



Developing a spatially informed matching workflow to link glacier names to unnamed geometries in the Randolph Glacier Inventory through external toponym sources

GEO 511 Master's Thesis

Author: Laura Bozzi, 19-916-568

Supervised by: Prof. Dr. Ross Purves, Prof. Dr. Michael Zemp, Dr. Ethan Welty

Faculty representative: Prof. Dr. Ross Purves

19.12.2025

Abstract

The Randolph Glacier Inventory (RGI) provides global glacier coverage but includes a vast number of unnamed glacier features, limiting its value for scientific communication and cultural preservation. This thesis presents a geometry-aware matching framework designed to automatically link RGI glacier polygons to possibly correct names using openly available, external data sources. The algorithm evaluates spatial similarity through overlap, containment and distance-based metrics, and generates ranked name suggestions for unnamed glaciers. It can also be applied to named glacier features for a comparison and disambiguation of name information.

The framework was applied to both a validation dataset of 254 named reference glaciers from the World Glacier Monitoring Service (WGMS) and 263,479 unnamed RGI glaciers. Validation results demonstrated high precision, with a mean average precision of 0.915 and ranked precision at one of 0.901. Globally, 46,468 glacier names were matched to unnamed RGI glaciers, of which over 12,417 matches exhibited very high geometric accuracy. Analysis revealed that the World Glacier Inventory contributes the largest number of new glacier names, followed by OpenStreetMap.

The results confirm that glacier geometries can be reliably linked to potential toponyms through an automated, spatially informed workflow applicable across diverse regions and geometry types. Furthermore, while external datasets can significantly enrich the RGI with additional names, the main limitation lies in the regional availability and completeness of reference data rather than in the matching method itself. Validation against WGMS ground truth data verified the robustness of the approach and supported the definition of geometric thresholds distinguishing correct from incorrect matches.

Although constrained by data availability and the experimental nature of the evaluation scheme, this research demonstrates the feasibility of automated name to geometry linkage on a global scale. Future work may enhance usability through a graphical user interface and by incorporating toponym data from historical documents and maps. Beyond its glaciological context, the study contributes to geographical information retrieval by advancing methods for spatially linking heterogeneous and incomplete geospatial datasets. Ultimately, the developed framework meets the WGMS's need for a solution to automatically manage missing names in the RGI while substantially reducing manual effort and providing a robust foundation for future data integration.

Keywords: glaciers, glaciology, toponyms, matching, entity linking, gazetteers

Acknowledgements

I hereby express my deepest thanks to Ross S. Purves, who supported me with knowledge on geographic information retrieval, insight about methodology and help me plan my milestones. He also realised when I needed more help and accordingly arranged the one or other more meeting.

I am also thankful to thank Michael Zemp, whose passion about glaciers is contagious! His knowledge about various glacier organisations and glaciers for case studies was essential in my work.

Moreover, sincere thanks to Ethan Welty for supporting me in code and data-related questions. He offered me a great knowledge base about GLIMS and RGI data structure, as well as many datasets in form of a Google Spreadsheet. This was invaluable information, and generating the matching framework would not have been possible without his great insights.

Furthermore, thanks to Nicolas Alou, who took the time to read my thesis and provide some important remarks.

Lastly, a heartfelt *mille grazie* goes to my partner Philipp for his constant and unwavering support. His knowledge and time taken for correction has also played a role in the hopeful success of my project. His reassurance and motivation sustained me throughout this year.

Contents

1	Introduction	1
2	Literature review	4
2.1	Background	4
2.1.1	Glacier organisations	4
2.1.2	Glaciers	5
2.1.3	Toponyms	6
2.1.4	Gazetteers	8
2.2	Reverse geocoding & toponym resolution	10
2.3	Spatial matching strategies	11
2.3.1	Intersection over Union (IoU)	13
2.3.2	On points and polygons	14
2.4	Research gap	15
3	Data & Database	16
3.1	Definition of <i>unnamed</i>	16
3.2	GLIMS & RGI	18
3.2.1	GLIMS	18
3.2.2	RGI	19
3.2.3	Comparison RGI version 6.0 to 7.0	21
3.3	Reference data	23
3.3.1	Glacier inventories	23
3.3.2	Gazetteers	24
3.3.3	Examples of regional data availability	25
3.3.4	Glacier name coverage in Switzerland	28
3.4	Database	29
3.4.1	Data preprocessing	29
4	Methods	33
4.1	Programmes and tools	33
4.2	Spatial extent and projection	34
4.2.1	Map projection	34
4.3	Metric behaviour	35
4.3.1	Absolute intersection	35
4.3.2	Relative intersection	35
4.3.3	Centroid distance	37
4.3.4	Edge distance	38
4.4	Unit test: evaluation of metric performance	40
4.4.1	Unit test design	40
4.4.2	Unit test results	41
4.4.3	Discussion of unit test	42
4.5	Matching algorithm	44
4.5.1	Input parameters	46
4.5.2	Metrics used	47
4.5.3	Scoring	50

4.5.4	Batch processing over many glaciers	52
4.5.5	Encoding	52
4.6	Validation methods	53
4.6.1	String matching	54
4.7	RGI regions methods	55
4.8	Evaluation	55
5	Results	56
5.1	Validation results	56
5.1.1	Validation: special cases	59
5.2	RGI regions results	62
5.3	Contribution of reference datasets	69
5.3.1	Global dataset contribution	69
5.3.2	Regional dataset contribution	70
5.4	Case studies	72
5.4.1	Case 1: Unnamed RGI polygon	72
5.4.2	Case 2: Comparison for a well-known glacier	73
5.4.3	Case 3: Unnamed polygon with clear match	75
6	Discussion	76
6.1	Research question 1: methodology	77
6.2	Research question 2: results	79
6.3	Research question 3: validation	82
6.4	Limitations	83
6.5	Outlook	85
7	Conclusion	86
8	Data and code availability	88
8.1	Code and environment	88
8.2	Database & data	88
9	Appendix	95
9.1	Tables of unit test	95
9.2	Results	96
9.2.1	RGI region results	96
9.2.2	Global contribution of reference datasets	97
9.2.3	National contribution of reference datasets	99

Glossary

CD Centroid Distance

ED Edge Distance

FoG Fluctuations of Glaciers

GLIMS Global Land Ice Measurements from Space

GTN-G Global Terrestrial Network for Glaciers

IoU Intersection over Union

NSIDC National Snow and Ice Data Center

OSM OpenStreetMap

RGI Randolph Glacier Inventory

WGMS World Glacier Monitoring Service

1 Introduction

Glacial ice covers about 10 percent of the Earth's surface, including glaciers, ice caps, and the ice sheets of Greenland and Antarctica (National Snow and Ice Data Center, 2025). However, glaciers are melting on a global scale, which has a large impact on meltwater as a water resource, e.g., in the Andes or Himalaya, as well as glacier's, especially Alpine glaciers contribution to global sea-level rise (Racoviteanu et al., 2009). Therefore, due to their changing and dwindling nature, glaciers as they exist today are highly relevant topic of research. Moreover, 2025 was declared the year of glacier preservation (WGMS, 2025b), emphasising the global importance of glacier monitoring. Yet most glaciers in the Randolph Glacier Inventory (RGI) lack a recorded name. In fact, of the 274,531 glacier entries in the RGI dataset, only 11,052 (4%) contain name information, leaving 263,479 glaciers unnamed.

This gap complicates scientific communication, impairs data integration and risks the loss of valuable cultural and historical knowledge. Although research can also be conducted with only the ID of a glacier, names carry considerable value for organisations such as the World Glacier Monitoring Service (WGMS), which rely on clear communication and user-friendly data presentation. For example, on the WGMS website, one can look at diverse glacier profiles, which are much more meaningful with a glacier name rather than only an ID. Given this need, the WGMS is seeking an appropriate method to automatically collect, manage and analyse the glacier name information from gazetteers and other openly accessible web sources. Consequently, this thesis addresses the overarching research question: How can glacier geometries in the Randolph Glacier Inventory be automatically matched with their correct glacier names using external, openly available sources?

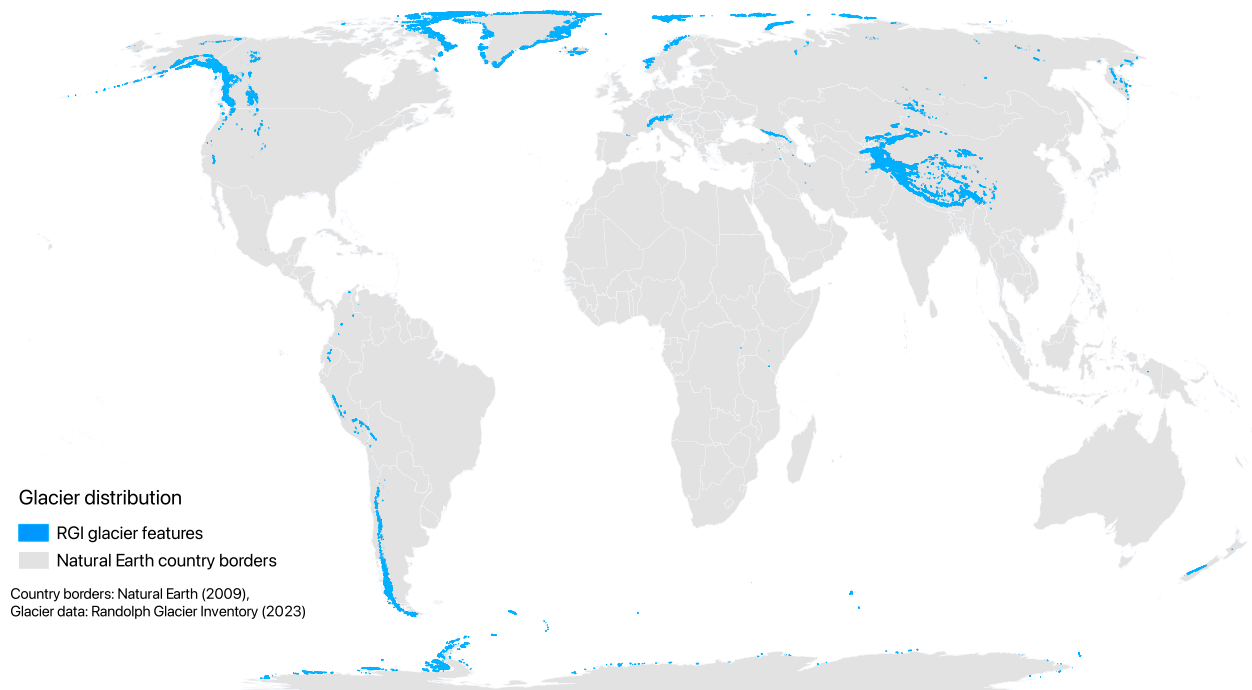


Figure 1.1: World map of all glaciers in the Randolph Glacier Inventory (RGI). Country borders with Natural Earth (Natural Earth, 2009), map projection: Eckert IV (ESRI:54012).

In response to this challenge, this thesis develops a geometry-aware matching workflow that identifies candidate names from fourteen polygon and ten point datasets, evaluates spatial similarity using intersection based metrics and distance measures, and presents ranked name suggestions for unnamed RGI glaciers. The workflow was implemented in Python and PostGIS, validated on a manually annotated sample, and initially tested in Switzerland before being generalised to global data.

To operationalise this broad question and turn it into a framework, three more specific research questions were formulated:

1. How can glacier geometries in the Randolph Glacier Inventory (RGI) be linked to glacier toponyms using an automated, spatially informed matching workflow?
2. To what extent can gazetteers and glacier inventories contribute additional glacier name information to the RGI, and how effectively can these heterogeneous sources be integrated?
3. How accurately can spatial entity-linking techniques based on overlap, containment, and distance metrics identify correct name-geometry pairs in a validation setting?

Essentially, this study addresses a coverage gap for glacier names in the RGI database and the matching of names to glaciers via their geometry and location. This project started in the assumption that every glacier has at least one name. However, as research and data exploration proceeded, it became evident that most glaciers are rather remote and not close to human settlement or scientific activity. What is more, some glaciers may have been given names in the course of history but never written down or only used in local language. Therefore, it cannot be assumed that every glacier has a least one name.

Any geographer learns of Tobler's first law of geography, which states that everything is related to everything, but close things are more related than distant things (Waters, 2017). In essence, this is what is taken as a starting point in this thesis' approach, namely the geographic closeness of data. Therefore, a "match" between features is when two things are so close they can be viewed as virtually the same entity.

While there is plenty of literature on glaciers, toponyms and geometry-aware linkage across heterogeneous data sources, no prior work was found to have combined all these topics. This thesis further touches on string matching, using it to validate the geometric matching algorithm. What is more, data availability and quality are of key interest, along with the distinction of polygon and point data - two geometry types very common in known gazetteers such as OpenStreetMap and GeoNames. This thesis contributes a geometry-aware, multi-source matching workflow for linking glacier names to unnamed RGI polygons. It further delivers a systematic validation of the workflow based on WGMS ground-truth data. Moreover, all reference data used for matching has been merged into a harmonised database, which will be made available for the WGMS to use.

This study also contributes to interdisciplinary research, bridging glaciology and geographic information science, more specifically geo-computation. Developing automated methods to link glacier names to their geometry will support organisations like the WGMS in improving data quality and usability. The research outcomes might also directly inform the design and improvement of gazetteers, such as GeoNames and OpenStreetMap (OSM), making them more robust for scientific applications. Finally, this thesis does not attempt to resolve historical or cultural naming disputes, but focuses on identifying the most spatially plausible reference name for each unnamed RGI glacier. It can further be applied to named glacier for resolving name ambiguities, which has not been done on the large scale within this thesis.

Previous work focused on spatial matching strategies will be reviewed, along with the ground-work laid by studies on toponyms and gazetteers. Further, glacier organisations and data they provide are introduced in more detail, before laying out the reference data used as toponym sources. Then a framework based on the relational matching approach by Walter and Fritsch (1999) is presented and adapted to glacier features. The results are presented graphically and numerically, before their implications are discussed, particularly in terms of the success of the matching workflow and the contribution of the toponym sources for spatial matching tasks. The work concludes with shortcomings of the reference data and possibly the framework, and proposes practical future research focused on addressing this.

2 Literature review

In this section, the theoretical groundwork for this thesis will be laid, providing insight into the current state of the research. Further, the most important organisations conducting glacier research and possible data sources of this thesis will be presented. Lastly, the research questions will be stated.

2.1 Background

2.1.1 Glacier organisations

The globally coordinated monitoring of glaciers began in 1894 and today includes more than 200,000 glaciers worldwide (WGMS, 2021). While glacier monitoring centred around glacier fluctuations, a need for a worldwide inventory arose. Hence, the *Temporal Technical Secretariat for the World Glacier Inventory* (TTS/WGI) was created in 1975 (Zemp, 2011). In 1999, the Global Land Ice Measurements from Space (GLIMS) initiative and database was launched with space-borne based images (Zemp, 2011). It now holds information on roughly 200,000 glaciers, covering an area of 750,000 km², which however does not include the ice sheets of Greenland and Antarctica (National Snow and Ice Data Center, 2024). Gärtner-Roer et al. (2022) mention around 215,000 glaciers in RGI version 6.0 - of which 40% have multi-temporal outlines. As GLIMS is multi-temporal, there are over 745,544 glacier analyses in total (status October 2025). More glacier inventories were created over the years, which will be discussed further on in this thesis. From the GLIMS database, a subset was created to achieve a global coverage of glacial ice around the year 2000, this is called the Randolph Glacier Inventory (RGI), and is a focal part of this thesis.

Global Terrestrial Network for Glaciers (GTN-G)

Run by the World Glacier Monitoring Service (WGMS), the GTN-G is an informal cooperation between and a framework created by the NSIDC and GLIMS (Zemp, 2011). Built up by the operational bodies of WGMS, NSIDC and GLIMS, its success is drawn through glacier monitoring and making data openly available (Zemp, 2011).

Above the mentioned operational bodies, the GTN-G has a Steering Committee, consisting of the Executive and the Advisory Boards. The GTN-G Executive Board develops and implements international strategies and standards for monitoring glaciers and ice caps, coordinates cooperation among the WGMS, NSIDC, and GLIMS, and compiles and distributes standardised data on glacier fluctuations. The Advisory Board provides guidance, consultation, and periodic evaluations of the Executive Board's work, advising on best practices and future directions for glacier and ice cap monitoring and data delivery to the scientific community.

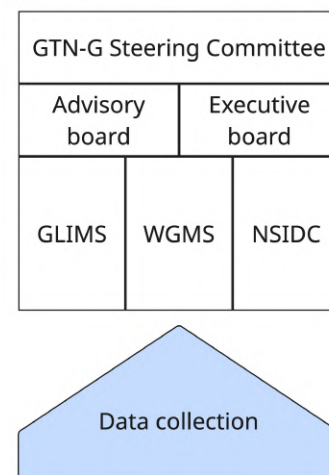


Figure 2.1: GTN-G central organisational structure, own illustration based on (Zemp, 2011, p.14)

While the first consists of a board member of the WGMS, GLIMS and NSIDC each, the latter consists of representatives of data users and producers, as well as from scientific organisations (Zemp, 2011). Overall, the GTN-G serves as a coordinating framework for various monitoring networks, bringing together existing and active observation services. Its purpose is to enhance information exchange while tackling challenges related to data accessibility, availability, and the standardisation of measurement techniques (WGMS, 2022b).

World Glacier Monitoring Service (WGMS)

The World Glacier Monitoring Service (WGMS) has been collecting and maintaining global glacier data since 1986 (Zemp, 2011). It emerged from earlier international glacier monitoring efforts that began in 1894. WGMS systematically records glacier changes in mass, volume, area, and length over time, as well as their global distribution. The key expertise of the WGMS lies on in-situ measurements collected by researchers in their collaboration network. This data is crucial for climate monitoring, hydrological modelling, and glaciological research. Operated under international scientific and UN organisations, WGMS is based at the University of Zurich and serves as a key resource for research, policy guidance, and capacity building (Zemp, 2011).

Global Land Ice Measurements from Space (GLIMS)

The *Global Land Ice Measurements from Space* (GLIMS) initiative is a universal effort to monitor and map the Earth's glaciers using satellite imagery from the ASTER (Terra) and LANDSAT instruments, along with historical records (GLIMS Consortium, 2005). Hence, the GLIMS inventory has a clear focus on remote sensing deduced data. Initiated in 1999, GLIMS is coordinated by a network of regional experts and has received contributions from over 60 institutions across 28 countries (Raup et al., 2007).

National Snow and Ice Data Center (NSIDC)

The overall coordination, technical development, and data management are handled by the US National Snow and Ice Data Center (NSIDC) in Boulder, Colorado (Raup et al., 2007). It is one of the three operational bodies of the GTN-G, next to WGMS and GLIMS (Zemp, 2011). Its focus lies on data management of glaciers. Any glacier data within the GTN-G is stored in relational databases either at the WGMS in Zurich and/or at the NSIDC in Boulder (Zemp, 2011).

