Environmental Research Letters

LETTER • OPEN ACCESS

Identifying multidisciplinary research gaps across Arctic terrestrial gradients

To cite this article: A-M Virkkala *et al* 2019 *Environ. Res. Lett.* **14** 124061

IOP Publishing Environ. Res. Lett. 14 (2019) 124061 https://doi.org/10.1088/1748-9326/ab4291

Environmental Research Letters

OPEN ACCESS

RECEIVED
18 March 2019

REVISED
22 August 2019

ACCEPTED FOR PUBLICATION
9 September 2019

PUBLISHED
16 December 2019

LETTER

Identifying multidisciplinary research gaps across Arctic terrestrial gradients

A-M Virkkala[1,4], A M Abdi[2,3], M Luoto[1] and D B Metcalfe[2]

[1] University of Helsinki, Department of Geosciences and Geography, Gustaf Hällströmin katu 2, FI-00014, Finland
[2] Lund University, Department of Physical Geography and Ecosystem Science, Sölvegatan 12, SE-22236, Sweden
[3] Lund University, Centre for Environmental and Climate Research, Sölvegatan 37, SE-22236, Sweden
[4] Author to whom any correspondence should be addressed.

E-mail: anna-maria.virkkala@helsinki.fi

Keywords: Arctic, tundra, high-latitude, environmental science, review, representativeness, sampling strategy

Supplementary material for this article is available online

Abstract

Global warming is driving environmental change in the Arctic. However, our current understanding of this change varies strongly among different environmental disciplines and is limited by the number and distribution of field sampling locations. Here, we use a quantitative framework based on multivariate statistical modeling to present the current state of sampling across environmental disciplines in the Arctic. We utilize an existing database of georeferenced Arctic field studies to investigate how sampling locations and citations of disciplines are distributed across Arctic topographical, soil and vegetation conditions, and highlight critical regions for potential new research areas in different disciplines. Continuous permafrost landscapes, and the northernmost Arctic bioclimatic zones are studied and cited the least in relation to their extent in many disciplines. We show that the clusters of sampling locations and citations are not uniform across disciplines. Sampling locations in Botany and Biogeochemistry cover environmental gradients the best, and Microbiology, Meteorology, Geosciences And Geographic Information Systems/remote Sensing/Modeling have the worst coverage. We conclude that across all disciplines, more research is needed particularly in the Canadian Arctic Archipelago, northern Greenland, central and eastern Siberia, and in some disciplines, in Canadian mainland, central Alaska, western Siberia and northern Taimyr region. We provide detailed maps of potential new sampling locations for each environmental discipline that consider multiple variables simultaneously. These results will help prioritize future research efforts, thus increasing our knowledge about the Arctic environmental change.

1. Introduction

Global warming is driving environmental change in the Arctic (IPCC 2013, AMAP 2017). This change encompasses profound shifts in soil conditions (Schuur *et al* 2015, Biskaborn *et al* 2019), species distributions (Pearson *et al* 2013, Myers-Smith and Hik 2018), and ecosystem functioning (Bond-Lamberty *et al* 2018, Keenan and Riley 2018). There is, however, large spatial variation in environmental conditions across the Arctic, and the response of different environments to climate warming can be highly variable (Phoenix and Bjerke 2016, Lara *et al* 2018). Therefore, sampling that adequately represents this variation is crucial to accurately understand ecosystem functioning across the Arctic as a whole.

There is a growing interest in efforts to synthesize the current extent of sampling locations and where new locations for environmental research are needed (Yang *et al* 2008, Hoffman *et al* 2013, Kumar *et al* 2016). The availability of high-resolution and spatially-explicit environmental data sets has greatly accelerated these efforts, because landscape variation of the Arctic can be better understood and readily visualized (Fick and Hijmans 2017, Hengl *et al* 2017). Resources and accessibility strongly constrain Arctic research, thus it is imperative to efficiently maximize scientific coverage of environmental conditions (IPCC 2013, Kulmala 2018). Sampling strategies and network representativeness mapping have been conducted for very specific fields (Kumar *et al* 2016), or at regional scales (Hoffman *et al* 2013), but not for entire disciplines at larger scales.

Here, we use a quantitative framework based on multivariate statistical modeling to present the current state of sampling across environmental science disciplines in the Arctic. We utilize an existing database of field studies across the Arctic that was developed by Metcalfe *et al* (2018). We build upon this earlier work with a more comprehensive and detailed investigation of how locations and citations within different disciplines are distributed across Arctic topographical, soil and vegetation conditions, and provide recommendations for potential new study areas in different disciplines. We focus our research on nine broad disciplines: Botany, Zoology, Microbiology, Soil Science, Biogeochemistry, Meteorology, Geosciences, Paleosciences, and Geographic Information Systems (GIS)/Remote Sensing (RS)/Modeling.

2. Methods

2.1. Literature review and database

The database of Arctic studies collected by Metcalfe *et al* (2018) consists of all primary field studies in the terrestrial Arctic published within the period of 1951–2015 with a minimum of one citation generated from keyword searches for 'arctic', 'subarctic' and 'sub-arctic' in the Web of Science. Some sampling locations from syntheses were included because their data remained unpublished. The Arctic was defined as all land north of the Arctic Circle (66.3 N). The total number of scientific articles and field sampling locations extracted were 1817 and 6237, respectively. From each article, geographic coordinates of field observations were extracted. Throughout the text, we use the term sampling location to describe field sampling locations that were reported for each field observation in a study. We extracted the coordinates that a paper presented, thus we used the effective resolution the authors chose in each paper. Sometimes a study included several field observations, but reported only one general sampling location. We also noted the primary discipline/s within environmental sciences featured in the article. These disciplines were then categorized into Botany, Zoology, Microbiology, Soil science, Biogeochemistry, Meteorology, Geosciences, Paleosciences, GIS/RS/Modeling, allowing each sampling location to belong to several disciplines simultaneously due to the multi-disciplinary nature of some studies. If a study had multiple sampling locations, article citations were divided by the number of locations to avoid replicating the total citation number for each location of the study. Citations for all articles are up to the year 2015.

2.2. Data extraction and preparation

Biogeophysical information for each sampling location was obtained using open-access spatial data that describe Arctic terrestrial systems (table 1). The geographic extent of these spatial data was limited to non-glaciated areas. Data extraction was performed from shapefiles and rasters in their original resolution (ca. 1 km) and projection (WGS 1984 or Lambert Azimuthal Equal Area projection) with *raster* package (Hijmans *et al* 2018) in R (R Core Team 2018). If a location was outside the geographic limits of the spatial data, the closest cell value was chosen instead. However, if a location was more than one degree latitude from the limits of the data (e.g. in central Greenland), it was given a 'No Data' value (11 sampling locations in the database). Thus, the final database that we used for the analysis consisted of 6226 sampling locations.

For the predictions, continuous spatial data were resampled to a 1 km resolution (0.0083°) with bilinear interpolation. All data were reprojected to WGS 1984 and cropped to the same extent. Resampling, projecting and cropping of spatial prediction datasets was done in ArcMap (ESRI 2018).

We used a list of INTERACT stations (https://eu-interact.org/field-sites/, appendix A9 is available online at stacks.iop.org/ERL/14/124061/mmedia) to visualize the distribution of existing infrastructure across the Arctic. INTERACT is an infrastructure project with a circumarctic network of 86 terrestrial stations in Arctic and alpine regions, offering information of and connections to stations. Out of these 86 stations, 34 were located within our study domain.

2.3. Data analysis

We studied the distribution of sampling locations and citations across topographical, soil and vegetation conditions, and environmental science disciplines to reveal understudied conditions across the Arctic. First, we analyzed differences in number of sampling locations or citations. We divided them by the spatial extent of the zone across bioclimatic zones, ecoregions, and permafrost zones, as we assume that sampling and citations should be proportionate to spatial extent of the condition to achieve a full understanding of environmental variability (Hirzel and Guisan 2002). Then, we examined the distribution of sampling locations and citations across MAGT (mean annual ground temperatures)—SOC (soil organic carbon stocks) and soil pH—NDVI (normalized difference vegetation index) realms showing the whole Arctic conditions, and conditions of the sampling locations and citations. To describe the Arctic conditions, we took a random sample ($n = 10\,000$) of the total pixels above the Arctic circle in the GIS data sets.

