Earth Syst. Sci. Data Discuss., author comment AC2
https://doi.org/10.5194/essd-2021-281-AC2, 2022

Reply on RC2

Femke van Geffen et al.

Author comment on "SiDroForest: a comprehensive forest inventory of Siberian boreal forest investigations including drone-based point clouds, individually labeled trees, synthetically generated tree crowns, and Sentinel-2 labeled image patches" by Femke van Geffen et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-281-AC2, 2022

Dear Reviewer, Thank you very much for your time and effort. The review we received is included below in regular text and the changes that are included in the manuscript and the responses are written below each point in bold and cursive starting with the word response.I, Femke van Geffen would like to, also on behalf of my team, thank the reviewers for their time and detailed comments on the manuscript and dataset. The comments greatly improve the work and are well appreciated. Kind regards, Femke van Geffen

Review:

The datasets presented seem to be quite useful for an important region where data is scarce for research. I believe these field survey data would be helpful for both small-scale and large-scale studies in understanding the circumboreal region. A few comments listed below for consideration.

Q1: Is the data set significant – unique, useful, and complete?

I think the data is unique and complete, though the authors can mention more about the usefulness of these datasets. For example, some scientific questions that these datasets can help address? What specific applications can be expected from these datasets? O***r how these datasets may help change the research landscape over time?***

Response: We highlight the usefulness of the data set with first information in the introduction and describe its uniqueness and suggestions to how to use the data in the discussion. The purpose of each data type is also mentioned in the product data sections. I have clarified the usefulness more in the current version of the paper, please see lines below:

1-The satellite data is processed to a high degree and prepared for machine learning tasks with labeled patches.

L56: "The fourth 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. "

2- The orthoimages provide geographical information on individual trees and shrubs that were recorded on each plot. These individuals have information such as tree height and species and can be used to calculate the projected cover of each species and give general insights into the vegetation dynamics.

L153: "Individual labelled trees surveyed during the fieldwork, including information on height, tree crown, and species. 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- The point clouds provide information on the 3D structure of the forest at each location. The point cloud products such as Canopy Height Model can be used to extract the height of all the trees on the plots, not just the recorded ones from fieldwork.

L36-41 "The first dataset provides Unmanned Arial Vehicle (UAV)-borne data products covering the vegetation plots surveyed during fieldwork (Kruse et al., 2021, https://doi.pangaea.de/10.1594/PANGAEA.933263). The dataset includes structure from motion (SfM) point clouds 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). The point cloud products provide information on the three-dimensional (3D) structure of the forest at each plot. "

4- The synthetic dataset provides 10000 images that were generated using an algorithm. The purpose for this is that they can now be used as input for a neural network which needs many images as input for training.

L51-55: "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). As machine learning algorithms need a large dataset to train on, the synthetic dataset was specifically created to be used for machine learning algorithms to detect Siberian larch species."

Q2: Is the data set itself of high quality?

Yes, the data is of high quality.

Presentation quality

I do not think the authors pointed out specific/suitable software for simple visualization and analysis as I typically need to install some software to view the data.

Response: Yes we did not specifically name suitable software. The Sentinel-2 and UAV orthoimages data can also be opened and viewed in any open source and commercial GIS and Remote Sensing software and for the point clouds AgiSoft or Cloudcompare can be used. All data can be loaded in R. Suggestions to software have been added to chapter 2 data and methods and to specific sections.

L165-170: " 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, 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 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."

Specific comments

(1) The abstract can be improved by why these datasets are needed.

Response: We enhanced the abstract text in making our statements clearer. Limited data on boreal forest structure are available, especially for this region (Eastern Siberia) and in quality with labels that can be used for machine learning purposes. Please find all the changes in the revised manuscript with tracked changes.

(2) The abstract seems too long and I would suggest further summarizing it and highlighting the datasets in brief.

Response: We have taken this suggestion into consideration. Please see new document with the shortened, edited abstract.