2.1.2 Glaciers

A glacier is commonly defined as a perennial body of land ice formed through the recrystallisation of snow and exhibiting evidence of past or present flow (Cogley et al., 2011; Zemp, 2011). This definition underlies most modern glacier inventories and provides the conceptual foundation for large-scale glacier mapping efforts such as the Global Land Ice Measurements from Space (GLIMS) initiative and the Randolph Glacier Inventory (RGI). While the physical definition of a glacier is relatively consistent across the literature, its implementation in inventories differs depending on mapping objectives, data sources and minimum size thresholds.

The Randolph Glacier Inventory adopts a minimum glacier area of 0.01km², following recommendations of the World Glacier Inventory (Pfeffer et al., 2014). This threshold aims to balance global completeness with practical limitations in mapping very small ice bod-

ies. However, some of the source inventories used in earlier RGI releases applied different minimum sizes, leading to regional inconsistencies in glacier inclusion. These differences directly affect the presence or absence of small glaciers in the RGI and contribute to the substantial number of unnamed glacier geometries that form the focus of this thesis.

Within the GLIMS framework, a glacier is defined as a perennial body of ice and snow observed at the end of the melt season, including all connected tributaries, accumulation areas, and debris-covered ice, while excluding exposed bedrock such as nunataks (Paul et al., 2009). A single GLIMS glacier entity may therefore represent a glacier complex that historically consisted of multiple named glaciers or tributaries. This aggregation practice improves consistency in geometric delineation but complicates the association between glacier geometries and traditional toponyms.

To address this, GLIMS provides a set of delineation and naming recommendations, including the consistent treatment of tributaries, stagnant ice, and snowfields, as well as the use of stable identifiers independent of names. Importantly, glacier names are regarded as auxiliary attributes rather than primary identifiers. Paul et al. (2009) emphasise that glacier names are not guaranteed to be unique, may vary in spelling or language, and can change over time. Furthermore, glacier names are described as useful but no mandatory attributes for the GLIMS inventory. As a result, modern inventories seem to prioritise geometric consistency and numeric identifiers over toponyms.

These practices have direct implications for automated name assignment. While glacier names remain essential for communication, historical reference, and regional studies, they are inconsistently recorded across inventories and gazetteers. Variations in glacier delineation, minimum size thresholds, and aggregation of tributaries further weaken the assumption of a one-to-one correspondence between a glacier name and a glacier geometry. This heterogeneity motivates the geometry based matching strategy adopted in this thesis, in which spatial agreement is treated as the primary criterion for linking unnamed RGI glacier geometries to candidate names from external datasets.

2.1.3 Toponyms

Against this background, the challenge of assigning names to unnamed glacier geometries is not only a technical problem of spatial correspondence, but also a conceptual one rooted in the nature of geographic names themselves. Glacier names are a specific form of toponyms, shaped by cultural practices, linguistic variation, and historical usage, and are therefore neither spatially nor temporally stable. Understanding how toponyms function, how they are recorded in gazetteers, and how they relate to geographic features is essential for interpreting the heterogeneity of glacier name data and for designing a robust matching strategy.

Place can be thought of as a relatively stable but not entirely unchanging portion of the Earth that people collectively recognise across different moments and contexts (Hastings, 2008). Toponyms, also known as placenames or geographic names, are the specific names assigned to a place, geographic feature or location (Hill, 2006). Naming a place is one of the first things that humans do in any environment, or a “naïve” geographic behaviour, thus domesticating it and creating a landscape out of a place (Hastings, 2008; Penko Seidl, 2019).

According to this definition, a glacier can thus also be considered as a place, and its name as a placename.

Comprehensive naming concepts exist everywhere in the world but local people have continuously named places regarding their importance or relevance (Penko Seidl, 2019). The author also mentions that researching the interconnection between toponyms and their place thus reveals information about how people perceive and comprehend a natural environment and how one adapts it their own need.

Depending on the purpose of the toponyms, one name may be sufficient, for example for information management, while for translation queries, all of the associated variants are useful (Hill, 2006). However, an information system should be capable of identifying a place regardless of the specific name variation provided by the user. In addition to the names themselves, gazetteers may contain supplementary details about place names, including the language and origin of the name, the historical periods in which it was used, transliterations, and the degree of certainty in linking a name to an ancient location (Hill, 2006). Normally, a placename in a gazetteer is composed of the toponym - the name itself - and the administrative modifier (Hill, 2006). For example, "Zurich, Switzerland" or "Mortersatsch Glacier" with respectively Switzerland and Glacier being the administrative modifier.

Toponyms should be unique and memorable within a certain space (Hill, 2006). Also, different types of toponyms exist - they can be descriptive or carry the name of the person who discovered the feature. However, with toponyms, also the problem of ambiguity arises. What is more, some, maybe most objects stay in a fixed place for their time being, however, glaciers are ever-changing, the global trend being that they diminish (Zemp, 2011). The factor of time is also important in the context of placenames (Hastings, 2008), when describing a place which changes over time such as a glacier, but also as different people and groups assign different names to the same object depending on their purpose (Hastings, 2008). Further challenges concerning placenames are posed by their classification, or feature type along with their geographic location (Hastings, 2008). The latter is the case when for example a glacier name is reduced to only point coordinates but its area in fact might be one square kilometre, where the point is set is not straight-forward and rarely documented well.

Beyond the name string itself, a toponym often implicitly conveys information about the type of geographic feature it refers to, which can be exploited in automated matching. As noted by Wu and Winter (2009), feature types tend to be more stable than names alone and can help constrain candidate matches when name variations occur. Identifying a feature as a glacier, for example, narrows the set of plausible sources even if the name is ambiguous or inconsistently recorded. However, feature types are not always encoded consistently across gazetteers and may differ in terminology or classification schemes (Wu & Winter, 2009).

2.1.4 Gazetteers

Gazetteers are geographic dictionaries or databases that provide information about placenames (toponyms) and their corresponding locations (Hill, 2006). In the context of geographic information science, a gazetteer typically includes the name of the place (e.g., Matterhorn), its coordinates (e.g., latitude and longitude), its feature type (e.g., mountain, glacier, city) and alternative names (e.g., different languages or historical versions). What is interesting in gazetteers is the fact that they can be produced incorporating bottom-up methods, that is, accepting and integrating data from multiple sources such as individuals (Acheson, De Sabbata & Purves, 2017). Both, GeoNames and OpenStreetMap include this participative or volunteered data along with authoritative data (Acheson, De Sabbata & Purves, 2017).

There are various types of gazetteers, some based on administrative records and others on user-generated data. Each serves specific purposes in geographical analysis and information retrieval, but no single gazetteer meets all needs. The study by Smart et al. (2010) introduces a mediation framework that facilitates access to and integration of multiple distributed gazetteers, creating a meta-gazetteer that enhances place name information. By combining data from different sources referring to the same geographic location, the approach applies multiple similarity metrics to identify matching toponyms (Smart et al., 2010). They analysed gazetteer quality in the UK, comparing between GeoNames, OpenStreetMap (OSM), Yahoo Where On Earth, Wikipedia and Ordnance Survey (OS). Of the mentioned, GeoNames had the overall most versatile suitability. For reverse geocoding specifically, GeoNames, Wikipedia and OSM were suitable (Smart et al., 2010). Glaciers belong to the natural feature types, which according to Acheson et al. (2017) tend to be less covered in gazetteers and originate from vague geometries. Furthermore, the coverage of several gazetteers sometimes overlaps in a specific geographic area and has no full agreement on the existing toponyms or geometries in the common region (Acheson, Villette et al., 2017).

Acheson et al. (2017) emphasise that while gazetteers are central to today's data-driven geographic research, their fitness for use is often overlooked. The authors demonstrate that global gazetteers such as GeoNames and Getty Thesaurus of Geographic Names (TGN) are uneven and biased with coverage patterns largely shaped by national boundaries, political and economic factors, and data availability. As a result, populated and developed regions dominate while natural features and less connected areas remain under-represented (Acheson, De Sabbata & Purves, 2017). This imbalance has implications for automated spatial analyses, where the uneven structure of gazetteers can reinforce existing geographic inequalities, amplifying coverage where data already exist and marginalising areas with sparse representation (Acheson, De Sabbata & Purves, 2017). Another analysis done by Graham and De Sabbata (2015) on populated places (with ≥ 1000 inhabitants), shows a clear trend that gazetteer data is not distributed evenly globally but rather clusters in some regions of the world and lacks in others. Their analysis however excludes placenames of natural features (Graham & De Sabbata, 2015).

Sigurðsson and Williams (2008) created an Icelandic glacier gazetteer as they set out with the aim of compiling all geographic names of the glaciers in Iceland and compile them into regional groups. In the end, they registered 269 glaciers, divided into eight regions

and listing all their toponyms, alternative and historic names, whilst the modern Icelandic name was the preferred and thus the main name. Their motivation for the work was the qualitatively insufficient and inconsistent choice of glacier toponyms on topographic maps published by the National Land Survey of Iceland. The spelling and point location of toponyms depended on the map product or was sometimes lacking fully (Sigurðsson & Williams, 2008). This problem is very similar to what can be observed in the GLIMS database, RGI and other glacier inventories. However, the authors address this problem by looking into historic and scientific literature, different maps of Iceland and scientific books and journal articles to find more glacier names. My approach will be limited solely to digital sources such as inventories and gazetteers. Thus, in aim the work by Sigurðsson and Williams (2008) corresponds to this thesis but not in the methodology.

With the rise of location-based services like GoogleEarth and MapQuest, digital gazetteers (DG) have become increasingly common and complex (Hastings, 2008). Their main function is to bridge informal references to places such as names and types with their precise geographic locations (Hastings, 2008). Often, information from different gazetteers must be compared or merged, which is addressed by Hastings (2008). They state that a key challenge is conflation: combining gazetteer data without losing the unique identity of each place. To address this, the author proposed a computational approach inspired by human reasoning, focusing first on geometries since places that do not overlap cannot be the same, then on type categories, and finally on names. This is key for the methods that will be applied in this thesis. By intersecting geometries and choosing only the ones that do intersect as a candidate, this first step corresponds to what Hastings states (2008).

In order to conflate polygon features, the article states that a bounding box can be drawn and compared or the polygon can be considered in its original form (Hastings, 2008). Their article introduces metrics that simulate human decision-making and describes an operational method for automating the conflation of DGs. Their approach is demonstrated through both conceptual and practical examples from the Lake Tahoe Basin in California and Nevada. Most importantly, Hastings (2008) mention that humans have little difficulty in deciding whether two mapped places describe the same thing. Even without the toponym, the form and location already tell a lot, along with elevation which can also be a discriminating factor (Hastings, 2008). Thus, most importantly, location and form will be taken into consideration to find two same glacier geometries.

A further study on the topic of reverse geocoding was conducted by Ma et al. (2016) which presented the method named Geohash-based reverse geocoding (GBRG), which is based on a Global Subdivision Model (GSM). In this method, the spatial objects are translated to Geohash codes, which act as location identifiers. In this step, line geometries, as well as area geometries are encoded with an adaptive Geohash method and lastly, a readable street address is obtained. In their study, PostGIS is used and the methods are applied to the city of Beijing, China (Ma et al., 2016). Their developed algorithm is efficient and achieved a high accuracy of 99.3%. However, their method showed to be less applicable for high latitudes, as the same level grids to encode the locations share the same latitude and longitude but would be distorted in higher latitudes (Ma et al., 2016). Beijing has the coordinates 39.9215, 116.4180, while for example the Aletsch Glacier in Switzerland has the coordinates 46.4889, 8.0531 (Google Maps, n.d.). Thus, the algorithm would maybe work for Switzerland but not further North.

2.2 Reverse geocoding & toponym resolution

This brought me to the topic of reverse geocoding, which can be defined as follows: given a physical location, for example in longitude/latitude coordinates, the aim is to map it to a meaningful, human readable string (Roth, 2015). Important to note is that the string is optimised for being read by humans, not computer software (Roth, 2015). For this thesis, the aimed for strings are glacier names. Ma et al. (2016) defined reverse geocoding as converting "a point location to a readable address or place name". Thus, the data input should be based on a database which can translate between different geographic data types, namely point objects, line objects and region objects (Ma et al., 2016). For geocoding and reverse geocoding, either an installed software or online services can be used (Li et al., 2017). Different suppliers provide reverse geocoding services such as Google, OpenStreetMap, Yahoo! and ESRI (Ajayakumar & Ghazinour, 2017) but also Wikipedia and GeoNames (Smart et al., 2010).

For instance, Kounadi et al. (2013) analysed the accuracy and privacy aspects associated with utilising various online reverse geocoding services in crime research. Also Ajayakumar and Ghazinour (2017) explored spatial privacy concerns in the context of reverse geocoding, using OpenStreetMap web services via geopy to process coordinates. Their study highlights privacy risks associated with geo-tagging and online check-ins, particularly on social media. By analysing X (formerly Twitter) check-in data, they demonstrate how easily accessible GIS tools and reverse geocoding services can expose users' spatial footprints. Additionally, they assess privacy policies and propose measures to enhance spatial privacy through policy improvements and algorithmic techniques (Ajayakumar & Ghazinour, 2017).

While reverse geocoding has been done by many researchers, most often the aim was to map addresses (Ma et al., 2016) or urban place names (Qiu et al., 2024). The latter research led to the algorithm called geoBERTTM, which aligns toponyms which refer to the same place in an urban setting, matches place names and thus mitigates ambiguity in urban toponyms (Qiu et al., 2024). Their framework is built so that the toponyms are first tokenised, then vectorised and finally a deep learning model is applied for the matching part (Qiu et al., 2024).

Kilinc et al. (2016) created a measure to match toponyms with the aim of matching different expressions which represent the same location, referred to DAS. This is a field of approximate string matching, and, according to (Kilinc, 2016), is a novelty when applied to toponyms. Their DAS algorithm is based on name, word and sentence similarity and is a hybrid measure between token- and edit-based feature categorisation. Their aim was to match official place names by different countries' National Geospatial-Intelligence Agency (NGA) to ambiguous toponyms in Turkey, Egypt, France, Germany, America and Russia. This is of interest to this thesis, as different languages are comprised. However, all location names with no alternative were deleted from the datasets, in order to only look at the ambiguous toponyms. Hence, an approximate string matching can be done to reduce duplicates and ambiguity in the case of glaciers with one or more names but it cannot be used for glacier entries which have an empty name field.

Similarly, Ardanuy et al. (2020) create and employ the DeezyMatch (deep fuzzy string matching) method. In the fuzzy string matching process, two main steps have shown to be widely

used: 1) pairwise classification for toponym matching, and 2) geographical candidate selection (Ardanuy et al., 2020). DeezyMatch is open-source and freely available in python libraries. It is based on a deep neural network for binary classification, using the Keras API for classifying toponym pairs. First, the algorithm is trained and validated with a dataset of strings. As a next step, the classified pairs from step 1, are given on to the candidate selection component which proposes the potential candidates for a toponym query. For this, DeezyMatch first computes embeddings for all gazetteer toponyms (once per model-gazetteer pair), then uses the same model to embed incoming query toponyms. Finally, it ranks candidates by embedding distance with a nearest-neighbour search and even noisy strings (e.g., variants like “Manchtftcr”) retrieve their correct matches (e.g., “Manchester”).

The DeezyMatch toponym matching, as well as the DAS and other string matching algorithms have the aim to assess whether two *strings* or *names* refer to the same place - however, the aim within this thesis is to match two *geometries* to each other, based on spatial similarity and location. Thus, not the whole algorithm or process described by rather the first step of selecting candidates based on location interesting for this thesis.

2.3 Spatial matching strategies

The aim of McKenzie et al. (2014) is to better utilise the variety of data sources within big data, as this variety is one of its advantages. The authors want to accomplish this by conflating information from several data sources, which includes the steps of establishing relations between the same features of different sources, known as matching, and further, attribute value merging. Once entities are matched, their attributes need to be merged using procedures that avoid logical contradictions. This second step involves understanding the nature and structure of the data to decide how to combine or resolve differences between the attribute values of the matched entities. Their matching procedure is based attributes of points of interest (POI) from Foursquare and Yelp, both platforms provide user-generated and place-related information, with Foursquare focusing on social check-ins and tips, and Yelp emphasising structured reviews and business details.

Walter and Fritsch (1999) propose a relational matching approach which integrates spatial data from different sources. The goal of their project is to take advantage of characteristics of more than one dataset, including improvement of the data quality, new applications for the dataset, higher reusability, or lastly cost minimisation of data acquisition (1999). Although their paper is not recent, the general matching methodology and framework remained the same and are of interest for this thesis. Their strategy consists of the five key steps outlined below (Walter & Fritsch, 1999).

1. Preprocessing: a global affine transformation is applied to minimise systematic spatial errors, supported by control points or automated feature matching. Further, topology checks and node simplifications are performed.
2. Computing potential matching pairs: buffering and buffer-growing techniques are used to identify all possible candidate pairs considering geometric proximity and topology, to create a comprehensive list of potential matches. Distance buffers are used for 1:1 and 1:n matching, while a buffer growing procedure is used for n:1 and n:m

matching. They state that only matches are considered which give added value in form of new information regarding the dataset.

3. Use of geometric constraints: statistical constraints such as limits on angle, shape and length differences are applied to reduce unlikely pairs from the candidate list, improving computational efficiency without significantly compromising matching accuracy.
4. Evaluation of matching pairs: each remaining pair is evaluated using a merit function based on mutual information from information theory, considering geometric and topological features and relations to quantify the information overlap.
5. Calculation of final matchings: an optimal combination of matching pairs is selected to maximise mutual information, with the search space managed using local clusters to make computation feasible while maintaining high match accuracy.

Walter and Fritsch (1999) applied their methods to linear elements such as streets and road networks, this cannot be directly applied to glacier polygons but sets a straight-forward example of how the process can be designed for this thesis. To minimise manual post-processing, the authors propose automatic quality checks based on information theory. These checks assess how much information is lost during matching to find uncertain matches (Walter & Fritsch, 1999). The authors measure overall quality by the average information loss, while each match is labelled as certain, mostly certain, or uncertain to help focus manual review where it's most needed. Lastly, after the automatic matching process and quality measures described above, a manual post-processing can occur (Walter & Fritsch, 1999). Their algorithm successfully matched on average 96.3% of the features, however, difficulties arose in regions with significantly different data acquisition.

