A scientometric review of permafrost research based on textual analysis (1948-2020)

Frédérique Bordignon

Published in Scientometrics. 2020. 10.1007/s11192-020-03747-4

A scientometric review of permafrost research based on textual analysis (1948-2020)

Frederique Bordignon
Ecole des Ponts, Marne-la-Vallée, France
frederique.bordignon@enpc.fr
0000-0002-4918-9137

Abstract

This article proposes an analysis of research dedicated to permafrost. Its originality is twofold: it covers a corpus (n=16,249) that has never been reviewed before and also makes use of a methodology based on successive textual analysis processes. With the text-mining of additional corpuses, we produce lists of qualified terms to fine-tune the indexing of the main corpus and isolate relevant terminology dedicated to infrastructure and soil properties. With these enrichments combined with other terminological extractions (such as place names recognition), we reveal the internal structure of permafrost research with the help of visual mapping and easily prove that permafrost research is multidisciplinary and multi-topical

The semantic map and the diachronic analysis of terms clusters show that the interest had turned since the 1980s towards the role of climate change but also on China's needs for its highway and railway construction sites. The very strong and growing impact of Chinese research, focused on the Tibetan area, is one of the highlights of our data. Furthermore, we propose a focus on infrastructure vulnerability and use soil properties as a proxy to measure the existing interactions between two distinct research communities. The results suggest that research has mainly focused so far on the feasibility of building on frozen ground and exploiting soils, but remains at an early stage of addressing the impact of global warming on infrastructure degradation and its resilience.

This study offers insights to permafrost experts, but also provide a methodology that could be reused for other investigations.

Keywords

Text-mining methods, Scientometrics, Semantic mapping, Lexical analysis, Permafrost, Infrastructure vulnerability

Availability of data and material

Data are available for download: http://dx.doi.org/10.17632/d8gvm96ykm.1

Introduction and related work

This article proposes an analysis of research dedicated to permafrost. Its originality is twofold insofar as it covers a corpus that has never been reviewed before and also makes use of textual analysis methods to do so.

We propose to examine a substantial corpus of 16,249 bibliographic records (1948-2020) about permafrost. According to the International Permafrost Association, permafrost is defined as ground (soil or rock including ice and organic matter) that remains at or below 0 °C for at least two consecutive years. It covers about 20% of the Earth's continental surface, or 25 million km^2, a quarter of the land area of the Northern Hemisphere (Wu et al. 2002). Most permafrost is found at high latitudes and in high mountain areas, mainly in the Arctic, Alaska, Canada, Siberia and the Alps.

Permafrost formation, its persistence and disappearance, its temperature and thickness and its distribution on earth are very sensitive to climate variations. Therefore, global warming has been causing permafrost to thaw (Zimov 2006). This causes soil bacteria to degrade organic matter that has remained in the ground for thousands of years, thereby releasing carbon dioxide and methane into the atmosphere. Climate change might also trigger dissociation of gas hydrates present in large quantities in permafrost, thus releasing methane to the atmosphere (Wooller et al. 2009). Overall, these greenhouse gases might, in turn, amplify global warming (positive-feedback loop). Increased temperatures and higher carbon dioxide concentration levels might also increase vegetal cover in permafrost areas and thus tend to reduce net greenhouse emissions to the atmosphere. Permafrost thawing already has other serious consequences: mercury in permafrost begins to escape and contaminate the food chain as it reaches the ocean (Sutherland et al. 2019), and buried viruses are discovered in the soil (Trubl et al. 2018). Eventually, permafrost thaw has negative impacts on the integrity of infrastructure including roads, railways, power transmission infrastructures, oil pipelines, mountain resort stations, and buildings. (Hjort et al. 2018) found that nearly 70% of current infrastructure in the Arctic is built on permafrost that is at risk of thawing by 2050. Their study showed that three-quarters of the population living in the arctic permafrost regions (i.e. about 3.6 million people) will be affected by this damage in the next 30 years because "a substantial proportion of the fundamental human infrastructure is potentially under risk: 48–87% (mean = 69%) of the current pan-Arctic infrastructure is located in areas where near-surface permafrost is projected to thaw by mid-century".

Permafrost is an inherently complex research topic: it is neither a material defined by its properties nor a well-defined geographical region; its instability is both the cause and the consequence of global warming and humans are alternately responsible and victims. And, its thawing is seen both as a socio-economic threat and as an opportunity for the exploitation and transport of new energy resources. As a consequence, there is a broad scientific and social interest in the knowledge of the cryosphere, which has yielded to research programs and strategies dedicated to its study and monitoring.

The study of permafrost is therefore at the forefront of geosciences research (Serrano Cañadas 2016). In addition to the IPCC *Special Report on the Ocean and Cryosphere in a Changing Climate* (2019), numerous review papers have been published (more than 400 in our corpus). Most of the time, they cover only part of the issue, e.g. a geographical area such as Tibet (Bibi et al. 2018; M. Yang et al. 2010), the Arctic (Ikram and Afzal 2019; Overland et al. 2019), Switzerland (Weber et al. 2019) or even solely a research station in Siberia (Leibman et al. 2015). Others only address some aspects of permafrost (e.g. physical or chemical features (Chang et al. 2016; Colombo et al. 2018), soil microbiology (Afouda et al. 2017; Canavan 2019), consequences of thawing and mitigation strategies (Aditya et al. 2017; Grosse et al. 2016; Vonk et al. 2015; Walvoord and Kurylyk 2016), exploration and exploitation of georesources (Li et al. 2016; L. Yang et al. 2019), the vulnerability of infrastructure (Ugwuishiwu et al. 2019), to mention only very recent examples). By proposing a method of analysis that allows the review of a much larger corpus of publications than those used in these studies, and especially without any predefined scope (neither geographical nor subject-specific), we offer an unprecedented overview of the knowledge associated with permafrost.

Only a few bibliometric studies have been carried out on the subject, one focusing on the community of permafrost scientists in the Iberian Peninsula (García-Hernández et al. 2019), one poster (Grosse and Lantuit 2008) analysing the PYRN-Bib database, a comprehensive bibliographic database on permafrost theses, another on social science publications related to the Arctic (Hua et al. 2012) and a last one on the sole geographical area of the Tibetan

Plateau (Xiao et al. 2017). There are also bibliometric works that have focused on part of the subject, such as (Côté and Picard-Aitken 2009) on Arctic research in Canada and (Aksnes and Hessen 2009) on structure and development of polar research. The most recent and also the most comprehensive study (Sjöberg et al. 2020) covers 2 decades (1998-2007 and 2008-2017). It is mainly a bibliometric study (citations, journals, research collaborations) with also a section dedicated to lexical analysis that our study will further develop. Indeed, in addition to providing a comprehensive and objective overview of the literature on permafrost, we propose a specific focus on research about the ways in which vulnerable infrastructure can adapt to climate change and be resilient, and we will use soil properties as a proxy to measure the existing interactions between two distinct research areas. This objective has led us to develop an original methodology to be included among the contributions of this work.

We suggest a lexicometric approach relying on lists of qualified terms to fine-tune the indexing of the corpus. These lists, derived from text-mining of additional corpuses, enable us to isolate relevant terminology dedicated to infrastructure and soil properties. By combining these enrichments with other terminological extractions (such as place names recognition), we aim at revealing the internal structure of permafrost research with the help of visual mapping. We present this methodology in detail so that it can be replicated for other studies.

The objectives of our study are therefore to propose a description of the research dedicated to permafrost since 1948, to verify if research on climate change and research on infrastructure mutually feed each other, and in parallel to propose an original methodology based on lexicometric principles.

Methods and data

To complete this comprehensive scientometric study, a large volume of scholarly publications must be considered. It is obviously impossible to proceed to a close reading of all these texts, nor just the Title, Abstract and Keywords metadata. Therefore, we adopt a distant reading approach (Moretti 2013) that enables us to grasp the expanding knowledge structure of permafrost studies in an objective way, using computational methods, starting with the text-mining of the elements at our disposal and then combining the results.

The objective is to produce a semantic map of the most important concepts, to identify clusters, and to ease the interpretation with specific terminology on the one hand and geographical areas on the other hand.

Terms extraction, clusterization and semantic mapping

We used the CorTexT platform[1] to perform our lexicometric analyses. This tool is based on language processing methods for the analysis and visualization of complex networks of concepts and has been used for many quantitative and qualitative analyses of the scientific literature (e.g. about food security (Cardon 2020), synthetic biology (Raimbault et al. 2016) or ecosystems services (Tancoigne et al. 2014).

(Callon et al. 1983) proposed co-word analysis to explore scientific fields and analyse their dynamics, that is to capture the frequency of pairs of words (or phrases). This theory is based on scientists' use of scientific publications as a vehicle for research ideas. When two terms appear in the same document and a fortiori in the same paragraph or sentence, it means they have an intrinsic relation. The basic data are therefore co-occurrence counts. By weaving these links between terms, we can map the semantic structure of a subject area with a network of concepts. In the network we generate, terms are linked with a strength proportional to a measure of similarity called *distributional measure* which means that two terms are all the closer when they co-occur with the same other terms (Weeds and Weir 2005).

