxiv.org/abs/2006.12485, 2020.

MacDonald, G. M., Kremenetski, K. V. and Beilman, D. W.: Climate change and the northern Russian treeline zone. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1501), 2283–2299, doi.org/10.1098/rstb.2007.2200, 2007.

Maguire, A. J., Eitel, J., Vierling, L., Boelman, N., Griffin, K., Jennewein, J. S. and Jensen, J. E.: ABoVE: Terrestrial Lidar Scanning Forest-Tundra Ecotone, Brooks Range, Alaska, 2016. ORNL DAAC, Oak Ridge, Tennessee, USA. doi.org/10.3334/ORNLDAAC/1782, 2020.

Mamet, S. D., Brown, C. D., Trant, A. J., and Laroque, C. P.: Shifting global *Larix* distributions: Northern expansion and southern retraction as species respond to changing climate. Journal of Biogeography, 46(1), 30–44. doi:10.1111/jbi.13465, 2019.

Miesner, Timon; Herzschuh, Ulrike; Pestryakova, Luidmila A; Wieczorek, Mareike; Kolmogorov, Alexei; Heim, Birgit; Zakharov, Evgenii S; Shevtsova, Iuliia; Epp, Laura Saskia; Niemeyer, Bastian; Jacobsen, Inga; Schröder, Julius; Trense, Darjona; Schnabel, Ellen; Schreiber, Xenia; Bernhardt, Nadine; Stuenzi, Simone Maria; Brieger, Frederic; Schulte, Luise; Smirnikov, Viktor; Gloy, Josias; von Hippel, Barbara; Jackisch, Robert; Kruse, Stefan (2022): Tree data set from forest inventories in north-eastern Siberia. PANGAEA, https://doi.pangaea.de/10.1594/PANGAEA.943547Miesner, Timon; Herzschuh, Ulrike; Pestryakova, Luidmila A; Wieczorek, Mareike; Kolmogorov, Alexei; Heim, Birgit; Zakharov, Evgenii S; Shevtsova, Iuliia; Epp, Laura Saskia; Niemeyer, Bastian; Jacobsen, Inga; Schröder, Julius; Trense, Darjona; Schnabel, Ellen; Schreiber, Xenia; Bernhardt, Nadine; Stuenzi, Simone Maria; Brieger, Frederic; Schulte, Luise; Smirnikov, Viktor; Gloy, Josias; von Hippel, Barbara; Jackisch, Robert; Kruse, Stefan (2022): Tree data set from forest inventories in north-eastern Siberia. PANGAEA, https://doi.pangaea.de/10.1594/PANGAEA.943547

Montesano, P. M., Neigh, C. S., Sexton, J., Feng, M., Channan, S., Ranson, K. J. and Townshend, J. R.: Calibration and validation of Landsat tree cover in the taiga−tundra ecotone. Remote Sensing, 8(7), 551, doi.org/10.3390/rs8070551, 2016.

1040 Montesano, P. M., Nelson, R. F., Dubayah, R. O., Sun, G., Cook, B. D., Ranson, K. J. R., and Kharuk, V.: The uncertainty of biomass estimates from LiDAR and SAR across a boreal forest structure gradient. Remote Sensing of Environment, 154, 398–407, doi.org/ 10.1016/j.rse.2014.01.027, 2014.

Neuville, R., Bates, J. S. and Jonard, F.: Estimating forest structure from UAV-mounted LiDAR point cloud using machine
1045 learning, Remote Sensing, 13(3), 352, doi.org/10.3390/rs13030352, 2021.

Panagiotidis, D., Abdollahnejad, A., Surový, P., and Chiteculo, V.: Determining tree height and crown diameter from high-resolution UAV imagery, International Journal of Remote Sensing, 38, 2392–2410, doi.org/10.1080/01431161.2016.1264028, 2017.