Ruiz-Lendínez et al. (2017) point out that many matching approaches focus on network matching such as Walter and Fritsch (1999) do, specifically on road networks. Also, they stress the fact that spatial data is essential for many areas such as engineering, natural sciences and land management. With the rise of geographic data, the level of quality in sources also becomes more heterogenous. Also volunteered geographic information has contributed to this increase of spatial data and its coverage. This is how the authors come to their initial point that different matching techniques are needed, not only such focusing on linear data, but also points and areas. Polygon to polygon matching, also called inter-element matching by the authors, is done with an algorithm which analyses area features from two geographic databases, while point to point matching, also called intra-element matching, is dependent on comparing the changes in their turning function, which comes from the field of shape recognition (Ruiz-Lendínez et al., 2017). The two main issues the authors look at are how the similarity between elements from two datasets can be measured, and how the optimal correspondence between vertices of the two matched features can be achieved. The first point is of interest for this thesis. In geographic information science, quantitative measures belonging to three different categories are used to evaluate the similarity; geometric, spatial relationship and attribute measures (Ruiz-Lendínez et al., 2017). Geometric are the location as well as the shape of an object, while the spatial relationship describes distance, direction and topological relations (Ruiz-Lendínez et al., 2017). For this thesis, location, shape and distance are of importance. Ruiz-Lendínez et al. (2017) further mention the limitations of the matching process' low level of automatisation, as well as the

production of false positives or false matching pairs when 1:N or N:M matching takes place. The authors try to prevent this by creating a Mean Accuracy Value (MAV) and working with thresholds. The MAV is exactly 1 (100%), where the matched polygons are exactly the same (2017).

2.3.1 Intersection over Union (IoU)

Intersection over union (IoU), also called Jaccard Index is one of the most important and broadly used metric used for similarity comparisons in object detection and instance or image segmentation (Rezatofighi et al., 2019; Zheng et al., 2021). It is used to measure the overlapping areas between two shapes, invariant to their scale as it includes a normalisation (Houde et al., 2025). An IoU result can range from 0 to 1, with 0 being no overlap at all and 1 a perfect overlap (Houde et al., 2025). Therefore, the higher the intersection over union, the better the alignment between two areal features.

The paper by Cho (2024) introduces weighted Intersection over Union (wIoU) as a novel evaluation metric for semantic segmentation, addressing limitations of traditional IoU, which treats all pixels equally and often under-represents boundary errors. The object boundaries within images led to high classification ambiguities. The wIoU enhances evaluation by assigning greater importance to pixels near object boundaries through a tunable parameter α . This allows it to flexibly capture both region-level accuracy and boundary precision in a single metric. The method has been validated across synthetic and real datasets, demonstrating its robustness and potential as both an evaluation tool and a loss function for training segmentation models (Cho, 2024). Their method can lead to a more balanced evaluation. By tuning α , wIoU can simulate traditional IoU (low α) or emphasise boundary performance (high α), thus offering a versatile evaluation tool. Furthermore, wIoU is more sensitive than conventional IoU and edge-based metrics in distinguishing subtle differences in segmentation quality, particularly where boundary errors dominate (Cho, 2024). It is especially helpful in cases where the object boundaries have rather complex shapes (Cho, 2024). While the goal of their paper and this project is slightly different, the boundary problem also arises when looking at glacier polygons. Namely, when two polygons lie close to each other or even overlap. When assigning a glacier name to an unnamed polygon, all the polygons closest to it which have names are considered. Thus, the intersection over union measure is meaningful as the total area is normalised between both geometries, which balances absolute and relative overlap. In contrast to the paper by Cho (2024), for this project, the containment is more important than the boundary. As the glacier boundaries change over the years and throughout the data anyway, the focus here shall lie on whether an query or input polygon is contained within or contains a reference glacier polygon. This will be elaborated in more detail in the methods part of this thesis.

While the application of the wIoU measure lies in image segmentation (Cho, 2024), another study by Hou et al. (2021) uses intersection over union (IoU) for object detection. The main idea is that bounding boxes with the same IoU can still have different spatial relationships, so the centre distance is used to guide better regression. Both approaches enhance the standard IoU by adding spatial distance considerations but they are used in different contexts. While not directly comparable in performance, they are conceptually related as they both address limitations of IoU by incorporating distance sensitivity.

On another count, Rezatofighi et al. (2019) presented a generalised version of the intersection over union (GIoU) function to overcome the shortcoming of IoU under the circumstance of non-overlapping bounding boxes. When two boxes don't overlap, the intersection area and thus the IoU is equal to zero. The GIoU adds centre point distance, providing a gradient and becoming non-zero even when IoU is zero, making it more effective for non-overlapping cases during training.

While the GIoU is an improvement to the traditional IoU when considering non-overlapping boundary boxes, it is slower due to needing more iterations (2019). Furthermore, when the bounding boxes or areal features are contained in one another, the GIoU result is the same as the result of the basic IoU function (2019). Zheng et al. (2019) recognise the problem of slow convergence and inaccurate regression and thus introduce two new variants to the intersection over union metric: distance (DIOU) and complete (CIOU). DIOU mitigates the issue of tightly clustered features by introducing the central point distance as an additional factor, which is minimised to achieve higher localisation accuracy. While CIOU includes overlap area, central point distance and aspect ratio which is supposed to result in a better performance and faster convergence, thus needing fewer iterations to align bounding boxes (Zheng et al., 2019).

Likewise, Houde et al. (2025) discuss the limitations of intersection over union (IoU) when applied to features that do not spatially overlap. In their study, several IoU-based loss functions were compared, including generalised IoU (GIoU), distance IoU (DIOU), complete IoU (CIOU), efficient IoU (EIoU), and a variant incorporating angular constraints. The metrics were evaluated using unmanned aerial vehicle imagery for tree detection tasks. Among the tested approaches, CIOU achieved the best overall performance, exhibiting the lowest validation loss and a high precision of 0.934 (Houde et al., 2025). GIoU and other extended variants also performed well, though with higher training and validation losses. While these modifications improve robustness in object detection tasks, particularly where bounding boxes do not overlap, they are primarily designed for image-based feature detection and are therefore not directly transferable to polygon-based glacier matching.

2.3.2 On points and polygons

What has to be added in the theme of data types, is the uneven information content which points and polygons convey. As glaciers are represented either as point or polygon feature, lines are not discussed in this thesis. Machado et al. mention that (2011) points (and rectangles) have limited information they carry, as opposed to more complex shapes such as polygons. Hill et al. (1999) also add to that, stating that many gazetteers at the time of writing contained only point locations, which don't represent the full extent of geographic place. The authors call for more data representations as detailed boundaries (Hill et al., 1999). Although their paper is from 1999, that is still valid 26 years later at the time of this thesis, as many datasets used in this project only had a point footprint. Goldberg et al. (2008, p.63) state the same limitations of points: "for example, overlap and containment cannot be derived from points". In many gazetteers, large vector features like rivers are simplified to a point at their delta, which overgeneralises the geometry and reduces the information gain of such data (Goldberg et al., 2008). This can be stated analogously when looking at glaciers, as a toponym point cannot distinguish between *glacier terminus A* and *the whole glacier complex B*.

2.4 Research gap

Reverse geocoding strategies as well as spatial matching techniques such as object detection and entity linking enable users to assign more information to a location or link two locations/geometries in order to augment information in either. Although there is vast research on the techniques themselves and many application on urban settings, natural features tend to be less covered in spatial matching research. What is more, already pointed out by Ruiz-Lendínez et al. (2017), spatial matching has yet to be applied on polygons in a big scale. Applied spatial matching to polygons of urban buildings in several Spanish cities. However, natural features are also lacking in their research. Thus Not only glaciers per se are not often studied in computational study fields, but also natural features in general. This could be seen in matching approaches, where mostly urban areas were analysed. In turn, while natural features in general are more covered in gazetteer and toponym studies (Acheson, Villette et al., 2017), glaciers tend to be forgone or not mentioned in such research. Therefore, spatial matching on natural features, specifically glaciers poses a clear research gap. Furthermore, a focus on glacier toponyms in computation research is also rather unprecedented. Lastly, when considering the broader research field a glacier offers, Hotelling et al. (2017, p. 2943) state that "mountain glaciers remain one of the most understudied, yet arguably most imperilled and rapidly changing, ecosystems on Earth." Concluding, the aim of the thesis sprung from the original need of the World Glacier Monitoring Service (WGMS) to collect, manage and analyse glacier information from openly available sources. In addition to this, the project serves to bridge the topics of spatial matching and entity linking, combining it with the topic of gazetteers, natural feature types and glacier toponyms in order to fill the research gap highlighted above.

3 Data & Database

This chapter describes the data basis, starting from the focus point of this thesis, the Randolph Glacier Inventory (RGI), which is derived from the Global Land Ice Measurements from Space (GLIMS) inventory. Both are a focal point of glacier mapping worldwide. Furthermore, the data used to source glacier names, here on called reference data, will be delineated, whereby referring to the used programmes and the applied methods.

3.1 Definition of *unnamed*

The criteria for determining whether a glacier is considered *named* or *unnamed* evolved over the course of this project. This was necessary because additional corner cases became apparent during algorithm testing, and because naming conventions vary substantially across regions and national inventory practices. As a result, both the Python workflow and the SQL queries were iteratively adapted to capture region-specific patterns of name usage and placeholder values. Below, several representative examples illustrate the classification logic.

Example named vs. unnamed:

- “CN5Y673P0077 Qiongkuermu Glacier” $\Rightarrow$ named; has spaces and a real word
- “CN5Y673P0088” $\Rightarrow$ unnamed; is only an alphanumeric ID, often in RGI region 13 (Central Asia)
- “(Severny Island)” $\Rightarrow$ unnamed; bracket-only with region, often in RGI region 09 (Russian Arctic)
- “Vodorazdelny” or “Gries” $\Rightarrow$ named; letters only, no digits
- “Parlung Zangbo No. 4” $\Rightarrow$ named; letters and digits, with spacing in-between

Certain regions displayed highly characteristic placeholder values for glaciers that did not carry official names. These patterns required region-specific handling in the SQL filtering logic. For example:

- Alaska (G-01): Unnamed_1, Unnamed_2, Unnamed_23, Unnamed_27, Unnamed_30, Unnamed_33, Unnamed_40, Unnamed_41, Unnamed_43
- Western Canada and USA (G-02): CA, CO, ID, MT, OR, WA, WY

These abbreviations appear to correspond to U.S. state codes and therefore do not constitute glacier names. They were explicitly included in the logic to prevent false classification as *named*. In the database, this logic was implemented as a generated boolean column `is_named` on `core.rgi_all`, which normalises the original name field and filters out a set of language- and region-specific placeholder patterns, while whitelisting a few documented alphanumeric glacier names.

Table 3.1: Number of glacier names which include digits, because of this they are flagged as *unnamed* glaciers. This list does not contain the total of unnamed RGI glaciers.

RGI region	Nr. of alphanumeric	Example
G-01 (Alaska)	9	Unnamed_23
G-07 (Svalbard)	10	NO4W00110521
G-12 (Caucasus)	6	432zh
G-13 (Central Asia)	21,510	CN5O270B0323
G-14 (SW Asia)	1,780	CN5Y656E0127
G-15 (SE Asia)	5	CN5Q222A0021
G-17 (S Andes)	32	HPS18

```
-- Generated boolean column for whether a glacier name entry is
   considered "named"

GENERATED ALWAYS AS (
  -- normalise name: lowercase, trim, remove angle brackets
  (lower(
    btrim(
      regexp_replace(coalesce(name, ''), ' [<>]', '', 'g'))))
    NOT IN (
      '', '-', '.', '_', 'none', 'null', 'nan', 'n/a', 'na', 's/n', '
      unknown', 'unnamed', 'namenlos', 'namenlos-i', 'namenlos-ii', 'nn
      ', 'nn,', 'n.n.', 'ohne_namen', 'ohne_name', 'ohne_bezeichnung',
      'keine', 'keine_angabe', 'no_name', 'not_named', 'unknown_glacier
      ', 'unnamed_glacier',
      -- region-specific placeholders
      'ca', 'co', 'id', 'mt', 'or', 'wa', 'wy')
    -- exclude patterns like 'Unnamed_23' or 'unnamed 1'
  AND NOT (
    lower(
      btrim(
        regexp_replace(coalesce(name, ''), ' [<>]', '', 'g')))) ~ '^
        unnamed[_[:space:]]*[0-9]+$')
    -- exclude bracket-only labels, e.g. '(Severnny Island)'
  AND NOT (
    btrim(coalesce(name, '')) ~ '^([^\]]*\])$')
    -- exclude code-like alphanumeric tokens (>=5 chars, must contain a
      digit),
    -- except for specific whitelisted cases
  AND NOT (lower(btrim(regexp_replace(coalesce(name, ''), ' [<>]', '',
    'g')))) ~ '^([?=.][0-9])[a-z0-9]{5,}$'
    AND lower(btrim(regexp_replace(coalesce(name, ''), ' [<>]', '', 'g'
    )))
    NOT IN ('toklat1', 'toklat3', 'tungalglatscher09n'))
) STORED;
```

An initial limitation of the strict classification rules is that they inadvertently flagged some glaciers which have a digit in the name: Toklat1, Toklat3 (in G-01 Alaska) and TUNGELGLETSCHER09n (in G-11 Central Europe). However, these were explicitly included to the named glaciers in a second step. Overall, the applied rules successfully captured the majority of unnamed glacier placeholders while still allowing documented alphanumeric names to be retained. Table 3.1 lists the number of entries classified as unnamed together with representative examples. The list is not exhaustive, but only lists names which fall into this category. It also highlights the vast number of this naming scheme in Central Asia.

3.2 GLIMS & RGI

3.2.1 GLIMS

By 2022, the Global Land Ice Measurements from Space (GLIMS) inventory included approximately 300,000 glacier outlines (Gärtner-Roer et al., 2022). Since June 2024, the database contains 726,529 glacier analyses (National Snow and Ice Data Center, 2024), each being a glacier outline at one point in time. Each glacier entity in the database includes multiple attributes; for this thesis, the relevant fields are: `glac_id` (the GLIMS glacier ID), `glac_name`, `wgms_id`, `local_id`, and `anlys_id` (the analysis ID for a specific glacier outline at a given time). The name attribute is of key focus in this thesis. The glacier names in GLIMS are compiled from existing inventories, published literature, and satellite-based mapping efforts. The nomenclature often follows the source inventories and geographic naming conventions used in the respective countries or regions (Zemp, 2011).

The GLIMS database supports applications such as glacier monitoring, hazard assessment, and detection of global glacial change (National Snow and Ice Data Center, n.d.). As glaciers respond sensitively to climate variability, they serve as valuable indicators of environmental change. GLIMS enables researchers to track glacier evolution from 1850 to the present, informing studies on water resources, ecosystems, hazards, and sea level rise. The database is publicly accessible via the GLIMS Glacier Viewer and offers data in various GIS-compatible formats, including Shapefile, KML, and GML (GLIMS Consortium, 2005; Raup et al., 2007). Until 2005, GLIMS allowed analysts to decide whether and how to divide branched ice masses into individual glaciers, acknowledging that different branches were often historically named and treated as separate glaciers.

GLIMS name dataset

Ethan Welty, the database manager at the World Glacier Monitoring Service (WGMS), provided me with the glacier names from GLIMS. The dataset is called `glims_names` and contains partly cleaned and dissolved glacier names for 16,602 features worldwide. Alike in the RGI and GLIMS datasets, name attribute is called `glac_name`, whereby unnamed glacier geometries were discarded. However, the purely numeric names were kept in place in case some of them correspond to a „Glacier No. X “style name such as in Russia. This is a simplified version of the vast GLIMS database, as only one polygon is given per named glacier outline, the other entities are not considered in this dataset. This dataset will also be used as a further name reference.

3.2.2 RGI

The Randolph Glacier Inventory (RGI) is a subset of the GLIMS inventory, containing glacier outlines measured around the year 2000. The inventory contains one entity per glacier with the thirteen core attributes: RGIId, GLIMSIId, RGIFlag, BgnDate, EndDate, CenLon, CenLat, 01Region, 02Region, Area, Zmin, Zmax, Zmed. The most current version is the version 7.0, released in 2023, the first version was released in 2012. Over the version history, the coverage has been improved, most glaciers now have exact dates and some attributes were reorganised within the database. The original motivation to create the inventory sprung from the preparation of the fifth IPCC report. In the RGI, not the glaciological or hypsometric measurements stand in focus but rather, the global completeness of glacier coverage is prioritised (RGI Consortium, 2017b). As little time was available for the first RGI version, shortcomings in the data quality were accepted. For example, the global to continental scale rather caters to broad analyses. However, with each version, the database was continuously improved (Gärtner-Roer et al., 2022). Which is also why the RGI database has extended its coverage with every new version but the glacier dates stay around the year 2000. Summed up, the RGI includes 274,531 glaciers, which cover an area of 706,744 km² (RGI Consortium, 2023a), not including the glaciers on the Antarctic Peninsula and Greenland ice shield but only the isolated glaciers and ice caps around the periphery (Gärtner-Roer et al., 2022). The data of the RGI can be downloaded as Shapefiles, which contain the glacier outlines as polygons in the WGS84 datum. Further, the covered area is organised in regions of first and second order (RGI Consortium, 2017b).

For this thesis, the RGI is the starting point of the analysis, which will be described further below (see Method). The focus hereby is on the glacier names. Primarily, the missing names, which in RGI are 96%, the definition of unnamed or missing can be found in 3.1. A closer look at the existing names, shows that they are not of consistent quality. There are glaciers with no name, an ID as a name, an ID followed by a name, e.g., in France, names which include umlauts and other non-ascii characters and are not displayed correctly. The main aim of this thesis will be to consult reference datasets in order to provide the most probable glacier name for a polygon or point input which has no name. However, there is also further room for improvement when considering the name quality, as they are not globally conform. Unfortunately, this goes beyond the scope of this project but can be pinpointed for further work.

The following table shows the total and named glacier numbers of worldwide glacier inventories Randolph Glacier Inventory (RGI), Global Land Ice Measurements from Space (GLIMS), Fluctuations of Glaciers (FoG) and World Glacier Inventory (WGI), along with the two gazetteers GeoNames and OpenStreetMap.

The 263,479 unnamed glacier features in the RGI version 7 data are of various sizes, ranging from 0.010 to 4,766.147 km². The average of unnamed glacier areas has a size of 1.672 km², with a standard deviation of 26.388 km². The full distribution can be seen in the histogram 3.1.