The terms extracted from the documents can then be gathered in clusters with an algorithmic classification that allows organizing the knowledge of the whole corpus in subsets or subnetworks. The clusters could correspond

[1] https://www.cortext.net/

to topics of interest that are intensively studied by researchers (He 1999). We can then observe their evolution over time and compare them with each other. We also suggest combining them with geographical areas: studied zones and studying countries.

Named entity recognition to retrieve geographical locations

As stated above, permafrost is defined as ground that remains at or below 0 °C for at least two consecutive years. It is therefore defined solely by temperature, not geographic location. Moreover, the definition of permafrost areas is descriptive (continuous, discontinuous, sporadic, or isolated); therefore, the boundary between any adjacent two permafrost zones is generally ambiguous (Zhang 2005). Permafrost distribution is a subject of study in itself. The permafrost areas can be studied very differently depending on the nature of the soil, their possible exploration for energy resources extraction, their population density and also their vulnerability to global warming. This is why it is interesting to see if permafrost research is region-specific.

CorTexT enables automated extraction of geographical place names to help us identify which areas are being studied. These place names are not only countries but also cities, lakes, mountains... We had to carry out a manual homogenization for some place names, i.e. we cleaned up the initial list by grouping some place names under the same label: for example, we maintained frequent place names that are not whole countries but part of a country, or on the contrary, we maintained a region stretching over several countries. Typically, we favoured the "Arctic" label to group together all the terms that mention it in conjunction with another area. For example, "Canadian Arctic" is listed under the "Arctic" label and all other Canadian locations are assigned the "Canada" label.

Meanwhile, with the Netscity tool (Maisonobe et al. 2019) and its capacity to parse affiliation lines, we retrieved the authors' country for all the publications in the corpus.

In the end, we can use this information to observe which countries are working on which areas and then cross-reference these data with thematic clusters. Of course a single publication can mention several studied zones and can have authors from different countries.

Figure 1 shows the whole process of data collection, their enrichment and the subsequent analyses.

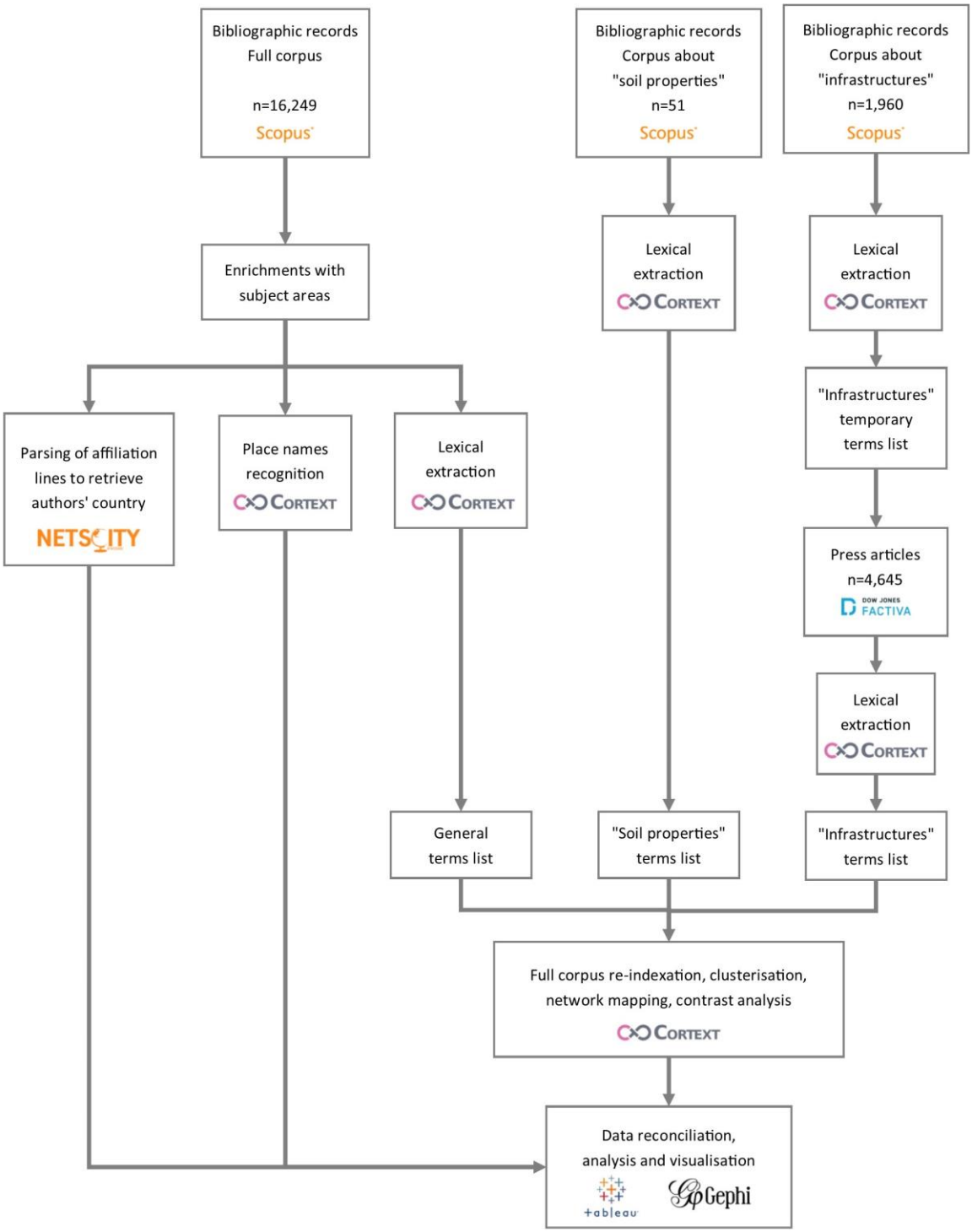

Figure 1. Workflow of the study design

Data collection and enrichments

Core data for global analysis

Delineating the corpus was easy because on the one hand, we have no predetermined criteria about disciplinary fields, journals, research-producing countries, geographical areas and on the other hand, although permafrost is a complex area of study, the word itself is neither polysemic nor does it have word variants in English. Hence we collected the bibliographic records with a very simple query in Scopus to retrieve all records containing "permafrost" in the Title, Abstract or Keywords fields: [TITLE-ABS-KEY (permafrost)]. We have favoured the use of Scopus over the Web of Science since we needed to maximize our chances of covering all the thematic fields and especially social sciences and economics which are known to be less well covered in the Web of Science in particular (Aksnes and Sivertsen 2019).
We exported on November 30th, 2019 the metadata of the 16,249 results, dating from 1948 to 2020 (publication date). Most of them are articles published in 2,111 different journals. Table 1 shows the distribution of document types.

Document type	**# publications**
Article	12,162
Conference Paper	2,761
Review	422
Book Chapter	422
Conference Review	104
Note	84
Book	63
Short Survey	42
Letter	37
Editorial	35
Erratum	32
Report	30
unknown	28
Article in Press	14
Abstract Report	7
Business Article	4
Data Paper	2
All	**16,249**

Table 1. Document types distribution

To broadly determine the subject area of those publications, we rely on the disciplinary classification of the journal or conference where they are published. We use the All Science Journal Classification (ASJC), a classification scheme assigning one or several subject area(s) to nearly 40,000 different sources. Thanks to the ASJC classification, we enriched 91% of the corpus references with at least one subject area and the associated upper field. Moreover, geographical terms are associated with each reference. Those data allow us to carry out the global analysis of the corpus.

Additional data for the focus on soil properties and infrastructure in permafrost areas

For the focus we wish to make on the specificity of construction in permafrost areas and the vulnerability of infrastructure to permafrost thawing, we need other data. We have to qualify the terms resulting from the lexical extraction to notably identify:

- those that are soil properties (e.g. *cemented soils, silty sand, water content estimation*),
- and those that fall within the vocabulary of civil engineering or denote infrastructures (e.g. *airport, road, railways, bridge, construction, foundation*).

Therefore, in addition to the global corpus on permafrost, we constituted two other corpuses which are only used for lexical extraction to have lists of standardized terms. There was too great a risk of forgetting some if it was done by hand. In order to best identify the terms expressing soil properties, an expert provided us with 51 references that contain a large number of them with a high degree of certainty. We carried out the lexical extraction with the script in CorTexT dedicated to terms extraction. The final list contains 124 terms revalidated by the expert.
We did not find a thesaurus comprehensive enough to provide a list of infrastructure types. We did not proceed in the same way as for soil properties because this is not a specialized scientific terminology. but a common vocabulary. (Roseau 2016) states that infrastructures share the same goal: to provide a basis by which the city and the territory are modernized and the term can refer as much to a port or river installation, to a map of high schools, to fortifications or even a tramway network. Therefore, we made up the terms list by proceeding in two steps and on two different corpuses. First, we built a corpus of scientific articles by querying Scopus using ISSNs of journals from subject areas[2] with a high probability of being infrastructure-related. We decided to retrieve 10,000 bibliographic records from Scopus. A first lexical extraction of terms denoting infrastructures was carried out with CorTexT. These terms were then used as a basis for a query in Factiva[3] to build up a corpus of 4,645 press articles. A lexical extraction is then performed on this corpus, and the list of extracted terms related to infrastructures is added to the first one. We obtain a list of 48 terms.
These lists are to be considered as outputs of this work for further studies (Bordignon 2020).

