ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Unraveling boreal forest composition and drivers across scales in eastern Siberia

To cite this article: Léa Enguehard *et al* 2024 *Environ. Res. Lett.* **19** 074050

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
1 March 2024

REVISED
16 May 2024

ACCEPTED FOR PUBLICATION
12 June 2024

PUBLISHED
25 June 2024

LETTER

Unraveling boreal forest composition and drivers across scales in eastern Siberia

Léa Enguehard[1,*], Stefan Kruse[1], Birgit Heim[1], Ingmar Nitze[2], Nicola Falco[3], Begum Demir[4] and Ulrike Herzschuh[1,5,6]

[1] Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam, Germany
[2] Permafrost Research, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam, Germany
[3] Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, United States of America
[4] Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany
[5] Institute for Environmental Science and Geography, University of Potsdam, Potsdam, Germany
[6] Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: lea.enguehard@awi.de

Keywords: boreal forest, Eastern Siberia, random forest, generalized linear model, Sentinel-2, Google Earth Engine, WorldClim

Abstract

The Siberian boreal forest is the largest continuous forest region on Earth and plays a crucial role in regulating global climate. However, the distribution and environmental processes behind this ecosystem are still not well understood. Here, we first develop Sentinel-2-based classified maps to show forest-type distribution in five regions along a southwest-northeast transect in eastern Siberia. Then, we constrain the environmental factors of the forest-type distribution based on a multivariate analysis of bioclimatic variables, topography, and ground-surface temperatur at the local and regional scales. Furthermore, we identify potential versus realized forest-type niches and their applicability to other sites. Our results show that mean annual temperature and mean summer and winter temperatures are the most influential predictors of forest-type distribution. Furthermore, we show that topography, specifically slope, provides an additional but smaller impact at the local scale. We find that the filling of climatic environmental niches by forest types decreases with geographic distance, but that the filling of topographic niches varies from one site to another. Our findings suggest that boreal forests in eastern Siberia are driven by current climate and topographical factors, but that there remains a portion of the variability that cannot be fully accounted for by these factors alone. While we hypothesize that this unexplained variance may be linked to legacies of the Late Glacial, further evidence is needed to substantiate this claim. Such results are crucial to understanding and predicting the response of boreal forests to ongoing climate change and rising temperatures.

1. Introduction

The boreal forest, one of the largest biomes on Earth representing roughly one-third of the world's total forested area, plays a key role in providing essential ecosystem services [1–4]. While evergreen coniferous trees dominate boreal forests of North America and western Eurasia, deciduous conifer forests of larch species prevail in eastern Eurasia [5]. This distribution is observed despite the similarity of climatic conditions across the boreal areas of the Northern Hemisphere, where environmental niches overlap [6]. The Siberian boreal forest is the largest continuous forested region on Earth and features ecosystem services that differ markedly from evergreen forests (e.g. carbon stocks) [6–8]. Larch species and permafrost form a unique eco-climate system, wherein permafrost provides water to the tree's roots, while larches regulate permafrost thawing [9]. Despite being a major ecosystem, the detailed distribution and the environmental drivers of boreal forest types in eastern Siberia remain poorly understood

[10, 11]. Therefore, the potential alteration to boreal ecosystems in the future is uncertain, which is a matter of local to global concern.

Research shows that the distribution of major plant functional types on global and continental scales has generally been assumed to be influenced by contemporary climate [12, 13]. Specifically, temperature has been shown to drive Holocene boreal forest dynamics at the regional scale [14]. Other studies propose that the historical constraints of the Late Glacial, or that the latter combined with contemporary climate greatly impact present-day boreal forest composition [5, 15]. Topographical variables, which can be used as proxies for hydrological processes and incoming radiation at the local scale, are a key control of plant community distribution [16–20] and possibly of the abundance of Siberian larch (*Larix*) [21]. Furthermore, topographic factors significantly affect soil moisture and nutrient availability impacting the productivity and carbon storage of boreal forests [22], and influence the recovery and succession of boreal forests following disturbances like wildfires [23]. A gap exists in understanding how various factors such as climate and topography influence forest-type distribution across spatial scales, ranging from local (meters) to regional (kilometers) scale. Disentangling the influence of topographic and climatic variables in boreal forests is needed to predict how these ecosystems may respond to climate change, develop effective conservation strategies, and assess the forest types' environmental niches.

Gaining a predictive understanding of how the eastern Siberian boreal forest will respond to climate change requires a comprehensive grasp of its spatial arrangement and the key factors controlling it across local and regional scales [24]. Potential niche is defined as the range of distribution that would be achieved if all dispersal constraints were overcome, while realized niche refers to the actual niche occupied by a species or forest type [15, 25]. Identifying realized and potential forest-type niches helps understand species distribution ranges and highlight potential limiting factors. Uncertainties arise regarding whether one forest type's niches can be transferred to another, and whether the environmental space could support different species. For instance, evergreen forests have expanded into larch refugia [26, 27], yet the availability of potential evergreen niches in eastern Siberia remains unsure. Such knowledge is crucial with progressing climate change, where species range and population size could shift under changing environmental conditions [28, 29]. It remains unclear whether and how the climatic and topographic factors of boreal forest types vary between central and northeast Siberia, and how the forest-type niches apply.

Satellite remote sensing is a valuable tool for understanding boreal forests, enabling various applications such as monitoring, mapping, and biomass estimations [30–33]. The growing accessibility of publicly available global multispectral satellite imagery, such as Sentinel-2 (S2) or Landsat missions, has expanded the horizon of broad-scale forest ecology studies [34–37]. S2's efficacy in forest mapping has been well-demonstrated in numerous studies with its multiple red edge and near-infrared (NIR) bands allowing the capture of detailed vegetation reflectance [38–40]. International programs provide global landcover maps [41, 42] of multiple forest types at a spatial resolution ranging from 100 to 300 m [41, 43]. Despite these advances, many applications in forest ecology demand finer detail than the currently available maps offer, particularly when attempting to elucidate the intricate drivers of forest types in boreal ecosystems, or estimating forest growing stock volumes [22, 44–46]. Currently, there is no high-resolution map focusing on mixed evergreen-summergreen boreal forests in eastern Siberia [47] which would highlight the overlapping niches between forest types.

To improve our understanding of the eastern Siberian boreal forest's spatial distribution and environmental factors, we first aim to provide the spatial distribution of summergreen and evergreen needleleaf forest types in eastern Siberia based on a remote sensing classification using late summer S2 imagery and training data from the field. Peak summertime series are commonly used to classify forests with optical satellite imagery, covering the maximum greenness foliage period [48–51]. In this study, we use the late summer reflectance to differentiate summergreen needleleaf from evergreen needleleaf forest types. The vegetation coloring period in late summer is particularly suited for differentiating forest types in Siberia as larch forest will turn into fall colors while evergreen will remain green. Secondly, we aim to identify the main drivers, considering topography and climate at the local and regional scale, governing the distribution of forest types with a multivariate modeling approach. Specifically, we use generalized linear models (GLM) to disentangle the importance of each driver. Finally, we explore the applicability of the identified environmental niches and assess their relevance in different areas, highlighting potential versus realized niche-filling using predictions. We hypothesize that climatic variables on a regional scale are more influential than local topography in shaping forest-type distribution in eastern Siberia, such as summergreen versus evergreen. We demonstrate our approach in five regions of eastern Siberia, where fieldwork was conducted between 2018 and 2021 along a southwest-to-northeast (SW-NE) transect.

2. Materials and methods

2.1. Study areas

This study focuses on boreal forests in Siberia covering a substantial geographical range between latitudes 59° N–68° N and longitudes 110°E–170°E. This SW-NE transect marks the transition from evergreen forest (*Picea, Pinus*) in western Yakutia to larch-dominated forest (*Larix*) in central and eastern Yakutia, and to taiga-tundra transition in the mountainous landscape of northern Chukotka. Our study encompasses five regions: Khamra (KH), Western Yakutia (WY), Central Yakutia (CY), Oymyakon region (OY), and Chukotka (CH) with forest plot data collected in 2018 and 2021 from across eastern Siberia (figure 1) [52–54].

Eastern Siberia is marked by one of the most extreme continental climates on Earth, with severe cold winters, warm summers, and short growing seasons [55]. Freezing temperatures occur for 6–8 months and snow can last several months [56]. Maximum temperatures reach +30 °C and minimum temperatures −50 °C. Precipitation is low (mean annual precipitation 340 mm) [57] with a maximum in July and minimum in February [58]. The boreal forest thus experiences very dry conditions [59–61].

Larches are deciduous needleleaf trees that mostly grow on continuous permafrost [62]. The growing season starts May/June, while the needles usually begin to senesce at the end of August to September depending on the location [63]. Eastern Siberia's topography varies greatly from low rocky plateaus and the central Yakutian lowlands in the west, via the Verkhoyansk Mountain Range peaking at 2400 m to the Chukotka mountain range in the east.

2.2. Forest-type mapping

2.2.1. Sentinel-2 mosaic and datasets for training and validation

Mapping forest types first required gathering and preparing suitable S-2 satellite data for classification. S2 atmospherically corrected late summer data were collected with a 30 km buffer around each plot of our study using the cloud-based platform Google Earth Engine (GEE) [64]. We selected a late-summer time series ranging from September 1st to October 15th, from 2018 to 2022. This time frame covers the field survey years [52, 53] and the fall coloring period emphasizing the differences between evergreen versus summergreen needleleaf trees. S2 images were pre-processed to produce mosaics for the subregions and generate training and validation datasets. These datasets are based on the forest surveys and assigned to seven forest-type classes (table 1, appendix A1.2).

2.2.2. Classification of forest types

Random Forest (RF) is a widely used machine-learning algorithm for vegetation mapping [39, 65–69], and was used as a classifier to predict the forest types. It had the best results when trained using 100 trees and a dataset split of 60/40% for training and validation, respectively (appendix A1.3). The quantitative accuracies reported in this paper rely on the independent validation set constituted by this reserved portion of the training data. Qualitative assessments of the classified regions of eastern Siberia are based on expert knowledge from field expeditions. Our forest-type maps of the subregions were qualitatively compared with pre-existing landcover maps—ESA World Cover and Copernicus Global Land Cover [70, 71]. A minimum mapping unit of 10 m was used, as it corresponds to the highest S2 band resolution. Figure 2 displays the processing workflow.

2.3. Identifying controlling factors for forest-type distribution

2.3.1. Climate and topographic controls

Identifying the controls over forest-type distribution at two different scales is necessary to understand forest-type distribution and realized niches better. Our approach aims to distinguish which climate variables are drivers at the regional scale (30 × 30 km), and which topographic variables at the local scale (3 × 3 km). Bioclimatic variables from WorldClim version 2.1 are 1 km resolution interpolated climate data averaged over 1970–2000 [57]. Bioclimatic variables are derived from the WorldClim monthly temperature and precipitation values to generate biologically meaningful variables The most relevant and least correlated bioclimatic variables were selected using a Principal Component Analysis (PCA), namely: *Annual Mean Temperature* (bio1), *Temperature Annual Range* (bio7), *Mean Temperature of Warmest Quarter* (bio10), *Mean Temperature of Coldest Quarter* (bio11), and *Precipitation of Driest Quarter* (bio17) (figure 3 and appendix A1.4). These selected variables represent a range of annual trends, seasonality, and extreme and limiting environmental factors. Additionally, the mean Ground Surface Temperature (GST) was used at 1 km resolution from 2018 to 2021 [72], corresponding to the years of fieldwork in eastern Siberia. Copernicus Global 30 m TanDEM-X resolution Digital Elevation Model (DEM) was employed to identify the local topographic controls over forest-type distribution, and accessed through GEE [73]. Topographic metrics that are considered proxies for hydrological processes and illumination from the DEM were derived, such as elevation (m), slope (°), Topographic Position Index (TPI) [74], and aspect (°) (appendix A1.5).

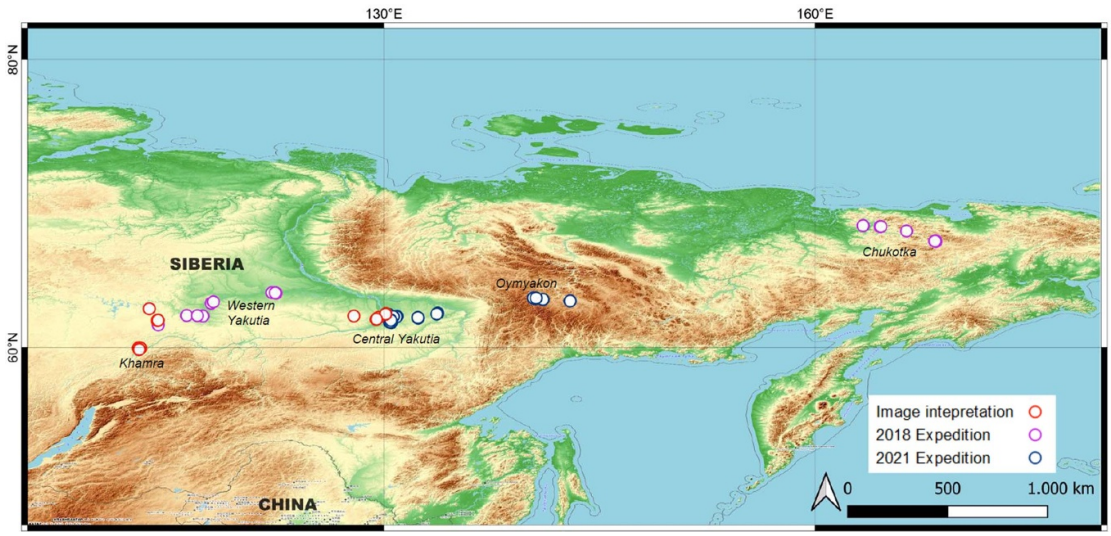

Figure 1. Area of study and location of field sites. 35 plots were visited in 2018 [52], mainly located in Chukotka and Western Yakutia. 44 plots were visited in 2021 [53], focusing on Central Yakutia and the Verkhoyansk mountains. We identified 24 additional vegetation plots through image interpretation of Sentinel-2 satellite imagery and Google Earth ®. Map data: © OpenStreetMap contributors, STRM.