Table 3.2: Glacier name coverage across worldwide glacier inventories and gazetteers

Dataset	Total glaciers	Unnamed glaciers	Named glaciers	% named	% named of RGI total
RGI (PLY)	274,531	263,479	11,052	4.03	4.03
GLIMS (PLY)	16,601	71	16,530	99.57	6.02
WGI (PT)	132,699	86,214	46,485	35.03	16.93
FOG (PT)	7,571	88	7,483	98.84	2.73
GeoNames (PT)	8,286	1	8,285	99.99	3.02
OSM (PLY & PT)	8,971	130	8,841	98.55	3.22

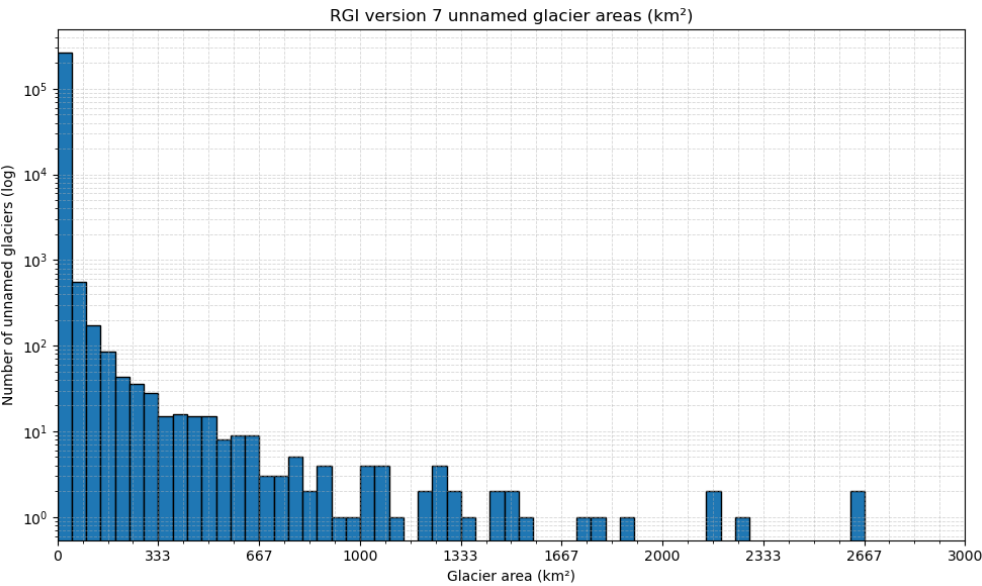


Figure 3.1: Histogram of RGI glacier polygon areas in km²

3.2.3 Comparison RGI version 6.0 to 7.0

RGI version 6.0 was published in July 2017, and in September 2023, the most recent version 7.0 followed it (RGI Consortium, 2023a). With this version, the glacier outlines have been improved in almost all regions. Additionally, all outlines and attributes now originate from GLIMS data (RGI Consortium, 2023a), whereas version 6.0 included some outlines from WGI and the Digital Chart of the World, where no other data was available (RGI Consortium, 2017b).

Table 3.3: Glacier counts and area across RGI versions (RGI Consortium, 2023b)

Version	Total glaciers	Total area
RGI 6.0	215,547	705,739 km ²
RGI 7.0	274,531	706,744 km ²

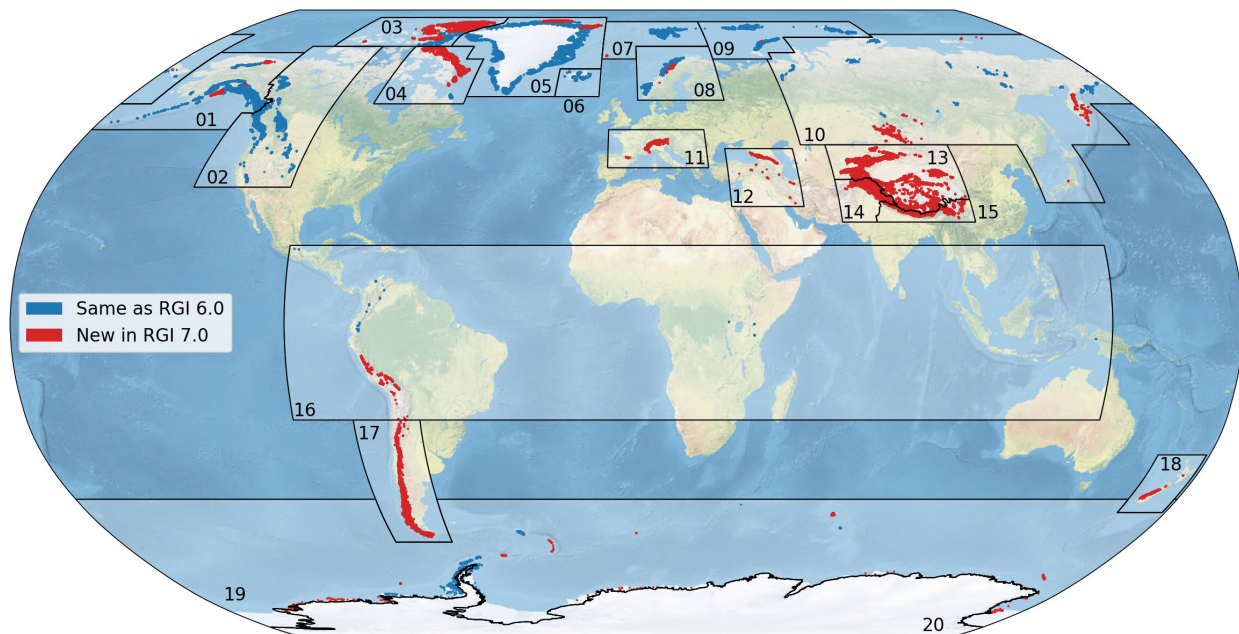


Figure 3.2: First-order regions of RGI v7.0 (RGI Consortium, 2023b)

The most important changes from RGI version 6 to 7, listed by the RGI Consortium (2023a) will be discussed here. Several region boundaries and region numbers have been adapted. One bounding box of the subregion 01-03 Alaska Peninsula, contained no more glaciers and thus was removed. The subregion 05-11 (Greenland Ice Sheet) was also removed, partly due to insufficient and coarse definition of the boundary and because the ice sheet itself is not included anyway. Also the Antarctic ice sheet has undergone data changes, namely was it split up into two first-order regions. Region 19 now contains only the islands in the periphery of Antarctica, and region 20 represents the Antarctic mainland, which encompasses the subregion "Antarctic Ice Sheet" but currently has no glaciers in the RGI dataset. Other region names have been harmonised to align the terms to how they're used in the documentation of the RGI. Technical changes to attributes contain: second-order regions are spelled out instead of using the abbreviations. Further, first and second-order regions

Table 3.4: First-order RGI v7 regions with full names, database codes, glacier counts, and total area, table from (RGI Consortium, 2023b)

Region	Full name	Count	Area (km ²)
01	Alaska	27,509	86,708
02	Western Canada and USA	18,730	14,521
03	Arctic Canada North	5,216	105,370
04	Arctic Canada South	11,009	40,538
05	Greenland Periphery	19,994	90,482
06	Iceland	568	11,060
07	Svalbard and Jan Mayen	1,666	33,959
08	Scandinavia	34,10	2,948
09	Russian Arctic	1,069	51,595
10	North Asia	7,155	2,643
11	Central Europe	4,079	2,124
12	Caucasus and Middle East	2,275	1,407
13	Central Asia	75,613	50,344
14	South Asia West	37,562	33,075
15	South Asia East	18,587	16,049
16	Low Latitudes	3,695	1,929
17	Southern Andes	30,634	27,674
18	New Zealand	3,018	886
19	Subantarctic and Antarctic Islands	2,742	133,432
20	Antarctic Mainland	0	0
Global		274,531	706,744

each now have a corresponding string code, which then is used to name the individual RGI files. In version 7, the WGMS_CODE was deleted from all files. Geographic changes to a first-order region were undergone for regions 13 (Central Asia) and 15 (South Asia East), where the boundaries were extended slightly to the East. Lastly, these changes ensure the agreement of the RGI regions with the Glacier Regions dataset of the Global Terrestrial Network for Glaciers (GTN-G). The map above shows the glacier regions in RGI 7.0, where 200,734 of the 274,531 outlines, equivalent to 73%, are derived from new inventories. This change lead to an improvement for 42% of the worldwide area (RGI Consortium, 2023b). New outlines in RGI 7.0 are coloured red, consistent outlines since RGI 6.0 in blue.

As for the glacier name attribute, the definitions from version 6 and 7 differ slightly, especially in precision of language and sources. In the documentary of the RGI version 6.0, the glacier name attribute is defined as:

”Name of the glacier, or the WGI or WGI-XF id code (modified after Müller et al., 1978) if available. Many glaciers do not have names, and coverage of those that do is incomplete. Of the 215,547 glaciers in RGI 6.0, 46,770 have information in their Name field, although for many the content is actually an id code. (The number of glaciers is 216,502 if Greenland glaciers strongly connected to the ice sheet are included.)” (RGI Consortium, 2017b, p. 17)

Whereas in the documentary of RGI version 7.0, the name attribute is only shortly defined as:

"glac_name long_name: Glacier name description: Glacier name (when available). datatype: str units: source: GLIMS rgi6_name: Name." (RGI Consortium, 2023a, p. 17)

3.3 Reference data

External sources, thus everything not RGI, will here on be described as reference data. Most of the used datasets have been provided in a spreadsheet by Ethan Welty. This list covers two worldwide gazetteers; GeoNames and OpenStreetMap, along with national toponym datasets usually from national mapping agencies (NMA). Further, worldwide glacier inventories like the WGI and FoG have been included. Lastly, there are some national glacier inventories that will be taken as reference data, but will not be closer described here, only listed in table 3.6. The integration of as many reference datasets was decided upon because literature and data research showed that no single database or gazetteer meets all needs (Smart et al., 2010).

3.3.1 Glacier inventories

WGI

Additionally to the the GLIMS database, the NSIDC hosts the World Glacier Inventory (WGI), covering data for over 130,000 glaciers, including glacier location, size, length, orientation, elevation, and classification. Compiled mainly from aerial photographs and maps, it provides a snapshot of global glacier distribution in the late 20th century (National Snow and Ice Data Center, 2023). Most glaciers have only one recorded entry, making it a reference rather than a continuous monitoring dataset. The WGI, originally published by the WGMS in 1989, can be searched by various attributes and downloaded as in a single ASCII text file format (National Snow and Ice Data Center, 2023).

FoG

The Fluctuations of Glaciers (FoG) database contains the glacier name in several languages. In this inventory, the mass balance and the length measurements have priority. The FoG series, published by the WGMS, provides standardised data on global glacier changes every five years. It includes information on glacier length, area, volume, and mass based on in-situ measurements, remote sensing, and reconstructions (Zemp, 2011). The goal is to offer a global overview, support further research, and encourage more extensive monitoring. The dataset is widely used in glaciology, climatology, hydrology, and quaternary geology (WGMS, 2022a). FoG publications are available for download, with some editions also offered in print. The most recent version 10 came out in 2012 and covers the years 2005-2010 (WGMS, 2022a). While specific naming sources are less explicitly detailed in the documentation, it can be inferred that the glacier names are also drawn from existing local, national, or international glacier datasets, scientific publications, or remote sensing products integrated into the database (Zemp, 2011). Glacier data and metadata were submitted amongst others by by M. Hoelzle (WGMS, 2012). In the FoG database report, the sources which are the individual observers and their sponsoring agencies are listed, these are numerous (WGMS, 2012). Thus, the FoG data is gathered from various parties.

SGI

The Swiss Glacier Inventory (SGI) 2016 is the most recent dataset from the Glacier Monitoring Switzerland (GLAMOS) organisation and contains 1,400 glacier entries with a surface area of roughly 961 km² (Linsbauer et al., 2021). 138 of the dataset's glaciers are unnamed. The SGI 2016 covers glacier areas, outlines and debris cover of all glaciers in Switzerland of the years 2013-2018. The glacier outline data is based on the Swiss Topographic Landscape model and if there was no name, the glacier names were derived from the swissNAMES3D dataset of 2020, both by the Federal Office of Topography Swisstopo (Linsbauer et al., 2021). The older inventories from GLAMOS are from the years 2010, 1973, 1931 and 1850 (Linsbauer et al., 2021). The authors mention that manual control by glaciologists is needed in order to create a glacier inventory which meets the field's standards, thus the glacier outlines were derived in cooperation with Swisstopo and GLAMOS, creating an interdisciplinary project (Linsbauer et al., 2021).

3.3.2 Gazetteers

GeoNames

GeoNames is a free gazetteer created by Unxos GmbH. Glaciers belong to the feature class H, including water bodies such as streams and lakes, and have the feature code "GLCR". GeoNames also includes alternative names for locations, for example historical or regional variations to a toponym (Ardanuy et al., 2020). GeoNames' characteristics relating to balance, data accuracy, completeness and origin is not very clear (Acheson, Villette et al., 2017). In terms of origin, GeoNames consists of data from a variety of sources, including official sources and individual contributors, but the sources for each feature are not specified in the free version (Acheson, Villette et al., 2017). This also leads to the fact that the precision of a record, such as a glacier point may vary depending on its source or be completely unknown (Acheson, Villette et al., 2017). In terms of glaciers, GeoNames includes 8,286 features, of which only one glacier is unnamed.

OpenStreetMap

Similar to GeoNames, OpenStreetMap (OSM) also has alternative names to the main toponym. Their name structure is very large, including different languages, which each are an attribute in the dataset. The tags:name is the main toponym, under tags:name:en or tags:name:de we often have an English or German variant (OpenStreetMap contributors, 2025). In Switzerland the toponyms are often given in French, Italian or Rumantsch too, depending on the popularity of the toponym. OpenStreetMap data contains 8,971 glacier entries globally, of which 8,841 (98.55%) have a name. OpenStreetMap is normally based on national mapping agency data as well as user contribution (Acheson, Villette et al., 2017; OpenStreetMap contributors, 2025).

3.3.3 Examples of regional data availability

To illustrate the heterogeneous data basis on which this thesis builds, figures 3.3 and 3.4 provide selected screenshots from two key reference sources, OpenStreetMap (OSM) and the WGMS Fluctuations of Glaciers (FoG) database. The examples are intentionally chosen from different regions of the Randolph Glacier Inventory (RGI) to highlight the strong spatial variability in glacier name availability.

In regions with dense mapping and long-standing glaciological monitoring traditions, such as Central Europe or Scandinavia, glacier names are widely available and consistently represented across multiple datasets. OSM polygons frequently carry explicit name attributes, while FoG and related inventories provide complementary point-based references. While the first screenshot, showing New Zealand glaciers, has dense information in its OSM basemap, the FoG lacks certain names still. What is not visible, is that the largest glaciers in the screenshot do have FoG names, i.e., Victoria, Fox and Franz Josef Glaciers. In the Sweden screenshot however, both, the OSM basemap along with the FoG points show good glacier name availability.

In contrast, remote or sparsely mapped regions, including parts of Central Asia, the Arctic and the Caucasus exhibit markedly lower glacier name coverage. In these regions, OpenStreetMap often lacks named glacier polygons altogether, while FoG and related inventories provide only sparse point-based references. This results in fragmented and uneven name availability, both within and across datasets.

Figure 3.3 and 3.4 provide illustrative examples of this regional heterogeneity. The screenshots are not intended as a quantitative assessment, but serve to visually demonstrate the uneven empirical foundation of global glacier name data. This heterogeneity directly motivates the multi-source matching strategy adopted in this thesis and helps explain the strong regional differences observed in the matching results presented in later chapters.

Regions with dense glacier name coverage

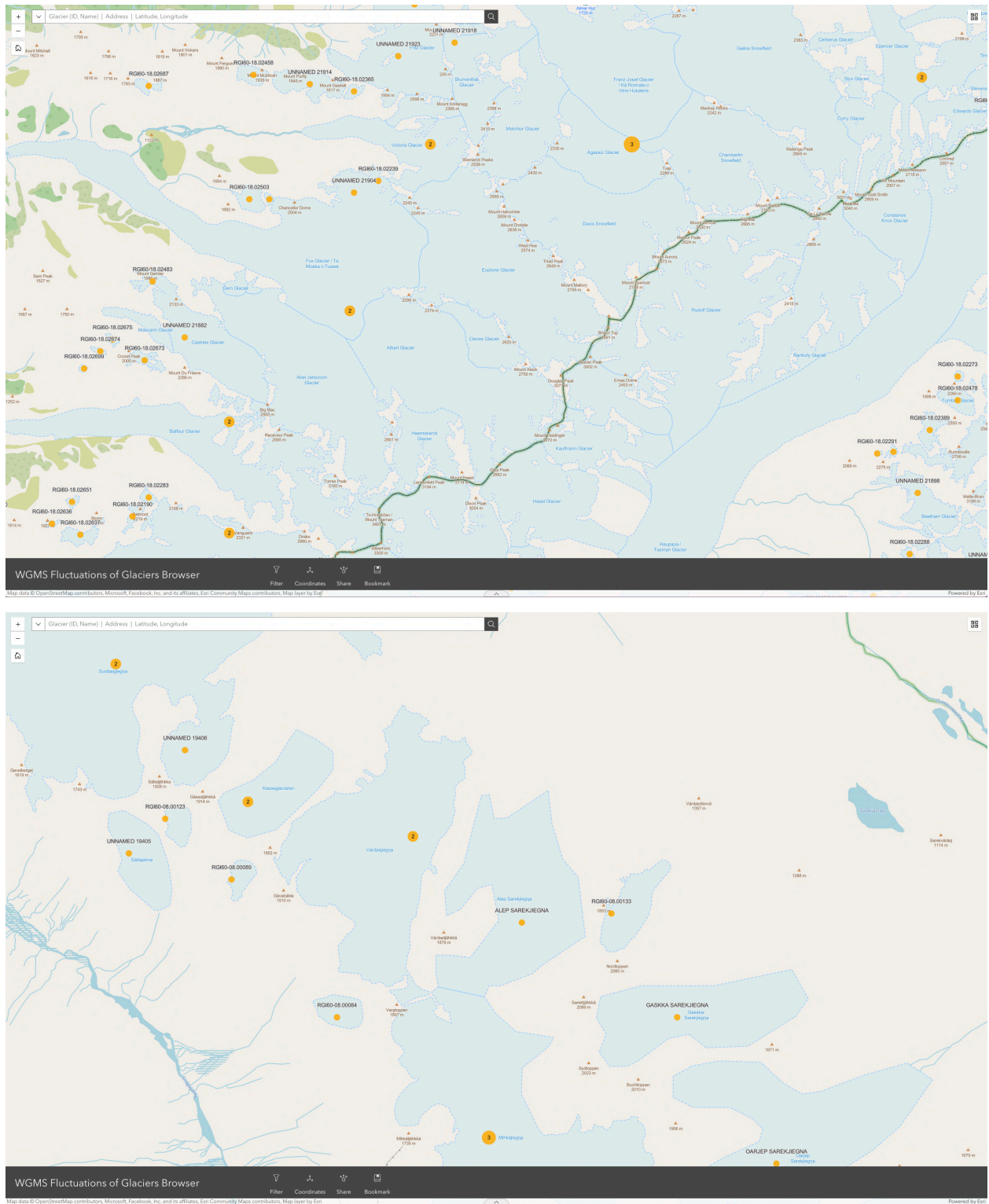


Figure 3.3: Examples of regions with dense and consistent glacier name coverage across reference datasets (OpenStreetMap and Fluctuations of Glaciers). Examples shown from New Zealand (RGI region 18) above and Sweden (RGI region 8) below (WGMS, 2025a).