These specialized terms, as well as those automatically retrieved by CorTexT during a first analysis of the full corpus, were then indexed in the textual elements of the 16,249 bibliographic records of the full corpus.

Results

With the results we present in this section, we provide some evidence that the method we have just described is relevant to characterize the literature produced on permafrost and that it can, therefore, be reused to describe any other field of research.

[2] The following subject areas of the ASJC classification have been used: "Geotechnical Engineering and Engineering Geology", "Civil and Structural Engineering", "Building and Construction" and "Architecture"
[3] Factiva is an online database of global news and business information.

Exponential growth in the number of publications about permafrost

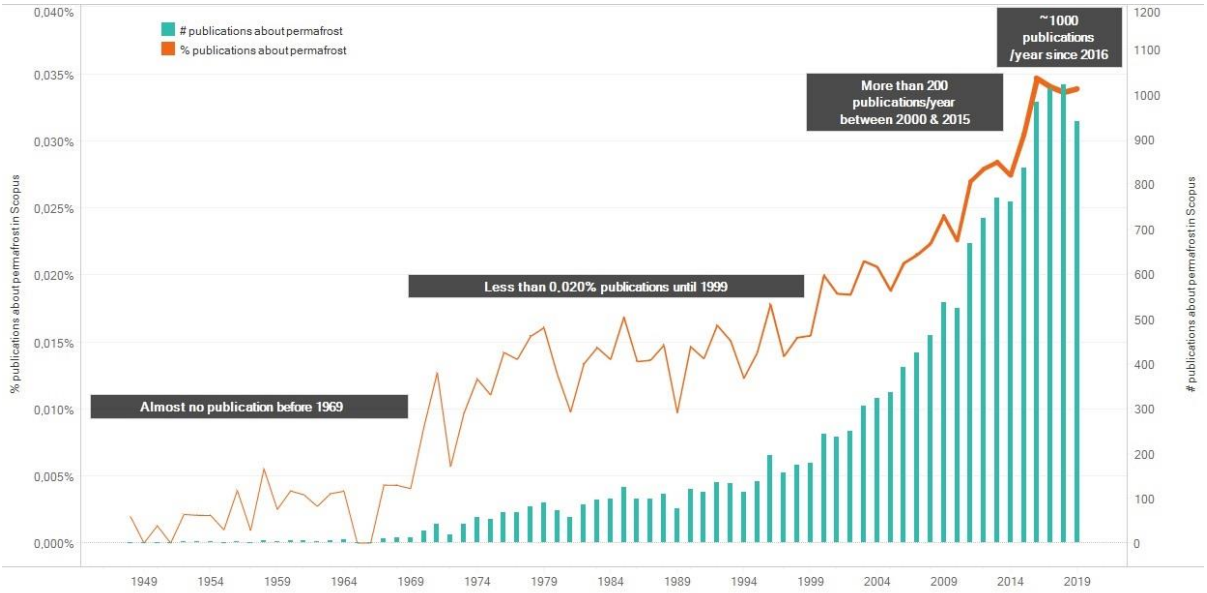

Figure 2. Annual absolute distribution and share of papers related to permafrost

Figure 2 shows exponential growth in both absolute volume and share of world research (indexed in Scopus) since 1948. Although the latter is, of course, minimal, a clear upward trend over time can be observed and shows a growing scientific interest: from 0.002% in 1948 to 0.034% in 2019.

The growth in the literature, shown in Figure 2, can be divided into 4 main periods:

- 1948-1969: hardly any publication before 1969,
- 1970-1999: less than 0.020% of world publications until 1999,
- 2000-2015: more than 200 publications/year between 2000 and 2015,
- 2016-2020: around 1000 publications/year since 2016

There is no discernible effect of the Polar Year (2007-2008), which is in any case in the middle of a period of strong growth. Moreover, the IPCC's special report on the cryosphere came out in 2019, so we do not have enough hindsight to see an impact.

Permafrost as a multidisciplinary and a multi-topical research subject

The distribution of publication sources within the ASJC classification shows that permafrost research is multi-disciplinary as it extends in all 27 fields of the ASJC classification and in 228 of the 307 subject areas. Table 2 shows how the corpus publications are distributed across the top 15 fields.

Field	# publications
Earth and Planetary Sciences	8,783
Environmental Science	3,821
Engineering	2,952
Agricultural and Biological Sciences	2,727
Social Sciences	1,319
Energy	1,178
Medicine	605
Physics and Astronomy	555
Computer Science	511
Materials Science	479
Immunology and Microbiology	422
Biochemistry, Genetics and Molecular Biology	375
Multidisciplinary	351
Chemical Engineering	266
Chemistry	245

Table 2. Distribution of the publications across the top 15 fields (ASJC classification)

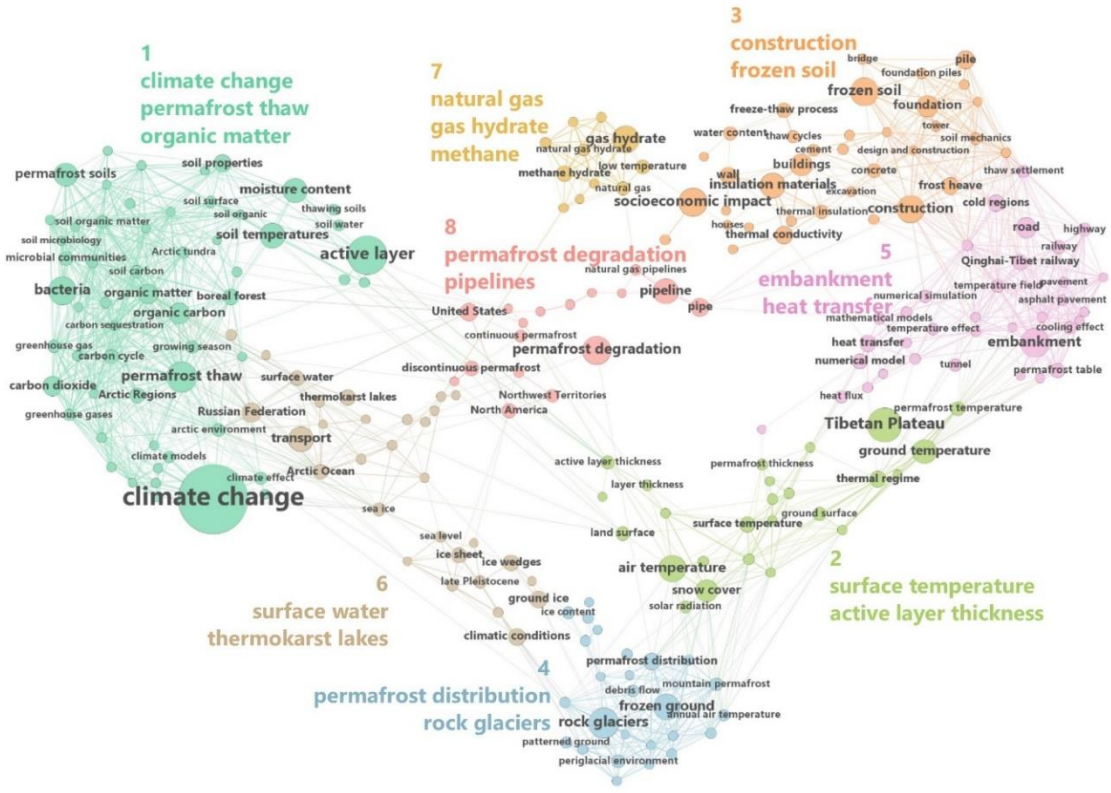

Figure 3. Co-occurrence map of extracted terms (n=239 nodes, node size = frequency, colour = cluster delineation)

The results of text-mining processes shows that permafrost research is multi-topical. Indeed, Figure 3 displays the semantic map of the whole corpus of publications. With the matrix of co-occurring terms, CorTexT generates a network of 239 nodes and 2,323 edges, and identifies 8 clusters (Table 3).

The network was spatialized using the Force Atlas algorithm (Jacomy et al. 2014): as long as it runs (in Gephi tool), the nodes repulse and the edges attract. Proximity between nodes results therefore from their connections

with their environment. The size of the nodes is related to their frequency. The colours are a function of a modularity algorithm that allows to identify clusters in an automated way. The clusters represent the topics of the whole corpus; they are sets of strongly related terms that contextualize each other's meaning. The 8 clusters of the network are labelled on the basis of their lexical contents: CorTexT suggests terms that are more specifically linked to a given cluster.