Table 1. The datasets used in the figures. Areas with permanent ice were removed from each dataset using the glacier outline data Natural Earth (Patterson and Kelos 2009).

Topic	Data set	Importance in the Arctic	Data set description	Extent and resolution	Access
Soil	Soil organic carbon stocks (SOC)	The Arctic SOC stocks are an important part of the carbon cycle (Hugelius *et al* 2014)	SOC stocks at 0–200 cm, in tons per ha. Based on statistical modeling of SOC observations.	Global, 1 km	Hengl *et al* (2014, 2017)
	Soil pH	pH is a proxy for nutrient concentrations of soils (Gough *et al* 2000)	Topsoil (0 cm) pH. Based on statistical modeling of pH observations.	Global, 1 km	Hengl *et al* (2014, 2017)
Permafrost	Mean annual ground temperatures (MAGT)	Soil temperatures drive multiple ecosystem processes (Groendahl *et al* 2007)	MAGT for 2000–2014. Based on statistical modeling of MAGT observations.	Circumpolar, 1 km	Aalto *et al* (2018a, 2018b)
	Permafrost zone	Degradation of permafrost can impose changes in e.g. biogeochemical cycles (Biskaborn *et al* 2019)	MAGT >0 °C no permafrost, -2–0 °C discontinuous, < -2°C continuous (Westermann *et al* 2015).	Circumpolar, 1 km	Aalto *et al* (2018a, 2018b)
Vegetation	Bioclimatic zones from Circumpolar Arctic Vegetation Map (CAVM)	Large-scale climate and vegetation patterns	Bioclimatic zones cropped to 66.3. area, a new sub-Arctic zone added south of the zones until 66.3 latitude.	Circumpolar	CAVM (2003), Walker *et al* (2005)
	Ecoregions	Ecologically uniform areas, reflect the distribution of biota	31 classes of ecologically and geographically defined areas in the Arctic.	Global	Olson *et al* (2001), The Nature Conservancy (2009)
	NDVI	Vegetation index describes vegetation productivity and carbon uptake (Tucker *et al* 2005, Street *et al* 2007)	Modis product (MOD13A2), June–July–August mean NDVI 2000–2014.	Global, 1 km	Didan (2015)
Topography	Digital elevation model (DEM)	Topography affects ecosystem processes (Sundqvist *et al* 2013)	GMTED2010 is an elevation dataset for global and continental scale applications.	Global, 1 km	Amatulli *et al* (2018), Danielson and Gesch (2011)
	Topographic wetness index (TWI)	Soil moisture impacts many ecosystem processes (Natali *et al* 2015)	TWI quantifies the influence of topography on hydrological processes. It is calculated from GMTED2010 (appendix A10).	Global, 1 km	Amatulli *et al* (2018), Danielson and Gesch (2011)

The aforementioned exploratory analysis was visualized in R using the *ggplot2* package (Wickham *et al* 2018).

We used statistical multivariate modeling to highlight areas lacking sampling locations when considering overall topographical, soil and vegetation variability. Estimating the representativeness of observation networks can be conducted in several ways (Kumar *et al* 2016). Previous research applied clustering analysis together with Euclidean distances to describe representativeness either with an ecoregion- or point-based approach (Hoffman *et al* 2013). Here, we used a generalized boosted regression model (GBM) from the boosted regression tree family to predict whether an area has environmental conditions that are represented by the current sampling network. GBM is a machine learning method based on an extension of AdaBoost algorithm (Freund and Schapire 1997) and gradient boosting machines (Friedman 2001). Data are split internally multiple times into training and evaluation sets, and trees are built recursively using the information from previous trees (Elith *et al* 2008). GBMs have been widely used in environmental science research (Marmion *et al* 2009, Buri *et al* 2017, Nussbaum *et al* 2018), because they consider interaction effects between predictors and can model non-linear relationships (Elith *et al* 2008). We used the 'Bernoulli' error distribution of the response variable as we were working with a binomial presence-absence data (1 = sampling location exists, 0 = sampling location is missing), and soil (SOC, pH, MAGT), vegetation (NDVI) and topography (DEM, TWI) as explanatory variables. Additionally, interaction depth was set to 3, number of trees to 200, and minimum number of observations in the terminal nodes of a tree to 10.

Since our database contains information about sampling locations only, we needed to artificially create locations with absence of sampling. We followed the methodology suggested by Barbet-Massin *et al* (2012) and created a random sample of terrestrial absence locations with same number of observations as our presence locations ($n = 6226$) with the *sp* package (Pebesma and Bivand 2018). No absences were created in areas with permanent ice. A 10 km buffer was created around the presence locations to avoid creating absences within their vicinity. Then, we obtained spatial data in these randomly sampled locations based on coordinate colocation. These were then combined with the literature database, which resulted in a data frame of 12 452 locations. The artificially created absences belonged to all disciplines. The predictors in the final data set did not suffer from high multicollinearities, as the correlations between the predictor variables was <0.75.

We ran the model with *gbm* (Greenwell *et al* 2019) both with the complete dataset of locations of all disciplines and separately for each discipline while consistently taking a random sample of absence data of the same size as the presence data ($n = 600$–4000). In general, machine learning models trained with larger data sets (e.g. in Botany) suffer less from overfitting the data, and are thus more reliable than models trained with a smaller data set. The model predicts both the presence-absence of sampling locations and the probabilities for the presence, of which the latter was used to describe the representativeness of sampling locations for each raster pixel across the whole Arctic. In the final map, high probabilities indicate a relatively good coverage of current sampling locations in similar conditions (1 = high probability that there is a sampling location in similar conditions), and low probabilities suggest lack of locations. This prediction cannot provide exact aerial estimates of under-sampled regions, but it provides a qualitative map to visually inspect the differences in representativeness across the Arctic.

We used cross-validation with 99 permutations and calculated the area under the curve (AUC) test statistic (Hanley and McNeil 1982) to evaluate model predictive performance with the *ROCR* package (Sing *et al* 2009). In the cross-validation procedure, a random sample of 70% of the data was used to test the model fit, and the remaining 30% were used to assess predictive performance. Test statistics were calculated after each permutation to evaluate the models. An AUC value of 1 represents perfect accuracy and 0.5 indicates that the model is no better than random.

3. Results

3.1. The extent of studies across the Arctic and in permafrost, CAVM and ecoregion classes

None of the disciplines have a uniform geographical distribution of sampling locations or citations across the Arctic (appendices A1 and A2). Most disciplines have the highest number of locations in a few regions in Alaska and Sweden, with some smaller clusters in northern Canada (Biogeochemistry, Zoology), central Siberia (Microbiology), western Russia (Botany, Biogeochemistry, Meteorology), and Svalbard (Soil science, Biogeochemistry, Meteorology). Some disciplines cover Alaska geographically relatively well (Botany, Paleosciences, GIS/RS/Modeling), and in others Canada is either sparsely covered or not covered at all (e.g. Microbiology, Meteorology, Geosciences). The citations do not follow the same pattern as sampling locations as they are even more concentrated within a few regions. In addition to Sweden and Alaska, there are highly localized clusters in Siberia (Botany, Zoology, Microbiology, Soil science, Biogeochemistry, Meteorology, Geosciences), Greenland (Microbiology, Soil science) and Canada (Botany, Biogeochemistry, GIS/RS/Modeling) that have a high number of citations.