Regions with sparse or heterogeneous glacier name coverage

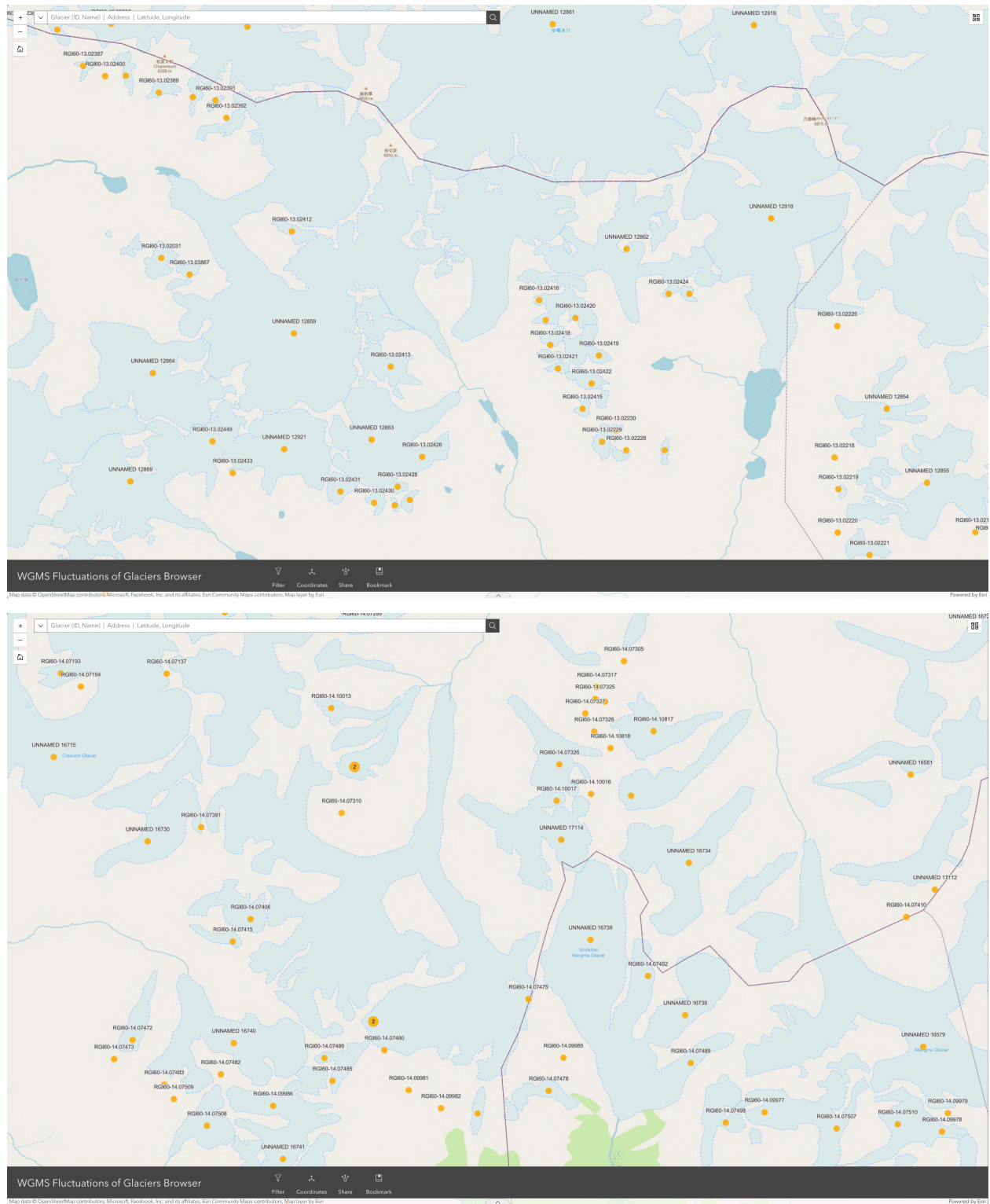


Figure 3.4: Examples of regions with sparse or heterogeneous glacier name coverage across reference datasets (OpenStreetMap and Fluctuations of Glaciers). Examples shown from China (RGI region 13) above and Pakistan (RGI region 14) below (WGMS, 2025a).

3.3.4 Glacier name coverage in Switzerland

To understand better the scope of glacier names data and examine the different datasets available, the data exploration has been done for Switzerland. Specifically, the individual datasets are assessed on how many glaciers they contain in total and how many how them have or do not have a name. To illustrate this better, the below table contains the datasets looked at in detail. These datasets have been preprocessed by filtering for glacier entries, which had to be done for OSM, swissNAMES3D & GeoNames. Clipping to the extent of Switzerland had to be computed for RGI, GLIMS, OSM, GeoNames, FoG. Only the SGI2016 dataset could be used as is. The same logic for unnamed glaciers was used as mentioned in 3.1, adapted and used in Python (Jupyter Lab with Python version 3.13.3, accessing the data in the PostgreSQL database (version 17.4))

Table 3.5: Overview of glacier name coverage across datasets
Switzerland, clipped to Swiss boundaries

Dataset	Total glaciers	Unnamed	Named
RGI v7 (rgi_ch)	1669	703	966
GLIMS (glims_names_ch)	1143	1	1142
OSM (osm_glacier_ch)	488	1	487
GLAMOS (sgi)	1400	138	1262
SwissNAMES3D (swissnames)	763	0	763
GeoNames (geonames_ch)	811	0	811
FoG (fog_ch)	1421	0	1421

The table shows repeatedly that many names within the RGI are missing, for Switzerland this includes almost half of the glacier features. GLIMS on the other hand shows many named glacier counts. What explains this, is that GLIMS (glims_names_ch) is a curated dataset with only named entries, thus one version of "unnamed" might have been overlooked. This was looked up, and it turns out to be the "Namenlos-II". The OpenStreetMap subset generally does not have many glacier features in Switzerland. However, it has names on all glacier polygons except one. The Swiss Glacier Inventory dataset excels at the named glacier count, and is partly based on earlier SGI versions, and where there was no name, on swissNAMES3D data of 2020 (Linsbauer et al., 2021). SwissNAMES3D, GeoNames and Fluctuations of Glaciers represent institutional datasets which tend to have high name coverage, with the first two being gazetteers and the latter being a glacier inventory. Also, GeoNames is normally based on the gazetteers by national mapping agencies (Unxos GmbH, 2025). However, the exact glacier feature counts differs between GeoNames and swissNAMES3D, but further sources are not clearly stated and thus remains unknown.

Lastly, FoG has a very high coverage of glacier names in Switzerland. The dataset has 1'412 named glaciers in Switzerland, which accounts for almost one fifth of the features in the worldwide dataset. This already highlights the problem of varying data availability depending on the country and/or remoteness of glaciers. Concluding, the data exploration for Switzerland shows that specialised, institutional datasets deliver the best coverage of glaciers in total as well as the most named entries - in this case, the data by Fluctuations of Glaciers and the Swiss Glacier Inventory 2016.

3.4 Database

A PostgreSQL database with the PostGIS extension was used to store, harmonise, and query all reference datasets required for the matching framework. The database, named `glaciers`, is structured into three schemas: `raw`, `core`, and `meta`. Each schema serves a distinct role in the data workflow, from initial ingestion to harmonised storage and efficient querying. Furthermore, a list of available datasets was provided to me by Ethan Welty, database manager at the World Glacier Monitoring Service. In the condensed table [3.6](#) below, all datasets are listed which have been loaded into the database. The database can be found on Zenodo ([INSERTLINK](#)) along with the code and used freely, citing this work.

3.4.1 Data preprocessing

Certain datasets required filtering because their contents were not exclusively glacier features. Two polygon datasets and two point datasets were therefore subset to glacier-only records. An example for `swissNAMES3D` is shown below:

```
CREATE TABLE raw.swissnames25_ply_glacier AS
SELECT *
FROM raw.swissnames25_ply
WHERE LOWER(TRIM(objektart)) IN ('gletscher');
```

This was done analogously for `nma_newzealand`, `nma_canada` and `geonames`. As for the table names, they consist of the country name and usually a prefix. NMA was chosen when the data came from a national mapping agency, GI for glacier inventories and no prefix for worldwide datasets such as `GeoNames`, `FoG`, `WGI`, `OSM` and `glims_names`. The exact names of database tables and their origin can be looked up in table [3.6](#). Overall, preprocessing is minimal and aims to preserve source fidelity.

Database schema

In a first step, all datasets were imported into `raw`, keeping their original structure and all entries. In a second step, two tables in the schema `core` were populated as harmonised tables. The two tables are called `glacier_polygons` and `glacier_points`, where the 14 polygon and 11 point datasets were imported respectively. In this schema, only named entries were imported, while adding a constant source label and mapping columns. Harmonisation included standardising attribute names, converting geometries to a consistent CRS, and ensuring each feature has a globally unique identifier. This guarantees that the matching algorithm can query all datasets in a uniform and predictable way. This harmonisation ensures that all datasets share a consistent structure, features can be filtered by source (e.g. `source='WGI'`), and the matching algorithm can query all reference datasets through two unified tables instead of 24 independent ones. Indices and views were further generated to ensure a smooth workflow. Lastly, the `meta` schema stores the WGMS validation dataset (`validation_rgi_poly`), used for the matching workflow evaluation. The meta schema was thought for all auxiliary datasets.

The Randolph Glacier Inventory (RGI) required a slightly different import procedure. The bundled download contains 19 sub-folders representing the RGI regions, each of which had to be imported separately into the `raw` schema. These were concatenated into a single table

`core.rgi_all`, preserving all attributes and geometries. Finally the RGI data can be accessed with materialised views, separated into named and unnamed entries.

Harmonised schema (core)

To enable efficient and uniform access during matching, all glacier features were subsequently transformed into two harmonised tables:

`core.glacier_polygons` (14 polygon datasets) and
`core.glacier_points` (11 point datasets).

Only named features were included in these harmonised tables, as the unnamed RGI polygons are used exclusively as the query set. Each feature in the harmonised tables contains the attributes as in [3.7](#).

Table 3.7: Example row from table `glacier_polygons` in schema `core`

Attribute	Value	Description
id	gi_austria:2015:41	source:edition:source_id
source	gi_austria	chosen name for dataset
source_id	41	ID originating from source dataset
name	Löffler Kees	glacier name in dataset
geom_wgs84	MULTIPOLYGON (((11.915166047182352 47.04439932893182, [...], 11.915166047182352 47.04439932893182)))	glacier geometry in EPSG:4326
geom_moll	MULTIPOLYGON (((940269.7191785573 5560536.164807859, [...], 940269.7191785573 5560536.164807859)))	glacier geometry in ESRI:54009
area_m2	579,687.2841304409	glacier size
attrs	"id": 41, "nr": 9059, "name": "Löffler Kees", "year": 2015, "objectid": 41, "shape_area": 579194.4275512695, "shape_leng": 5154.93726882654	all attributes of original dataset
edition	2015	year of dataset; if given a range, the most recent year was taken

Each feature in the harmonised tables is assigned a stable identifier `id`, composed of `source`, `edition`, and `source_id`. However, some sources lacked a distinct and unique ID. Where the `source_id` was not unique, a hashed and reproducible identifier for every feature was generated. In practice, this hashing procedure was implemented only for three datasets: Svalbard (polygons), GLIMS names (polygons), and GNS World (points). This was used to update the `source_id` field, serving as the primary stable identifier throughout the workflow.

Table 3.6: Overview of polygon and point datasets integrated into database (DB) and used as reference datasets

Name	DB table name	Year	Reference	Region
Polygon datasets				
OpenStreetMap	osm_ply	2025	(OpenStreetMap contributors, 2025)	World
Austrian Glacier Inventory GI 4	gi_austria	2015	(Buckel & Otto, 2018)	Austria
GLIMS names	glims_names	2025	(Welty & WGMS, 2025)	World
swissNAMES3D	nma_switzerland	2025	(Bundesamt für Landestopografie swisstopo, 2025)	Switzerland
Swiss Glacier Inventory 2016	gi_switzerland	2016	(Linsbauer et al., 2021)	Switzerland
Catalog of Russian Glaciers	gi_russia	2019	(Institute of Geography of the Russian Academy of Sciences, 2021)	Russia
Second Glacier Inventory of China	gi_china	2011	(Liu et al., 2012)	China
Svalbard glacier inventory	gi_svalbard	2020	(Kohler et al., 2021)	Svalbard
Italian Glacier Inventory	gi_italy	2007	(Salvatore et al., 2015)	Italy
Inventory of Glaciers and Perennial Snowfields of the USA	gi_usa	2020	(Fountain et al., 2022)	United States
Hydrografi Direkt	nma_sweden	2025	(Ministry of Rural Affairs and Infrastructure: Lantmäteriet, 2025)	Sweden
NZ Place Names – Polygons	nma_newzealand	2025	(Land Information New Zealand, 2015)	New Zealand
Inventario Nacional de Glaciares (Peru)	gi_peru	2020	(SIG INAIGEM, 2023)	Peru
Glaciers of Central Karakoram National Park	gi_pakistan_karakoram	2014	(ICIMOD RDS, 2020)	Pakistan
Point datasets				
Registry of place names in Norwegian polar areas	nma_svalbard	2023	(Norwegian Polar Institute, 2011)	Svalbard
Greenland Glacier Names	nma_greenland	2015	(Bjørk et al., 2015)	Greenland
Canadian Geographical Names Database	nma_canada	2016	(GNBC Secretariat, Canada Centre for Mapping and Earth Observation, 2011)	Canada
Icelandic Meteorological Office: Glaciers	gi_iceland	2023	(Icelandic Meteorological Office, 2023)	Iceland
GeoNames	geonames	2025	(Unxos GmbH, 2025)	World
World Glacier Inventory	wgi	2012	(NSIDC National Snow & Ice Data Center, 2012)	World
OpenStreetMap (points)	osm_pnt	2025	(OpenStreetMap contributors, 2025)	World
Fluctuations of Glaciers	fog	2025	(World Glacier Monitoring Service (WGMS), 2025)	World
SCAR Composite Gazetteer of Antarctica	scar	2025	(Scientific Committee on Antarctic Research, 2014)	Antarctica
Geographic Names Server	gns_world	2025	(National Geospatial-intelligence agency, 2025)	World (excl. USA, Antarctica)

Coordinate reference systems and geometry handling

To ensure globally comparable spatial metrics, particularly intersection over union, overlap ratios, and distances, all geometries were stored in two coordinate systems: WGS84 (EPSG:4326) as the original CRS from the source dataset and Mollweide (ESRI:54009) as an equal-area projection optimised for global analysis. A new column `geom_moll` was created in each harmonised table. The same procedure was applied to `core.glacier_points` and `core.rgi_all`. Spatial indices were created on all geometry columns to accelerate candidate retrieval. The original WGS84 columns were standardised as `geom_wgs84` for consistent access in Python.

Materialised RGI views

Materialised views are faster than using views or accessing a table directly (The PostgreSQL Global Development Group, 2025). Which is why, as the RGI contains more than 275,000 glacier polygons and is queried repeatedly during matching, two indexed materialised views were created: `core.rgi_all_named_moll` and `core.rgi_all_unnamed_moll`. These separate named and unnamed features and greatly reduce query time for the 263,479 unnamed RGI input polygons used in the final matching task.

4 Methods

The subsequent chapter contains the detailed steps of how the glacier data is processed and an evaluation of different metrics. The central part will lie on the generation and design of a matching algorithm and finally how the matching results will be evaluated, including string matching and precision calculation. This workflow directly addresses the research question one; How can glacier geometries in the Randolph Glacier Inventory (RGI) be linked to glacier toponyms using an automated, spatially informed matching workflow?

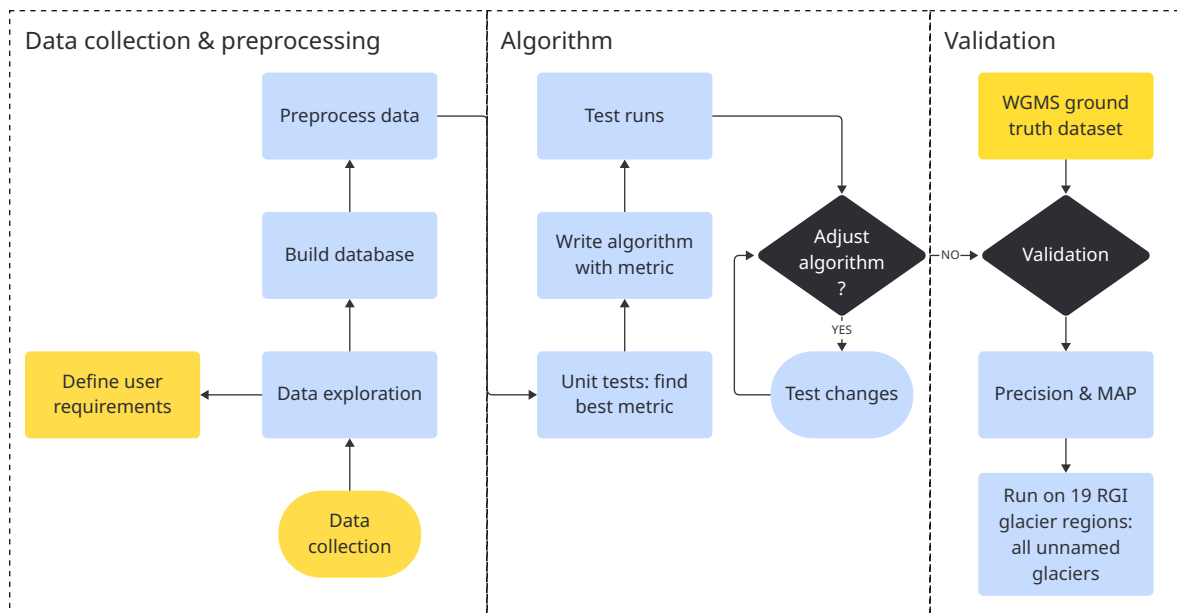


Figure 4.1: Schematic overview of complete methodology

4.1 Programmes and tools

In order to have a visual overview of the glacier inventories, gazetteers and national mapping agency data, a simple data exploration was undergone. The initial, visual data exploration was completed in QGIS version 3.40.12. The [GLIMS Glacier Viewer](#) website was further used for inspect GLIMS, RGI and WGI data, for visual validation and plausibility checks. PostgreSQL 17.4 with PostGIS (extension) and a DBeaver graphic user interface were used for storing the data, as well as restructuring the secondary datasets and merge them into the same schema, divided into points and polygons, which then can be used as sources for the matching candidates. The PostGIS extension was added for spatial data, indexing and visualisations, which was also used in combination with QGIS - either to load data into the database, or to visually inspect it. Python version 3.13.3 within JupyterLab (version 4.4.2) were used for the precise data exploration [3.3.4](#). Different libraries were used, see section dependencies [8](#) for exact versions of the Python libraries. All scripts and code used for data preprocessing, matching, and analysis are available on GitHub under the following link: https://github.com/laurui/Masterthesis_25/tree/main/mscthesis.