Clusters	**# publications**
1 Climate change, permafrost thaw & organic matter	7,986
2 Surface temperature & active layer thickness	7,199
3 Construction & frozen soil	6,347
4 Permafrost distribution & rock glaciers	5,576
5 Embankment & heat transfer	4,398
6 Surface water & thermokarst lakes	3,658
7 Natural gas, gas hydrate & methane	1,551
8 Permafrost degradation & pipelines	1,286

Table 3. List of clusters and their number of publications

In Figure 3, nodes are the terms that constitute the basic unit for cluster formation. Of course, a publication may contain several of these terms that do not necessarily belong to the same cluster. As CorTexT enables the assignment of several clusters to a single publication, we can calculate their pairwise intersections and their Jaccard index[4] are displayed in Figure 4. This allows us to have the volume of publications shared by each cluster with all the others and to derive the importance of their relationship at publication level and not only at terms level.

[4] The Jaccard index measures similarity between finite sample sets; it is the size of the intersection divided by the size of the union of the sets. We used the online calculator available here: http://www.molbiotools.com.

		2	3	4	5	6	7	8
		Surface temperature & active layer thickness (n=7199)	Construction & frozen soil (n=6347)	Permafrost distribution & rock glaciers (n=5576)	Embankment & heat transfer (n=4398)	Surface water & thermokarst lakes (n=3658)	Natural gas, gas hydrate & methane (n=1551)	Permafrost degradation & pipelines (n=1286)
1	Climate change, permafrost thaw & organic matter (n=7986)	4094 0.3691	2663 0.2282	2750 0.2543	1845 0.1751	2279 0.2434	668 0.0753	520 0.0594
2	Surface temperature & active layer thickness (n=7199)		2862 0.2679	2991 0.3057	2655 0.2969	1862 0.2070	391 0.0468	466 0.0581
3	Construction & frozen soil (n=6347)			1878 0.1870	2510 0.3048	1183 0.1341	623 0.0856	761 0.1107
4	Permafrost distribution & rock glaciers (n=5576)				1476 0.1737	1305 0.1646	280 0.0409	248 0.0375
5	Embankment & heat transfer (n=4398)					889 0.1240	348 0.0621	502 0.0969
6	Surface water & thermokarst lakes (n=3658)						376 0.0778	306 0.0660
7	Natural gas, gas hydrate & methane (n=1551)							245 0.0945

Figure 4. Clusters pairwise intersections (number of publications and Jaccard index) coloured by Jaccard Index

Cluster 1 (Climate change, permafrost thaw & organic matter) is the largest in terms of the number of bibliographic records (n=7,986) and also in terms of the number of concepts it comprises (57 nodes) linked on average to more than 10 others. This makes it the densest cluster of the network. It deals with the impact of climate change and more precisely of thawing on the soil nature and microbiology (terms: *environmental factors*, *thawing soils*, *active layer*, *organic matter*, *bacteria*, *greenhouse gases*, *soil temperature*, *soil water*). It also contains the climate models that integrate these parameters and the reference to carbon cycle (terms: *greenhouse effects*, *carbon cycle*, *climate systems*, *climate models*).

Cluster 2 (Surface temperature & active layer thickness) is dominated by case studies dedicated to permafrost active layer, that is the surficial layer above permafrost which thaws during summer. Its thickness varies according to surface temperature and snow cover (terms: *thermal regime*, *solar radiation*, *air temperature*, *snow cover*). Numerous studies have been published about those phenomena in the Tibetan Plateau, firstly in the early 1980s to meet the need for the reconstruction of the Qinghai-Tibet highway (3,901 km between Beijing and Lhasa, asphalted in 1985) and in the 2000s thereafter for the construction of the Qinghai-Tibet railway (1,956 km connecting Xining (Qinghai Province) to Lhasa, inaugurated in 2006). Figure 4 shows that Cluster 1 (Climate change, permafrost thaw & organic matter) has the highest rate of similarity with Cluster 2.

Cluster 3 (Construction & frozen soil) is dominated by papers about special design and construction techniques that are required for building on permafrost. Those techniques have to avoid disturbing the thermal balance that preserves the frozen ground and anticipate the consequences of global warming. In all logic, this cluster is partly based on terms relating to infrastructure and civil engineering (terms: *houses*, *bridge*, *pile*, *foundation*, *buildings*, *concrete*). It is important to point out that *socioeconomic impact* is part of this cluster.

Cluster 4 (Permafrost distribution & rock glaciers) is about the mapping of permafrost distribution across the globe and represents a wide array of geophysical studies. Characterizing permafrost distribution and dynamics is important because it helps to estimate ground ice storage and annual water discharge rate for example, and also

to anticipate slope stability problems and thaw-induced landslides (terms: *permafrost occurrence, alpine areas, rock glaciers, geological surveys, ground-penetrating radar*). It is most strongly linked with Cluster 2 (2991/5576 papers, see Figure 4).

Cluster 5 (Embankment & heat transfer) includes another part of the list of terms about infrastructures and civil engineering (terms: *road, railway, highway, tunnel, embankment, pavement)*. Embankment stability and consolidation have been drawing increasing attention due to greater permafrost degradation risk. On that matter, many studies were carried out to assess the safety and efficiency of the Qinghai-Tibet railway construction.

Cluster 6 (Surface water & thermokarst lakes) reflects a large number of heterogeneous publications which nevertheless have in common that they address the role of surface water with studies on the hydrological cycle in cryospheric-dominated watersheds, dissolved organic carbon in river water, permafrost meltwater release or geochemical reactions variations during the ice-free season (terms: *thermokarst lakes, surface water, ice sheet, sea level)*.

Cluster 7 (Natural gas, gas hydrate & methane) represents studies that have been conducted since the 1960s on energy recovery issues, with the potential exploitation of natural gas resources, and mostly gas hydrate reservoirs (terms: *energy resource, hydrate dissociation, natural gas, methane hydrate, gas production)*. On another note, studies focusing on estimating the amount of greenhouse gases (carbon dioxide and/or methane) released from organic matter decomposition or gas hydrate dissociation are rather found in Cluster 1, notably because they are related to works on carbon cycle.

Cluster 8 (Permafrost degradation & pipelines) is the smallest one. It focuses on oil exploration and transportation, including works on the thermal interaction between underground gas pipeline and surrounding permafrost (terms: pipelines, pipe, permafrost degradation, continuous *permafrost, discontinuous permafrost*). Clusters 7 & 8 both deal with energy-related works (production and transport of energy resources) and “socioeconomic impact” acts as a link between Cluster 3 and those two.

Clusters evolution influenced by global warming and China's socio-economic needs

Figure 5 shows how the composition of permafrost research has evolved with an area bump chart displaying both magnitude and rank for the 8 clusters. The frequency count is normalized as a percentage for each year and allows to track the relative weight of each cluster. Figure 5 helps to capture discontinuities in cluster distribution:

- First of all, we can clearly see the progression of Cluster 1 (Climate change, permafrost thaw & organic matter) since the mid-1970s to become the most important in proportion from the mid-1990s with an ever-increasing share.
- On the contrary, Cluster 8 (Permafrost degradation & pipelines) and therefore studies related to energy transportation issues have received much less attention since the end of the 1980s. Surprisingly, Cluster 7 (Natural gas, gas hydrate & methane) is in parallel stable over the whole period.
- Cluster 3 (Construction & frozen soil), after having been dominant until 1990, lost importance in scientific production in favour of environmental issues. However, there has been a resurgence of interest since 2005.
- There has also been a decrease in the share of interest in permafrost distribution, Cluster 4.
- Finally, Cluster 5 (Embankment & heat transfer) is characterized by 2 peaks of interest in 1975 and in 2000-2005, corresponding to the Qinghai-Tibet highway and railway construction sites.

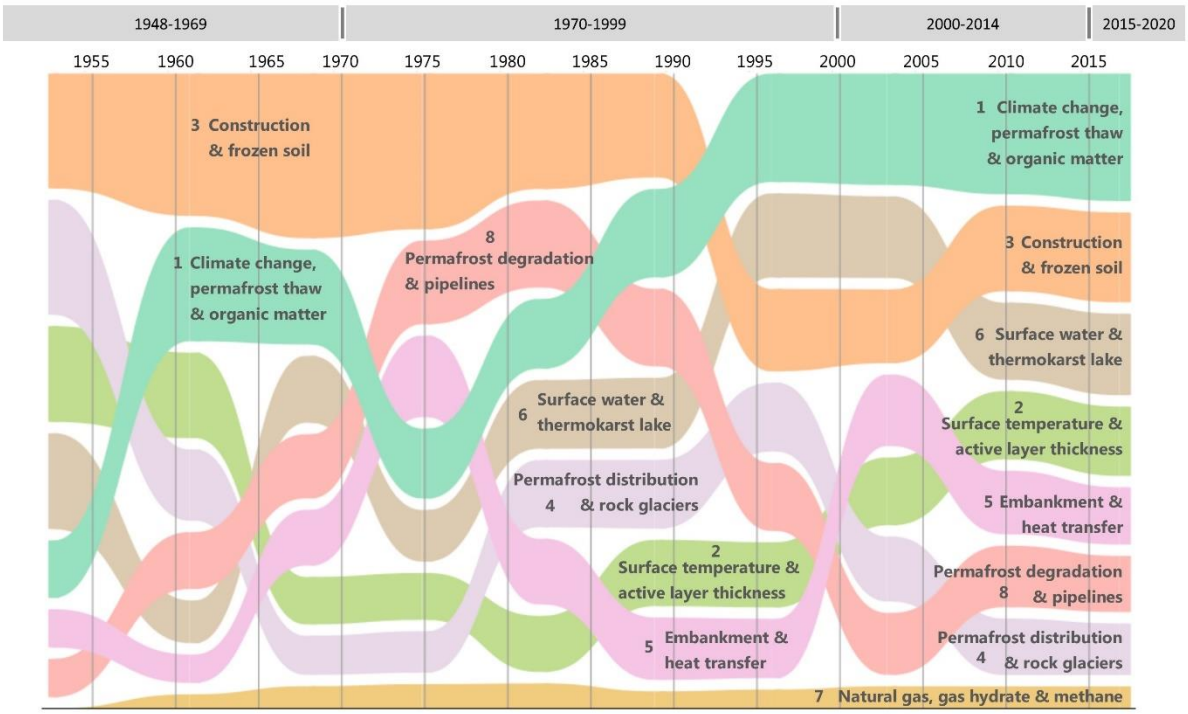

Figure 5. Evolution of clusters share over time (normalization of frequency count)