1050
R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2020.

Rees, W. G., Hofgaard, A., Boudreau, S., Cairns, D. M., Harper, K., Mamet, S. and Tutubalina, O.: Is subarctic forest advance
1055 able to keep pace with climate change? Global Change Biology, 26(7), 3965–3977, doi.org/10.1111/gcb.15113, 2020.

Schepaschenko, D., Moltchanova, E., Fedorov, S., Karminov, V., Ontikov, P., Santoro, M., See, L., Kositsyn, V., Shvidenko, A., Romanovskaya, A., Korotkov, V., Lesiv, M., Bartalev, S., Fritz, S., Shchepashchenko, M. and Kraxner, F.: Russian forest sequesters substantially more carbon than previously reported, Scientific Reports, 11, 12825, doi.org/10.1038/s41598-021-
1060 92152-9, 2021.

Schepaschenko D., Chave J., Phillips O.L., Lewis S.L., Davies S.J., Réjou-Méchain M., Sist P. and Scipal K.: The Forest Observation System, building a global reference dataset for remote sensing of forest biomass, Scientific Data, 6(1), 198, doi.org/10.1038/s41597-019-0196-1, 2019.

1065
Schepaschenko, D., Shvidenko, A., Usoltsev, V., Lakyda, P., Luo, Y., Vasylyshyn, R., Lakyda, I., Myklush, Y., See, L., McCallum, I., Fritz, S., Kraxner, F and Obersteiner, M.: A dataset of forest biomass structure for Eurasia, Scientific Data, 4, 170070. doi.org/10.1038/sdata.2017.70, 2017.

1070   Shevtsova, I., Herzschuh, U., Heim, B., Kruse, S., Schröder, J., Troeva, E., Pestryakova, L. A., and Zakharov, E. S.: Foliage projective cover of 57 vegetation sites of central Chukotka from 2016, PANGAEA, doi.org/10.1594/PANGAEA. 908570, 2019.

Shevtsova, I., Heim, B., Kruse, S., Schröder, J., Troeva, E., Pestryakova, L. A., Zakharov, E. S. and Herzschuh, U.: Strong 1075   shrub expansion in tundra-taiga, tree infilling in taiga and stable tundra in central Chukotka (north-eastern Siberia) between 2000 and 2017. Environmental Research Letters, 15(9), doi.org/10.1088/1748-9326/ab9059, 2020a.

Shevtsova, I., Kruse, S., Herzschuh, U., Schulte, L., Brieger, F., Stuenzi, S. M., Heim, B., Troeva, E. I., Pestryakova, L. A. and Zakharov, E. S. (2020 b): Foliage projective cover of 40 vegetation sites of central Chukotka from 2018. PANGAEA, 1080   doi.pangaea.de/10.1594/PANGAEA.923664, 2020b.

Shevtsova, I., Kruse, S., Herzschuh, U., Schulte, L., Brieger, F., Stuenzi, S. M., Heim, B., Troeva, E. I., Pestryakova, L. A. and Zakharov, E. S.: Total above-ground biomass of 39 vegetation sites of central Chukotka from 2018. PANGAEA, doi.pangaea.de/10.1594/PANGAEA.923719, 2020c.

1085

Shevtsova, I., Herzschuh, U., Heim, B., Schulte, L., Stünzi, S., Pestryakova, L. A., Zakharov, E. S. and Kruse, S.: Recent above-ground biomass changes in central Chukotka (Russian Far East) using field sampling and Landsat satellite data. Biogeosciences, 18, 3343–3366, doi.org/10.5194/bg-18-3343-2021, 2021.

1090   Simard, M., Pinto, N., Fisher, J.B. and Baccini, A.: Mapping Forest canopy height globally with spaceborne lidar, Journal Geophys. Res, 116, G04021, doi:10.1029/2011JG001708, 2011.

Sumbul. G., Charfuelan, M., Demir, B. and Markl. V.: BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. IEEE International Geoscience and Remote Sensing Symposium, 5901–5904, Yokohama, Japan, July 1095   2–August 2, 2019.