4.2 Spatial extent and projection

As the RGI database has worldwide spatial coverage, the matching algorithm is aimed for a global application too. The initial data exploration, unit testing and first matching attempts were tested on Swiss data. In a further step, more regions, defined by bounding boxes, were defined to apply and test the methods. The regions were chosen to represent different glacier types, latitudes as well as data conditions and language scripts. These regions could not be chosen randomly, as glacier presence is not given everywhere, therefore, the bounding boxes for further testing were confined to Norway/Sweden, Pakistan and Austria respectively. Lastly, the validation was done for a worldwide ground truth dataset of 254 named glaciers, as well as the final matching, computed on all the unnamed RGI glacier polygons, also spread out over the whole world. For this reason, it is important that a suitable map projection is chosen.

4.2.1 Map projection

For the map projection, the requirements are to be able to carry out calculations on distance and areas, also, that it can be applied on the whole Earth without too many distortions. As the final matching will be done with 263,471 globally distributed glaciers, it is crucial that the script runs efficiently. However, what is won in computation time, might be lost in a certain distortion. For the part of the distance and area calculations for the matching candidates, it is important to choose a projected coordinate reference system (CRS) (Purves et al., 2018). Paul et al. (2009) also emphasised the importance of a fitting metric projection, as glacier areas in GLIMS are recorded in square kilometres with three decimal places. The analysis within the matching function is based on distances as well as areas. As can be seen in figure 1.1, glaciers are distributed throughout the whole world, however, the calculations within the matching function happens per glacier in one area, centred through the glacier of interest. Around the input glacier, a geometric buffer is created as to reduce the area of analysis and the computation time, therefore the different glacier regions are not compared at the same time but after one another. This workflow speaks for an azimuthal projection, which can be adapted per input glacier. Thus, a Lambert Azimuthal equal-area projection was tried out, centred and re-projected on each input. While this would have lead to results with almost no distortion, the process was too computationally intensive and had to be re-evaluated. Thus, the sites [ProjectionWizard](#) and [Compare Map Projections](#) were used again to find the best fitting projection. As the matching algorithm uses both distance and area metrics, a compromised projection between equidistant and equivalent had to be found. Candidates for this were the Mollweide (ESRI:54009), Hammer (ESRI:53044) and Aitoff (ESRI:53043 / ESRI:54043).

Finally, the area correctness is of more importance than distance because polygons contain more information and are used to compute intersection over union and overlap fractions. Furthermore, the final application of the process is being done with RGI polygon inputs only, the factor of equal area is key. Also for distances, 500-2000m will have some level of distortion, but as a compromise this will be accepted. This is why, the Mollweide projection was chosen. The Mollweide projection is equal area, has the poles as points and is available as a CRS in PostgreSQL.

4.3 Metric behaviour

To understand how different spatial metrics such as distance, area and perimeter behave on real glacier geometries, a preliminary analysis was conducted using the RGI data clipped to Switzerland and the SGI2016 dataset. The aim of this step is not yet to evaluate the matching success, but to characterise the general distribution, variability, and potential of area- and distance-based metrics. This analysis provides the conceptual foundation for the subsequent unit test (section 4.4), where the metrics are evaluated in a controlled, validation based experiment.

Across all analyses in this section, spatial relationships were computed between *unnamed* RGI polygons and *named* reference polygons from the SGI2016 dataset. For point-based datasets, analogous distance-based metrics were calculated with the Fluctuations of Glaciers (FoG) dataset.

4.3.1 Absolute intersection

At first, the absolute area of intersection was inspected. This answers the question: how many glaciers overlap in any way with another glacier? Absolute intersection quantifies the shared area between two glacier polygons and thus represents the most direct form of spatial overlap. It is useful for identifying potential matches but is strongly influenced by glacier size. The minimal overlap area was set to 0.0000005 km^2 or 0.5 m^2 . With these settings, 490 out of the total 703 unnamed RGI glaciers in Switzerland find an overlap with a named glacier from the Swiss Glacier Inventory 2016 (SGI2016). The histogram shows that the intersection area of most is between $0.0000005\text{-}0.5 \text{ km}^2$, roughly 120 overlaps have areas between $0.65\text{-}1.05 \text{ km}^2$.

Table 4.1: Summary statistics for absolute area of intersection between RGI and SGI2016

Statistic	count #	mean	std	min	25%	50%	75%	max
Value (# km^2)	490/703	0.11	0.39	0.00	0.01	0.02	0.07	6.59

Overall, the absolute intersection provides an intuitive first indicator of spatial correspondence. However, the absolute intersection areas cannot be compared directly to one another. For example, if two very large glacier polygons intersect, the overlap tends to be much larger than when two small glaciers intersect. Therefore, when larger absolute intersection areas are scored better than smaller ones, small glaciers would be penalised. Hence, its usefulness is limited by strong area-dependence, motivating the use of a normalised overlap metrics.

4.3.2 Relative intersection

While the absolute intersection is interesting at first, it also reveals that depending on the glacier size, the absolute intersections cannot be compared with one another. This is why, a normalised metric such as the relative intersection or overlap is more meaningful. Relative intersection normalises the shared area by the size of the involved glaciers, enabling comparisons across glaciers of different magnitudes.

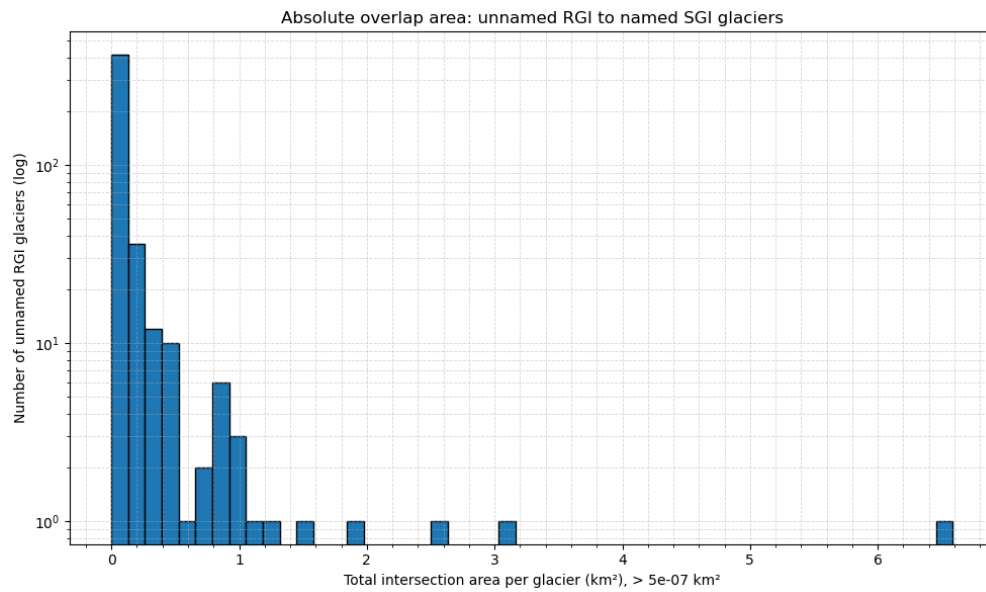


Figure 4.2: Histogram of absolute area of intersection between RGI and SGI2016

Table 4.2: Summary statistics for relative area of intersection between RGI and SGI2016

Statistic	count #	mean	std	min	25%	50%	75%	max
Value (# %)	490/703	63.54	24.02	0.01	50.03	69.71	80.81	100.00

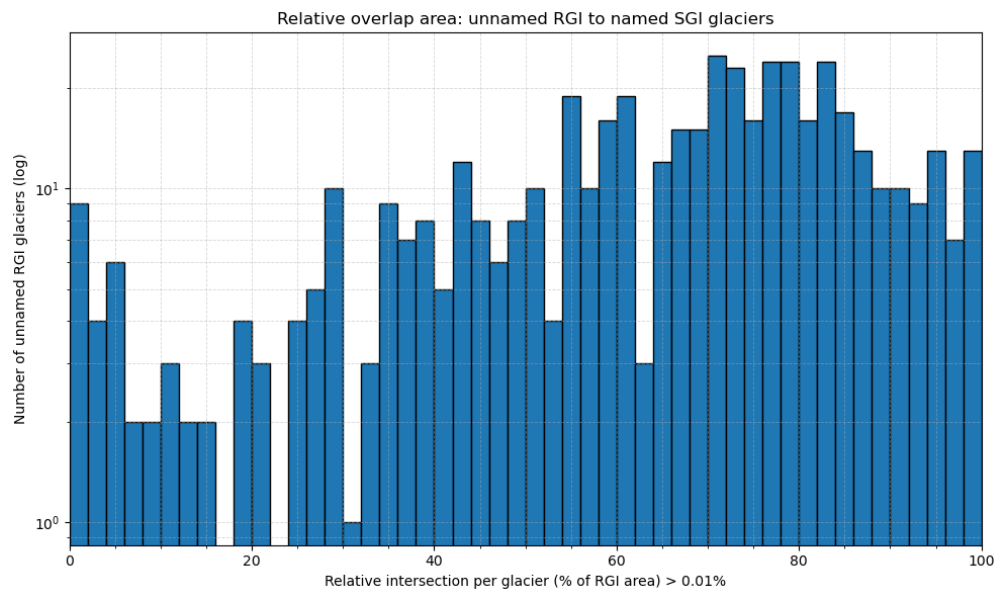


Figure 4.3: Histogram of relative area of intersection between RGI and SGI2016

In Switzerland, 703 RGI glacier polygons lack a name. Of these, 490 (69.7%) overlap at least one SGI glacier by more than 0.0000005 km² (see 4.2) and by more than 0.01% of their area. However, the distribution of relative overlaps is very different to the left-skewing absolute overlaps. While the highest counts lie between 0-2.5%, the percentage areas of overlap practically cover the full span, with a large number lying above 50%, and actually a quarter of all intersections covering 77-100% of the unnamed RGI polygon. This shows, that there is a great added value of taking the SGI dataset as a name source, and that many so far unnamed RGI glacier polygons can be provided with a probable name, based on their relative intersection.

4.3.3 Centroid distance

Centroid distance measures the Euclidean distance between glacier centroids (or between a point and a polygon centroid). It is particularly important for datasets that provide only point geometries.

A distance based metric is needed to evaluate the matching of glacier points, where there are no polygons, such as for GeoNames and FoG. In the case of polygons. The centroid was created with GeoPandas ".geometry.centroid" function, and in the case of points, the point itself is being used to calculate the distance. For this reason, not only the SGI (polygon), but also the FoG (point) histogram is given for comparison.

Table 4.3: Summary statistics for centroid distance (CD) between RGI and SGI2016

Statistic	count #	mean	std	min	25%	50%	75%	max
Value (# m)	703/703	188.64	1082.87	0.00	0.00	0.00	127.99	18448.31

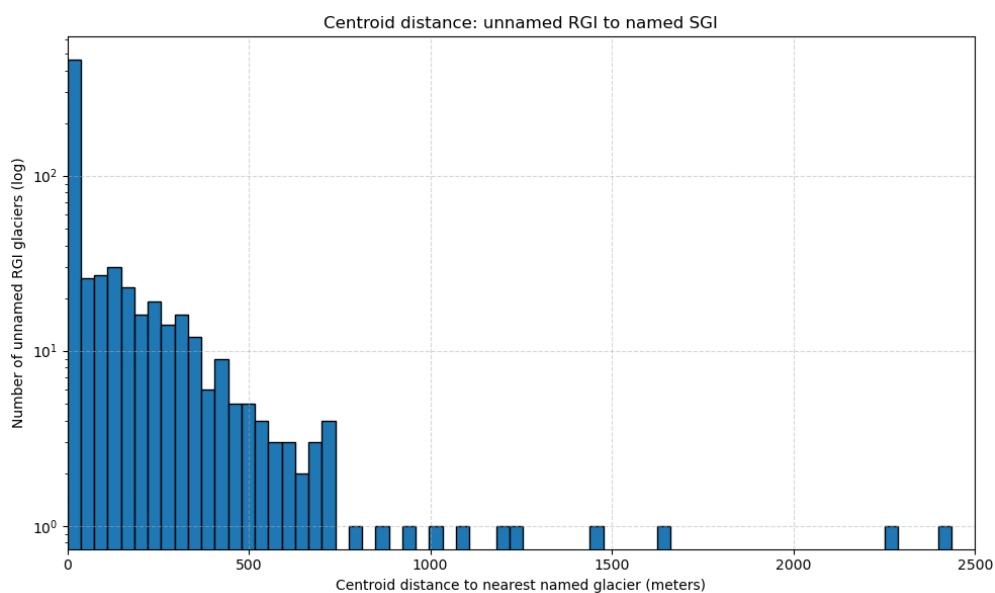


Table 4.4: Histogram of centroid distance between RGI and SGI2016

While for the intersection measures more overlap means a better match, distance measures result in a better match, the closer the centroids. Thus, the strong left-skew in both SGI and FoG distance distributions indicates that many unnamed RGI polygons lie very close to named reference features, suggesting that centroid distance is a stable and discriminative proximity measure, especially where polygon overlap cannot be computed.

Table 4.5: Summary statistics for centroid distance (CD) between RGI and FoG

Statistic	count #	mean (m)	std	min	25%	50%	75%	max
Value (# m)	703/703	524.58	1172.18	6.30	109.93	362.67	593.47	19358.40

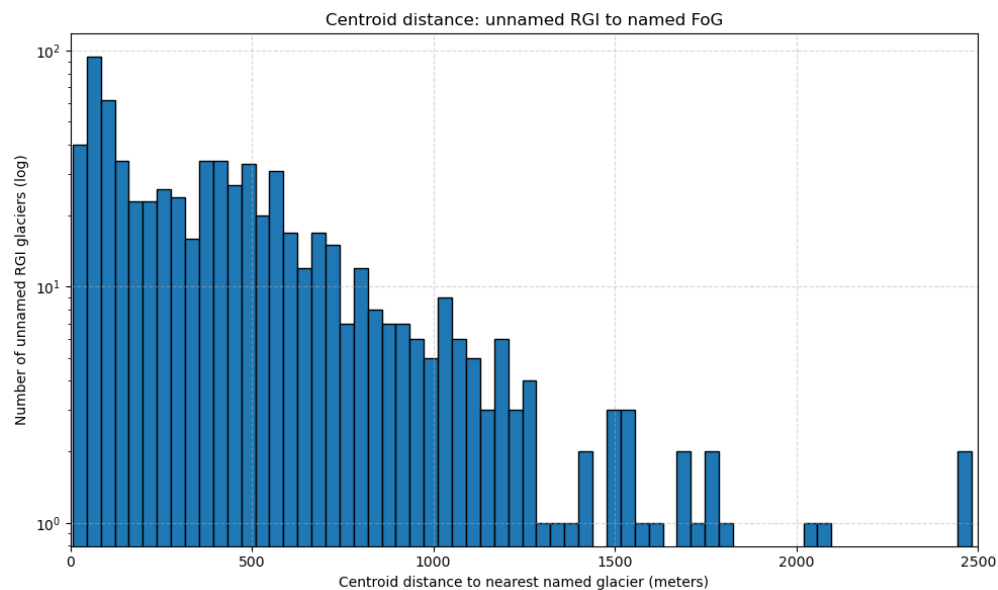


Table 4.6: Histogram of centroid distance between RGI and FoG

4.3.4 Edge distance

In the case of glacier polygons, the edge distance seems more meaningful than the centroid distance. The centroid of a glacier is very susceptible to change, depending on the glacier geometry. For example, the Gorner Glacier in Valais, Switzerland can have a variety of polygon geometries. In the RGI dataset, the Gorner Glacier comprises a large area of 56.604 km² (as of: 13.08.2003), which in OpenStreetMap is divided up into ten different, smaller polygons, only one of them called Gorner Glacier. Thus, if a glacier is one large polygon or many individual parts can influence the position of the centroid, whereas the edge distance is less susceptible. This is why, another distance based metric, the edge distance is computed. The edge distance measures the Euclidean distance between polygon boundaries (or boundary to point distance) and is less sensitive to internal geometry differences than centroid distance.

The edge distance is calculated with the GeoPandas ".sjoin_nearest" function, where the edge distance to the nearest polygon is saved in a column called "distance_col". Also in this case, the shorter the distance, the better the possible match. Similar to the CD histogram,

the biggest part of the unnamed glaciers, has a named glacier very close to it, or even overlapping, where the ED is 0m. In fact, over 600 glacier counts are in the first bin, being 0-37m roughly. Overall, the most counts are within the first 250m, which is a good sign that the SGI dataset brings added value and matching of more names is probable.

Table 4.7: Summary statistics for edge distance (ED) between RGI and SGI2016

Statistic	count #	mean	std	min	25%	50%	75%	max
Value (# m)	703/703	154.85	1071.72	0.00	0.00	0.00	37.55	18315.01

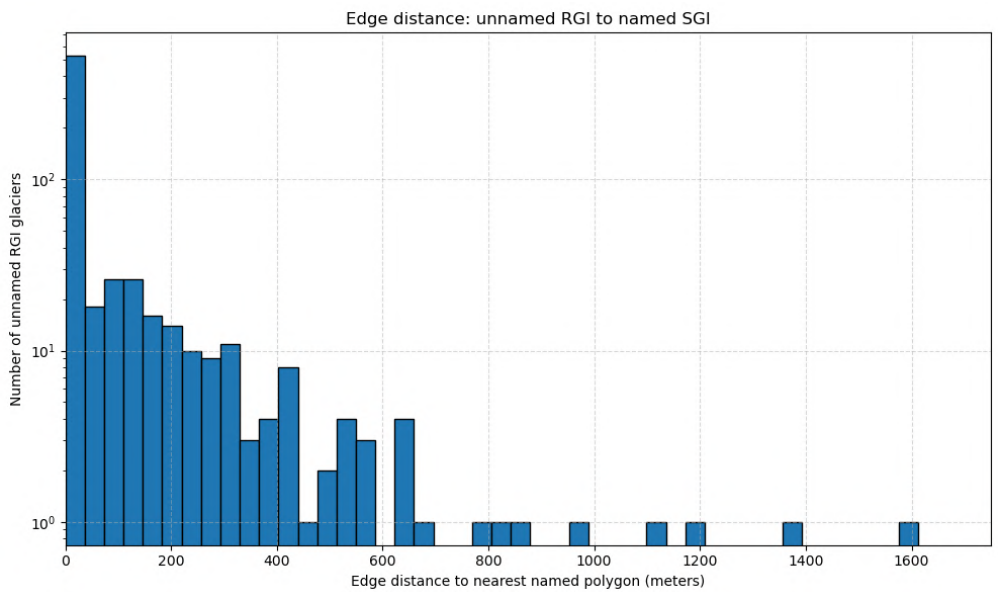


Table 4.8: Histogram of edge distance between RGI and SGI2016

The frequency of very small edge distances, including many values of zero metres, confirms that SGI polygons frequently overlap, touch or nearly touch their unnamed RGI counterparts. This reinforces edge distance as a reliable indicator of geometric proximity.