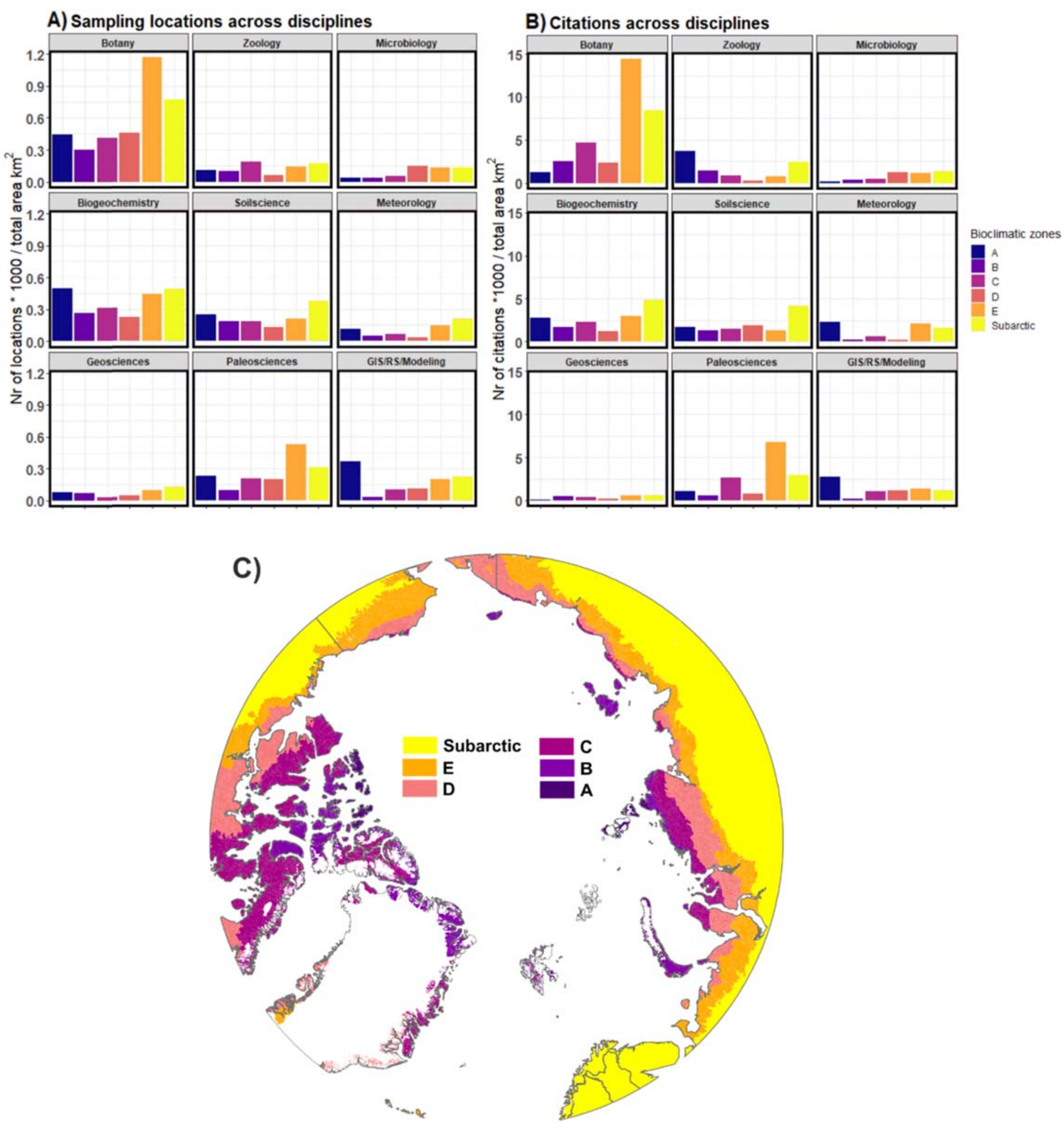

Figure 1. (A) Number of sampling locations and (B) number of citations per bioclimatic zone area and across environmental research disciplines. (C) Map of bioclimatic zones across the Arctic (Walker *et al* 2005). The letters indicate zones where the mean temperature (°C) in July is: **A** = 1–3; **B** = 4–5; **C** = 6–7; **D** = 8–9, **E** = 10–12; **sub-Arctic** = > 12. The Arctic was defined as all land north of the Arctic Circle (66.3 N).

There are differences in the proportional number of sampling locations and citations across Arctic bioclimatic zones (figure 1, appendix A3). The warmest zones, sub-Arctic and zone E (mean July temperatures >10 °C), are studied and cited the most in relation to their extent, particularly in Botany, and Paleosciences. The least amount of sampling locations and citations per unit area are located in zones A, B, C, and D, which represent the coldest climatic conditions (mean July temperatures 1–9 °C). However, in some disciplines (e.g. Biogeochemistry and GIS/RS/Modeling), the coldest zone A is well studied and cited in relation to its extent. Some disciplines (e.g. Botany and Paleosciences) display particularly large differences in sampling locations and citation per unit area among bioclimatic zones. Indeed, the large peak in sampling locations (corrected by the spatial extent of the zone) in zone E seems to originate mainly from these two fields (appendix A3), whereas in other disciplines, the proportional number of sampling locations is more uniform across the zones. In a few cases, despite the low number of sampling locations in a specific zone, there are relatively high citations (e.g. Zoology, Meteorology in zone A) or barely any citations at all (e.g. Geosciences in zone A).

The number of sampling locations relative to ecoregion area is variable across the Arctic (figure 2, appendix A5), ranging to high (e.g. Scandinavian Montane Birch Forest and Grasslands), medium (e.g. Interior Alaska Taiga) and low (e.g. Chukchi Peninsula Tundra). However, for citations, this pattern is even more biased to a few highly cited areas (e.g. Alaska-Yukon Arctic, Scandinavian Montane

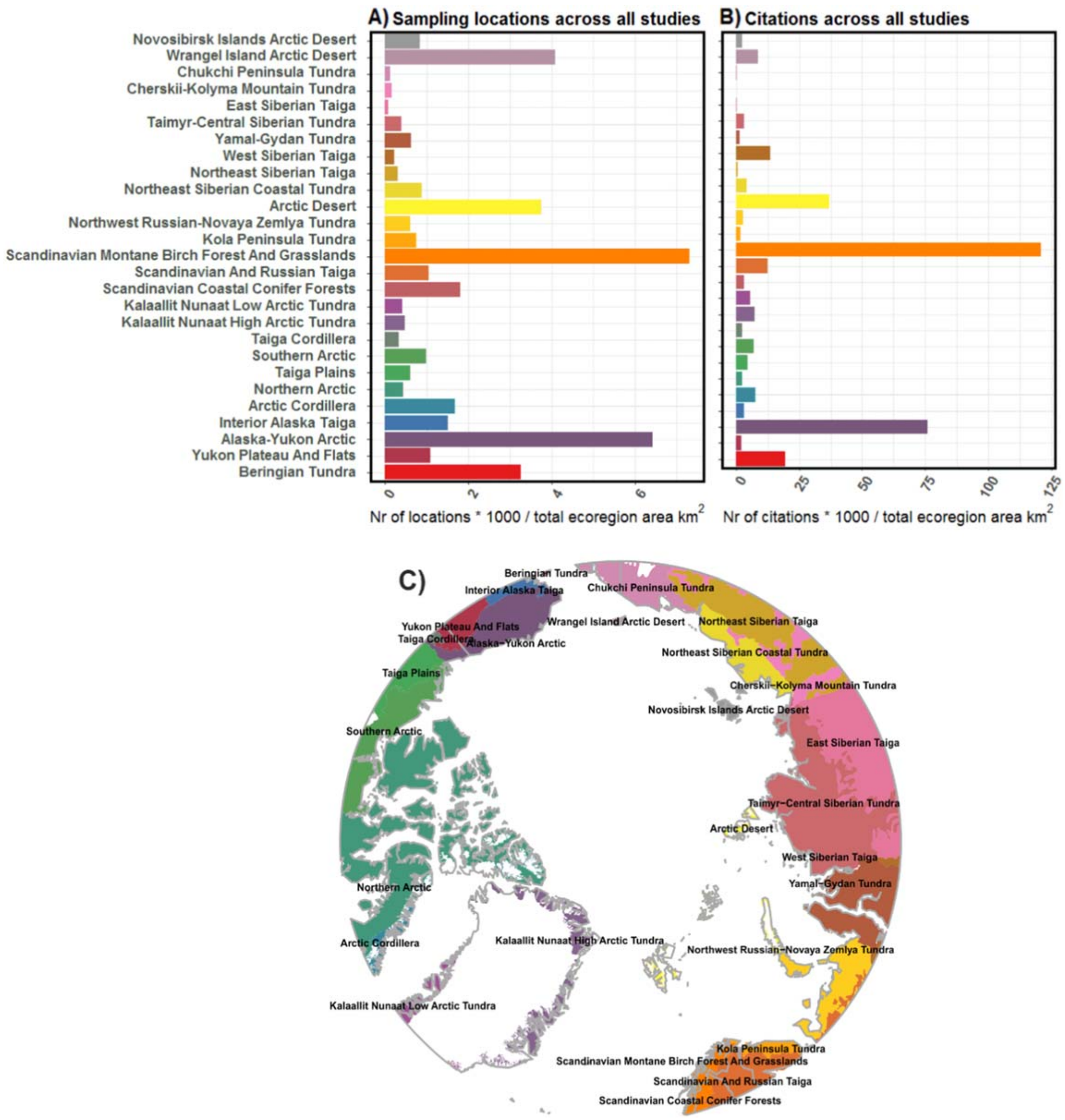

Figure 2. (A) Number of sampling locations and (B) number of citations per ecoregion area and across all studies in the Arctic. (C) Ecoregions across the Arctic (Olson *et al* 2001). The Arctic was defined as all land north of the Arctic Circle (66.3 N).