Research focus changes over time and is uneven across geographical areas

Thanks to CorTexT *Name Entity Recognizer* script, we observed the most frequently cited geographical areas in the corpus and whether the studied areas have changed over the years. Figure 6 clearly shows a growing interest in permafrost areas in China, namely Qinghai-Tibet, which ends up being cited in 12% of the publications in 2019, compared to 1,35% in 2001. Over the whole period, the Arctic is the most studied zone, with an increase in 20 years, from 22% in 2001 to almost ⅓ of the publications dedicated to it in 2019.

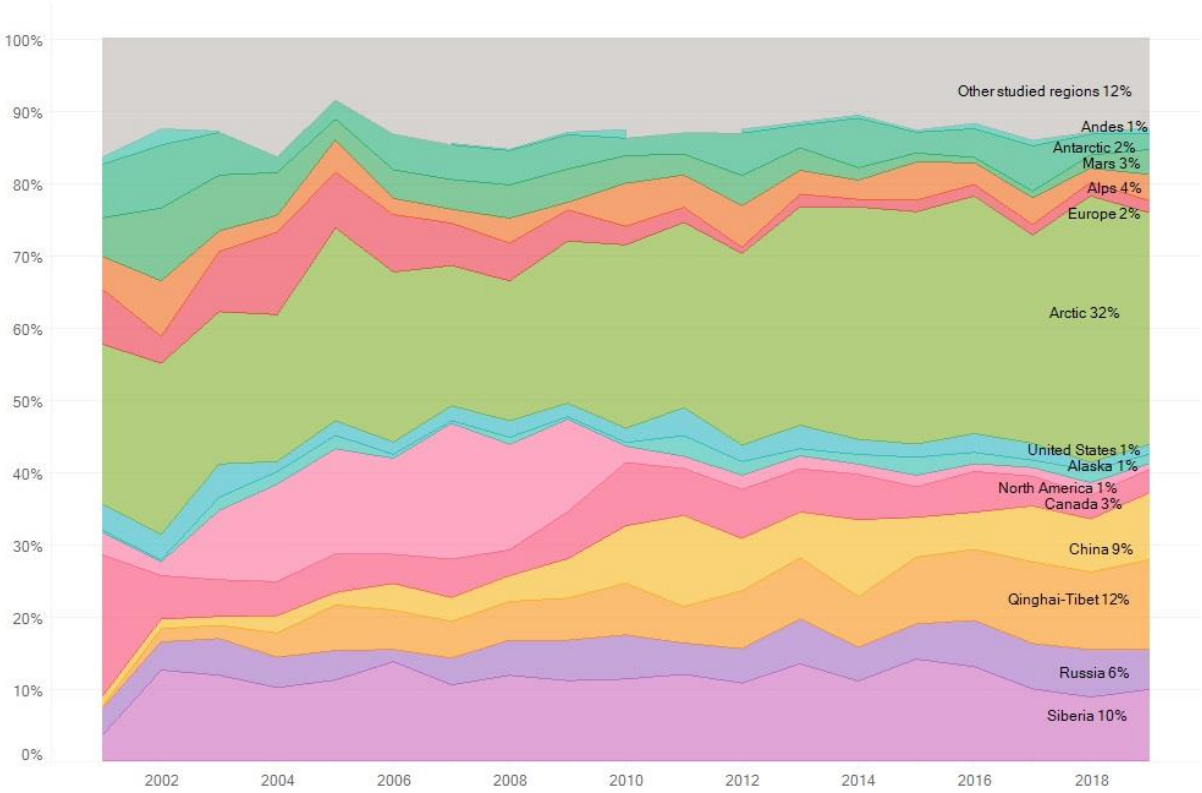

Figure 6. Evolution of studied zones over time (2000-2019)

Thanks to Netscity, we could retrieve the authors' countries and generate a matrix (Figure 7) that compares the studied regions with the studying countries and also shows their respective evolutions over time.
As observed previously, the Arctic is the most studied area, yet mostly by only 3 countries: Russia, the United States and Canada.

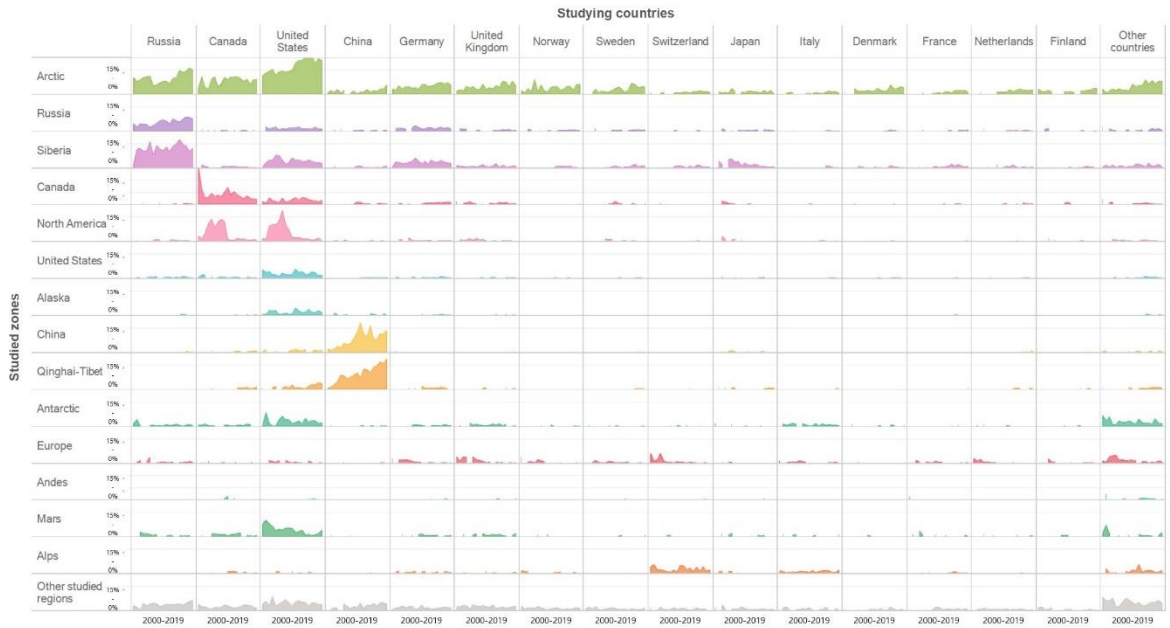

Figure 7. Evolution of studied zones by studying countries over time (2000-2019)

Each country obviously tends to study the regions on its own territory: Siberia by Russia, Alaska by the United States, the Alps mainly by Switzerland and also by Italy, hardly by France. The United States are the most diversified in studying the broadest range of different terrains in addition to the Arctic. On the other hand, China is focused on its territory and even more particularly on the Tibetan area. Lastly, despite up to nearly 10% of publications in 2002, the interest for Mars and Antarctic is decreasing. These two regions were mainly investigated by the United States, which have shifted their attention over the years to focus on the Arctic.
Both figures (Figure 6 and Figure 7) show that over the period 2003-2009, the North America zone has been the focus of particular attention, and that was mainly from Canada and the United States. We have no definite explanation for this and can only formulate 2 hypotheses:

- an important research program might have been funded over this period with the focus on this research area, but we did not find any record of it, and especially it would have been joint between the United States and Canada;
- the term “North America” might be a lexical shift, presumably due to a trend in the scientific community at the time, which made this area stand out more clearly. Indeed, the label “North America” is not fed with any other lexical variant. It is, therefore, literally the original expression "North America" that was frequently used at that time.

Figure 8 shows which are the studied geographic regions according to the clusters. For all clusters, most publications mentioned the Arctic. Leaving the Arctic aside, we can then identify some geographical features per cluster. The studies conducted on the Qinghai-Tibet region significantly contribute to the Embankment & heat transfer (18%), Surface temperature & active layer thickness (16%) and Construction & soil (12%) clusters. This is consistent with the demand for the expertise needed for the Qinghai-Tibet railway and highway projects. No other clear focus is observed inside the Construction & soil cluster.
The Surface water & thermokarst lakes cluster is largely fed by studies in North America, Canada and also Russia and Siberia. The United States and Alaska are the preferential areas for the pipeline cluster. Finally, 15% of the Natural gas, gas hydrate & methane cluster consist of publications about regions in China (other than Qinghai-Tibet) and also Canada.