Table 1. Total number of visited and image interpretation plots, number of pixels pixels, associated detailed class names and class labels [54].

Detailed class name	Class label	Expedition plot	Image interpretation	Total polygons
Larix woodland	Sparse Larch	19	0	19
Open *Larix* forest	Medium Larch	12	0	12
Closed *Larix* forest	Dense Larch	16	0	16
Needleleaf Evergreen forest	Evergreen	9	18	27
Mixed broadleaf and needleleaf Summergreen forest	Mixed Summergreen	12	0	12
Mixed broadleaf and needle leaf Summergreen and Evergreen forest	Mixed Summergreen-Evergreen	3	6	9
Burned or bare	Burned or bare	8	0	8
Total polygons	—	79	24	103
Total pixels	—	2,844	864	3,708

2.3.2. Multivariate analysis

A common regression method to identify important predictor variables is GLM [75–80]. We predicted the forest-type presence or absence in the environment by applying GLMs with a binomial distribution and a logistic link function (appendix A1.6).

The models were developed for the three classes of main interest and best classification accuracy: *Sparse Larch, Dense Larch*, and *Evergreen*. Acknowledging regional differences in the biophysical drivers and related spatial autocorrelation, one model per class per subregion was set (figure 4) for the regional and the local scale, i.e. 30 fitted GLMs. The statistical analyses were performed using R statistical software [81], and the GLMs created using the *glm* function in the *lme4* package [82]. Model fitness was assessed using the *DHARMa* package [83]. The pseudo-R-squared (pR^2) values were computed using the *pR2* function in the *pscl* package to account for the variance explained by the models [84].

For the regional models, the forest-type, bio1, bio7, bio10, bio11, bio17, and GST were extracted with a 1 × 1 km grid over the entire subregion area, i.e. the 30 km buffer around each plot. For each regional model, the predicted variable is forest type, which was binarized such that the pixel is equal to 0 if the class is absent and 1 if present, with a dataset for each class. The predictor variables are the above-mentioned bioclimatic variables and GST which have been identified as potentially important regional drivers of forest-type distribution. For the local models, elevation, slope, TPI, and aspect were used with a 0.1 × 0.1 km grid over a 3 km grid around each plot. This choice of reducing the size and resolution of the grid is to help understand the local effects of topographic variables. As Sato and

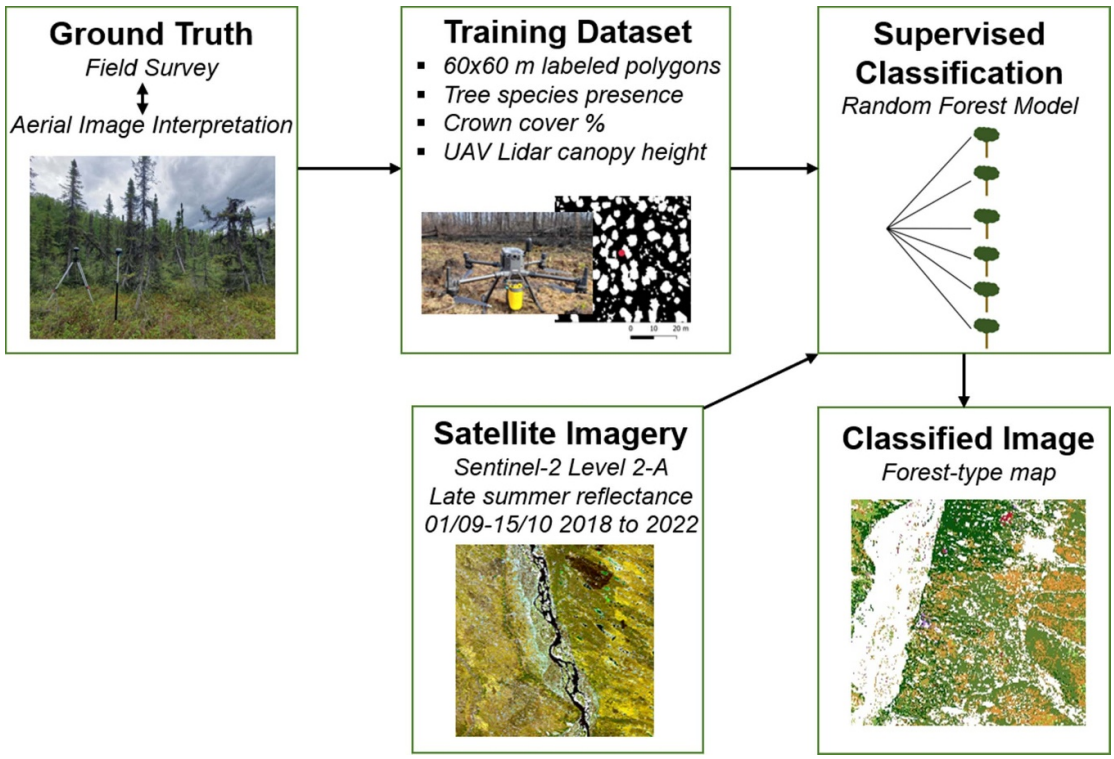

Figure 2. Classification schematic. We combined S2 late summer surface reflectance (2018–2022) with *in-situ* and UAV-based forest plot data, and image interpretation to define the training and validation datasets. We implemented a Random Forest regression to obtain classified maps of different forest-type categories. Red Green Blue satellite image is from the Copernicus Sentinel Data [2023].

Kobayashi suggest [21], topographic properties at the 100 m scale were used as a proxy for local hydrological conditions. For each local model, the predicted variable is forest type, binarized, and the predictor variables are the topographic variables. For both model setups, the predictor variables were centered by subtracting the mean before being included in the models due to the different ranges of values of the predictors, which improves the interpretation of the estimates and numerical stability. We evaluated the results by looking at the estimated values and associated *p*-values.

2.3.3. Potential versus realized niches predictions

The previous GLM analysis focused on identifying the realized forest types' niches based on abiotic factors (topography and climate). Subsequently, we determined whether the realized abiotic forest-type niches are relevant from one region to another along the SW-NE transect. If the correlation between the predicted and realized niches is low, other phenomena like biotic interactions could substantially drive the forest types [85].

To identify if the predicted versus realized environmental niches match, the potential niche of each class was predicted using the dataset of each of the five subregions with every fitted GLM. We call the predicted niche the output of the predicted model, and the realized niche the classified maps previously obtained. The R *predict* function and the Pearson correlation coefficient were used to evaluate the compatibility of the predicted and realized environmental niches across all sites.

3. Results

3.1. Estimation of forest-type distribution

Five major forest-type regional maps were produced, built with 30 km buffers around individual forest plots that were clustered in the five subregions. An overall accuracy of approximately 78% and f-scores varying between 35% and 92% for individual classes were obtained (table 2). Our results show that *Medium Larch* is the most frequent forest type within the entire mapped area (excluding masked areas) representing 20.8% of the area. The *Evergreen* class represents 19.3% of the mapped area and is especially dominant in the southwest of the study area and along the Lena River. *Dense Larch* constitutes 17.5% of the regional maps and dominates central and western Yakutia, followed by *Mixed Summergreen-Evergreen* covering 12.8% and mainly situated in the Khamra region. Finally, *Mixed Summergreen, Sparse Larch*, and *Burned or bare* represent roughly 9%–10% each. *Sparse Larch* is mainly found in the eastern Chukotka part and at higher elevations, while

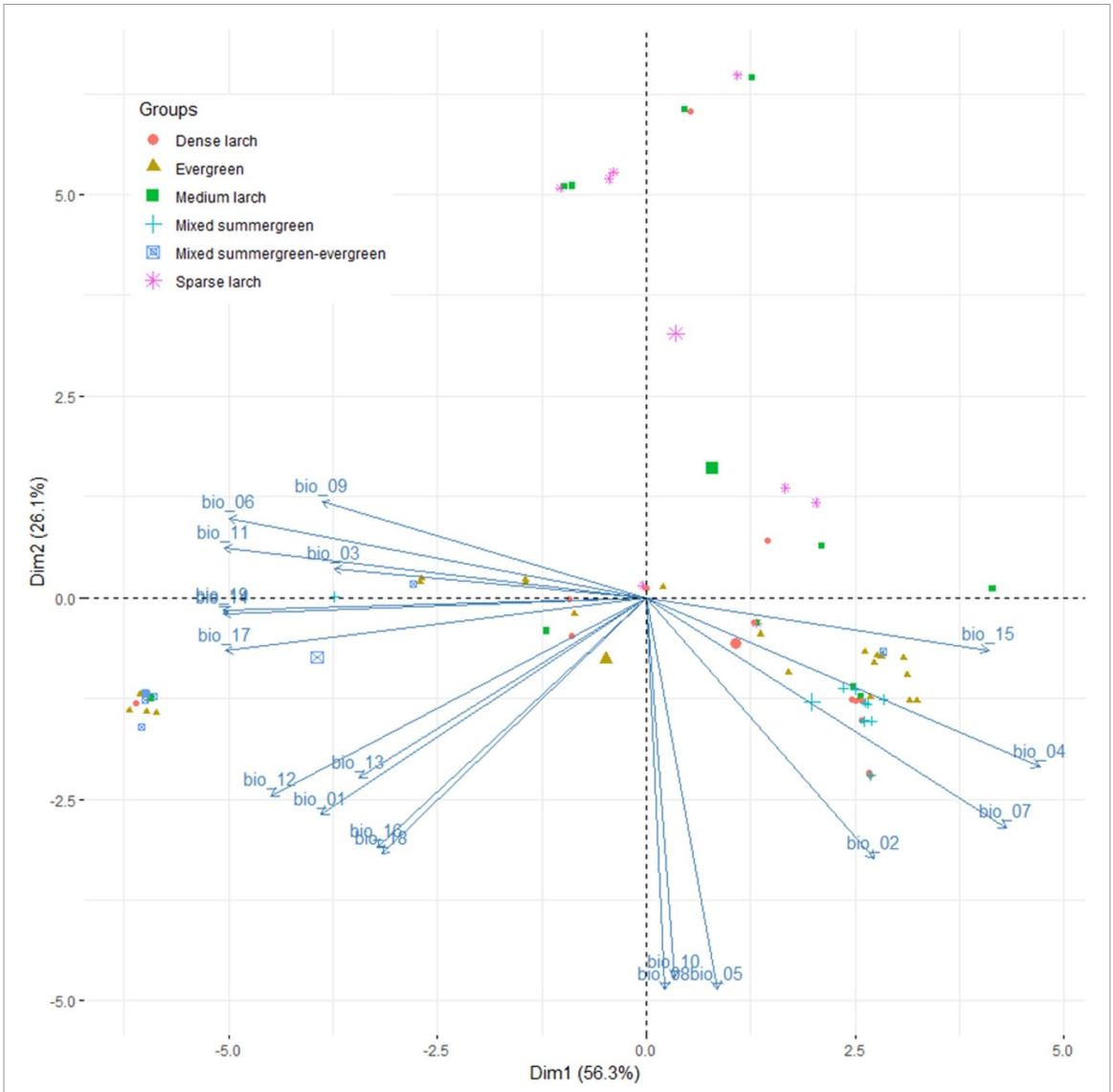

Figure 3. Principal component analysis of the bioclimatic variables related to the forest type classes. Each arrow corresponds to a single bioclimatic variable, the variable name is explained in appendix A1.4.

Mixed Summergreen is centered in central and western Yakutia (figure 4).

Medium Larch and *Mixed Summergreen* have the lowest accuracies (table 2) as their reflectance spectra were in-between classes, and fewer training/validation plots were available for these classes (appendix A2.1). For most classes, the precision scores are higher than the recall, implying that the classification tends to be cautious in predicting the positive class. We performed a qualitative accuracy assessment and found that our maps provide more thematic details than pre-existing landcover maps (appendix A2.2).

3.2. Controls for forest-type distribution: multivariate modeling of environmental drivers

A GLM was fitted for each region and class ($N = 30$) to highlight the importance of the environmental variable in predicting the presence/absence of the class. Overall, the regional GLMs explain more variance than the local models, suggesting that local topographic metrics alone do not explain an important part of the forest-type distribution compared to the climate variables in the regional models alone. However, the local models show the relative importance among the topographical metrics. The *pseudo-*R^2 varied from 0.001 to 0.75, ranging from poor predictive power to very high (appendix A2.3). Low pR^2 models are usually where the forest type is not prevalent in the region, inducing few presence values of the predicted variable (e.g. *Evergreen* in Chukotka, or *Sparse Larch* in Khamra). Conversely, high pR^2 occurs in models where the class is predominant in the region (e.g. *Sparse Larch* in Chukotka, *Dense Larch* in Western Yakutia).