4.4 Unit test: evaluation of metric performance

Unit testing originates in software development and focuses on verifying the behaviour of the smallest independent part of a program (The MathWorks Inc., n.d.). Per definition, a unit can be any self-contained element such as a function, method, class, or script and is tested separately from the rest of the system. Moreover, a unit test can be done in early stages of developing an application, to ensure code quality and check whether each part works as intended (The MathWorks Inc., n.d.). Applied on the workflow in this thesis, a unit describes any distance, area or perimeter metric possible to be used for matching glacier geometries. Specifically, the metrics are: absolute and relative area of intersection, centroid distance (CD), edge distance (ED) and intersection over union (IoU).

Building on the exploratory metric analysis, a controlled unit test was designed to formally evaluate how well each metric predicts the correct glacier match. This section and test provides the empirical basis for selecting the metrics integrated into the final matching algorithm. The process helps at answering the research question 3: How can entity linking (or object detection) techniques be employed to accurately match glacier toponyms from open web sources to their corresponding glacier outlines in the RGI database, ensuring correct spatial association based on location and spatial proximity? As listed in the literature review (section 2.3), there are several metrics which can be employed, either focusing on shape, distance and/or location. Applied to this chapter the essence of the research question at hand is: which metric best predicts a correct geometry and thus name match?

4.4.1 Unit test design

Geometric matching

To determine which spatial metrics are most effective for geometry matching, a controlled unit test was designed. This experiment systematically evaluates distance and area-based metrics across multiple reference datasets using a ground-truth subset of named RGI glaciers in Switzerland. For this, a sample of 30 named Swiss glaciers was randomly selected from the RGI dataset with a seed (random_state=42) for reproducibility. The following metrics were used to match the best candidate to each glacier in the sample set: absolute intersection area (in km²), relative intersection area, intersection over union (IoU), centroid distance (CD) and edge distance (ED). At this point, this is a 1:1 match, meaning the best possible candidate based on each metric is assigned to each input glacier. This was computed for each metric and each reference dataset respectively.

String matching

Following, the output of each metric was validated with an approximate string matching algorithm. In order to evaluate the match, the RGI name (input) and top candidate name (output) had to be checked for similarity and then classified as correct or false.

Hill (2006) states that placenames in gazetteers are usually composed of the toponym - the name itself - and the administrative modifier; in section 2.1.3 the example of "Morteratsch Glacier" was given, where Glacier is the administrative name. Wu and Winter (2009) further emphasise on the importance of a toponym's feature type, which narrows down the potential matches. As only glaciers are analysed in this project, the string matching should

focus on the main part of the name and ignore the feature type descriptions such as glacier, Kees, vadrett, etc.

For this, a string matching algorithm was generated, where the names were normalised, tokenised and finally compared. What is being compared is the core of the name, for example, for Grosser Aletschgletscher is split up into Grosser | Aletsch | gletscher prefix | core | glacier term. Prefix and glacier term are then removed and the core token is normalised to "aletsch". This core name was then analysed with a fuzzy string matching, using the fuzz package from the rapidfuzz Python library. To generate a score, the two core names were assessed with the `fuzz.token_set_ratio()`, which was then printed in the output table for manual checking. The name score threshold was set to 37 (out of 100), in order label all correct name matches as correct, this score was set on the go, with the sole aim of discriminating correct and false matches. As the final goal is to produce an algorithm which matches geometries without names, the name matching score in the precision test is not meaningful but only used as a control measure. Furthermore, manual control of the matched sample set was necessary and accounted for any classification mistakes. Lastly, the manually checked table could be used to calculate the top-1 precision. That describes the correct matches divided by the total matches, or the true positives (TP), divided by the true positives plus the false positives (FP): $P@1 = TP / (TP + FP)$. This design allows isolating the performance of individual metrics while controlling for dataset differences and ensuring reproducibility.

The final goal of this unit test, is to find out which geometric constraint performs best in order to apply the metric to the unnamed glaciers of the RGI dataset, and other data inputs possibly in future work. Thus, the evaluation of each metric's precision will follow below. The test was computed for a sample set in Switzerland, therefore, the reference datasets included: swissNAMES3D (Bundesamt für Landestopografie swisstopo, 2025), Swiss Glacier Inventory 2016 (Linsbauer et al., 2021), OpenStreetMap (OpenStreetMap contributors, 2025), Fluctuations of Glaciers 2025 (WGMS, 2022a) and GeoNames (GeoNames, 2025). The unit test was also computed with a GLIMS dataset, which acts as a check but cannot be considered conclusive, as the RGI names are taken from the GLIMS inventory. Listed and discussed below are only the performance results of the polygon dataset Swiss Glacier Inventory 2016 (SGI2016) and the point dataset Fluctuations of Glaciers (FoG), the full results can be found in the appendix (section 9.1). These were chosen to be displayed, as the SGI2016 is an institutional, well documented dataset with a high number of named glaciers in Switzerland. This can also be said for the FoG, which has a high named glacier density in Switzerland, but not worldwide - it is given as comparison to evaluate the metrics centroid and edge distance on point data.

4.4.2 Unit test results

The following tables show and summarise metric performance ($P@1$) across the datasets SGI2016 and FoG, providing a direct comparison of spatial metrics under real-world naming conditions. Furthermore, the basic statistics are included. In the tables, the abbreviations represent: **AI** (Absolute intersection), **RI** (Relative intersection), **IoU** (Intersection over union), **CD** (Centroid distance) and **ED** (Edge distance).

While all detailed tables of the unit test are included in the appendix (see section 9.1), the precision values are inserted and averaged below.

Table 4.9: Location matching of RGI and SGI2016 polygon candidates

Metric	min	max	mean	median	P@1
AI	0.01 km ²	4.83 km ²	0.46 km ²	0.08 km ²	0.90
RI	0.09	0.95	0.68	0.76	0.90
IoU	0.01	0.82	0.53	0.58	0.90
CD	8.67 m	1169.2 m	151.06 m	53.24 m	0.87
ED	0.00 m	0.00 m	0.00 m	0.00 m	0.80

Table 4.10: Location matching of RGI and FoG point candidates

Metric	min	max	mean	median	P@1
CD	19.91 m	1074.6 m	209.24 m	108.47 m	0.90
ED	0.00 m	416.21 m	51.75 m	0.00 m	0.90

Table 4.11: Summary of P@1 precision for Swiss reference datasets (RGI sample, $n = 30$).

Metric	swissNAMES3D	SGI2016	OSM	GLIMS	FoG	GeoNames	Average P@1
AI	0.60	0.90	0.30	1.00	-	-	0.70
RI	0.60	0.90	0.30	1.00	-	-	0.70
IoU	0.57	0.90	0.30	1.00	-	-	0.69
CD	0.57	0.87	0.27	1.00	0.90	0.57	0.70
ED	0.53	0.80	0.33	1.00	0.90	0.57	0.69

4.4.3 Discussion of unit test

Although the unit test provides useful insight into metric behaviour, the precision values across metrics are very similar. This can be seen, as in average (see table 4.11), all metrics perform alike one another. This outcome is expected for Switzerland, where glacier inventories are vast, geometries are of high quality, and spatial agreement between RGI and reference datasets is typically strong. In such a data-rich region, several reasonable spatial metrics converge towards similar predictions, and the precision at one is not sensitive enough to capture small differences in ranking quality within a sample of only 30 glaciers.

Furthermore, the unit test evaluates each metric in isolation, whereas the actual matching algorithm applies an additional spatial filtering step. These filters remove most implausible candidates before scoring, meaning that the operational behaviour of the algorithm is not fully reflected by the unit-test setup. Consequently, the precision test should be interpreted as a comparative indication of the general metric suitability, rather than as a strict optimisation criterion.

Despite similar P@1 values, intersection over union remains the conceptually strongest area-based metric for global glacier matching. This is also supported by the literature review. IoU normalises over both geometries, mitigating the bias towards large glaciers (unlike the absolute area). It further penalises small incidental overlaps, addressing issues seen in the relative intersection where tiny glaciers can dominate. Different versions of the IoU metric are widely used in spatial object detection and entity linking, offering a theoretically

grounded similarity measure. Captures shape and footprint agreement, which is especially important for fragmented glaciers such as the Gorner Glacier.

Edge distance (ED) complements the IoU in situations where:

- the reference data are points (GeoNames, FoG, WGI),
- polygons are missing, fragmented, or inconsistently delineated,
- centroid behaviour becomes unstable for elongated or irregular glaciers.

ED is less sensitive to internal geometry structure. The metric behaviour analysis (section 4.3.4) shows that the vast majority of unnamed RGI polygons lie within very short edge distances (0–250 m) of a named glacier polygon, confirming ED as a reliable proximity indicator.

Thus, IoU and ED offer different but complementary strengths. Where the first describes shape and area similarity, the latter approximately shows nearest-neighbour values. This combination is well suited for a global-scale matching where both polygons and points serve as candidate inputs. In summary, even though the precision test shows only moderate differences between the individual metrics, the combined findings from metric behaviour, literature research, and user requirements justify the selection of IoU and edge distance as the key similarity measures in the final matching workflow. IoU offers a balanced and scale-independent overlap metric for polygon data, while edge distance ensures stable distance-based matching for both polygon and point datasets. Together, and supported by preliminary spatial filtering, these metrics form a robust and interpretable basis for global glacier name attribution.

4.5 Matching algorithm

The matching algorithm takes an unnamed RGI glacier as input, identifies nearby candidate glaciers from multiple reference datasets, specifically, 14 polygon and 10 point datasets (described in section 3), and ranks them according to their spatial agreement in shape and location. Internally, it standardises geometries, applies spatial pre-filtering, and lastly assigns a confidence score based on intersection over union (IoU) and edge distance (ED) which is consistently defined for both polygon and point geometries.

Concerning the structure of the matching process, the general methodology is designed after the example of Walter and Fritsch (1999), consisting of five steps and adapted to glacier polygon input:

1. Preprocessing: This is mainly computed in the database. Filtering named from unnamed data entries for the Randolph Glacier Inventory as well as the external name sources is a key step.
2. Computing potential matching pairs: Because the matching at hand is a 1:N match, e.g., one input polygon and five candidates, a distance buffer as in Walter and Fritsch (Walter & Fritsch, 1999) is the second step. This reduces the area of interest, the potential candidates and thus the computation time.
3. Use of geometric constraints: Based on the unit testing (see 4.4) and exploratory study, the intersection over union (IoU), as well as the edge distance will be of key focus.
4. Evaluation of matching pairs: Scoring points and polygon matches separately but on comparable scales. A scoring matrix has been generated for this use.
5. Calculation of final matchings: Due to the fine-tuned confidence score, a final ranking is enabled, comparable to how a human would rank the candidates.

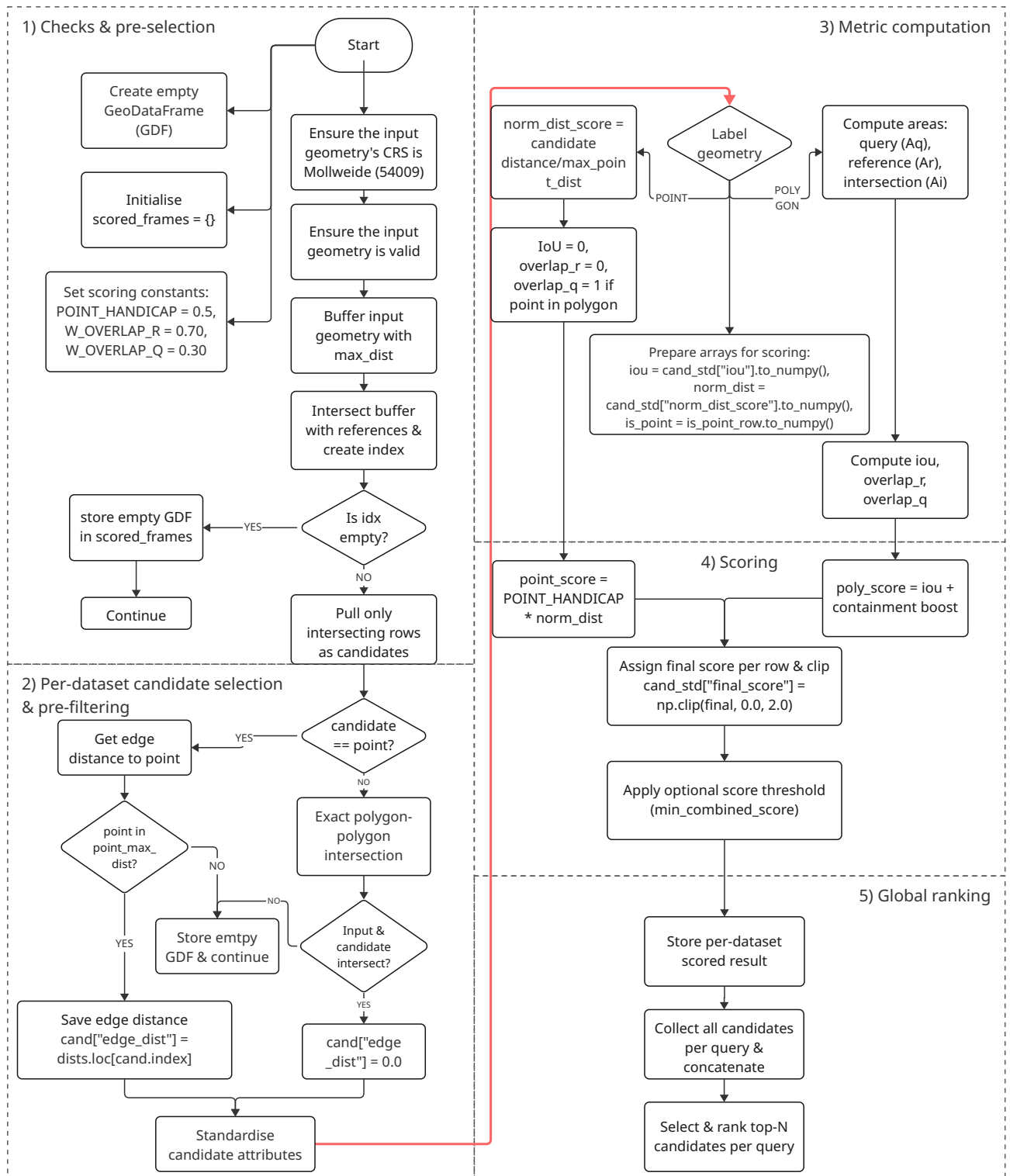


Figure 4.4: Flowchart depicting the steps of the main matching function, divided into its thematic blocks.

4.5.1 Input parameters

The input parameters are not few, but many are given as default values, which can be changed if wanted to. Following is a table of the parameters including their default value and a short description.

Table 4.12: Input parameters of the matching algorithm (`match_multi` function)

Parameter	Default value	Description
<code>input_geom</code>	—	Shapely geometry, GeoSeries/GeoDataFrame (first row used), or GeoJSON-like dict. If the object does not carry a CRS, an <code>input_crs</code> must be provided.
<code>input_crs</code>	—	CRS of the input geometry (e.g., "EPSG:4326"). Required when <code>input_geom</code> is plain Shapely or raw GeoJSON without a CRS.
<code>datasets</code>	—	Refers to the list of dataset specifics. Each <code>gdf</code> is a GeoDataFrame of candidate outlines.
<code>max_dist: float</code>	2000	Overall search radius in meters. Grows input geometry by <code>max_dist</code> meters in all directions.
<code>point_max_dist: float</code>	500	Point search radius in meters; also used to normalise point edge distances into <code>norm_dist_score</code> in $[0, 1]$.
<code>min_combined_score: Optional[float]</code>	None	If set, filters out candidates whose <code>combined_score</code> lies below this threshold.
<code>top_n_per_dataset: int</code>	5	Number of top-scoring candidates to keep per dataset before concatenation.
<code>top_n_overall: int</code>	10	Number of top-scoring candidates to return per input geometry after concatenating all datasets.

The input geometry can be given as a polygon or point. Specifically, there are four possibilities for giving the input: 1) coordinates of a simple polygon or point, 2) reading the geometry from a file, in the format of a GeoJSON for example, 3) by a row index, and 4) by inserting the ID number of a glacier of interest and indicating the geometry row, which has to be done also for case 3. Further, the CRS of the input geometry has to be indicated.

The default values for the search radius are chosen based on the histogram of unnamed RGI glacier areas [3.1](#), where the average size of an unnamed glacier is 1.672km^2 , thus growing an input area by 2 km to all sides will include enough candidates, while reducing computation time drastically. The candidates are again filtered for keeping only intersecting geometries.

4.5.2 Metrics used

Inside `match_multi` function, the metrics are computed in several steps, outlined in the following list:

1. The input geometry is reprojected to the Mollweide CRS and corrected for any geometric invalidities.
2. A buffer around the input (with radius `max_dist`) defines a spatial window.
3. For each dataset:
 - (a) A spatial index query keeps only candidates that intersect the buffer.
 - (b) If the dataset consists only of points, candidates are further restricted to edge distances $\leq$ `point_max_dist` from the input polygon.
 - (c) If the dataset contains polygons, candidates must intersect the input polygon.
4. For polygon-to-polygon cases:
 - (a) The intersection area is computed.
 - (b) From this, the code derives IoU, overlap of the query (`overlap_q`), and overlap of the reference (`overlap_r`), all clipped to $[0, 1]$.
5. For mixed cases (point–polygon or polygon–point), IoU and overlaps are set to zero or simple coverage flags, and the score is driven mainly by distance.
6. For points, `norm_dist_score` is derived from `edge_dist` and `point_max_dist`, assigning higher scores to closer points.

In summary, polygons receive a full set of geometric overlap metrics (IoU and two overlap ratios), while points receive a distance-based score scaled with a handicap. The individual metrics and their computation are described in more detail in the following paragraphs.