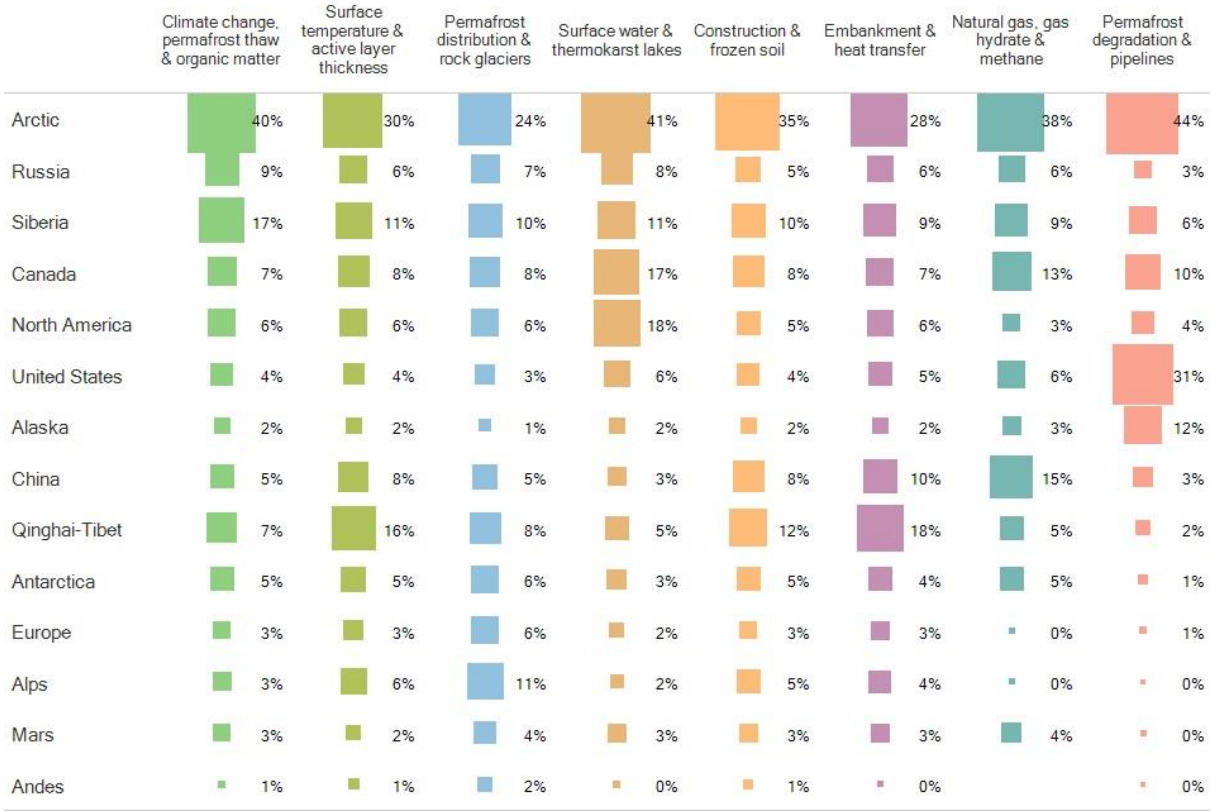

Figure 8. Distribution matrix of clusters and studied zones – For each cluster, percentage of publications mentioning each studied zone

Finally, by cross-referencing all our data, we are able to focus on the Arctic zone and show that the 8 Arctic countries, that is those claiming Arctic territories (i.e.: Norway, Sweden, Finland, Russia, the United States, Canada, Denmark, and Iceland) do not have different research targets than all the other non-Arctic countries. Figure 9 shows on which clusters Arctic research is distributed.

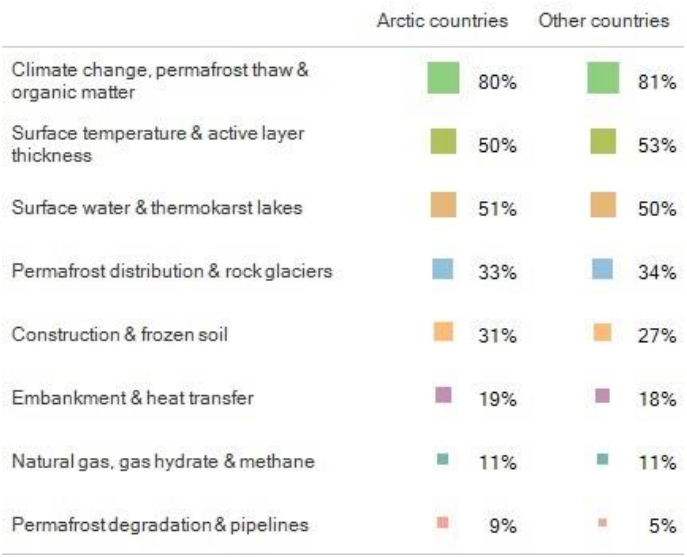

Figure 9. Arctic countries vs others: percentage of publications about the Arctic zone contributing to each cluster

Focus on soil properties and infrastructure in permafrost areas

Finally, we would like to demonstrate that the combination of classical text-mining processes with the use of carefully tailored lists of qualified terms opens up opportunities to identify gaps within the knowledge map of a specific topic. As an illustration, we chose to focus on infrastructures in permafrost areas. The idea is to measure how research was conducted on a complex topic that, *a priori*, involves specificities related to construction and maintenance of infrastructures on permafrost, including not only the description of soil properties and their characterization but also the evolution of these latter with climate change to assess the potential vulnerability of infrastructure. On one hand, permafrost thaw-induced degradation of soil properties might damage infrastructures built on it, and a proper estimation of these effects is valuable in helping owners and governments to anticipate increased maintenance costs. On the other hand, the knowledge on permafrost soils acquired by the geotechnical community over the years could benefit more globally to environmental scientists interested in permafrost. In other words, the objective of this particular focus is to determine how or if research on climate change and research on infrastructure mutually feed each other.

First, we compare the frequency of terms retrieved for Clusters 1 (Climate change, permafrost thaw & organic matter) and 3 (Construction & frozen soil).

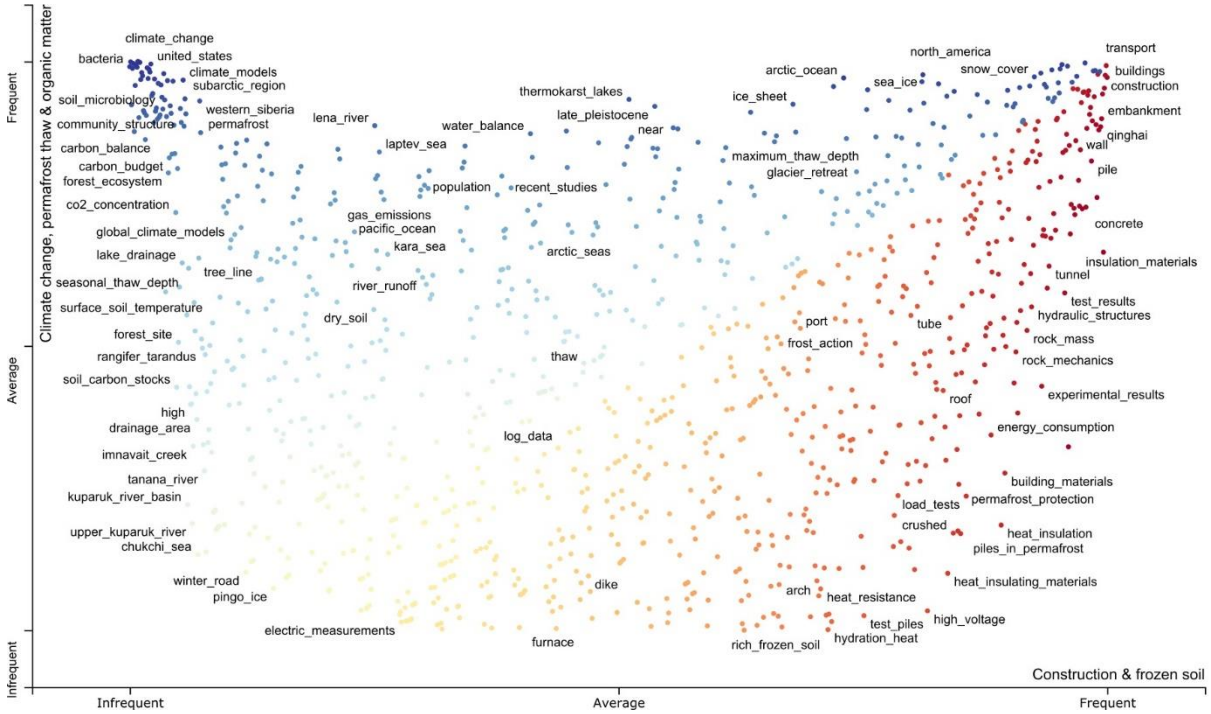

Figure 10. Contrast analysis of terms from Clusters 1 (blue) and 3 (red). Colour intensity increases with frequency.