3.2.1. Regional models for forest-type distribution

The magnitude of the model estimates (i.e. log odds) shows the importance of the predictors. Here, mean annual temperature and mean temperature of the warmest quarter and of the coldest quarter are the

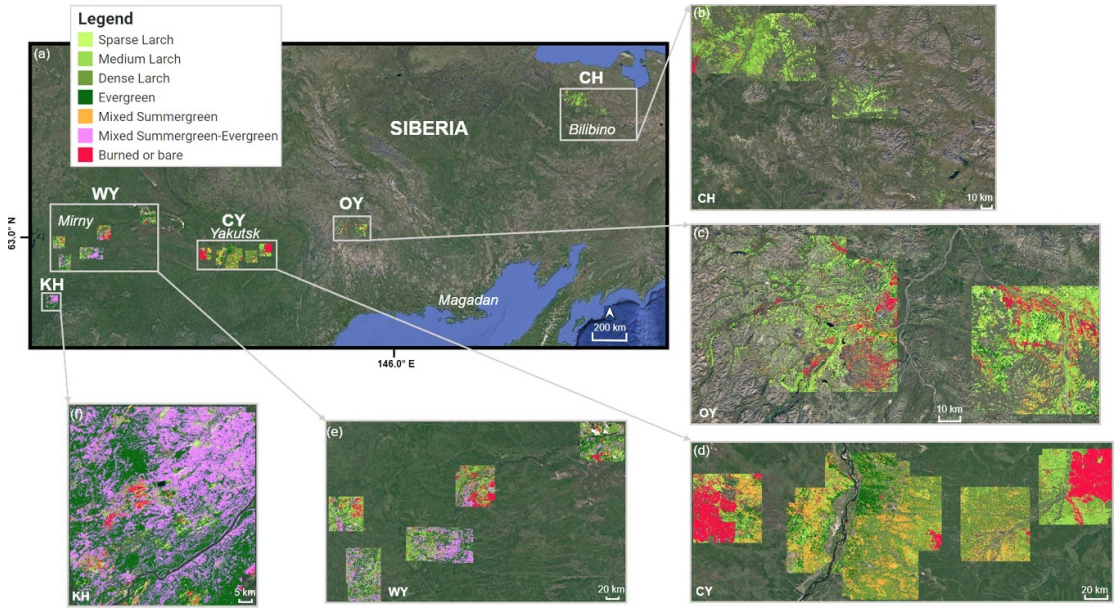

Figure 4. Classification maps of the estimated forest-type distribution. (a) Overview map with geographic regions of interest: (b) CH Chukotka, (c) OY Oymyakon, (d) CY Central Yakutia, (e) WY Western Yakutia, and (f) KH Khamra. Map data: Google, ©2024 NASA, TerraMetrics.

Table 2. Performance of the forest-type classification showing overall accuracy, Kappa coefficient, precision, recall, and f-score. Classes have high accuracies except Mixed Summergreen and Medium Larch.

Class label	F-score	Precision	Recall
Sparse Larch	88.27	81.32	96.53
Medium Larch	35.37	42.27	30.40
Dense Larch	77.07	63.03	99.16
Evergreen	92.28	91.53	93.04
Mixed Summergreen	44.51	65.21	33.78
Mixed Summergreen-Evergreen	78.47	79.02	77.93
Burned or bare	68.87	97.64	53.20
Accuracies			
Overall accuracy		78.17	
Kappa		72.82	

most influential predictors for all three classes across all regions (figure 5) with coefficients varying from −25 to +25 (compared to −0.4 to +0.8 for precipitation of driest quarter).

The estimates suggest that increasing temperature in the warmest or coldest quarter is more favorable for the occurrence of *Sparse Larch* and *Evergreen* compared to *Dense Larch*. In the Oymyakon region, *Dense Larch* shows strong negative log odds to the warmest quarter temperature. Conversely, the mean annual temperature has consistently significant negative values for *Sparse Larch* and *Evergreen*, indicating that an increase in mean annual temperature reduces the likelihood of *Sparse Larch* and *Evergreen*. GST is not statistically significant except for *Evergreen* where we observe a positive relationship from NE (Oymyakon) to SW (Khamra), meaning increasing GST is more favorable to *Evergreen along* this gradient. The annual temperature range seems influential for *Sparse Larch* and *Evergreen*, where we observe a stronger positive relationship for *Sparse Larch* and a shift from negative to positive log odds for *Evergreen* from Central Yakutia to the Khamra region in southwest Yakutia.

Furthermore, our results show that precipitation of the driest quarter (bio_17), representing the coldest winter months, varies significantly across regions. For both *Sparse Larch* and *Evergreen*, we observe a decrease in the log odds from Chukotka to central Yakutia and an increase to the west to the Khamra region. These results suggest that from Chukotka to Central Yakutia, both classes tend to be located where winter precipitation is higher than from Yakutia to Khamra. Winter precipitation in the three coldest months is not an important predictor of *Dense Larch*, except in western and central Yakutia where the log odds are positive.

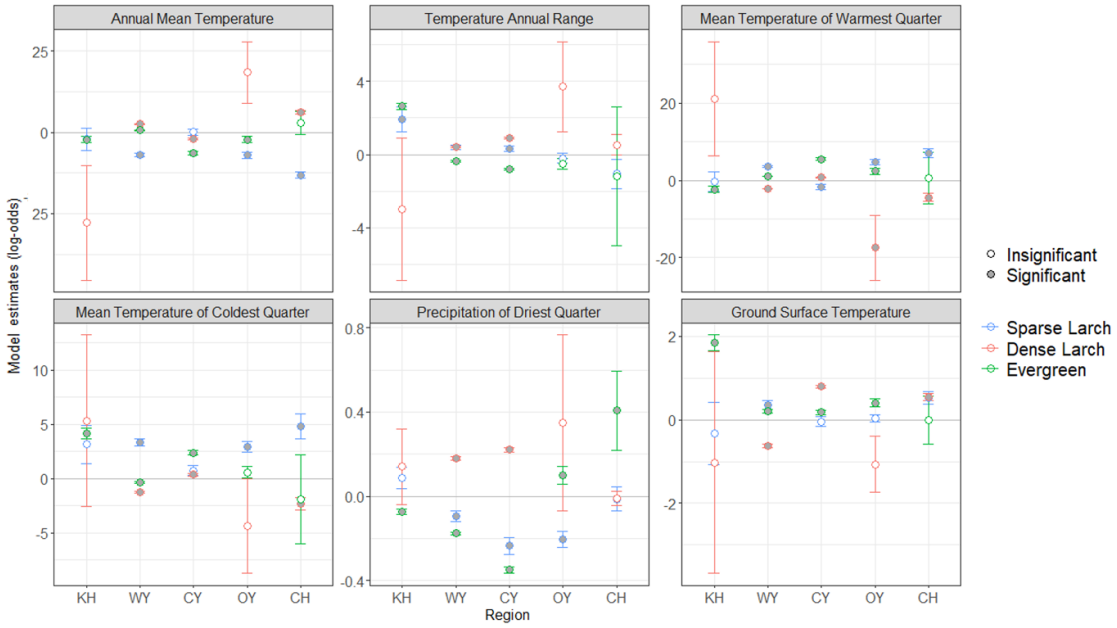

Figure 5. Results of 15 regional generalized linear models including annual mean temperature (bio1), temperature annual range (bio7), mean temperature of warmest quarter (bio10), mean temperature of coldest quarter (bio11), precipitation of driest quarter (bio17), and ground surface temperature (GST) within a 30 km radius around each vegetation plot, and sampled at 1 km. Each point corresponds to the output of the model, i.e. the log odds. The *y*-axis is the estimated log-odds output of the model, and the *x*-axis the region. If the predictor variable was not statistically significant ($p > 0.05$), the circle is unfilled. KH: Khamra, WY: Western Yakutia, CY: Central Yakutia, OY: Oymyakon, and CH: Chukotka.

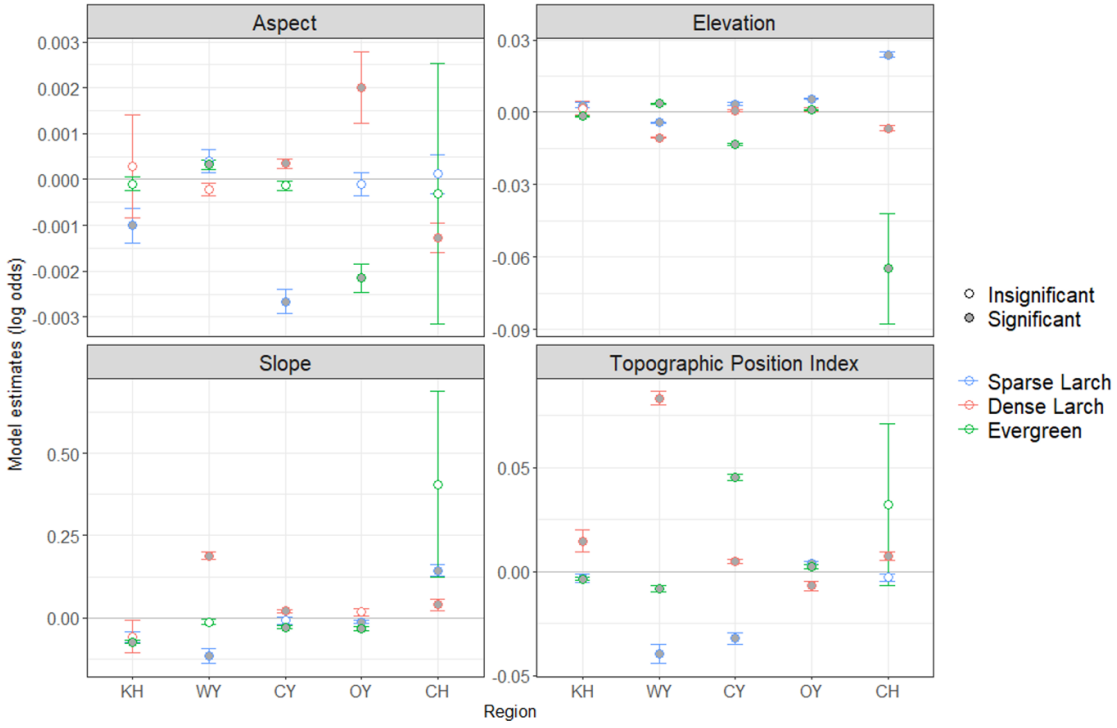

Figure 6. Results of the 15 local generalized linear models including aspect, elevation, slope, and topographic position index, within a 3 km radius around each vegetation plot, and sampled at 0.1 km. Each point corresponds to the output of the model, i.e. the log odds. If the predicting variable was not statistically significant ($p > 0.05$), the circle is unfilled. KH: Khamra, WY: Western Yakutia, CY: Central Yakutia, OY: Oymyakon, and CH: Chukotka.

3.2.2. Local topographic models for forest-type distribution

The results of the local models suggest that slope, followed by TPI are the most influential predictors in the models for all three classes and regions (figure 6). According to the modeling results, an increase in slope augments the likelihood of *Dense Larch* presence versus *Evergreen* and *Sparse Larch*, both of which

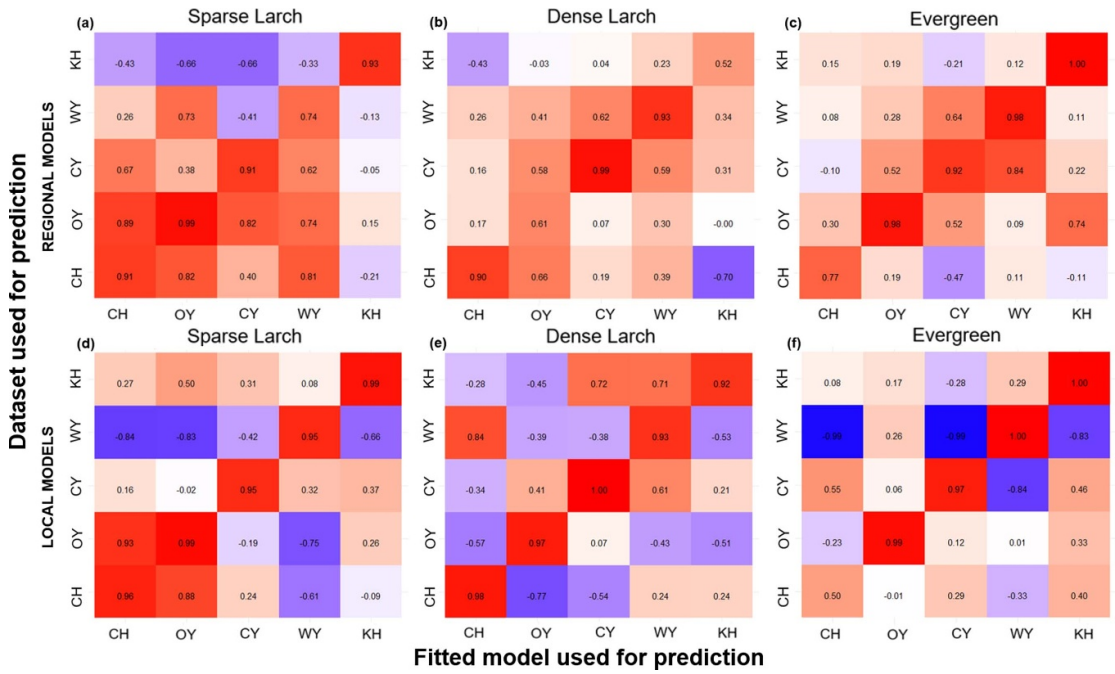

Figure 7. Correlation matrices between models and regions showing the Pearson correlation coefficients. The *x*-axis shows the different fitted models, and the *y*-axis shows the different datasets used for prediction. The top row is of regional models and the bottom row of local models. The columns show forest-type class: (left) *Sparse Larch*, (middle) *Dense Larch*, and (right) *Evergreen*. Red colors indicate positive correlations, and blue negative. The darker the color the stronger the correlation. KH: Khamra, WY: Western Yakutia, CY: Central Yakutia, OY: Oymyakon, and CH: Chukotka.

have negative estimates in Chukotka. Consequently, *Dense Larch* is more likely to be situated on steeper slopes than *Sparse Larch* and *Evergreen* except for in the Chukotka region.