Birch Forest and Grasslands) with a few ecoregions having higher citations as only a few studies can increase their relative number of sampling locations due to their small extent. The proportionally highest number of sampling locations and citations is found in areas without permafrost (figure 3, appendix A4). The low number of observations in the continuous permafrost zone is apparent particularly in Zoology, Microbiology, Meteorology, Geosciences and GIS/RS/Modeling.

3.2. The extent of studies in topographical-soil-vegetation realm

The sampling locations cover the Arctic MAGT-SOC and pH-NDVI realms to some extent, but the clustering of locations and citations to a few conditions is high (figure 4). In the MAGT-SOC realm, the two dark clusters of locations in figure 4(B) (cluster 1 MAGT −8 to −4; SOC 1000–1500, cluster 2 MAGT −1 to +2; SOC 400–600) do not converge with the larger cluster over the entire Arctic conditions in figure 4(A) (MAGT −15 to −5; SOC 300–1200), thus sampling locations miss the main MAGT-SOC cluster. Citations are even more clustered to a few pixels (figure 4(C)). In the pH-NDVI realm, low productivity (NDVI < 0.25) and low pH (pH < 6) areas are particularly under-sampled and cited, and the lower cluster in figure 4(D) (pH 6–7; NDVI 0–0.25) is omitted by the sampling locations and citations (figures 4(E), (F)). The same overall pattern of frequently studied conditions is apparent for all disciplines (appendices A6 and A7), but the conditions are not covered as well by the disciplines (larger gaps found e.g. MAGT −15 to −10, −5–0, SOC

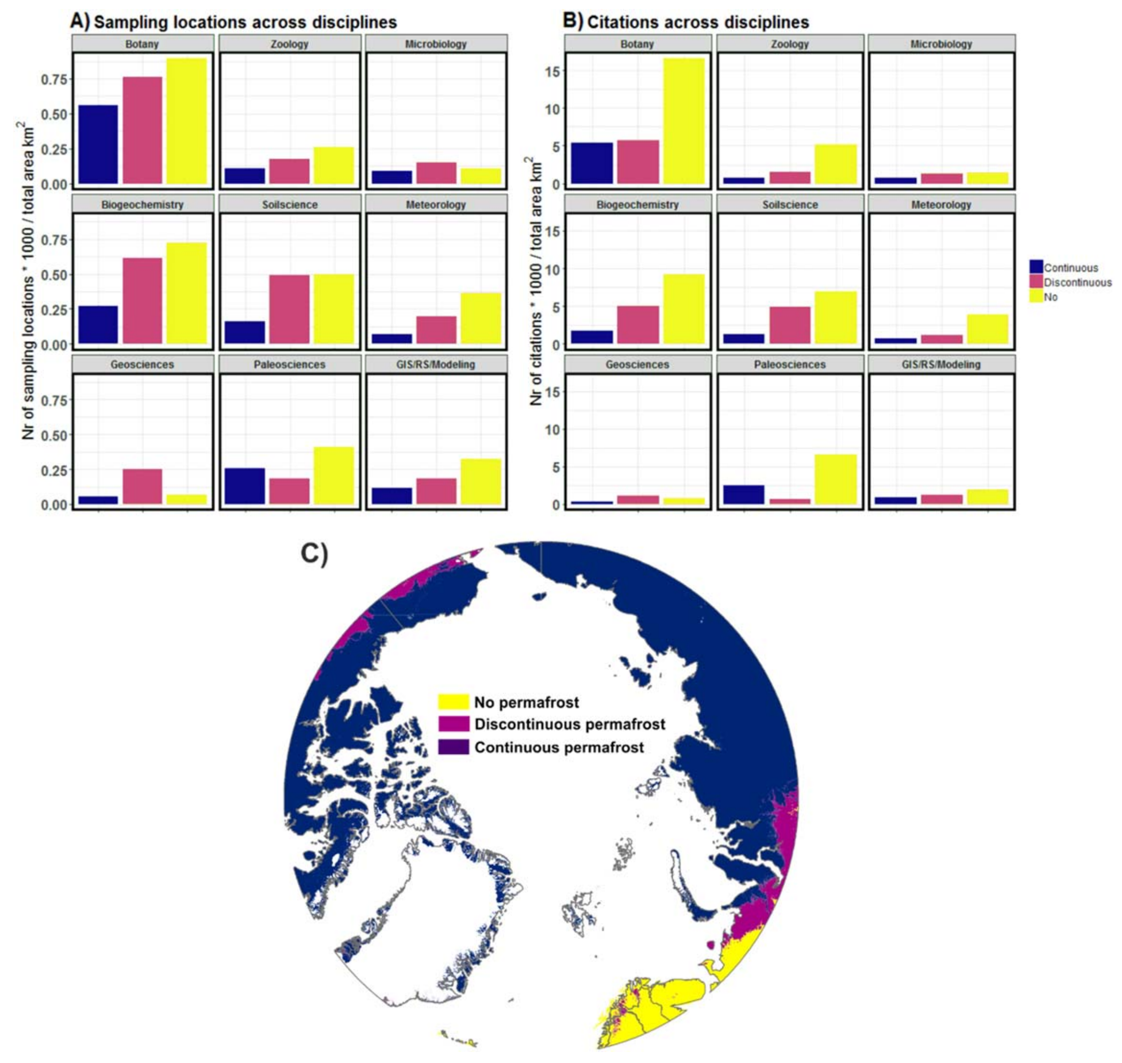

Figure 3. (A) Number of sampling locations and (B) number of citations per permafrost zone area and across environmental research disciplines. (C) Map of permafrost zone across the Arctic (Aalto *et al* 2018a, 2018b). The Arctic was defined as all land north of the Arctic Circle (66.3 N).

2000–4000) except in Botany, Biogeochemistry, Soil science and Paleosciences.

3.3. The current extent of sampling locations

The mean AUC value of the GBM models varied between 0.75 and 0.85 (appendix A8), thus the predictive performance was good and the models can reliably predict the representativeness of Arctic sampling locations. A probability map of coverage across all disciplines is shown in figure 5. Most of Alaska, Fennoscandia, southern parts of Greenland, and smaller areas in western and northeastern Russia have high probabilities (thus are well covered), whereas the Canadian Arctic Archipelago, some parts of northern mainland Canada, northern Greenland and easternmost and central Siberia, and Siberian Taimyr region in the north are under-studied. Siberia has a patchy distribution of probabilities with smaller high probability clusters scattered across the region. Although the maps of separate disciplines follow the same large-scale patterns in representativeness, there are differences across the disciplines (figure 6). First, the highly sampled regions are found in northern Fennoscandia and northern or southern Alaska, but additional regions are not evenly distributed. For example, in some disciplines southern Fennoscandia (e.g. Meteorology, GIS/RS/Modeling), entire Alaska (Botany, Paleosciences), Siberian region south from Taimyr (Botany, Microbiology, Geosciences, Paleosciences, GIS/RS/Modeling) and eastern Russia (Microbiology, Geosciences) have high probabilities. Second, the lowest probabilities, shown in the darkest color, are found in the Canadian Arctic Archipelago, northern Greenland, northern Taimyr region, central and eastern Siberia, but additional understudied regions are found in central Alaska (e.g. Microbiology, Meteorology, Geosciences), southern Fennoscandia (e.g. Botany, Microbiology, Paleosciences), in the entire Taimyr region (e.g. Biogeochemistry,