Abstract

The SiDroForest data collection is an attempt to remedy the scarcity of forest structure data in the circumboreal region by providing adjusted and labelled tree level and vegetation plot level data for machine learning and upscaling purposes. We present datasets of vegetation composition and tree and plot level forest structure for two important 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 data collection consists of four datasets that contain different complementary data types that together support in-depth analyses from different perspectives of Siberian Forest plot data for multi-purpose applications.

i) The first dataset provides Unmanned Arial Vehicle (UAV)-borne data products covering the vegetation plots surveyed during fieldwork (Kruse et al., 2021, https://doi.pangaea.de/10.1594/PANGAEA.933263). The dataset includes structure from motion (SfM) point clouds 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). The point cloud products provide information on the three-dimensional (3D)

structure of the forest at each plot.

ii) The second dataset contains spatial data in the form of point and polygon shape files of 872 labelled individual trees and shrubs that were recorded during fieldwork at the same vegetation plots (van Geffen et al., 2021c, https://doi.pangaea.de/10.1594/PANGAEA.932821). The dataset contains information on tree height, crown diameter, and species type. These tree- and shrub-individual labelled point and polygon shape files were generated on top of the UAV RGB orthoimages. The individual 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 current state of individual trees and 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). As machine learning algorithms need a large dataset to train on, the synthetic dataset was specifically created to be used for machine learning algorithms to detect Siberian larch species.

iv) The fourth 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. The detailed 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, the third and fourth datasets are prepared 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 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 SiDroForest data collection adds unique insights into remote hard to reach circumboreal forest regions.

(3) I am not convinced how this field level data can help studies in this region. Only for optimized data containing annotated vegetation categories. What is the motivation for doing these machine learning applications? DO we need to look at data from a temporal perspective?

Response: The idea is that more detailed vegetation information can be extracted from this dataset per plot. Also, there is information on the vegetation structure of each plot in the SfM point clouds.

The machine learning applications for tree crowns created a set with n = 10 0000 crowns that can be used to automatically detect larches in UAV RGB images of forest plots.

It depends on the application if we should include temporal perspectives. The temporal information in the current dataset is the three seasons of satellite data included in the Sentinel-2 part of the dataset. At the current state this data collection is now an important and unique openly published data collection, a snapshot for the time window around 2018 for the taiga-tundra transition zone in Chukotka and the evergreen-summergreen transition zone in Central Yakutia. if in the future again plots of the same region will be assessed, change can be detected. This time stamp in 2018 can also support the assessment of land surface satellite products of the second late decade of this century compared to later satellite acquisitions in future and earlier satellite acquisitions in the past.

(4) The introduction should be shortened as well to highlight the contribution of datasets to the scientific literature or specific questions/challenges to be addressed

Response: Thank you for your feedback. We have rewritten the introduction and shortened it considerably (please see the track change file).

(5) Figure 1 suggested sampled sites are very limited. Can the authors justify why these sites were selected rather than other sites?

Response: The sites were selected based on the two important transition zones in these areas. The summergreen-evergreen and the Tundra-Taiga transition zone. The sites included in this dataset are from one very extensive, 2 month long expedition in 2018 where the team traveled from the City of Yakutsk following the only possible summer road in the direction of Mirny and then Lensk. We then followed the bioclimatic gradient leading to the easternmost larch dominated forests and westernmost mixed-species forests in the Lake Khamra region, this expedition part that assessed xxx forest plots took 4 weeks and was very efficient. From the Tundra-Taiga transition zone there is only summer road close to River Kolyma to Bilibino and then the expedition team needed to take helicopters to the mountainous tundra and forest tundra regions of Lake Rauchagytgyn and Lake Ilirney. This part of the expedition also took 4 weeks and was very time intensive with a lot of plots assessed (n= x). The sites were selected beforehand based on satellite imagery using the NDVI maximum summer values and change as well as disturbance products like Hansen et al. forest loss/gain for selecting various accessible sites from the road. This information has now also been added to the paper to make it more clear why we selected these sites. The 2018 two month-long expedition was a very extensive expedition covering a variety of environments with an unusual high number of field plots with successful UAV acquisitions.

(6) I am not sure whether the codes for generating these datasets are available as it is not mentioned explicitly in the manuscript.

Response: The code to generate the Synthetic tree crowns is referred to in the text, Kelley (2019) and the link to the github with the code is in the reference list: GitHub repository: https://github. com/akTwelve/cocosynth, 2019. This code can be downloaded and used to make the synthetic dataset. All code is in

Python.

The individual trees and the Sentinel-2 were generated by hand and with free remote sensing and GIS (QGIS) software now put in in chapter 2. The SfM point clouds were made with Agisoft software, described in chapter 2. The tree crown detection was implemented in R which we describe and refer to Brieger et al. (2019) where it is described in much detail.