The outcomes of the TPI analysis indicate that *Sparse Larch* tends to manifest at relatively lower elevations, and *Dense Larch* at relatively higher elevations (e.g. ridges), excluding the Oymyakon region where local elevation variations are more pronounced than in the other regions. Elevation displays weak coefficients, but *Evergreen* shows a tendency to be located at a lower absolute elevation. Aspect exhibits considerable variation between regions and classes; however, the robustness of the results is not evident in the magnitude and statistical significance of the estimates.

3.2.3. Potential versus realized niches

The applicability of the predicted (potential) versus realized environmental niches along the SW-NE transect in eastern Siberia was assessed. The potential niche of each forest-type class was predicted using the dataset of each subregion with its fitted model, corresponding to the realized niche, and visualized in space (figures 7 and 8).

For *Sparse Larch*, the regional models appear to effectively capture its ecological niche across the regions of interest, featuring notably high positive correlation coefficients (figures 7(a) and 8(a)). This suggests that the realized and potential regional climate niches of *Sparse Larch* align well across regions. For *Dense Larch* and *Evergreen*, the regional models appear to effectively capture their ecological niche only when the model and dataset are geographically close together (figures 7(b) and (c)). For example, the regional model for *Dense Larch* trained in Oymyakon with the Western Yakutia dataset shows a low Pearson coefficient (0.30; figure 7(c)), which could imply that the climate features are different and inapplicable between regions or that there is a mismatch between the realized and potential niches. Additionally, forest types like *Evergreen* in Chukotka or *Sparse Larch* in Khamra are not common and therefore lead to low correlation coefficients.

Local models show varying results for all classes, with both positive and negative Pearson coefficients (figures 7(d), (f) and 8(b)). This suggests that, based on local topography, the potential ecological niches of the three classes are not filled, and/or that the topographic variables are not compatible from one region to another.

4. Discussion

Our study shows that the distribution of the forest types *Sparse Larch*, *Dense Larch*, and *Evergreen* is mainly influenced by climatic drivers on the regional scale rather than topographic boundary conditions on the local scale. Regional models that include

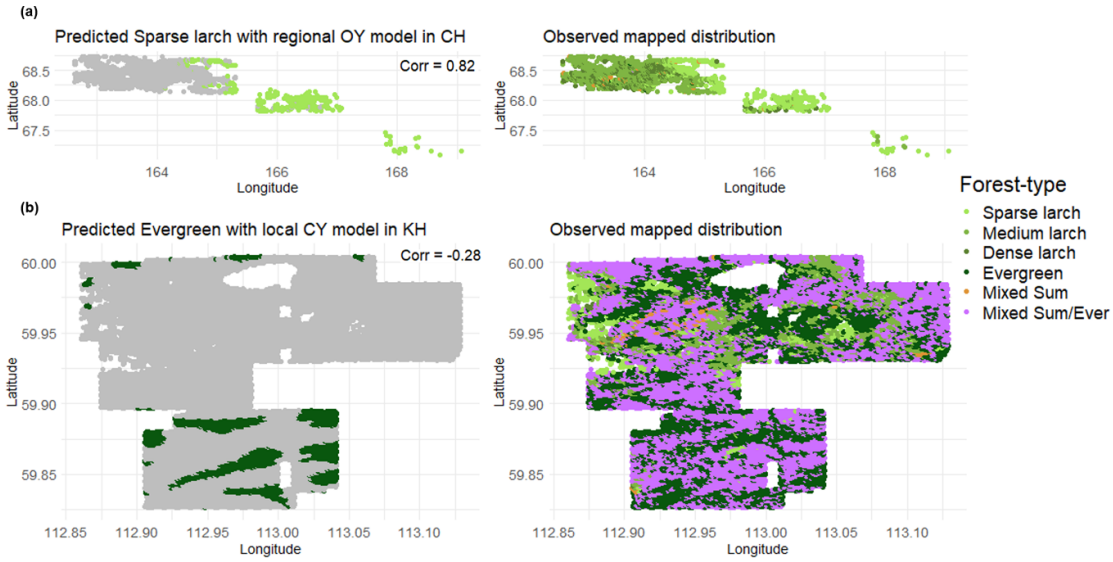

Figure 8. Predicted class distributions in geographical space compared to their observed distribution, with associated correlation coefficient. Left panels are: (a) predicted *Sparse Larch* distribution using the regional Oymyakon model for the Chukotka region; and (b) predicted *Evergreen* distribution using the local Central Yakutia model in Khamra. In grey is where the forest type is absent. The right panels correspond to the mapped class distribution. Mixed Sum: Mixed Summergreen, Mixed Sum/Ever: Mixed Summergreen-Evergreen.

only selected bioclimatic variables and GST generally have higher predictive power than local models that include the topographic variables only. These findings suggest that the regional climate variables are more significant drivers of forest-type distribution than local topographic variables, supporting our initial hypothesis. Nevertheless, the predictive power of some models is low, probably reflecting the complexity of the processes behind boreal vegetation dynamics [14, 86]. Another main finding of this study is the positive correlation between potential and realized forest-type niches, which weakens with increasing distance from one region to another along the SW-NE transect.

4.1. Air temperature variables drive forest types' distribution at the regional scale

Using climatic drivers in GLMs on the regional scale identified the predominant factors of forest types with air temperature variables as the main driver of forest type in eastern Siberia. Air temperature variables are significantly more important than precipitation variables and GST in explaining the distribution of the three forest types. These results are consistent with a paleoclimate study that showed the importance of temperature as a driver of forest dynamics in western Siberia across the past 9000 years [14]. A biome reconstruction study based on global fossil pollen records finds that vegetation change in the Northern Hemisphere may be limited by moisture changes, meaning the effect of temperature and precipitation combined [87]. Moreover, we show that mean annual temperature has the most influence, followed by mean temperature of the warmest quarter and mean temperature of the coldest quarter. Our results suggest that an increase in mean annual temperature (keeping the other variables stable) decreases the likelihood of occurrence of Sparse Larch and Dense Larch. In the Khamra region, we find that Evergreen responds positively to an increase in the coldest temperature but negatively to an increase in the warmest temperature. We notice similar patterns between mean annual temperature and the warmest and coldest temperatures for all forest types. Wilmking *et al* [88], suggest that, depending on the precipitation regime and the age of the forest, increasing temperature could positively or negatively influence tree growth. Eastern Siberia being a high-latitude region is subject to unprecedented rising temperatures due to climate change [89–92]. Our GLM results suggest that when mean annual temperatures increase in the taiga-tundra ecotone of Chukotka, where the transition zone is diffuse [93], less *Sparse Larch* is expected, but there is a higher likelihood of *Dense Larch*. These results could imply a latitudinal treeline migration and forest densification similar to processes currently observed in Chukotka further to the south [94]. A part of the regional models' variance remained unexplained, indicating the influence of additional factors on forest-type distribution, such as topography.

4.2. Slope as an indicator of moisture drives forest types' distribution at the local scale

Local models including topographic variables at the local scale, explained the respective influence of slope, TPI, aspect, and elevation on forest types distribution. Our local modeling approach indicates that

slope and TPI are the main local topographic drivers of the three forest types. The results highlight the preference for *Dense Larch* to be found on steeper slopes compared to *Sparse Larch* and *Evergreen*. The TPI findings suggest that *Sparse Larch* typically occupies the lower elevations of its local surroundings. In contrast, *Dense Larch* is positioned at the higher elevations of its surroundings. Sato and Kobayashi [21] find that larch forests are controlled by topography-mediated hydrology and that they prefer patches with lower inundation risks. Our finding that dense larch forest tends to be located on steeper slopes than sparse larch forests may be explained in terms of hydrology with higher soil water drainage and hence lower risks of flooding compared to lower elevation or flatter lands. Furthermore, hilly terrain provides more space to grow than valley bottoms, and the latter contains fine-grained water-saturated sediments which could explain the use of the environmental space on slopes. Even though past studies proved slope orientation to be an important driver of larch abundance [21], our study does not find aspect to be a significant driver of the three forest types. Another study [95] has found that, as a result of topographically induced differences of solar radiation, Siberian larch was limited to north-facing slopes in the northern Mongolian mountain taiga. When considering our modeling approach and large study area, the variability in aspect across classes and regions is not significant. While aspect could potentially influence forest distribution, its importance diminishes greatly when considering slope, TPI, and elevation in the GLM, especially slope. Khamra, Western Yakutia, and Central Yakutia have low topography, and therefore no major influence of aspect is expected. In addition, the influence of topography on boreal forest distribution might be reduced or negated by the impact of fire or other disturbances that could locally modify forest distribution. Larch density responds to microsite conditions such as water tracks and natural levees [81], but capturing such intricate topographical patterns remains challenging due to the moderate resolution of the topographic data (GLO-30 m). The multivariate analysis of both local and regional models helped define the realized niches of *Sparse Larch, Dense Larch*, and *Evergreen*.

4.3. Realized and potential environmental niches diverge with increasing distance

The fitted local and regional models were applied to other regions to investigate the applicability of the identified realized environmental niches. For *Dense Larch* and *Evergreen*, we find that regional models are well represented at the landscape scale, and when the classes are sufficiently present in the area. This suggests that the potential and realized habitats of closely located regions, such as Central Yakutia and Western Yakutia, align effectively. The question arises as to why, with increasing distance, the realized and potential species distributions do not appear to converge. Pearson and Dawson [85], discuss how a low realized versus potential ratio might indicate different phenomena such as biotic interactions, edaphic or other non-climatic environmental factors, and modeling misspecifications. Some *Larix* species have overlapping bioclimatic niches [96] which, in turn, affect the forest-type niche. For example, in the Khamra and Western Yakutia regions, Larix and Evergreen also occur in mixed classes (*Mixed Summergreen-Evergreen; Mixed-Summergreen*) and therefore Larix and Evergreen are present but, in the prediction, the realized versus potential niches are not aligned because we map them by their endmember classes of *Evergreen* or *Larix* only. However, if we integrate the *Mixed Summergreen-Evergreen* and the *Mixed-Summergreen* classes in space, abundant Larix and Evergreen is filling its niches.

Svenning and Skov [15] argue that the low filling of the potential range is mainly due to dispersal limitations, which reflects historical constraints from the post-glacial recolonization from ice-age refugia. Additionally, Herzschuh [5] proposes that the boreal tree refugia are governed by the environmental conditions experienced during the Last Glacial. Specifically, deciduous and evergreen boreal forests were determined by different conditions during the Last Glacial resulting from genetic constraints and indirectly from the glacial climate [5]. Additional environmental factors, such as fire history at a site can affect the recovery and trajectories of boreal forests [97, 98], especially at Lake Khamra in the southern part of our study where wildfire history has been documented over the last two millennia [99]. Evergreen tree taxa, over the past millennia to recent years, have progressively transgressed in eastern Siberia from SW to the NE summergreen larch refugia [26, 27]. Our results predict that these taxa have potential climatic niches towards the East (figure 7(c)). If the transgression continued, Evergreen would theoretically have environmental space to grow in this region, where potential niches could be filled, provided other factors such as competition with larch and poor dispersal are not limiting. We show that the climate- and topography-driven models explain some, but not all, of the forest-type distribution. These findings prompt consideration of the hypotheses put forth by Svenning and Skov [15] and Herzschuh [5], and we suggest that both the Late Glacial environmental factors and current climate and topography play a role in shaping the present-day distribution of boreal forests.

4.4. Methodological improvements and limitations

This paper is based on state-of-the-art methods, such as RF supervised classification and GLM. Reliable forest-type maps were produced based on S2 late summer reflectance composites trained with forest

plot data from our 2018 and 2021 expeditions in eastern Siberia. These maps provide an assessed distribution of forest types, which, to our knowledge, was not previously available due to a lack of research focusing on this region and due to missing optimization of the available global landcover maps for this region.

Using late summer reflectance proves especially effective for high-latitude areas where obtaining cloud-free time series satellite images can be challenging, and the distinction between evergreen and summergreen is low during the peak of photosynthetic productivity. Here, we optimized the acquisition specifically for late summer, necessitating fewer images compared to full-year time-series approaches. This method helped us robustly map three spectral endmembers that are highly distinguishable: *Sparse Larch, Dense Larch*, and *Evergreen*. However, we encountered difficulties with classes that spectrally fell in-between the reflectance value ranges of the end members, as their spectral values show a wide range and a similar reflectance shape (appendix A2.1). For example, the *Medium Larch* class represents reflectance spectra fluctuating between *Sparse Larch* and *Dense Larch*. Our labeling approach, based on *in-situ* and drone-derived crown cover percentage may also be suboptimal for mapping this class, as continuous values rather than class-level labels might be better.

From a qualitative assessment, we identified similarities in spatial patterns of forest distribution with other global landcover maps. Our maps demonstrate higher thematic accuracy and diversity of classes, especially for regions where important forest types were not mapped or are missing in the global landcover maps. This reliable mapping stems from our focus on optimizing the map for the forest growing in eastern Siberia where limited reference data are available in contrast to regions like Western Europe [100]. Still, despite having a substantial number of field plots available, challenges persisted in mapping forest types due to the scarcity of some classes, and generally insufficient labeled data. Augmenting the dataset with additional ground-truthed data would undoubtedly enhance the precision of our classification. Access to more publicly available labeled data is imperative for further refining remote sensing classification for upscaling to a broader scale.