Figure 10 shows that there is not a convergence of all the terms as far as their frequency of appearance per cluster is concerned: the figure reveals that terms related to soil properties and microbiology are only frequent in Cluster 1 and not at all in Cluster 3. We can confirm this idea with a simplified version of the semantic map (Figure 11) and use soil properties related terms as a proxy to measure the existing interactions between two distinct research communities.

Figure 11 shows indeed that terms referring to infrastructure or civil engineering concepts mainly appear in Clusters 3 and 5. Cluster 3 also contains some terms related to soil properties but these are mostly in Cluster 1. This mapping shows a clear demarcation between these two lists of terms and indicates a weak interaction. This distance between the terms associated with climate change and those used in civil engineering raises the question of whether the consequences of global warming are well enough taken into account for the construction of resilient infrastructures or the assessment of the vulnerability of existing infrastructure.

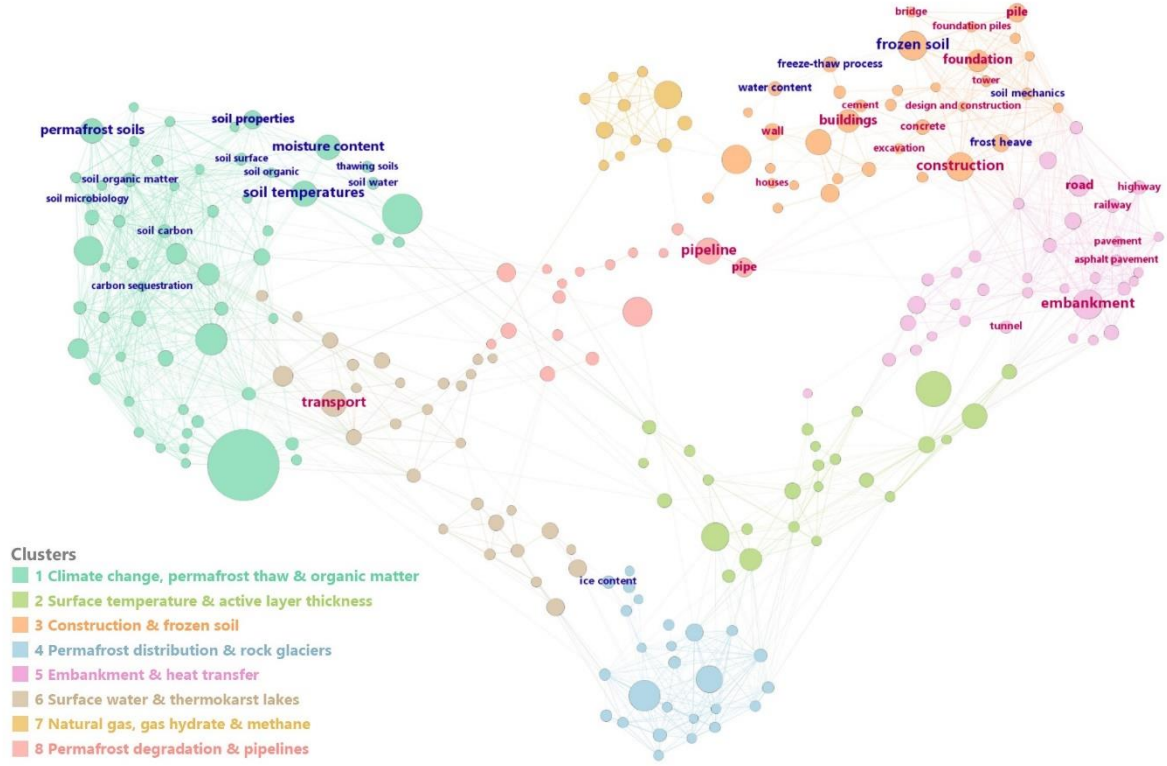

Figure 11. Simplified co-occurrence map of extracted terms - Contrast between soil properties terms (blue) vs infrastructures terms (red) - (n=239 nodes, node size = frequency, colour = cluster delineation)

Discussion and conclusion

The analysis of bibliometric indicators (journal classification and number of publications) and the semantic map we suggested shows that the scientific production associated with permafrost is increasingly substantial, and is also multidisciplinary and multi-topical. With the diachronic analysis of clusters, we observed that the interest had turned logically, since the 1980s towards the role of climate change but also on China's needs for its railway and highway construction sites. The very strong and growing impact of Chinese research, focused mainly on the Tibetan area, is one of the highlights of our data.
Otherwise, the Arctic is, unsurprisingly, the most studied target area, and is now mentioned in ⅓ of the publications; next are Siberia and Russia, with a constant interest over time and also a balanced contribution to all research topics. The United States are well diversified in terms of the areas they study, and at the same time, they inherently constitute an important studied area for research about oil transportation.

Furthermore, the climate change positive feedback loop in degrading permafrost ecosystems is widely supported by the scientific community and microbiological degradation is known to lead to increased carbon and methane emissions. The cluster representation we proposed shows the intricacy of climate change and microbiology terms embedded in the same cluster (Climate change, permafrost thaw & organic matter). On the other hand, while the impact of permafrost thawing on infrastructures is also known (with high confidence according to the IPPC report), studies on the topic do not include this "part of the loop". Indeed, with the focus we proposed to identify the terms of civil engineering and soil properties, we pointed out a demarcation. While our study easily proves that permafrost research is multidisciplinary and multi-topical, it also shows that all findings and new knowledge do not interact as much as they could or should. This suggests that research has mainly focused so far on the

feasibility of building on frozen ground and the possible use of these soils, but remains at an early stage of addressing the impact of global warming on infrastructure degradation and its resilience. This is clearly an opportunity for further research in order to find appropriate local solutions and avoid heavy costs and dramatic consequences for communities. This research gap revealed by our data is confirmed by the limited importance of socioeconomic impacts in the publications of our corpus. The term "socioeconomic impacts" does, in fact, appear in the semantic mapping: the node is in the cluster related to the construction, at the border of the 2 clusters dealing with energy resources. However, there is no detail, i.e. terms suggesting human relocation, private and public budgeting, analysis of cost damages do not appear. This does not mean that they do not exist at all, but simply that they are too infrequent to be displayed in the graph, whereas they are present in the IPCC report on the cryosphere. This cannot be attributed to the fact that Scopus is a generalist bibliographic database where social sciences journals are less well represented insofar as our corpus includes 1,319 publications in social sciences, mainly in the geography, planning and development subject area, which is not marginal.

Although the corpus is quite large, the text-mining processes we performed enabled us to go into great details and to provide an accurate review of permafrost literature. It is not possible to compare our results with other studies since there are no similar ones, not even reviews over a large enough corpus. Nevertheless, we can assume that the findings would be even better if full-texts and not only Title-Abstract-Keywords metadata were taken into account. This could also have avoided a possible bias due to the fact that abstracts have evolved over time, and in particular have proved to become more informative (Ermakova et al. 2018). But on the other hand, the undertaking of collecting all the full-texts would have been daunting.

We hope this study will not only offer insights to permafrost experts, but also provide a methodology that could be reused for other investigations.

Reference list

Aditya, L., Mahlia, T. M. I., Rismanchi, B., Ng, H. M., Hasan, M. H., Metselaar, H. S. C., et al. (2017). A review on insulation materials for energy conservation in buildings. *Renewable and Sustainable Energy Reviews*, *73*, 1352–1365. https://doi.org/10.1016/j.rser.2017.02.034

Afouda, P., Dubourg, G., Labas, N., Raoult, D., & Fournier, P. E. (2017). Draft genome sequence of Agrococcus baldri strain Marseille-P2731. *Genome Announcements*, *5*(10). https://doi.org/10.1128/genomeA.00015-17

Aksnes, D. W., & Hessen, D. O. (2009). The Structure and Development of Polar Research (1981–2007): a Publication-Based Approach. *Arctic, Antarctic, and Alpine Research*, *41*(2), 155–163. https://doi.org/10.1657/1938-4246-41.2.155

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, *4*(1), 1–21. https://doi.org/10.2478/jdis-2019-0001

Bibi, S., Wang, L., Li, X., Zhou, J., Chen, D., & Yao, T. (2018). Climatic and associated cryospheric, biospheric, and hydrological changes on the Tibetan Plateau: a review. *International Journal of Climatology*, *38*, e1–e17. https://doi.org/10.1002/joc.5411

Bordignon, F. (2020). Data for: "A scientometric review of permafrost research based on textual analysis (1948-2020)." Mendeley Data. https://doi.org/10.17632/d8gvm96ykm.1

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Canavan, B. C. (2019). Opening Pandora's Box at the roof of the world: Landscape, climate and avian influenza (H5N1). *Acta Tropica*, *196*, 93–101. https://doi.org/10.1016/j.actatropica.2019.04.021