Thanh Noi, P., and Kappas, M.: Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery, Sensors, 18, doi.org/10.3390/s18010018, 2018.

1100   Troeva, E. I., Isaev, A. P., Cherosov, M. M., and Karpov, N. S. (Eds.). The Far North: Plant Biodiversity and Ecology of Yakutia (Vol. 3). Springer Science & Business Media, 2010.

van Geffen, F., Brieger, F., Pestryakova, L. A., Herzschuh, U. and Kruse, S.: SiDroForest: Synthetic Siberian Larch Tree Crown Dataset of 10 000 instances in the Microsoft's Common Objects in Context dataset (coco) format. PANGAEA, doi.org/10.1594/PANGAEA.932795, 2021a.

van Geffen, F., Geng, R., Pflug, B., Kruse, S., Pestryakova, L. A, Herzschuh, U. and Heim, B.: SiDroForest: Sentinel-2 Level-2 Bottom of Atmosphere labelled image patches with seasonal information for Central Yakutia and Chukotka vegetation plots (Siberia, Russia). PANGAEA, doi.pangaea.de/10.1594/PANGAEA.933268, 2021b.

van Geffen, F., Schulte, L., Geng, R., Heim, B., Pestryakova, L. A., Herzschuh, U. and Kruse, S.: SiDroForest: Individual-labelled trees acquired during the fieldwork expeditions that took place in 2018 in Central Yakutia and Chukotka, Siberia. PANGAEA, doi.pangaea.de/10.1594/PANGAEA.932821, 2021c.

Wang, D., Wan, B., Liu, J., Su, Y., Guo, Q., Qiu, P. and Wu, X.: Estimating aboveground biomass of the mangrove forests on northeast Hainan Island in China using an upscaling method from field plots, UAV-LiDAR data and Sentinel-2 imagery, International Journal of Applied Earth Observation and Geoinformation, 85, 101986, doi.org/10.1016/j.jag.2019.101986, 2020.

Pete Warden's Blog, How Many Images Do You Need to Train A Neural Network? https://petewarden. com/2017/12/14/how-many-images-do-you-need-to-train-a-neural-network/, accessed 10-05-2021.

Weinstein, B. G., Marconi, S, Bohlman, S., Zare, A., and White, E.: Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. Remote Sensing, 11, 1309, doi.org/10.3390/rs11111309, 2019.

Weinstein, B. G., Marconi, S, Bohlman, S., Zare, A., Singh, A., Graves, S. J. and White, E.: A remote sensing derived data set of 100 million individual tree crowns for the National Ecological Observatory Network, eLife 10:e62922, doi.org/10.7554/eLife.62922, 2021

Wittke, S., Yu, X., Karjalainen, M., Hyyppä, J. and Puttonen, E.: Comparison of two-dimensional multitemporal Sentinel-2 data with three-dimensional remote sensing data sources for forest inventory parameter estimation over a boreal forest. International Journal of Applied Earth Observation and Geoinformation, 76, 167–178, doi.org/10.1016/j.jag.2018.11.009, 2019.

QGIS 3. 10 Development Team, QGIS Geographic Information System. Open-Source Geospatial Foundation Project. http://qgis.osgeo.org, 2019.

Xiao, K., Engstrom, L., Ilyas, A. and Madry, A.: Noise or signal: The role of image backgrounds in object recognition. arXiv, arXiv:2006.09994, 2020.

1140 Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X. and Yan, G.: An Easy-to-Use Airborne LiDAR Data Filtering Method Based on Cloth Simulation. Remote Sensing, 8, 501, doi.org/10.3390/rs8060501, 2016.

Zhao, T., Yang, Y., Niu, H., Wang, D. and Chen, Y.: Comparing U-Net convolutional network with mask R-CNN in the performances of pomegranate tree canopy segmentation. In Multispectral, Hyperspectral, and Ultra-spectral Remote Sensing
1145 Technology, Techniques and Applications VII, 10780, 107801J, International Society for Optics and Photonics, 2018.