Our study focused on bioclimatic and topographic variables, yet other factors such as natural disturbances like wildfire are also thought to shape the distribution of boreal forests in eastern Siberia [101–103]. For example, a study highlighted that fire appears as the main factor controlling the dominance of deciduous over evergreen forests in Siberia, with different forest types having different thresholds of fire tolerance [104]. The authors [104] suggest that evergreen conifers are the natural late-successional species both in central and eastern Siberia. Other forest disturbances caused by insect outbreaks or wind can also impact the dynamics and hence the distribution of forest types [105]. Our study did not focus on such factors though, being centered on bioclimatic and topographic variables. However, as our models reveal unexplained variance, it prompts speculation that factors such as fire or disease may play a significant role.

5. Conclusion

We investigated the spatial distribution and drivers of boreal forests in eastern Siberia, along a SW–NE transect. We provided high-resolution maps of summergreen and evergreen needleleaf forest types in eastern Siberia based on S2 imagery and field data. We determined the main drivers governing the distribution of forest types, considering climate and topography. Lastly, we identified potential and realized environmental niches of the forest types and assessed their spatial applicability.

Our results demonstrate that our classification approach achieved higher accuracies for *Sparse Larch, Dense Larch*, and *Evergreen* than the other classes due to their distinct reflectance spectra. The late summer reflectance allowed the *Evergreen* and *Larch* classes to be well differentiated, enabling the possibility of investigating their drivers at a fine resolution. The forest-type driver's analysis indicated temperature is a key regional driver, with topography playing a role at the local scale. Additionally, the realized and potential climate niches of *Dense Larch* and *Evergreen* align well when they are spatially close, but do not appear to converge as the distance increases, indicating potential dispersal limitations. Our evaluation reveals a notable unexplained portion of the variability in certain models, hinting at other influential factors. We propose that both historical legacy environmental factors and current climate and topography contribute to shaping the present-day distribution of boreal forests.

Considering the ongoing climate change and rising temperatures in the Arctic, we anticipate that these changes will significantly impact forest types. Potential environmental niches could be filled in different regions, or species may undergo reshuffling within their niche space [106]. Our findings, along with other studies, suggest that the effects of climate change will lead to a northward advance of the treeline ecotone, or its densification [107–110]. Such phenomena will impact the climate-feedback interactions with boreal forests (i.e. albedo decrease), which can significantly contribute to global warming as Earth-system modeling studies revealed [111]. Consequently, understanding the composition and drivers of these unique high-latitude ecosystems becomes essential in evaluating the present condition of the forests and anticipating their responses to climate change.

Data availability statement

The data that support the findings of this study are openly available. The training and validation dataset for the Random Forest classification is available at https://doi.org/10.1594/PANGAEA.964699 and cited in the references as Enguehard *et al* [54]. Additionally, the produced maps are available as GeoTIFF at https://doi.org/10.1594/PANGAEA.967133.

Acknowledgments

We acknowledge the project funding ERC Consolidator Grant Glacial Legacy (772852), and BMWi AI-vergreens. Further, we acknowledge funding for our equipment for the project 'Potsdamer InnoLab für Arktisforschung' No. F221-08-AWI/001/002 from the Brandenburg Ministry for Science, Research and Culture (MWFK). We thank AWI intern Peter C Frandsen who worked on shapefiles construction, and Cathy Jenks for proofreading.

Appendix

A1. Appendix methods

A1.1. Sentinel-2 satellite mosaic generation and extraction of forest-covered area

We used a S2 late-summer time surface reflectance ranging from September 1st to October 15th, 2018–2022. S2's multispectral instrument gathers data in 13 spectral bands at a spatial resolution ranging from 10 to 60 m with a temporal resolution of 3–5 d. The S2 bands used in our study to compile the temporal composite from 2018 to 2022 are summarized in table A1.1. We added the NDVI derived from the NIR and the red band as the ratio of the difference between B8 and B4 to the sum of the two reflectances [112]. From the S2 time-series collection, we used only images with less than 10% cloud coverage. We masked the remaining ones using the S2 qualitative assessment layer (band QA60), applying a bit mask for cloud and cirrus pixels bitmask (Bit 10: opaque clouds, Bit 11: cirrus clouds). As our study focused on forests, we masked non-forested areas using a satellite-derived global tree canopy cover product from Hansen *et al* [113]. Finally, mosaics of the regions linked to the field plots were created with the S2 image collection using the 25th-percentile pixel value of each band which specifically provides atmospheric noise reduction and is robust to outliers compared to the median pixel [114–116].

We constructed a forest map using the Hansen tree canopy cover map for the year 2000 defined as canopy closure in percentage for all vegetation taller than 5 m in height [113]. This product is available at 30.92 m nominal resolution with a value range from 0% to 100%. We checked with different thresholds for the treeline regions where we have field data and expert knowledge, and set our mask to the minimum cover value, 1%, meaning all areas with less than 1% tree canopy cover map were masked. We also tested NDVI peak summer thresholds to serve as an active tree mask, however, the threshold masks were very noisy, with no optimal threshold for all maps, and we did not find a reasonable threshold to extract all sparse forest types in the Oymykyon mountain region. The ABoVE tree canopy cover dataset [117] presents better tree coverage patterns with fewer artifacts than the Hansen dataset [113]. The ABoVE dataset should be preferred if forest canopy cover percentage is the focus of the research. For the application as a forest mask, the 1% Hansen tree cover threshold is rather similar in the boreal domain (figure A1.1) but more restrictive when excluding shrubland in the treeline ecotone (figure A1.1(a)). We consider that using the Hansen tree canopy cover map for the year 2000 is a robust approach, because, in general for all the maps, forest change in the past 20 years has predominantly occurred due to fires. In this case, we could modify the state of the Hansen tree mask by mapping the recent fire scars in the classification using the class 'burned or bare' to produce the actual state of the forest-covered area.

A1.2. Sentinel-2 labeling of the training and validation dataset

The forest-type classification's training and validation dataset were based on the pre-processed S2 collection, the forest plot, and the image interpretation sites. 60 m × 60 m (36 pixels) polygons centered on the 30 m × 30 m field plots were built, as they were set in homogeneous environments, and assigned a single forest-type label to all pixels contained within each of the 79 sites [54]. Each of the 79 forest plots was labeled with the forest-type based on field data of tree species and crown cover percentages using the method as follows:

- When the vegetation plot had one species of tree, the plot was assigned a label as this species only. For example, a plot with 60% *Larix* crown cover was designated 'Larch'.
- When the plot had two species with one comprising less than 10%, it was assigned a label as the dominant species. For example, a plot with 60% *Larix* with 5% *Pinus* was designated 'Larch'.
- When the plot had multiple tree species with similar cover (relative difference <20%), it was assigned a label of mixed forest. There were two cases: '*Mixed Summergreen*' (e.g. *Larix* and *Betula*), and '*Mixed Summergreen-Evergreen*' (e.g. *Larix* and *Picea*).

Additionally, we calculated the plot crown cover percentage based on LiDAR point clouds recorded in 2021 with a Yellowscan Mapper carried by an M300 DJI drone, and structure from motion from the

Table A1.1. Sentinel-2 bands used for the classification and their description.

Band name	Wavelength S2A/S2B (nm)	Description
B2	496.6/492.1	Blue
B3	560/559	Green
B4	664.5/665	Red
B5	703.9/703.8	Red Edge 1
B6	740.2/739.1	Red Edge 2
B7	782.5/779.7	Red Edge 3
B8	835.1/833	NIR
B8A	864.8/864	Red Edge 4
B11	1613.7/1610.4	SWIR 1
B12	2202.4/2185.7	SWIR 2

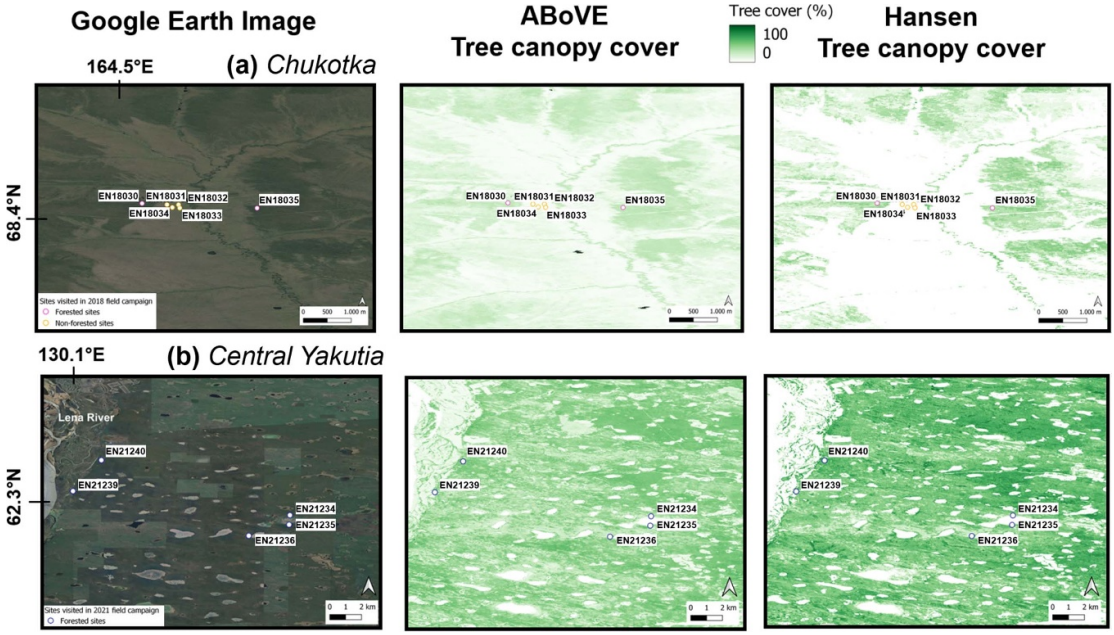

Figure A1.1. Comparison of Hansen 2000 and ABoVE for the year 2020 tree cover percentage in two regions of interest. (a) Chukotka, in the treeline region. (b) Central Yakutia, west of the Lena River in the Alas region. In the treeline region of Chukotka (a), ABoVe shows higher tree percentages where the sites EN18031 to EN18034 are located. However, we collected field data from these sites that show this area is only low shrub/tundra and non-forested as of the summer of 2018 [52]. In the lowland regions of Central Yakutia (b), Hansen and ABoVe match well. Map data: Google, ©2024 TerraMetrics. Adapted from [117]. CC BY 4.0. and Map data: Hansen/UMD/Google/USGS/NASA.

built-in RGB camera carried by the DJI Phantom 4 recorded in 2018 [118]. Due to the abundance of larch stands in eastern Siberia and having access to the *in-situ* estimated and UAV-based crown cover data, we expanded three larch categories into *Sparse Larch* (crown cover $< 20\%$), *Medium Larch* (crown cover between 20 and 50%), and *Dense Larch* (crown cover $> 50\%$). The crown cover percentage was calculated from Lidar pointclouds where we extracted the mean tree crown cover percentage (for trees > 2 m with a resolution of 3 cm) from a Canopy Height Model for the plot area. Both UAV and field crown cover estimations had similar values varying between 0% (meaning no tree), and 100% (meaning entirely covered by tree canopy). If the two values were dissimilar, we chose the *in-situ* value.

To enrich and balance the dataset, we added a total of 24 image interpretation plots (of a similar 60×60 m size) corresponding to the '*Evergreen*' and '*Mixed Summergreen-Evergreen*' labels that were chosen with expert knowledge using the NDVI band of our collection and Google Earth© imagery compared with spatially closely visited field sites at the same period. The '*Burned or bare*' label was assigned to recent or older burnt forest plots. In total, we defined six forest classes and one class for barren or recently burned as labels: *Sparse Larch, Medium Larch, Dense Larch, Evergreen, Mixed Summergreen, Mixed Summergreen-Evergreen*, and *Burned or Bare*. These classes are the basis of the forest-type classification (table A1.2, section 2.2.1 in the main text, replicated below).

Table A1.2. Total number of visited and image interpretation plots and pixels and associated detailed class name and short class labels [54].

Detailed class name	Class label	Expedition plot	Image interpretation	Total polygons
Larix woodland	Sparse Larch	19	0	19
Open *Larix* forest	Medium Larch	12	0	12
Closed *Larix* forest	Dense Larch	16	0	16
Needleleaf Evergreen forest	Evergreen	9	18	27
Mixed broadleaf and needleleaf Summergreen forest	Mixed Summergreen	12	0	12
Mixed broadleaf and needle leaf Summergreen and Evergreen forest	Mixed Summergreen-Evergreen	3	6	9
Burned or bare	Burned or bare	8	0	8
Total polygons	—	79	24	103
Total pixels	—	2844	864	3708

Table A1.3. Bioclimatic variables from WorldClim v2.0. In bold and italic are the variables used in the study.

Bioclimatic variable	Short name
Annual mean temperature	bio1
Mean diurnal range	bio2
Isothermality	bio3
Temperature seasonality	bio4
Max temperature of warmest month	bio5
Min temperature of coldest month	bio6
Temperature annual range	bio7
Mean temperature of wettest quarter	bio8
Mean temperature of driest quarter	bio9
Mean temperature of warmest quarter	bio10
Mean temperature of coldest quarter	bio11
Annual precipitation	bio12
Precipitation of wettest month	bio13
Precipitation of driest month	bio14
Precipitation seasonality	bio15
Precipitation of wettest quarter	bio16
Precipitation of driest quarter	bio17
Precipitation of warmest quarter	bio18
Precipitation of coldest quarter	bio19

A1.3. Classification of forest types

Random Forest is an ensemble classifier consisting of multiple decision trees constructed using randomly selected training datasets and random subsets of predictor variables. The results from each tree are aggregated, and an individual pixel's classification is decided based on the class that emerges as the most frequent choice across all decision trees. To avoid autocorrelation, we sampled pixels from the distinct 60 × 60 m polygon extent of the forest plots for validation and training sets, so that both sets are not from the same vegetation plot. To tackle unbalanced classes, we applied stratified sampling to maintain a proportional representation of classes within both training and validation datasets [119].