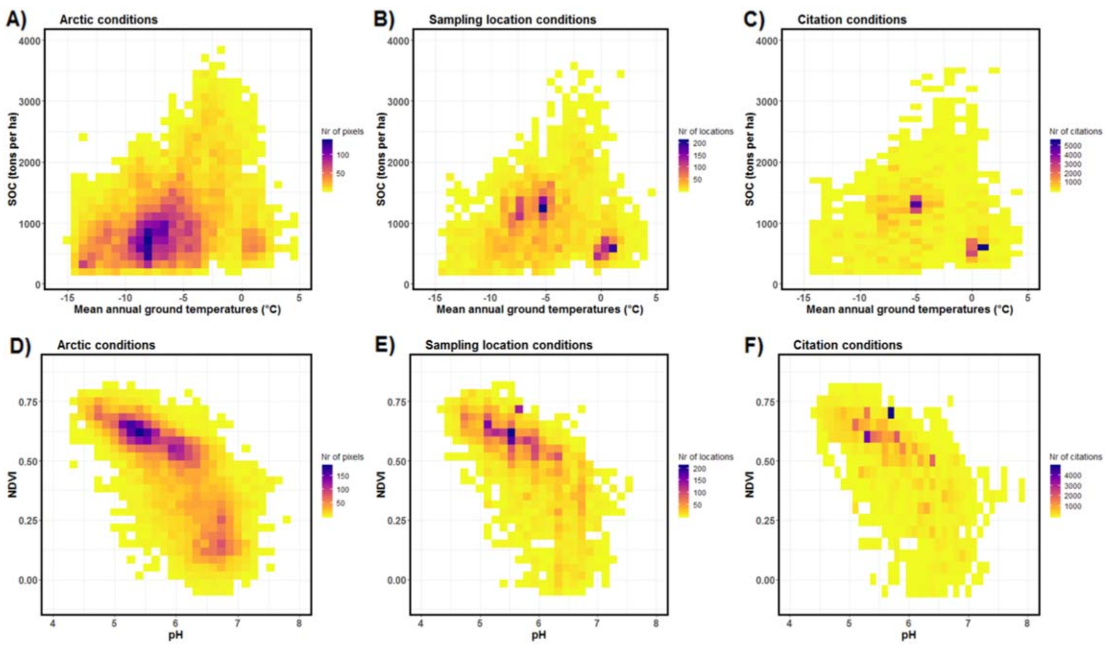

Figure 4. The MAGT-SOC realm (A)–(C) of the whole Arctic (A), sampling locations (B), and citations (C). The pH-NDVI realm (D)–(F) of the whole Arctic (D), sampling locations (E), and citations (F). The Arctic was defined as all land north of the Arctic Circle (66.3 N).

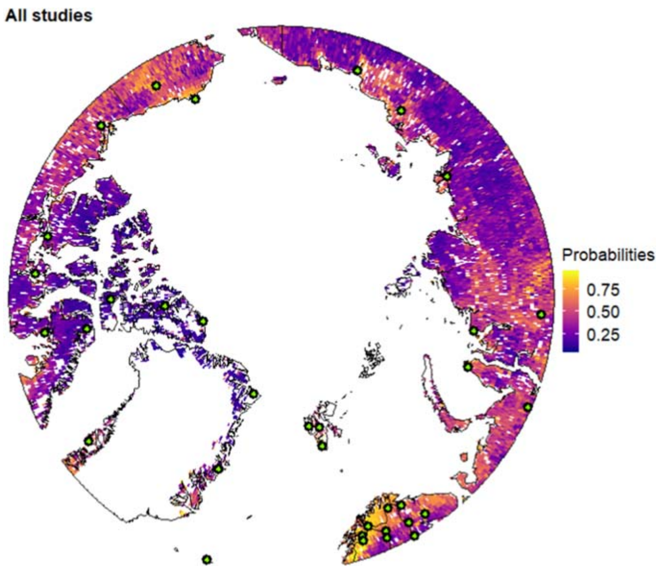

Figure 5. Probability map of sampling locations across all studies. High probabilities in yellow signify areas where conditions are well-studied, while low probabilities in violet signify areas where conditions lack sampling locations. The Arctic was defined as all land north of the Arctic Circle (66.3 N). The green points represent INTERACT stations (https://eu-interact.org/field-sites/) north from the Arctic Circle ($n = 34$). INTERACT is an infrastructure project with a circumarctic network of terrestrial stations in Arctic and alpine regions, offering access to stations.

Meteorology) or western Siberia (e.g. Botany, Zoology, Meteorology, Paleosciences). Third, medium probabilities are found for example in western Russia (e.g. Biogeochemistry, Soil science), northern mainland Canada (e.g. Zoology, Soil science, GIS/RS/Modeling) or eastern Greenland (e.g. Zoology, GIS/RS/Modeling). There are many INTERACT stations located across high, medium and low probability regions (figure 5).

4. Discussion

This study reveals Arctic terrestrial conditions and regions that are currently under-investigated and require targeted empirical research. In the following parts, we give a brief introduction of Arctic environmental variability in current and future climate, provide suggestions of new study areas for different

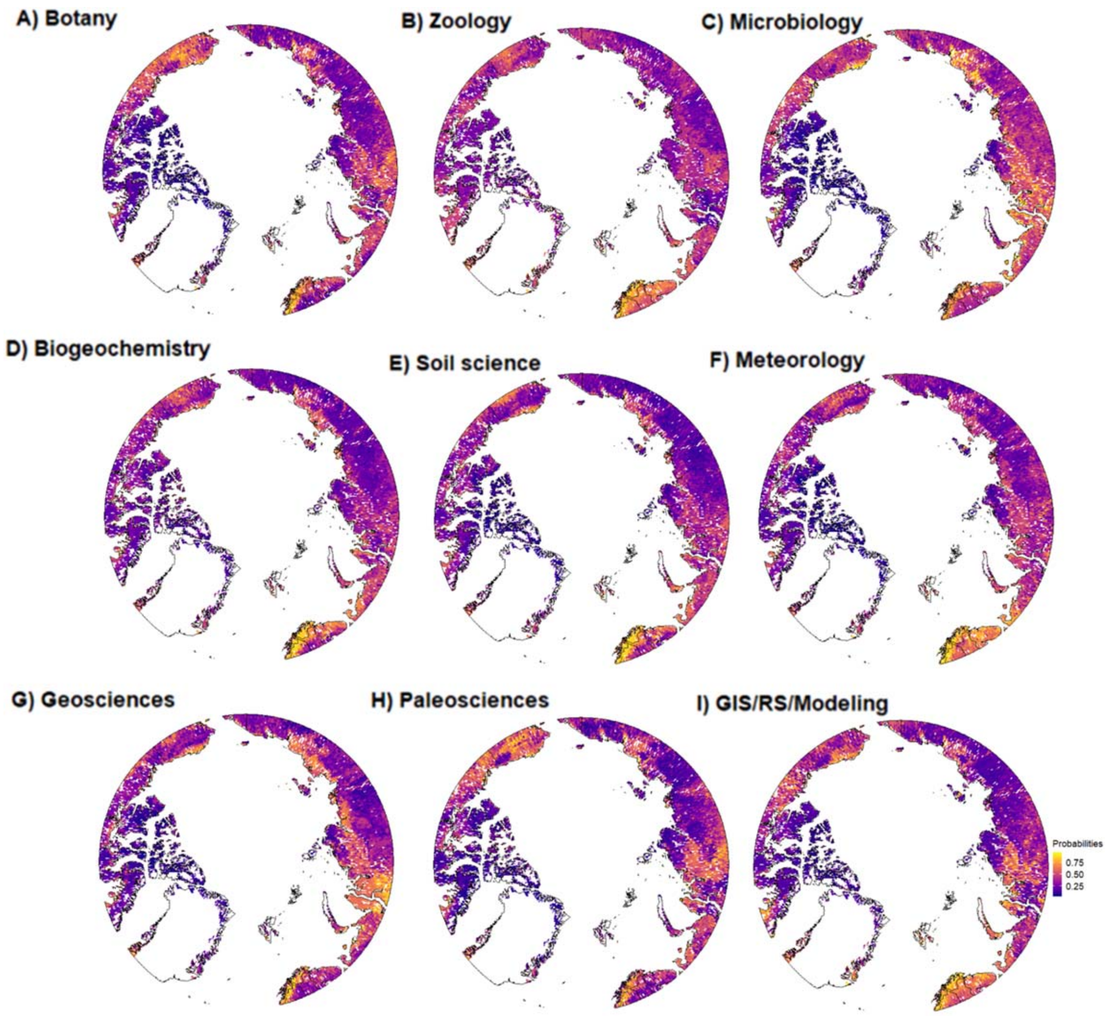

Figure 6. Probability map of sampling locations across all disciplines. High probabilities in yellow signify areas where conditions are well-studied, while low probabilities in violet signify areas where conditions lack sampling locations. The Arctic was defined as all land north of the Arctic Circle (66.3 N).

disciplines (table 2), and compare our results with other studies dealing with the representativeness of sampling locations. We acknowledge that the reasons scientific studies are conducted are more complicated than the location. However, our study focusing on sampling locations is an important step towards a better understanding of the status of Arctic environmental science research.