Cardon, V. (2020). Bounded Futures: Growing a Boundary Foreknowledge Infrastructure in Food Security Research. *Science, Technology and Society*, *25*(1), 38–66. https://doi.org/10.1177/0971721819889918

Chang, Q., Sun, Z., Ma, R., Wang, X., & Long, X. (2016). A review of groundwater flow and its interaction with surface water in permafrost region. *Advances in Science and Technology of Water Resources*, *36*(5), 87–94. https://doi.org/10.3880/j.issn.1006-7647.2016.05.016

Colombo, N., Salerno, F., Gruber, S., Freppaz, M., Williams, M., Fratianni, S., & Giardino, M. (2018). Review: Impacts of permafrost degradation on inorganic chemistry of surface fresh water. *Global and Planetary*

Change, *162*, 69–83. https://doi.org/10.1016/j.gloplacha.2017.11.017
Côté, G., & Picard-Aitken, M. (2009). *Arctic research in Canada a bibliometric analysis*. www.science-metrix.com. Accessed 15 December 2019
Ermakova, L., Bordignon, F., Turenne, N., & Noel, M. (2018). Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences. *Frontiers in Research Metrics and Analytics*, *3*. https://doi.org/10.3389/frma.2018.00016
García-Hernández, C., Ruiz-Fernández, J., & Serrano-Cañadas, E. (2019). Social network analysis in Geosciences: scientific collaboration between periglacial scholars in the Iberian Peninsula. *Cuadernos de Investigación Geográfica*. https://doi.org/10.18172/cig.3939
Grosse, G., Goetz, S., McGuire, A. D., Romanovsky, V. E., & Schuur, E. A. G. (2016). Changing permafrost in a warming world and feedbacks to the Earth system. *Environmental Research Letters*, *11*(4). https://doi.org/10.1088/1748-9326/11/4/040201
Grosse, G., & Lantuit, H. (2008). PYRN-Bib 3.1: The Permafrost Young Researchers Network Bibliography of Permafrost-Related Theses. *EPIC3Permafrost Young Researchers Network.1, 3, 49 p.*
He, Q. (1999). Knowledge Discovery Through Co-Word Analysis. *Library Trends*, *48*(1), 133–159.
Hjort, J., Karjalainen, O., Aalto, J., Westermann, S., Romanovsky, V. E., Nelson, F. E., et al. (2018). Degrading permafrost puts Arctic infrastructure at risk by mid-century. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07557-4
Hua, W., Yuan, S., Yan, M., & Li, Y. (2012). A quantitative analysis of Arctic related articles in the humanities and social sciences appearing in the world core journals. *Scientometrics*, *91*(3), 703–718. https://doi.org/10.1007/s11192-012-0690-0
Ikram, M. T., & Afzal, M. T. (2019). Aspect based citation sentiment analysis using linguistic patterns for better comprehension of scientific knowledge. *Scientometrics*, *119*(1), 73–95. https://doi.org/10.1007/s11192-019-03028-9
Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679
Leibman, M., Khomutov, A., Gubarkov, A., Mullanurov, D., & Dvornikov, Y. (2015). The research station "Vaskiny Dachi", Central Yamal, West Siberia, Russia - A review of 25 years of permafrost studies. *Fennia*, *193*(1), 3–30. https://doi.org/10.11143/45201
Li, X. Sen, Xu, C. G., Zhang, Y., Ruan, X. K., Li, G., & Wang, Y. (2016). Investigation into gas production from natural gas hydrate: A review. *Applied Energy*, *172*, 286–322. https://doi.org/10.1016/j.apenergy.2016.03.101
Maisonobe, M., Jégou, L., Yakimovich, N., & Cabanac, G. (2019). NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities. In *International Conference on Scientometrics and Informetrics (ISSI 2019), Sep 2019, Rome, Italy*.
Moretti, F. (2013). *Distant Reading*. Verso.
Overland, J., Dunlea, E., Box, J. E., Corell, R., Forsius, M., Kattsov, V., et al. (2019). The urgency of Arctic change. *Polar Science*, *21*, 6–13. https://doi.org/10.1016/j.polar.2018.11.008
Raimbault, B., Cointet, J.-P., & Joly, P.-B. (2016). Mapping the Emergence of Synthetic Biology. *PLOS ONE*, *11*(9), e0161522. https://doi.org/10.1371/journal.pone.0161522
Roseau, N. (2016). Pouvoirs des infrastructures. *Histoire Urbaine*, *45*(1), 5–16. https://doi.org/10.3917/rhu.045.0005
Serrano Cañadas, E. (2016). Periglaciarismo y Permafrost. *Polígonos. Revista de Geografía*, *28*(28), 15. https://doi.org/10.18002/pol.v0i28.4283
Sjöberg, Y., Siewert, M. B., Rudy, A. C. A., Paquette, M., Bouchard, F., Malenfant-Lepage, J., & Fritz, M. (2020). Hot trends and impact in permafrost science. *Permafrost and Periglacial Processes*, ppp.2047. https://doi.org/10.1002/ppp.2047
Sutherland, W. J., Broad, S., Butchart, S. H. M., Clarke, S. J., Collins, A. M., Dicks, L. V., et al. (2019). A Horizon Scan of Emerging Issues for Global Conservation in 2019. *Trends in Ecology & Evolution*, *34*(1), 83–94. https://doi.org/10.1016/j.tree.2018.11.001
Tancoigne, E., Barbier, M., Cointet, J. P., & Richard, G. (2014). The place of agricultural sciences in the literature on ecosystem services. *Ecosystem Services*, *10*, 35–48. https://doi.org/10.1016/j.ecoser.2014.07.004
Trubl, G., Jang, H. Bin, Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., et al. (2018). Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems*, *3*(5). https://doi.org/10.1128/mSystems.00076-18
Ugwuishiwu, B. O., Echiegu, E. A., Okorie, E. T., & Nwakaire, J. N. (2019). Climate change and vulnerablelity of engineering infrastructures: A critical review. In *2019 ASABE Annual International Meeting*. American Society of Agricultural and Biological Engineers. https://doi.org/10.13031/aim.201900069
Vonk, J. E., Tank, S. E., Bowden, W. B., Laurion, I., Vincent, W. F., Alekseychik, P., et al. (2015). Reviews and

Syntheses: Effects of permafrost thaw on arctic aquatic ecosystems. *Biogeosciences Discussions*, *12*(13), 10719–10815. https://doi.org/10.5194/bgd-12-10719-2015

Walvoord, M. A., & Kurylyk, B. L. (2016). Hydrologic Impacts of Thawing Permafrost-A Review. *Vadose Zone Journal*, *15*(6). https://doi.org/10.2136/vzj2016.01.0010

Weber, S., Beutel, J., Da Forno, R., Geiger, A., Gruber, S., Gsell, T., et al. (2019). A decade of detailed observations (2008–2018) in steep bedrock permafrost at the Matterhorn Hörnligrat (Zermatt, CH). *Earth System Science Data*, *11*(3), 1203–1237. https://doi.org/10.5194/essd-11-1203-2019

Weeds, J., & Weir, D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, *31*(4), 439–475. https://doi.org/10.1162/089120105775299122

Wooller, M. J., Ruppel, C. D., Pohlman, J. W., Leigh, M. B., Heintz, M., & Anthony, K. W. (2009). Permafrost gas hydrates and climate change: Lake-based seep studies on the Alaskan north slope. *Fire in the Ice: NETL Methane Hydrate Newsletter*, *9*(3), 6–9. http://pubs.er.usgs.gov/publication/70190381

Wu, Q., Liu, Y., Zhang, J., & Tong, C. (2002). A review of recent frozen soil engineering in permafrost regions along Qinghai-Tibet Highway, China. *Permafrost and Periglacial Processes*, *13*(3), 199–205. https://doi.org/10.1002/ppp.420

Xiao, C., Tian, L., Wu, Q., Zhang, D., Zhang, T., Wu, T., et al. (2017). Global Cryosphere Evolution and Land Surface Processes on the Tibetan Plateau. In *The Geographical Sciences During 1986—2015* (pp. 263–279). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-1884-8_11

Yang, L., Liu, Y., Zhang, H., Xiao, B., Guo, X., Wei, R., et al. (2019). The status of exploitation techniques of natural gas hydrate. *Chinese Journal of Chemical Engineering*, *27*(9), 2133–2147. https://doi.org/10.1016/j.cjche.2019.02.028

Yang, M., Nelson, F. E., Shiklomanov, N. I., Guo, D., & Wan, G. (2010). Permafrost degradation and its environmental effects on the Tibetan Plateau: A review of recent research. *Earth-Science Reviews*, *103*(1–2), 31–44. https://doi.org/10.1016/j.earscirev.2010.07.002

Zhang, T. (2005). Influence of the seasonal snow cover on the ground thermal regime: An overview. *Reviews of Geophysics*, *43*(4), RG4002. https://doi.org/10.1029/2004RG000157

Zimov, S. A. (2006). Climate change: Permafrost and the Global Carbon Budget. *Science*, *312*(5780), 1612–1613. https://doi.org/10.1126/science.1128908