Several parameters were used to quantitatively assess the classified map, including precision, recall, overall accuracy, and F1 score which are commonly used to assess unbalanced datasets. Precision (consumer's accuracy) measures the proportion of correctly identified positive cases from all cases identified as positive, whereas recall (producer's accuracy) measures the proportion of true positives that were correctly identified by the model The F1 score provides a balance between precision and recall, and the overall accuracy gives information about the whole classification result [120, 121].

A1.4. Principal component analysis of the bioclimatic variables

Bioclimatic variables were derived from the monthly averaged temperature and precipitation values to generate more biologically meaningful variables. We identified five groups that were the least correlated. One bioclimatic variable per group was selected for our analysis. We chose from group one bio11, from group 2 bio17, from group 3 bio1, from group 4 bio10 and from group 5 bio7 (table A1.3).

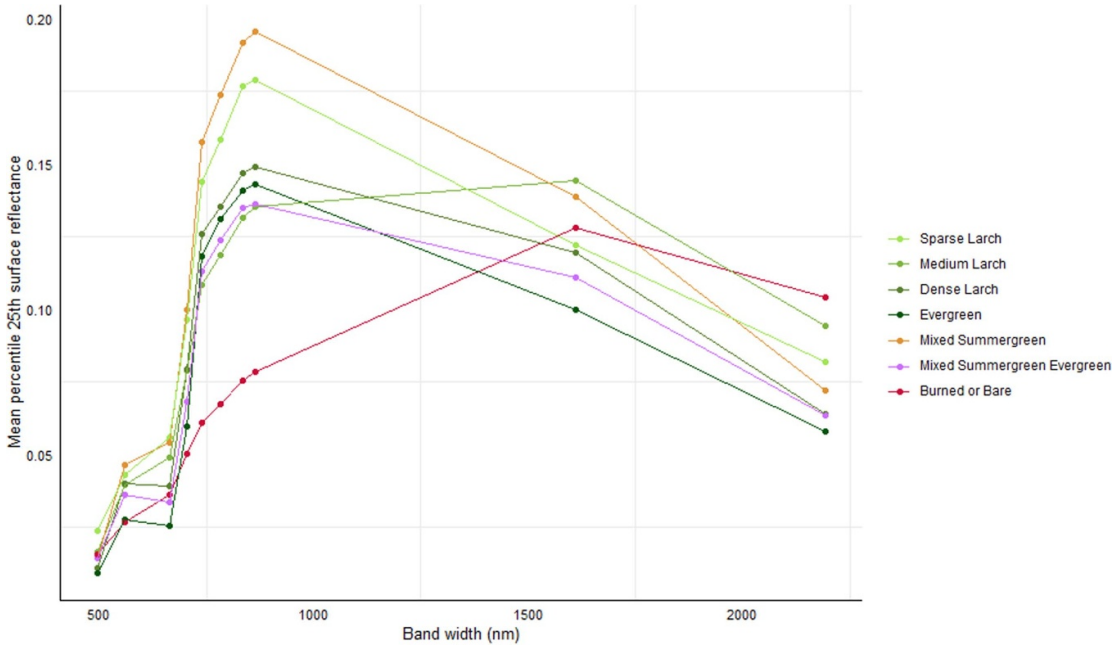

Figure A2.1. Mean percentile of the 25th late summer surface reflectance pixel value for the training pixels of each mapped class.

A1.5. Topographic controls of forest types

The slope was derived with the *ee.Terrain.slope* function in GEE, and represents the degree of inclination of the surface. The TPI was computed using the DEM *focal_mean* function in GEE, which is where the mean elevation within a specific neighborhood of each pixel and centered on that pixel is then subtracted from the original elevation of that pixel. In our case, we used a 15-pixel neighborhood (15 m $\times$ 30 m $\times$ 30 m) to get landscape variations, and we standardized the TPI to a mean of zero and a standard deviation of 1 to avoid a large range of TPI values. The TPI represents the relative elevation compared to its neighborhood, therefore positive TPI represents a ridge and negative values a depression. Finally, the aspect was computed with the *ee.Terrain.aspect* function in GEE, and represents the orientation of the terrain with 0 being North, 90 East, 180 South, and 270 West.

A1.6. Multivariate analysis

GLM is a multivariate statistical method that allows the formation of regression between a response variable and explanatory (predictor) variable(s), and unlike traditional regressions, its parameters do not need to be normally distributed and no assumptions are made about their error distribution [122, 123]. We used GLM with a binomial distribution and a logistic link function, with a binary response where to 1 = presence of the class and 0 = absence. Thus, the GLM returns logistic regression coefficients representing the log odds, as follows [122]:

$$\gamma i = \log\left(\frac{pi}{1-pi}\right) \quad (1)$$

where γi is the linear model, pi is the probability of an event occurring with a given location i. The linear model formed is then a logistic regression of the success or failure of a given binary variable. Therefore, each estimated coefficient is the expected change in the log odds of the class of interest being present (predicted variable = 1) for a unit increase in the corresponding predictor variable holding the other predictor variables constant at a certain value.

Even though probabilities are easier to understand than log odds, we did not convert log odds to probabilities. Log odds have negative or positive values, which reflect the nature of the relationship between the predicted variable and the predictors (a positive coefficient implying a positive relationship and a negative log odds for a negative relationship). Additionally, the magnitude of the log odds informs about the importance of the predictor. Information about the magnitude and direction of the relationship is lost when converting log odds to probabilities. For example, a log odd of +20 would have the same probability as log odds of +4, but the latter shows a weaker relationship to the predictor.

A2. Appendix results

A2.1. Spectral reflectance per class

The averaged late summer surface reflectance spectra show distinct ranges of values for the seven classes (larch classes versus evergreen) (figure A2.1). *Medium Larch* and *Mixed Summergreen-Evergreen* show less distinct surface reflectance curves than the classes *Sparse Larch, Dense Larch,* and *Evergreen* and their spectral values are more diffuse between larch and

evergreen classes. *Evergreen* shows a 'green vegetation' reflectance spectrum with a green peak due to high absorption in the red band, but relatively low NIR reflectance, whereas the summergreen classes show only low photosynthetic activities in late summer by low absorption in the red band but are still distinct by high reflectance in the NIR. The B*urned or bare* class shows a flat reflectance in the visible wavelength range and a high reflectance in the SWIR.

A2.2. Spatial qualitative accuracy assessment

We compared our classified map with two other landcover products: Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 9 (LC100), and ESA WorldCover 10 m v200 (ESA WC) with respectively 100 and 10 m spatial resolution [70, 71]. We selected one subset in each study area (KH, WY, CY, OY, CH) and analyzed the spatial land cover patterns with expert knowledge (figure A2.2). In every region, our map provided significantly more detailed information compared to ESA WC, which exhibits far fewer land cover classes. Below are our observations concerning the comparison with the LC100 map.

(a) **Khamra region (KH):** LC100 provides a detailed landcover map, with evergreen forest spatial patterns similar to our map. However, we mapped more *Evergreen* forests close to Lake Khamra, which we visited in 2018 allowing us to ground-truth the data [52, 124]. Additionally, LC100 presents several pixels as closed forest or as an undefined class while we map detailed mixed classes.

(b) **Western Yakutia (WY):** Our forest-type map provides more detail than LC100, identifying newly burnt areas and different types of forests. LC100 shows mainly closed deciduous forests while we map different deciduous and mixed forest classes.

(c) **Central Yakutia (CY):** LC100 shows mainly closed deciduous needleleaf forests, but we provide more details notably with the correct presence of *Evergreen* forests on both sides of the Lena River, as assessed by field data in 2018 and 2021 [52, 53, 124]. Our forest-type map also shows newly burnt or bare areas.

(d) **Oymyakon region (OY):** LC100 maps a significant amount of undefined closed forest, while we show more sparse and dense larch, as well as identified burnt or bare areas.

(e) **Chukotka (CK):** The region around Lake Illyrney appears with more forest in the SW with LC100, but not around the lake, where shrubs are identified. We identified mainly *Sparse Larch* around the lake, where we have ground-truthed samples that match our classified map [52, 124]

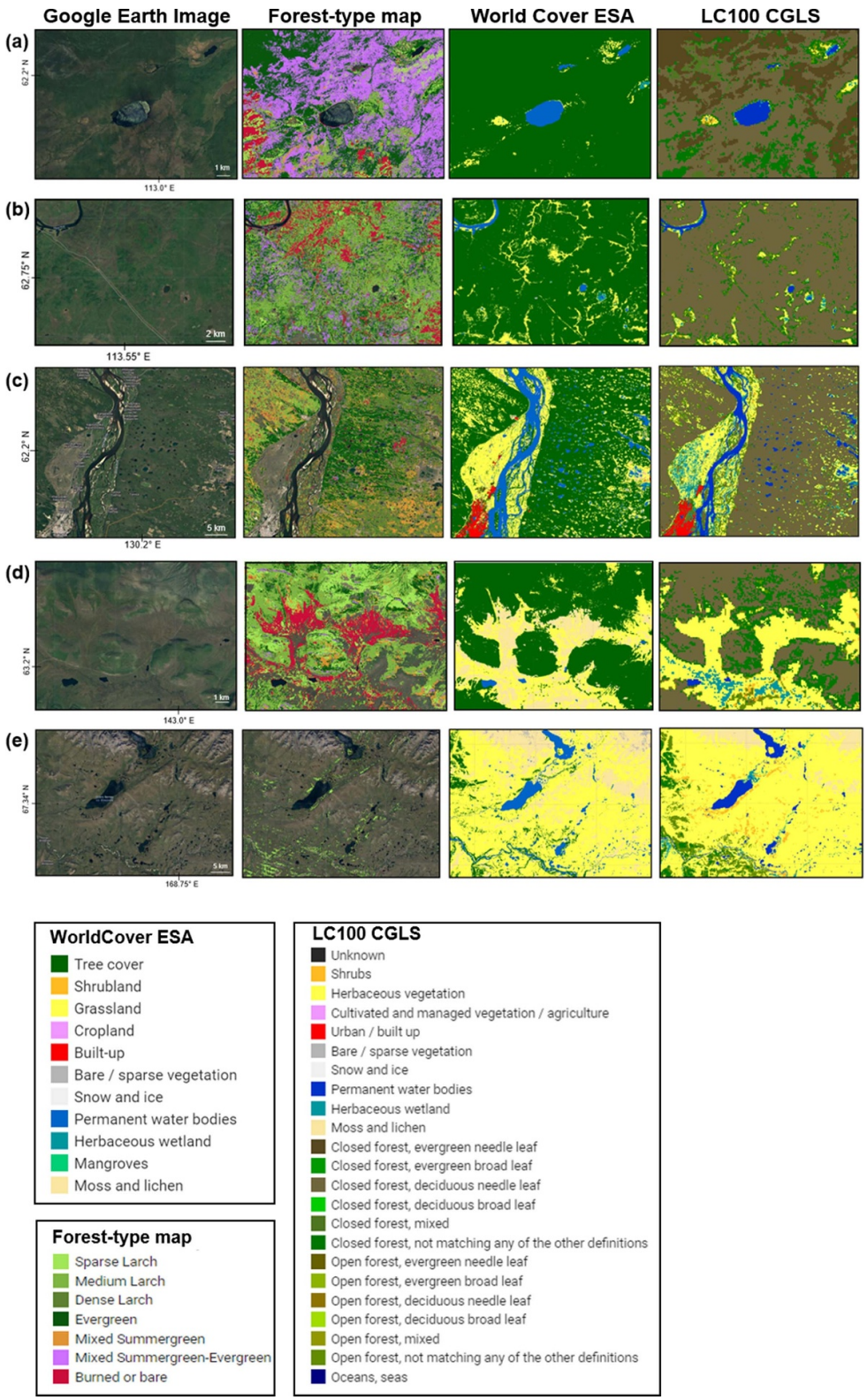

Figure A2.2. Comparison of our classified forest type map results with two global land cover products (ESA World cover and LC100 CGLS) in regions of expert knowledge. (a) Lake Khamra region in southwest Yakutia, (b) Mirny region in Western Yakutia, (c) Lena River and the western and eastern banks in Central Yakutia (d) Verkoyansk mountains close to Oymyakon, and (e) Lake Illyrney in the Chukotka region. Map data: Google, ©2024 TerraMetrics. Map data: Google, ©2024 TerraMetrics. Reproduced from [71]. CC BY 4.0. Reproduced from [70]. CC BY 4.0.

A2.3. Model performance

Table A2.3 shows the *pseudo-*R^2 (pR^2) of each model, based on three different calculation methods from the *pscl R package*: McFadden, Maximum likelihood, and Cragg and Uhler's. A model is commonly well-fitted if pR^2 ranges between 0.2 and 0.4.

Table A2.3. Pseudo-R^2 of each model.