4.1. High-priority areas for terrestrial environmental field research

High-Arctic bioclimatic zones A, B and C (mean July temperatures 1–7 °C), located mainly in Canadian Arctic Archipelago, northern Greenland and in a few northernmost regions across Siberia, are currently understudied in almost all disciplines. In bioclimatic zone A, vegetation is mostly barren with some lichens, mosses, and graminoids and cushion forbs, whereas in bioclimatic zone B mosses, herbaceous plants and prostrate shrubs have higher abundance. In bioclimatic zone C, vegetation cover is higher and consists of prostrate shrubs generally taller than in the bioclimatic zone B (Walker *et al* 2005). Some regions in the high-Arctic have a high number of endemic plant species (e.g. Ellesmere and northern Greenland) or rare endemic vascular plants (e.g. in northern Taimyr region) (Talbot *et al* 1999, Daniëls *et al* 2013). Thus, targeted sampling of these harsh and barren environments in the future is crucial to better understand Arctic ecosystem functioning.

Some of the ecoregions having low proportional number of sampling locations are also found within these high-Arctic zones (Northern Arctic; zones A, B, C, and D, Taimyr-Central Siberian Tundra; all zones, East Siberian Taiga; sub-Arctic, Chukchi Peninsula Tundra; zones D, E, and Northeast Siberian Taiga; sub-Arctic). These findings together with the pH-NDVI realm results show that highly productive environments are also understudied. Low and high productivity environments both contain areas with high coverage of thermokarst or large yedoma deposits, thus they are important from the abiotic perspective as well (Schuur *et al* 2015, Strauss *et al* 2017).

Table 2. Understudied conditions and areas across Arctic disciplines. High-priority areas are highlighted in bold in understudied areas. The fourth column lists INTERACT stations (https://eu-interact.org/field-sites/) within or in a close proximity to the understudied areas. The unit for soil organic carbon stocks (SOC) is tons of carbon per ha, and for mean annual ground temperatures (MAGT) degrees in Celsius. NDVI (normalized difference vegetation index) is a unitless index describing vegetation productivity.

Discipline	Understudied conditions	Understudied areas	Potential INTERACT stations in understudied areas
Botany	Bioclimatic zones A, B, C, D; SOC 3000–4000; pH > 6.5	**Canadian Arctic Archipelago,** northern Canada, **northern Greenland, northern Taimyr, central and eastern Siberia,** southern Fennoscandia, western Russia	Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Canadian High Arctic, M'Clintock, Cen Ward Hunt, Igarka Geocryology, Villum, Kolari, Värriö, Khibiny
Zoology	Bioclimatic zones A, B, D; continuous permafrost zone; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.3	western Alaska, **Canadian Arctic Archipelago,** northern Canada, **northern Greenland,** northern Taimyr, **central, western and eastern Siberia**	Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Cen Ward Hunt, Igarka Geocryology, Western Arctic
Microbiology	Bioclimatic zones A, B, C; continuous permafrost zone; MAGT −15 to −10, 0–5; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.4	central Alaska, **Canadian Arctic Archipelago,** northern Canada, **northern Greenland**, Svalbard, **northern Taimyr, central and** eastern **Siberia**	Igloolik, Cen bylot island, Fishline Mars, Polar Environment Atmospheric Research Lab, Cen Ward Hunt, Kolari, Värriö, Western Arctic, Toolik, Villum, Zackenberg
Biogeochemistry	Bioclimatic zones B, C, D; continuous permafrost zone; SOC 2000–4000; pH > 6.5	central Alaska, **Canadian Arctic Archipelago, northern Canada, northern Greenland, northern Taimyr, central,** western **and eastern Siberia**, southern Fennoscandia	Igloolik, Cen bylot island, Fishline Mars, Polar Environment Atmospheric Research Lab, Cen Ward Hunt, Igarka Geocryology, Villum, Kolari, Värriö, Zackenberg, Western Arctic
Soil science	Bioclimatic zones B, C, D, E; continuous permafrost zone; MAGT −15 to −10, 0–5; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.3	**central Alaska, Canadian Arctic Archipelago,** northern Canada, **northern Greenland**, **central,** western **and eastern Siberia**, western Russia, southern Fennoscandia	Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab Canadian High Arctic research station, M'Clintock, Cen Ward Hunt, Igarka Geocryology, Villum, Kolari, Värriö, Zackenberg, Western Arctic, Khibiny
Meteorology	Bioclimatic zones A, B, C, D; continuous permafrost zone; MAGT −15 to −10, 0–5; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.4	central Alaska, **Canadian Arctic Archipelago,** northern Canada, **northern Greenland**, **northern Taimyr, central,** western **and eastern Siberia**	Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Canadian High Arctic, M'Clintock, Cen Ward Hunt, Igarka Geocryology, Villum, Zackenberg, Willem Barentsz, Western Arctic
Geosciences	Bioclimatic zones A, B, C, D, E, sub-Arctic; continuous permafrost and no permafrost zone; MAGT −15 to −10, 0–5; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.4	**central Alaska, Canadian Arctic Archipelago, northern Canada, northern Greenland,** northern Taimyr, **central and eastern Siberia,** Svalbard, southern Fennoscandia, western Russia	Toolik, Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Canadian high Arctic research station, M'Clintock, Cen Ward Hunt, Igarka Geocryology, Villum, Zackenberg, Western Arctic Kolari, Värriö
Paleosciences	Bioclimatic zones A, B, C, D; MAGT: 0–5, SOC: 3000–4000; pH < 6 and pH > 6.5; NDVI < 0.3	southeastern Alaska, **Canadian Arctic Archipelago,** northern Canada, **northern Greenland, northern Taimyr, central and eastern Siberia**, southern Fennoscandia, Svalbard, western Russia	Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Canadian high Arctic research station, M'Clintock, Cen Ward Hunt, Villum, Samoylov, Zackenberg, Khibiny, Kolari, Värriö
GIS/RS/Modeling	Bioclimatic zones B, C, D; continuous permafrost zone; MAGT −15 to −10, 0–5; SOC 2000–4000; pH < 6 and pH > 6.5; NDVI < 0.4	central Alaska, **Canadian Arctic Archipelago**, northern Canada, **northern Greenland, northern Taimyr, central,** western and **eastern Siberia**	Igloolik, Cen Bylot, Fishline Mars, Polar Environment Atmospheric Research Lab, Canadian high Arctic research station, M'Clintock, Cen Ward Hunt, Igarka Geocryology, Villum, Samoylov, Khibiny

Shifts in Arctic vegetation are expected due to global warming (Tape *et al* 2006, Myers-Smith *et al* 2011, Pearson *et al* 2013), thus understanding the current vegetation status will provide the baseline for future projections. Vegetation cover is predicted to increase in the barren environments in response to warming. Graminoid- or shrub dominated ecosystems are replaced by trees, or graminoids or prostrate shrubs are predicted to be replaced by dwarf and low shrubs. The regions we identify as under-sampled are predicted to have varying patterns of vegetation change. In the Taimyr region, trees and low shrubs, in eastern coastal Siberia particularly trees, and in easternmost Siberia dwarf and low shrubs are predicted to increase their distribution (Pearson *et al* 2013). In Canada, the projected vegetation shift is highly variable with trees advancing in shrubby areas and low shrubs replacing dwarf shrubs, and dwarf shrubs and graminoids replacing prostrate shrubs (Pearson *et al* 2013). In northern Greenland, no dramatic changes in shrub distributions are expected (Pearson *et al* 2013). Given the paucity of sampling across these environments, further studies are needed to build upon these preliminary conclusions.

Mean annual ground temperatures, which impact permafrost distribution, are also changing (Biskaborn *et al* 2019) in under-sampled regions. Aalto *et al* (2018b) showed that the highest increases in mean annual ground temperatures by 2080 would occur in Taimyr and east of Taimyr and in a few areas in the Canadian Arctic. Major declines in permafrost extent are expected to occur in the areas surrounding the Taimyr region, northeastern Siberia and in West Greenland (Aalto *et al* 2018b). Permafrost soils store large SOC stocks that are high in the under-sampled regions in western Canada and some parts of the Canadian archipelago (Hugelius *et al* 2014, Hengl *et al* 2017), though permafrost extent is not predicted to decrease in these areas as rapidly as for example in the Taimyr region (Aalto *et al* 2018b). More observations are needed from permafrost areas with MAGT ranging between −4 °C and −1 °C with high SOC stocks (2000–4000 t ha^{-1}) as these conditions could represent the tipping point of permafrost thaw driving a positive carbon cycle feedback.