Model name	McFadden	Maximum likelihood	Cragg and Uhler's	
model_loc_c0_Chuk	0.507 635 925	0.409 840 84	0.634 288 727	Local model
model_loc_c2_Chuk	0.021 402 023	0.018 715 284	0.031 917 95	Regional model
model_loc_c3_Chuk	0.127 859 064	0.003 055 098	0.129 197 295	
model_loc_c0_Ulu	0.095 937 937	0.118 370 565	0.161 921 732	
model_loc_c2_Ulu	0.015 5346	0.003 399 695	0.017 270 083	
model_loc_c3_Ulu	0.016 420 959	0.015 147 689	0.025 0269	
model_loc_c0_Yak	0.020 230 599	0.005 753 78	0.023 185 81	
model_loc_c2_Yak	0.001 104 611	0.001 444 766	0.001 979 459	
model_loc_c3_Yak	0.045 057 724	0.048 312 294	0.072 454 479	
model_loc_c0_WesYak	0.027 837 418	0.008 833 662	0.032 365 355	
model_loc_c2_WesYak	0.110 787 748	0.091 221 87	0.157 748 303	
model_loc_c3_WesYak	0.020 159 659	0.025 391 417	0.035 227 321	
model_loc_c0_Kha	0.007 135 776	0.002 299 296	0.008 338 921	
model_loc_c2_Kha	0.016 050 132	0.000 871 038	0.016 482 601	
model_loc_c3_Kha	0.011 695 565	0.015 183 814	0.020 808 386	
model_reg_c0_Chuk	0.649 817 089	0.467 281 394	0.752 963 428	
model_reg_c2_Chuk	0.104 702 97	0.108 049 289	0.162 606 467	
model_reg_c3_Chuk	0.072 102 58	0.003 130 197	0.073 566 113	
model_reg_c0_Ulu	0.115 924 572	0.125 970 888	0.183 371 546	
model_reg_c2_Ulu	0.124 309 763	0.003 946 543	0.126 048 038	
model_reg_c3_Ulu	0.028 477 539	0.024 224 054	0.041 960 423	
model_reg_c0_Yak	0.070 004 475	0.012 113 989	0.075 812 259	
model_reg_c2_Yak	0.119 133 106	0.150 010 618	0.201 508 958	
model_reg_c3_Yak	0.242 247 25	0.174 448 029	0.319 052 646	
model_reg_c0_WesYak	0.129 669 278	0.023 325 497	0.140 167 383	
model_reg_c2_WesYak	0.194 459 858	0.160 177 483	0.270 346 243	
model_reg_c3_WesYak	0.090 432 731	0.103 885 79	0.147 844 071	
model_reg_c0_Kha	0.033 159 579	0.004 712 887	0.035 493 755	
model_reg_c2_Kha	0.150 536 495	0.002 479 774	0.151 593 065	
model_reg_c3_Kha	0.098 273 102	0.126 640 256	0.169 331 086	

ORCID iDs

Léa Enguehard https://orcid.org/0000-0002-2144-8264
Stefan Kruse https://orcid.org/0000-0003-1107-1958
Birgit Heim https://orcid.org/0000-0003-2614-9391
Nicola Falco https://orcid.org/0000-0003-3307-6098

References

[1] Brandt J P, Flannigan M D, Maynard D G, Thompson I D and Volney W J A 2013 An introduction to Canada's boreal zone: ecosystem processes, health, sustainability, and environmental issues *Environ. Rev.* **21** 207–26
[2] Pan Y *et al* 2011 A large and persistent carbon sink in the world's forests *Science* **333** 988–93
[3] Gauthier S, Bernier P, Kuuluvainen T, Shvidenko A Z and Schepaschenko D G 2015 Boreal forest health and global change *Science* **349** 819–22
[4] Bonan G B, Pollard D and Thompson S L 1992 Effects of boreal forest vegetation on global climate *Nature* **359** 716–8
[5] Herzschuh U 2020 Legacy of the Last Glacial on the present-day distribution of deciduous versus evergreen boreal forests *Glob. Ecol. Biogeogr.* **29** 198–206
[6] Rogers B M, Soja A J, Goulden M L and Randerson J T 2015 Influence of tree species on continental differences in boreal fires and climate feedbacks *Nat. Geosci.* **8** 228–34
[7] Shuman J K, Shugart H H and O'Halloran T L 2011 Sensitivity of Siberian larch forests to climate change *Glob. Change Biol.* **17** 2370–84
[8] Hugelius G *et al* 2014 Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps *Biogeosciences* **11** 6573–93
[9] Zhang N, Yasunari T and Ohta T 2011 Dynamics of the larch taiga–permafrost coupled system in Siberia under climate change *Environ. Res. Lett.* **6** 024003
[10] Moncrieff G R, Hickler T and Higgins S I 2015 Intercontinental divergence in the climate envelope of major plant biomes *Glob. Ecol. Biogeogr.* **24** 324–34
[11] Zhu D *et al* 2015 Improving the dynamics of Northern Hemisphere high-latitude vegetation in the ORCHIDEE ecosystem model *Geosci. Model Dev.* **8** 2263–83
[12] Pan Y, Birdsey R A, Phillips O L and Jackson R B 2013 The structure, distribution, and biomass of the world's forests *Annu. Rev. Ecol. Evol. Syst.* **44** 593–622
[13] Woodward F I, Lomas M R and Kelly C K 2004 Global climate and the distribution of plant biomes *Phil. Trans. R. Soc.* B **359** 1465–76
[14] Kuosmanen N, Seppä H, Reitalu T, Alenius T, Bradshaw R H W, Clear J L, Filimonova L, Kuznetsov O and Zaretskaya N 2016 Long-term forest composition and its drivers in taiga forest in NW Russia *Veg. Hist. Archaeobot.* **25** 221–36
[15] Svenning J and Skov F 2004 Limited filling of the potential range in European tree species *Ecol. Lett.* **7** 565–73
[16] Falco N, Wainwright H, Dafflon B, Léger E, Peterson J, Steltzer H, Wilmer C, Rowland J C, Williams K H and Hubbard S S 2019 Investigating microtopographic and soil controls on a mountainous meadow plant community using high-resolution remote sensing and surface geophysical data *J. Geophys. Res. Biogeosci.* **124** 1618–36

[17] Niu Y, Zhou J, Yang S, Chu B, Ma S, Zhu H and Hua L 2019 The effects of topographical factors on the distribution of plant communities in a mountain meadow on the Tibetan Plateau as a foundation for target-oriented management *Ecol. Indic.* **106** 105532
[18] Xu Y, Franklin S B, Wang Q, Shi Z, Luo Y, Lu Z, Zhang J, Qiao X and Jiang M 2015 Topographic and biotic factors determine forest biomass spatial distribution in a subtropical mountain moist forest *For. Ecol. Manage.* **357** 95–103
[19] Wang B, Zhang G and Duan J 2015 Relationship between topography and the distribution of understory vegetation in a Pinus massoniana forest in Southern China *Int. Soil Water Conserv. Res.* **3** 291–304
[20] Lan G, Hu Y, Cao M and Zhu H 2011 Topography related spatial distribution of dominant tree species in a tropical seasonal rain forest in China *For. Ecol. Manage.* **262** 1507–13
[21] Sato H and Kobayashi H 2018 Topography controls the abundance of siberian larch forest *JGR Biogeosci.* **123** 106–16
[22] Larson J, Wallerman J, Peichl M and Laudon H 2023 Soil moisture controls the partitioning of carbon stocks across a managed boreal forest landscape *Sci. Rep.* **13** 14909
[23] Liu Z and Yang J 2014 Quantifying ecological drivers of ecosystem productivity of the early-successional boreal *Larix gmelinii* forest *Ecosphere* **5** 1–16
[24] Zimmermann N E and Kienast F 1999 Predictive mapping of alpine grasslands in Switzerland: species versus community approach *J. Veg. Sci.* **10** 469–82
[25] Gaston K J 2003 *The Structure and Dynamics of Geographic Ranges* (Oxford University Press)
[26] He Y, Huang J, Shugart H H, Guan X, Wang B and Yu K 2017 Unexpected evergreen expansion in the siberian forest under warming hiatus *J. Clim.* **30** 5021–39
[27] Baisheva I *et al* 2024 Late Glacial and Holocene vegetation and lake changes in SW Yakutia, Siberia, inferred from sedaDNA, pollen, and XRF data *Front. Earth Sci.* **12** 1354284
[28] Maharjan S K, Sterck F J, Raes N, Zhao Y and Poorter L 2023 Climate change induced elevational range shifts of Himalayan tree species *Biotropica* **55** 53–69
[29] Chapin F S, Callaghan T V, Bergeron Y, Fukuda M, Johnstone J F, Juday G and Zimov S A 2004 Global change and the boreal forest: thresholds, shifting states or gradual change? *Ambio* **33** 361–5
[30] Pickell P D, Coops N C, Gergel S E, Andison D W and Marshall P L 2016 Evolution of Canada's boreal forest spatial patterns as seen from space *PLoS One* **11** e0157736
[31] Berner L T and Goetz S J 2022 Satellite observations document trends consistent with a boreal forest biome shift *Glob. Change Biol.* **28** 3275–92
[32] Orndahl K M, Macander M J, Berner L T and Goetz S J 2022 Plant functional type aboveground biomass change within Alaska and northwest Canada mapped using a 35-year satellite time series from 1985 to 2020 *Environ. Res. Lett.* **17** 115010
[33] Margolis H A, Nelson R F, Montesano P M, Beaudoin A, Sun G, Andersen H-E and Wulder M A 2015 Combining satellite lidar, airborne lidar, and ground plots to estimate the amount and distribution of aboveground biomass in the boreal forest of North America *Can. J. For. Res.* **45** 838–55
[34] Lechner A M, Foody G M and Boyd D S 2020 Applications in remote sensing to forest ecology and management *One Earth* **2** 405–12
[35] Xie Y, Sha Z and Yu M 2008 Remote sensing imagery in vegetation mapping: a review *J. Plant Ecol.* **1** 9–23
[36] Roy D P *et al* 2014 Landsat-8: science and product vision for terrestrial global change research *Remote Sens. Environ.* **145** 154–72
[37] Berra E F and Gaulton R 2021 Remote sensing of temperate and boreal forest phenology: a review of progress, challenges and opportunities in the intercomparison of *in-situ* and satellite phenological metrics *For. Ecol. Manage.* **480** 118663
[38] Hościło A and Lewandowska A 2019 Mapping forest type and tree species on a regional scale using multi-temporal Sentinel-2 data *Remote Sens.* **11** 929
[39] Hemmerling J, Pflugmacher D and Hostert P 2021 Mapping temperate forest tree species using dense Sentinel-2 time series *Remote Sens. Environ.* **267** 112743
[40] Abdi A M 2020 Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data *GISci. Remote Sens.* **57** 1–20
[41] Buchhorn M, Lesiv M, Tsendbazar N-E, Herold M, Bertels L and Smets B 2020 Copernicus global land cover layers—collection 2 *Remote Sens.* **12** 1044
[42] ESA 2017 Land cover CCI product user guide version 2
[43] Defourny P *et al* 2023 Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics (in preparation)
[44] Kuuluvainen T 1994 Gap disturbance, ground microtopography, and the regeneration dynamics of boreal coniferous forests in Finland: a review *Ann. Zool. Fenn.* **31** 35–51
[45] Johnson C E, Ruiz-Méndez J J and Lawrence G B 2000 Forest soil chemistry and terrain attributes in a catskills watershed *Soil Sci. Soc. Am. J.* **64** 1804–14
[46] Rees W G, Tomaney J, Tutubalina O, Zharko V and Bartalev S 2021 Estimation of boreal forest growing stock volume in Russia from Sentinel-2 MSI and land cover classification *Remote Sens.* **13** 4483
[47] Bartsch A, Höfler A, Kroisleitner C and Trofaier A 2016 Land cover mapping in northern high latitude permafrost regions with satellite data: achievements and remaining challenges *Remote Sens.* **8** 979
[48] Wan H, Tang Y, Jing L, Li H, Qiu F and Wu W 2021 Tree species classification of forest stands using multisource remote sensing data *Remote Sens.* **13** 144
[49] Wang X, Shi J, Wang C, Gao C and Zhang F 2023 Remote sensing inversion and mapping of typical forest stand age in the loess plateau *Remote Sens.* **15** 5581
[50] Stavrakoudis D, Dragozi E, Gitas I and Karydas C 2014 Decision fusion based on hyperspectral and multispectral satellite imagery for accurate forest species mapping *Remote Sens.* **6** 6897–928
[51] Bolyn C, Lejeune P, Michez A and Latte N 2022 Mapping tree species proportions from satellite imagery using spectral–spatial deep learning *Remote Sens. Environ.* **280** 113205
[52] Kruse S *et al* 2019 Russian-german cooperation: expeditions to Siberia in 2018
[53] Morgenstern A *et al* 2023 Russian-German cooperation: expeditions to Siberia in 2021
[54] Enguehard L, Heim B, Kruse S, Frandsen P, Jackisch R and Herzschuh U 2024 Labelled Siberian boreal forest types shapefiles based on field surveys (https://doi.org/10.1594/PANGAEA.964699)
[55] Ohta T 2010 Hydrological aspects in a Siberian larch forest *Permafrost Ecosystems* ed A Osawa, O A Zyryanova, Y Matsuura, T Kajimoto and R W Wein (Springer Netherlands) pp 245–69
[56] Burton P J *et al* 2010 *Forests and Society—Responding to Global Drivers of Change G* Mery *et al* ed (International Union of Forest Research Organizations)
[57] Fick S E and Hijmans R J 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas *Int. J. Climatol.* **37** 4302–15
[58] Ohta T *et al* 2008 Interannual variation of water balance and summer evapotranspiration in an eastern Siberian larch forest over a 7-year period (1998–2006) *Agric. For. Meteorol.* **148** 1941–53
[59] Fujinami H, Yasunari T and Watanabe T 2016 Trend and interannual variation in summer precipitation in eastern Siberia in recent decades *Int. J. Climatol.* **36** 355–68

[60] Fukutomi Y, Igarashi H, Masuda K and Yasunari T 2003 Interannual variability of summer water balance components in three major river basins of Northern Eurasia *J. Hydrometeorol.* **4** 283–96
[61] Serreze M C and Etringer A J 2003 Precipitation characteristics of the Eurasian Arctic drainage system *Int. J. Climatol.* **23** 1267–91
[62] Osawa A and Zyryanova O A 2010 Introduction *Permafrost Ecosystems* ed A Osawa, O A Zyryanova, Y Matsuura, T Kajimoto and R W Wein (Springer Netherlands) pp 3–15
[63] Isaev A P *et al* 2010 Vegetation of yakutia: elements of ecology and plant sociology E I Troeva, A P Isaev, M M Cherosov and N S Karpov ed *The Far North* vol 3 (Springer Netherlands) pp 143–260
[64] Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R 2017 Google Earth Engine: planetary-scale geospatial analysis for everyone *Remote Sens. Environ.* **202** 18–27
[65] Bourgeau-Chavez L, Endres S, Battaglia M, Miller M, Banda E, Laubach Z, Higman P, Chow-Fraser P and Marcaccio J 2015 Development of a bi-national great lakes coastal wetland and land use map using three-season PALSAR and landsat imagery *Remote Sens.* **7** 8655–82
[66] Fu B, Wang Y, Campbell A, Li Y, Zhang B, Yin S, Xing Z and Jin X 2017 Comparison of object-based and pixel-based Random Forest algorithm for wetland vegetation mapping using high spatial resolution GF-1 and SAR data *Ecol. Indic.* **73** 105–17
[67] Saini R and Ghosh S K 2018 Exploring capabilities of Sentinel-2 for vegetation mapping using random forest *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLII–3** 1499–502
[68] Mohammadpour P, Viegas D X and Viegas C 2022 Vegetation mapping with random forest using Sentinel 2 and GLCM texture feature—a case study for Lousã region, portugal *Remote Sens.* **14** 4585
[69] Breiman L 2001 Random Forest *Machine Learning* (Dordrecht Kluwer Academic Publishers) pp 5–32
[70] Buchhorn M, Smets B, Bertels L, Roo B, Lesiv M, Tsendbazar N E, Herold M and Fritz S 2020 Copernicus global land service: land cover 100m: collection 3: epoch 2019: globe (version V3.0.1)
[71] Zanaga D *et al* 2022 ESA WorldCover 10 m 2021 (version v200) (https://doi.org/10.5281/zenodo.7254221)
[72] Westermann S *et al* 2024 ESA permafrost climate change initiative (permafrost_cci): permafrost ground temperature for the Northern Hemisphere (https://doi.org/10.5285/20ec12f5d1f94e99aff2ed796264ee65)
[73] European Space Agency 2022 Copernicus DEM GLO-30: global digital elevation model (https://doi.org/10.5270/ESA-c5d3d65)
[74] Weiss A 2001 Topographic position and landforms analysis Poster presentation
[75] Guirado E, Delgado-Baquerizo M, Benito B M, Molina-Pardo J L, Berdugo M, Martínez-Valderrama J and Maestre F T 2023 The global biogeography and environmental drivers of fairy circles *Proc. Natl Acad. Sci. USA* **120** e2304032120
[76] Chefaoui R M, Casado-Amezúa P and Templado J 2017 Environmental drivers of distribution and reef development of the Mediterranean coral *Cladocora caespitosa Coral Reefs* **36** 1195–209
[77] Terribile L C, Diniz-Filho J A F, Rodríguez M Á and Rangel T F L V B 2009 Richness patterns, species distributions and the principle of extreme deconstruction *Glob. Ecol. Biogeogr.* **18** 123–36
[78] Adámek M, Jankovská Z, Hadincová V, Kula E and Wild J 2018 Drivers of forest fire occurrence in the cultural landscape of Central Europe *Landsc. Ecol.* **33** 2031–45
[79] Hedwall P-O, Uria-Diez J, Brunet J, Gustafsson L, Axelsson A-L and Strengbom J 2021 Interactions between local and global drivers determine long-term trends in boreal forest understorey vegetation *Glob. Ecol. Biogeogr.* **30** 1765–80
[80] Nelder J A and Wedderburn R W M 1972 Generalized linear models *J. R. Stat. Soc.* **135** 370–284
[81] R Core Team 2023 _R: a language and environment for statistical computing_ *R Foundation for Statistical Computing 2023*
[82] Bates D, Mächler M, Bolker B and Walker S 2015 Fitting linear mixed-effects models using lme4 *J. Stat. Softw.* **67** 1–48
[83] Hartig F 2022 DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models
[84] Jackman S 2020 pscl: classes and methods for r developed in the political science computational laboratory
[85] Pearson R G and Dawson T P 2003 Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* **12** 361–71
[86] Scheffer M, Hirota M, Holmgren M, Van Nes E H and Chapin F S 2012 Thresholds for boreal biome transitions *Proc. Natl Acad. Sci. USA* **109** 21384–9
[87] Cao X, Tian F, Dallmeyer A and Herzschuh U 2019 Northern Hemisphere biome changes (>30°N) since 40 cal ka BP and their driving factors inferred from model-data comparisons *Quat Sci Rev* **220** 291–309
[88] Wilmking M, Sanders T G M, Zhang Y, Kenter S, Holzkämper S and Crittenden P D 2012 Effects of climate, site conditions, and seed quality on recent treeline dynamics in NW Russia: permafrost and lack of reproductive success hamper treeline advance? *Ecosystems* **15** 1053–64
[89] Overpeck J *et al* 1997 Arctic environmental change of the last four centuries *Science* **278** 1251–6
[90] Serreze M C, Walsh J E, Iii F S C, Osterkamp T, Dyurgerov M, Romanovsky V, Oechel W C, Morison J, Zhang T and Barry R G 2000 Observational evidence of recent change in the northern high-latitude environment *Clim. Change* **46** 159–207
[91] Previdi M, Smith K L and Polvani L M 2021 Arctic amplification of climate change: a review of underlying mechanisms *Environ. Res. Lett.* **16** 093003
[92] Soja A J, Tchebakova N M, French N H F, Flannigan M D, Shugart H H, Stocks B J, Sukhinin A I, Parfenova E I, Chapin F S and Stackhouse P W 2007 Climate-induced boreal forest change: predictions versus current observations *Glob. Planet. Change* **56** 274–96
[93] Montesano P M, Neigh C S R, Macander M, Feng M and Noojipady P 2020 The bioclimatic extent and pattern of the cold edge of the boreal forest: the circumpolar taiga-tundra ecotone *Environ. Res. Lett.* **15** 105019
[94] Shevtsova I, Heim B, Kruse S, Schröder J, Troeva E I, Pestryakova L A, Zakharov E S and Herzschuh U 2020 Strong shrub expansion in tundra-taiga, tree infilling in taiga and stable tundra in central Chukotka (north-eastern Siberia) between 2000 and 2017 *Environ. Res. Lett.* **15** 085006
[95] Klinge M, Schneider F, Dulamsuren C, Arndt K, Bayarsaikhan U and Sauer D 2021 Interrelations between relief, vegetation, disturbances, and permafrost in the forest-steppe of central Mongolia *Earth Surf. Process. Landf.* **46** 1766–82
[96] Schulte L, Li C, Lisovski S and Herzschuh U 2022 Forest-permafrost feedbacks and glacial refugia help explain the unequal distribution of larch across continents *J. Biogeogr.* **49** 1825–38
[97] Glückler R, Gloy J, Dietze E, Herzschuh U and Kruse S 2024 Simulating long-term wildfire impacts on boreal forest structure in Central Yakutia, Siberia, since the last glacial maximum *Fire Ecol.* **20** 1
[98] Shvetsov E G, Kukavskaya E A, Buryak L V and Barrett K 2019 Assessment of post-fire vegetation recovery in Southern Siberia using remote sensing observations *Environ. Res. Lett.* **14** 055001

[99] Glückler R, Herzschuh U, Kruse S, Andreev A, Vyse S A, Winkler B, Biskaborn B K, Pestryakova L and Dietze E 2021 Wildfire history of the boreal forest of south-western Yakutia (Siberia) over the last two millennia documented by a lake-sediment charcoal record *Biogeosciences* **18** 4185–209
[100] Tsendbazar N-E *et al* 2020 Copernicus global land service: land cover 100m: version 3 globe 2015–2019: validation report *Zenodo* (https://doi.org/10.5281/ZENODO.3938974)
[101] Loranty M M, Alexander H D, Kropp H, Talucci A C and Webb E E 2021 Siberian ecosystems as drivers of cryospheric climate feedbacks in the terrestrial arctic *Front. Clim.* **3** 730943
[102] Loranty M M, Lieberman-Cribbin W, Berner L T, Natali S M, Goetz S J, Alexander H D and Kholodov A L 2016 Spatial variation in vegetation productivity trends, fire disturbance, and soil carbon across arctic-boreal permafrost ecosystems *Environ. Res. Lett.* **11** 095008
[103] Loranty M M *et al* 2018 Reviews and syntheses: changing ecosystem influences on soil thermal regimes in northern high-latitude permafrost regions *Biogeosciences* **15** 5287–313
[104] Schulze E-D, Wirth C, Mollicone D, von Lüpke N, Ziegler W, Achard F, Mund M, Prokushkin A and Scherbina S 2012 Factors promoting larch dominance in central Siberia: fire versus growth performance and implications for carbon dynamics at the boundary of evergreen and deciduous conifers *Biogeosciences* **9** 1405–21
[105] Gromtsev A 2002 Natural disturbance dynamics in the boreal forests of European Russia: a review *Silva Fenn.* **36** 41–55
[106] Antão L H *et al* 2022 Climate change reshuffles northern species within their niches *Nat. Clim. Change* **12** 587–92
[107] Wieczorek M, Kruse S, Epp L S, Kolmogorov A, Nikolaev A N, Heinrich I, Jeltsch F, Pestryakova L A, Zibulski R and Herzschuh U 2017 Dissimilar responses of larch stands in northern Siberia to increasing temperatures-a field and simulation based study *Ecology* **98** 2343–55
[108] Pearson R G, Phillips S J, Loranty M M, Beck P S A, Damoulas T, Knight S J and Goetz S J 2013 Shifts in Arctic vegetation and associated feedbacks under climate change *Nat. Clim. Change* **3** 673–7
[109] Kruse S, Gerdes A, Kath N J, Epp L S, Stoof-Leichsenring K R, Pestryakova L A and Herzschuh U 2019 Dispersal distances and migration rates at the arctic treeline in Siberia—a genetic and simulation-based study *Biogeosciences* **16** 1211–24
[110] Lloyd A H and Fastie C L 2002 Spatial and temporal variability in the growth and climate response of treeline trees in Alaska *Clim. Change* **52** 481–509
[111] Stocker B D, Roth R, Joos F, Spahni R, Steinacher M, Zaehle S, Bouwman L, Ri X and Prentice I C 2013 Multiple greenhouse-gas feedbacks from the land biosphere under future climate change scenarios *Nat. Clim. Change* **3** 666–72
[112] Rouse J W, Haas R H, Schell J A, Deering D W and Harlan J C 1974 Monitoring the vernal advancement and retrogradation (greenwave effect) of natural vegetation (Remote Sensing Center Texas A&M University College Station)
[113] Hansen M C *et al* 2013 High-resolution global maps of 21st-century forest cover change *Science* **342** 850–3
[114] Benhammou Y, Alcaraz-Segura D, Guirado E, Khaldi R, Achchab B, Herrera F and Tabik S 2022 Sentinel2GlobalLULC: a Sentinel-2 RGB image tile dataset for global land use/cover mapping with deep learning *Sci. Data* **9** 681
[115] Corbane C, Politis P, Kempeneers P, Simonetti D, Soille P, Burger A, Pesaresi M, Sabo F, Syrris V and Kemper T 2020 A global cloud free pixel- based image composite from Sentinel-2 data *Data Briefs* **31** 105737
[116] Simonetti D, Pimple U, Langner A and Marelli A 2021 Pan-tropical Sentinel-2 cloud-free annual composite datasets *Data Briefs* **39** 107488
[117] Feng M, Sexton J O, Wang P, Channan S, Montesano P M and Wagner W, Wooten M and Neigh C S 2022 ABoVE: Tree Canopy Cover and Stand Age from Landsat, Boreal Forest Biome, 1984-2020 (https://doi.org/10.3334/ORNLDAAC/2012)
[118] Brieger F, Herzschuh U, Pestryakova L A, Bookhagen B, Zakharov E S and Kruse S 2019 Advances in the derivation of Northeast Siberian forest metrics using high-resolution UAV-based photogrammetric point clouds *Remote Sens.* **11** 1447
[119] Sadaiyandi J, Arumugam P, Sangaiah A K and Zhang C 2023 Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset *Electronics* **12** 4423
[120] Barsi Á, Kugler Z, László I, Szabó G and Abdulmutalib H M 2018 Accuracy dimensions in remote sensing *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLII–3** 61–67
[121] Taubenbock H, Esch T, Felbier A, Roth A and Dech S 2011 Pattern-based accuracy assessment of an urban footprint classification using TerraSAR-X data *IEEE Geosci. Remote Sens. Lett.* **8** 278–82
[122] Zuur A F, Ieno E N, Walker N J, Saveliev A A and Smith G M 2009 GLM and GAM for absence–presence and proportional data *Mixed Effects Models and Extensions in Ecology with R* pp 245–59 (Springer)
[123] Atkinson P M, German S E, Sear D A and Clark M J 2003 Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression *Geogr. Anal.* **35** 58–82
[124] Miesner T *et al* Tree data set from forest inventories in north-eastern Siberia 2022 (https://doi.org/10.1594/PANGAEA.943547)