Several INTERACT stations are located in the under-sampled areas in the Canadian Arctic, northern Greenland, Taimyr region and northeastern Siberia (figure 5). These could inform future field sampling campaigns to gain a better understanding of the wide variability of Arctic ecosystem functioning. Although we focus on under-sampled areas, we want to highlight that well-sampled regions are also undergoing rapid changes in the future (e.g. changes in permafrost extent in northern Fennoscandia, the advancement of trees in Alaska as shown in Pearson *et al* 2013, Aalto *et al* 2018b). Sampling locations within these regions has been, and will continue to be, extremely important to gain deeper insight of how Arctic environments are changing.

4.2. Comparison with other reviews

Our work estimates the spatial representativeness of sampling locations across multiple scientific disciplines while accounting for citations of the locations. We identify four representativeness categories for the disciplines: (1) well-sampled and well-cited areas (e.g. northern Alaska and Fennoscandia in all disciplines), (2) under-sampled and under-cited areas (Taimyr, Canadian archipelago in all disciplines), (3) under-sampled and well-cited areas (some patchy locations in eastern Siberia in Botany, Microbiology, Zoology, Soil science, and Biogeochemistry), and (4) well-sampled and under-cited areas (southern Arctic Alaska in Botany, Paleosciences, GIS/RS/Modeling). Despite a small number of locations per zone relative to their spatial extent, a few combinations of area and discipline were relatively well-cited (e.g. bioclimatic zone A in Zoology, West Siberian Taiga in all studies).

Our results are mostly consistent with the findings of previous review works from more specific research topics that have also illustrated how spatial variability has not been fully captured in sampling locations. Martin *et al* (2017) discovered large experimental and observational evidence gaps for shrubification studies in the Circumpolar Arctic region over the Eurasian Arctic, particularly in the Taimyr region, and also in the eastern coastal Siberia and the northernmost islands of high-Arctic Canada. Martin *et al* (2017) argued that 65% of the observations originated within the warmest parts of the Arctic tundra (bioclimatic zone E, where average July temperatures are above 9 °C) and that controls and mechanisms in colder regions are overlooked. Vilmi *et al* (2017) found that some regions in Alaska, northern provinces and territories of Canada, and Russia have not been comprehensively studied in plant species richness research and the only region that was relatively well-studied was Fennoscandia. Our work shows additional well studied conditions in botanical research in some parts of Greenland, Svalbard and Siberia, and research gaps particularly in northern Canada and the Canadian Arctic Archipelago, northern Greenland, northern Taimyr region, and central and eastern Siberia.

A review on microbial biogeography in Arctic soils discussed the distribution of soil bacterial diversity studies in the Arctic (Malard and Pearce 2018) and discovered that the number and distribution of studies is sparse. However, the distribution of sampling locations was different from our study. The number of studies was highest for the Canadian Arctic, which had four intensively studied sites. Whereas Alaska, Greenland, Svalbard, Fennoscandia and Russia had only one intensively studied area for each domain. Our review shows a rather patchy map for microbial sampling location representativeness, which most importantly highlights the need for more research in both high latitudes and high elevations across all continents.

Our results also corroborate a biogeochemical review on growing-season CO_2 flux chamber studies that highlighted the need for more research in extreme conditions (e.g. low temperatures, high topsoil pH, high SOC stocks) (Virkkala *et al* 2018). They also showed that Alaska and Fennoscandia were studied the most, and the Canadian Arctic Archipelago and Siberia were understudied. Here, we suggest that there are some conditions even within these well-studied regions, in central Alaska or southern Fennoscandia, that are understudied in Biogeochemistry. A study of representativeness of eddy covariance FLUXNET sites discovered that the Taimyr region together with its surrounding eastern coastal region and western peatland region, some parts of northwestern Canada and the Canadian Arctic Archipelago, and western Norway were the least represented (Kumar *et al* 2016). As we found in this study, the representativeness was moderate in some parts of the Canada. However, our findings suggest that western Norway is relatively well-sampled in Biogeochemistry, and the whole Canadian Arctic is underrepresented.

We are not familiar with Arctic representativeness studies from other disciplines, although there are reviews that summarize the current state of the disciplines with some examples from across the Arctic (e.g. a remote sensing review focusing on Alaska by Stow *et al* 2004).

4.3. Limitations of our approach

We acknowledge limitations both in our literature review, discipline classifications and modeling methods. First, we did not search non-English scientific literature, which might explain some of the spatial research gaps, particularly in Siberia. Second, the classification of studies into disciplines was based on expert assessment of 20 individual researchers. Our aim was to investigate broad disciplines and not delve deeper into the level of sub-disciplines in order to keep our message clear. Third, the location accuracy and the number of sampling locations reported varied across the studies. Finally, our modeling method does not allow us to calculate the total area of missing sampling locations, but it is rather a tool to visualize patterns and differences in representativeness across the Arctic.

5. Conclusions

This study captures the extent of sampling locations and citations in a spatially-explicit manner across broad Arctic environmental disciplines and terrestrial gradients. The resultant high-resolution maps that consider multiple environmental conditions simultaneously expose potential new sampling locations for each environmental discipline. This enables us to consider if and how an uneven distribution of sampling locations translates into gaps in knowledge across environmental gradients. We summarize understudied conditions and areas for each environmental discipline and provide a list of high-priority areas that are of particular importance to understand Arctic terrestrial ecosystem functioning in a changing climate (table 2).

There are vast areas in the Arctic that are lacking sampling locations and citations. High-priority future study areas in terms of current and future soil and vegetation conditions are found in several regions across the Arctic, with an emphasis on high-Arctic regions. The Canadian Arctic Archipelago is a high-priority research area as it will face vegetation shifts, changes in MAGT and contains large SOC stocks. The second high-priority research area is central Siberia which is experiencing permafrost loss and this will continue in the future. Additional important areas are northern Taimyr and eastern Siberia, which are experiencing advancement of shrubs and trees and loss of permafrost due to a rapid increase in MAGT, northwestern Greenland, where permafrost extent is predicted to change, and central parts of Arctic Alaska, which is experiencing large vegetation shifts. Our results should help prioritize future research efforts, thus increasing our knowledge about the Arctic environmental change.

Acknowledgments

AMV was funded by Alfred Kordelin Foundation, The Finnish Cultural Foundation, and Väisälä fund. AMV and ML were funded by the Finnish Academy (project number 286950). DBM was supported by an European Research Council Consolidator grant (ECOHERB, 682707) and an Action Group grant (F 2016/668) awarded by the Lund University Strategic Research Area 'Biodiversity and ecosystem services in a changing climate'. We want to thank Juha Aalto for processing the Modis NDVI image and Pekka Niittynen and Heidi Mod for the discussions related to pseudo-absences and modeling.

Data availability statement

The data from the literature review, which was used in all data analyses, is openly available at the following link: https://figshare.com/s/cee6070c4598c4d85700. The representativeness maps that support the findings of this study are openly available at https://doi.org/10.25412/iop.9162191.

ORCID iDs

A-M Virkkala https://orcid.org/0000-0003-4877-2918
A M Abdi https://orcid.org/0000-0001-6486-8747
M Luoto https://orcid.org/0000-0001-6203-5143
D B Metcalfe https://orcid.org/0000-0001-8325-9269

References

Aalto J *et al* 2018a Data from: statistical forecasting of current and future circum-Arctic ground temperatures and active layer thickness *Dryad Digit. Repository* (https://doi.org/10.5061/dryad.886pr72)

Aalto J *et al* 2018b Statistical forecasting of current and future circum-Arctic ground temperatures and active layer thickness *Geophys. Res. Lett.* **45** 4889–98

AMAP (ed) 2017 *Snow, Water, Ice and Permafrost in the Arctic (SWIPA)* (Oslo, Norway: Arctic Monitoring and Assessment Programme (AMAP)) p 269

Amatulli G *et al* 2018 A suite of global, cross-scale topographic variables for environmental and biodiversity modeling *Sci. Data* **5** 180040

Barbet-Massin M *et al* 2012 Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* **3** 327–38

Beven K J and Kirkby M J 1979 A physically based, variable contributing area model of basin hydrology *Hydrol. Sci. Bull.* **24** 43–69

Biskaborn B K *et al* 2019 Permafrost is warming at a global scale *Nat. Commun.* **10** 264

Bond-Lamberty B *et al* 2018 Globally rising soil heterotrophic respiration over recent decades *Nature* **560** 80–3

Buri A *et al* 2017 Soil factors improve predictions of plant species distribution in a mountain environment *Prog. Phys. Geogr.* **41** 703–22

CAVM T 2003 Circumpolar Arctic vegetation map *Conservation of Arctic Flora and Fauna (CAFF) Map No. 1* (Anchorage, Alaska: US Fish and Wildlife Service) (http://data.arcticatlas.org/geodata/cp/cp_biozone_la.html)

Danielson J and Gesch D 2011 *Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)* US Geological Survey Open-File Report 2011–1073 p 26

Daniëls F J A *et al* 2013 Plants *Arctic Biodiversity Assessment. Status and Trends in Arctic Biodiversity. Conservation of Arctic Flora and Fauna, Akureyri* ch 9, p 310–53

Didan K 2015 MOD13A2 MODIS/terra vegetation indices 16-Day L3 global 1km SIN grid V006 [data set] NASA EOSDIS LP DAAC (https://doi.org/10.5067/MODIS/MOD13A2.006)

Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees *J. Animal Ecol.* (https://doi.org/10.1111/j.1365-2656.2008.01390.x)

ESRI 2018 *ArcGIS Desktop: Release 10* (Redlands, CA: Environmental Systems Research Institute)

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

Freund Y and Schapire R E 1997 A decision-theoretic generalization of on-line learning and an application to boosting *J. Comput. Syst. Sci.* **55** 119–39

Friedman J H 2001 Greedy function machine: a gradient boosting machine *Statistics* **29** 1189–232

Gough L *et al* 2000 Vascular plant species richness in Alaskan arctic tundra: the importance of soil pH *J. Ecol.* **88** 54–66

Greenwell B *et al* 2019 Package ' gbm ' (https://CRAN.R-project.org/package=gbm)

Groendahl L, Friborg T and Soegaard H 2007 Temperature and snow-melt controls on interannual variability in carbon exchange in the high Arctic *Theor. Appl. Climatol.* **88** 111–25

Hanley J A and McNeil B J 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology* **143** 29–36

Hengl T *et al* 2014 SoilGrids1 km—global soil information based on automated mapping *PLoS One* **9** e105992

Hengl T *et al* 2017 SoilGrids250m: global gridded soil information based on machine learning *PLoS One* **12** 1–40

Hijmans R J *et al* 2018 *Package 'Raster': Geographic Data Analysis and Modeling*

Hirzel A and Guisan A 2002 Which is the optimal sampling strategy for habitat suitability modelling *Ecol. Modelling* **157** 331–41

Hoffman F M *et al* 2013 Representativeness-based sampling network design for the State of Alaska *Landscape Ecol.* **28** 1567–86

Hugelius G *et al* 2014 Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps *Biogeosciences* **11** 6573–93

IPCC 2013 Climate change 2013: the physical science basis *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* ed T F Stocker *et al* (Cambridge and New York: Cambridge University Press) p 1535

Keenan T F and Riley W J 2018 Greening of the land surface in the world's cold regions consistent with recent warming *Nat. Clim. Change* **8** 825–8

Kulmala M 2018 Build a global Earth observatory *Nature* **553** 21–3

Kumar J *et al* 2016 Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements *Earth Syst. Sci. Data Discuss.* 1–25

Lara M J *et al* 2018 Reduced arctic tundra productivity linked with landform and climate change interactions *Sci. Rep.* **8** 2345

Malard L A and Pearce D A 2018 Microbial diversity and biogeography in Arctic soils *Environ. Microbiol. Rep.* **10** 611–25

Marmion M *et al* 2009 The performance of state-of-the-art modelling techniques depends on geographical distribution of species *Ecol. Modelling* **220** 3512–20

Martin A C *et al* 2017 Shrub growth and expansion in the Arctic tundra: an assessment of controlling factors using an evidence-based approach *Environ. Res. Lett.* **12** 085007

Metcalfe D B *et al* 2018 Patchy field sampling biases understanding of climate change impacts across the Arctic *Nat. Ecol. Evol.* **2** 1443–8

Myers-Smith I H *et al* 2011 Shrub expansion in tundra ecosystems: dynamics, impacts and research priorities *Environ. Res. Lett.* **6** 045509

Myers-Smith I H and Hik D S 2018 Climate warming as a driver of tundra shrubline advance *J. Ecol.* **106** 547–60

Natali S M *et al* 2015 Permafrost thaw and soil moisture driving CO_2 and CH_4 release from upland tundra *J. Geophys Res.: Biogeosci.* **120** 525–37

Nussbaum M *et al* 2018 Evaluation of digital soil mapping approaches with large sets of environmental covariates *SOIL* **4** 1–22

Olson D M *et al* 2001 Terrestrial ecoregions of the world: a new map of life on earth *BioScience* **51** 933–8

Patterson T and Kelos N 2009 Natural Earth: free vector and raster map data at 1:10 m, 1:50 m, and 1:110 m scales (https://naturalearthdata.com)

Pearson R G *et al* 2013 Shifts in Arctic vegetation and associated feedbacks under climate change *Nat. Clim. Change* **3** 673–7

Pebesma E and Bivand R 2018 Package ' sp ' (https://CRAN.R-project.org/package=sp)

Phoenix G K and Bjerke J W 2016 Arctic browning: extreme events and trends reversing arctic greening *Global Change Biol.* **22** 2960–2

Quinn P *et al* 1991 The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models *Hydrol. Process.* **5** 59–79

R Core Team 2018 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing)

Schuur E A G *et al* 2015 Climate change and the permafrost carbon feedback *Nature* **520** 171–9

Sing T *et al* 2009 Package 'ROCR' (https://CRAN.R-project.org/package=ROCR)

Stow D A *et al* 2004 Remote sensing of vegetation and land-cover change in Arctic Tundra ecosystems *Remote Sens. Environ.* **89** 281–308

Strauss J *et al* 2017 Deep Yedoma permafrost: a synthesis of depositional characteristics and carbon vulnerability *Earth Sci. Rev.* **172** 75–86

Street L E *et al* 2007 What is the relationship between changes in canopy leaf area and changes in photosynthetic CO_2 flux in arctic ecosystems? *J. Ecol.* **95** 139–50

Sundqvist M K, Sanders N J and Wardle D A 2013 Community and ecosystem responses to elevational gradients: processes,

mechanisms, and insights for global change *Annu. Rev. Ecol. Evol. Syst.* **44** 261–80
Talbot S S *et al* 1999 *Atlas of Rare Endemic Vascular Plants of the Arctic. Conservation of Arctic Flora and Fauna* (CAFF) Technical Report No. 3 US Fish and Wildlife Service, Anchorage, Alaska p 73
Tape K, Sturm M and Racine C 2006 The evidence for shrub expansion in Northern Alaska and the Pan-Arctic *Glob. Change Biol.* **12** 686–702
The Nature Conservancy 2009 Global Ecoregions (http://maps.tnc.org/gis_data.html)
Tucker C J *et al* 2005 An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data *Int. J. Remote Sens.* **26** 4485–98
Vilmi A *et al* 2017 Geography of global change and species richness in the North *Environ. Rev.* **25** 184–92
Virkkala A-M *et al* 2018 The current state of CO_2 flux chamber studies in the Arctic tundra : a review *Prog. Phys. Geogr.* **42** 162–84
Walker D A *et al* 2005 The circumpolar Arctic vegetation map *J. Vegetation Sci.* **16** 267–82
Westermann S *et al* 2015 A ground temperature map of the North Atlantic permafrost region based on remote sensing and reanalysis data *Cryosphere* **9** 1303–19
Wickham H *et al* 2018 Package 'ggplot2' (https://CRAN.R-project.org/package=ggplot2)
Yang F *et al* 2008 Assessing the representativeness of the AmeriFlux network using MODIS and GOES data *J. Geophys. Res.: Biogeosci.* **113** 1–11