Edge distance

During the discussion of the results with my supervisors, it became evident that an edge-to-edge distance measure is more appropriate than centroid distance. The reason is that centroid distance does not become zero when two polygons overlap or when one polygon fully contains the other - situations that represent the strongest possible spatial match. In contrast, the edge distance is zero as soon as two glacier geometries touch or overlap, which aligns better with the intended scoring logic. Consequently, the workflow uses the minimum edge-to-edge distance for polygon to polygon comparisons, whereas

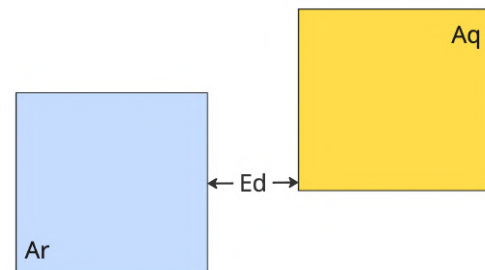


Figure 4.5: Schematic illustration of the edge-to-edge distance (ED)

for polygon to point cases, it measures the distance from the polygon boundary to the point; and for point to point cases, it reduces to the Euclidean distance between the two points.

Intersection over union

Instead of solely looking at the intersection over union, it became valuable to inspect its component parts. For this, the reference area (A_r), the query area (A_q), and the intersection area (A_i) are defined. Thus, IoU is:

$$\text{IoU} = \frac{A_r \cap A_q}{A_r \cup A_q} \quad \text{and in calculation terms:} \quad \text{IoU} = \frac{A_i}{A_r + A_q - A_i}$$

The IoU measure therefore comprises three individual variables, which in turn enables one to create containment or overlap measures along the idea of IoU, two being:

$\text{overlap}_{\text{query}}$ and $\text{overlap}_{\text{reference}}$.

If both are ≈ 1 then it indicates a strong match. If only one of the measures ≈ 1 , it would mean a containment situation. Specifically, $\text{overlap}_q = 1$ indicates that the intersection area lies fully within the queried polygon, and $\text{overlap}_r = 1$ means that the A_i lies fully within the reference glacier polygon if both are of similar sizes.

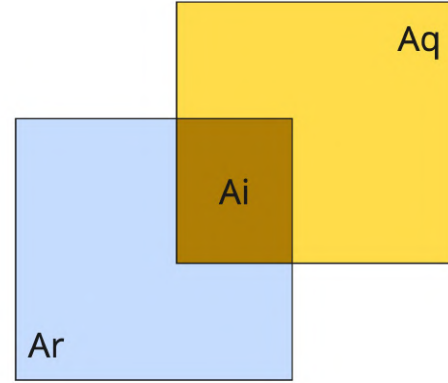


Figure 4.6: IoU and its geometric components: reference area (A_r), query area (A_q), and intersection area (A_i)

Query overlap:

For every candidate/reference geometry r , compute the fraction of the query that overlaps it:

$$\text{overlap}_{\text{query}} = \frac{\text{area}(q \cap r)}{\text{area}(q)} \quad \text{or} \quad \text{overlap}_q = \frac{A_i}{A_q}$$

Reference overlap:

For every candidate/reference geometry r , compute the fraction of the reference (candidate) covered by the query: $\text{overlap}_{\text{reference}} = \frac{\text{area}(q \cap r)}{\text{area}(r)}$ or $\text{overlap}_r = \frac{A_i}{A_r}$

How to read the IoU metrics:

- $\text{IoU} \approx 1.0$ only when features nearly coincide.
- $\text{overlap}_q \approx 1, \text{overlap}_r < 1 \Rightarrow$ query contains the reference (parent–child case).
- $\text{overlap}_q < 1, \text{overlap}_r \approx 1 \Rightarrow$ query is contained in the reference (sub-glacier case).
- $\text{overlap}_q \approx \text{overlap}_r \approx 1 \Rightarrow$ outlines represent the same glacier.
- both overlaps small but $\text{edge_dist} \approx 0 \Rightarrow$ outlines merely touch (IoU near zero).

This decomposition of IoU enables the workflow to identify both strong matches and containment relationships. It also allows ordering candidates by overlap_q (query mostly inside candidate) or by overlap_r (candidate mostly inside query). The two situations can be observed in the following matching examples.

Table 4.13: Large input RGI glacier Gorner in Switzerland with sub-glaciers as candidates $\Rightarrow \text{overlap}_r \approx 1$

Candidate name	dataset	geom_type	edge_dist	iou	overlap_q	overlap_r	combined_score
Gornergletscher	GI_SWITZERLAND	MultiPolygon	0.0	0.6972	0.7100	0.9748	1.5925
Gornergletscher	NMA_SWITZERLAND	MultiPolygon	0.0	0.2578	0.2592	0.9787	1.0206

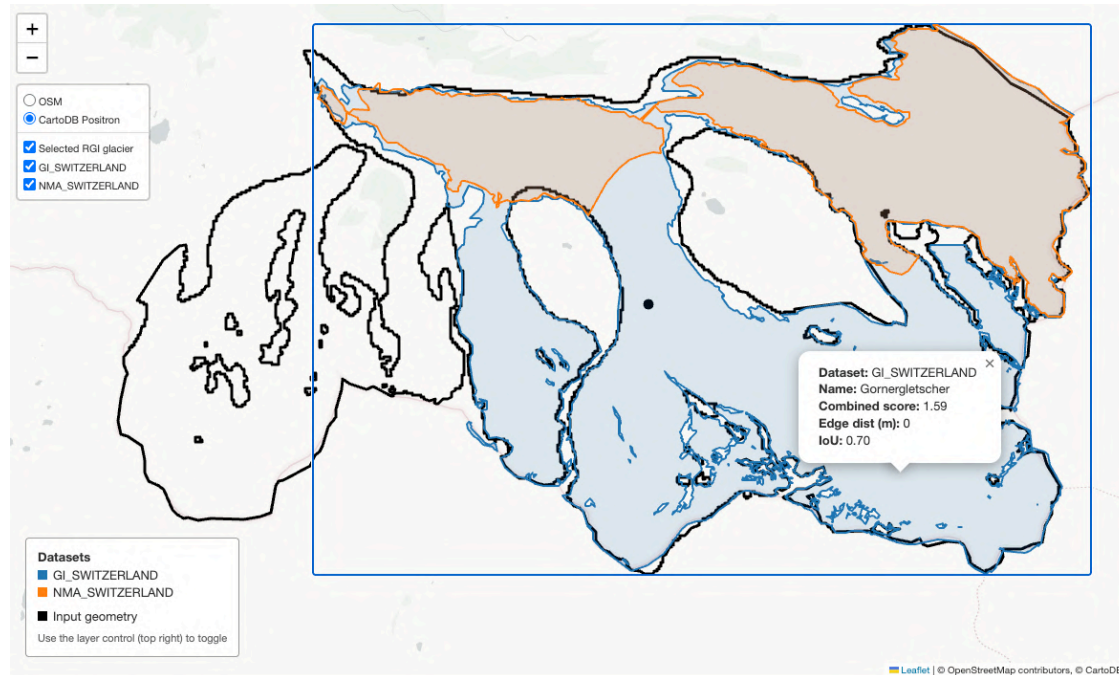


Figure 4.7: Gorner Glacier in Switzerland. RGI polygon in black, Swiss Glacier Inventory 2016 (GI_SWITZERLAND) polygon in blue and swissNAMES3D (NMA_SWITZERLAND) in orange. Shows the situation of large input glacier with complex sub-glaciers: input fully contains the candidate $\Rightarrow \text{overlap}_r \approx 1$

Table 4.14: Small input RGI glacier Kvislajökull in Iceland with ice dome Hofsjokull as candidate $\Rightarrow \text{overlap}_q \approx 1$

Candidate name	dataset	geom_type	edge_dist	iou	overlap_q	overlap_r	combined_score
Hofsjokull	GLIMS_NAMES	MultiPolygon	0.0	0.1164	1.0000	0.1164	0.4979

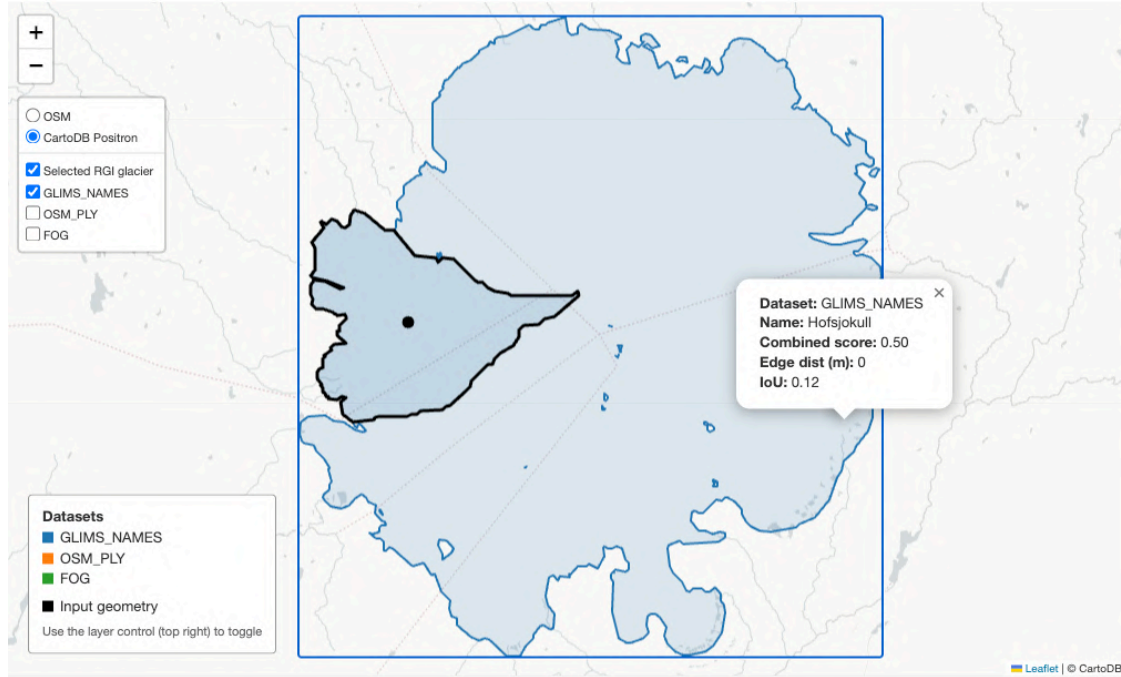


Figure 4.8: Kvislajökull Glacier in Iceland. RGI polygon in black and GLIMS names polygon in blue. Query polygon fully in candidate $\Rightarrow \text{overlap}_q = 1$

4.5.3 Scoring

The scoring component of the matching algorithm assigns a unified similarity value to all candidate geometries returned during the spatial pre-filtering stage. Because the reference datasets contain both polygon and point footprints, and because these geometries encode fundamentally different amounts of spatial information (see section 2.3.2), the scoring functions for polygons and points are designed separately but mapped onto a common scale. The goal of the scoring scheme is to prioritise candidates that show strong geometric consistency through overlap, containment, and shape similarity, while still allowing point data to contribute when polygonal information is unavailable. The final scoring design therefore separates polygon and point logic while normalising both onto a shared 0–2 scale to permit joint ranking. Scoring the matching candidates is necessary in order to produce a ranking, when the candidates per query are more than one. The scoring also needs to reflect how well the candidate geometrically matches the input, especially in the final case, where the match is computed for unnamed glacier polygons, and the geometry is lastly the only thing to assess whether a match is good or not. Thus, the score can be seen as a confidence score of the geometric matching between query input and candidate. Further, each requirement that became evident in the process is stated in the following scoring matrix.

Table 4.15: Criteria and their role in the polygon and point scoring framework

Criterion	Symbol	Definition	Polygons: role / weight	Points: role / weight
Geometry type / reliability	–	Polygons represent glacier extent, points mark approximate locations	Polygons as is	Points receive a handicap $h_{\text{point}} < 1$
Area agreement	IoU	Intersection over union between input polygon and candidate polygon	Primary driver of polygon score	Not used (IoU = 0)
Containment (reference $\subset$ query)	overlap_r	Fraction of candidate polygon area contained in the input polygon	High weight in containment score	Optional minor effect
Containment (query $\subset$ reference)	overlap_q	Fraction of input polygon area contained in the candidate polygon	Low weight (avoids large domes dominating)	Optional minor effect
Distance to glacier edge	norm_dist	Normalised distance from point to polygon (1 inside; outside: $1 - d/\text{max_dist}$)	Not used for polygons (set to 0)	Primary driver of point score

In summary, the matching score ranges from 0–2 for polygons and 0–0.5 for points. It is important to note that the scoring is experimentally designed and not probability-based. Therefore, it has strong geometric meaning but no statistical calibration.

The match function was applied to single polygon cases, covering different types of containment. During this process, the weights for the overlap metrics (W_{OVERLAP_R} , W_{OVERLAP_Q}), along with the handicap for points (POINT_HANDICAP) were fine-tuned. Either can be adapted by users. For this reason, the scoring constants are set to the very top of the matching function, so that it can be easily found and adapted. The point handicap represents the maximum score a point at $\text{norm_dist} = 1$ can achieve. This is due to the fact that the final score for points is made up of the $\text{point_handicap} * \text{norm_dist}$. Further, the weights for containment were set as such, that the case of “candidate polygon is part of query polygon” (high overlap_r , depicted in 4.7) gets rewarded more than the case “candidate polygon contains the query polygon” (high overlap_q , depicted in 4.8). However, in total, these weights amount to 1, which ensures a containment boost also for in-between cases.

Polygons and points operate on different but intentionally designed ranges. The ranges are intentionally asymmetric: polygons encode area agreement and containment, while points encode only proximity. Scoring them on identical numeric ranges would imply equivalent information content, which in turn would misrepresent their differing spatial precision. It also matches the general methodological concept in geographic information science: heterogeneous evidence sources $\Rightarrow$ heterogeneous scoring scales $\Rightarrow$ combined ranking.

Structure of input - output

In summary, the single data frames being fed to the algorithm consist of the columns:

Query: query_geom, query_name, query_id

Reference: geometry, name, id

While the results of the process are given in a CSV, consisting of the columns:

Query: query_id, query_name

Reference: id, name, dataset

Metrics & scoring: edge_dist, norm_dist, iou, overlap_q, overlap_r, combined_score, geom_type

4.5.4 Batch processing over many glaciers

The function `match_multi` operates on a single input geometry. To apply the same procedure to many unnamed RGI glaciers, it is wrapped in two batch routines: `run_batch` (within one region) and `run_regions` (across all RGI regions).

Per-region batch matching (`run_batch`)

The function `run_batch` takes a GeoDataFrame of RGI input glaciers (typically all unnamed glaciers in one RGI region) and applies `match_multi` to each feature in turn. First, it ensures that the inputs are stored in the equal-area Mollweide projection (ESRI:54009), so that distances and areas are measured in metres. The reference datasets are passed in once and reused for all queries to avoid redundancy. For every glacier polygon, `run_batch` calls `match_multi` with a fixed search radius and distance threshold, keeps the top k candidates per glacier, and annotates the results with the original RGI identifier and name (fields `query_id` and `query_name`) and an optional region label. Glaciers for which no candidate is found within the search radius are counted as "misses". All per-glacier candidate tables are then concatenated into a single result GeoDataFrame in Mollweide CRS, which forms the basis for the later validation and coverage analyses.

Global run over all RGI regions (`run_regions`)

The function `run_regions` extends this logic to all RGI glacier regions. For each region index r , it uses a helper to load the corresponding unnamed RGI polygons, normalises their CRS to Mollweide, and then calls `run_batch` with a pre-built list of polygon and point reference datasets. The results for each region are written to disk both as FlatGeobuf files (including geometry) and as CSV files (attributes only). This produces one result file per RGI region, which can later be concatenated for global statistics or inspected region by region.

4.5.5 Encoding

All datasets were encoded in UTF-8. Furthermore, the .csv output also has the fixed encoding of UTF-8, along with the FlatGeobuf output file, where the encoding is by default UTF-8. All names that now result in illegible or wrong letters, stem from the original file download and could not be improved. Specifically, the Italian glacier inventory (Salvatore et al., 2015), which has question marks in 17 cases, e.g., "Ghiacciaio di Aouilli+?" . The names look the same on their website (CGI Italian Glaciers BETA) and were not cleaned in this process.

4.6 Validation methods

A dataset provided by Michael Neuwirth (WGMS) contains the ground truth of 254 glaciers from all over the world. These are called reference glaciers, comprising glaciers the WGMS uses for its website and which have mass balance information. The widespread spatial distribution leads to a broad variation of covered regions and languages, which is a good trial run for the whole matching process. The provided dataset (called `glaciernames_validation`) includes the attributes: `country`, `names`, `latitude`, `longitude` and `outline_id`. The latter is equal to the `rgi_id` in the RGI database, and with this, the features from `glaciernames_validation` were joined to their respective RGI geometry in SQL. This resulted in a ground truth dataset with the attributes: `names`, `outline_id` and geometry in form of RGI multi-polygons. The aim of this validation is to compute the precision of the matching algorithm and generate statistics about the numbers.

The RGI polygons were given as input into the match function, which computed the best five glacier candidate for each input polygon. Since the RGI is a subset from the Global Land Ice Measurements from Space (GLIMS) inventory, the latter was excluded from the reference datasets to prevent self-matching and ensure the independence of the validation process.

The parameters for the matching function applied to the WGMS validation dataset were chosen as follows:

```
code:
val_results = run_batch(
    inputs_gdf=val_inputs,
    datasets=datasets_sel,          # glims_names excluded for
        validation
    region_label="VALIDATION",     # optional tag
    max_dist=2000,
    top_n_overall=5,
    min_combined_score=None,
    point_max_dist=500.0)
```


4.6.1 String matching

This process is analogous to the string matching in section 4.4.1 and will therefore not be described in detail. The only difference in the string matching workflow between the unit test and the validation is that the proposed candidates are assigned correct (= 1) or false (= 0), instead of a fuzzy score. In a next step, the string matching classification is manually corrected. Based on this, the evaluation values are computed and given in the results section. Two examples of cases with patterns that occurred repeatedly during the manual classification check:

- If the ground truth name has a suffix, like East/West or a number, OpenStreetMap or Glims Glacier Viewer was consulted. If there was no other glacier part in the vicinity - this was classified as correct, for example, query name: Glaciar Antisana 15 α | Glaciar Antisana 15 Alpha; candidate name: Nevado Antisana
- If the correct name is in brackets, this was also checked in the Glims Viewer and adapted accordingly, this following case was set to correct: query name: Hurd Glacier; candidate name: Sally (Hurd) Glacier

Further classification examples will be given below. The following lines show candidates, which were classified as false (0), but were manually changed to correct (1). These special cases were not recognised by the string matching algorithm. Often it concerns the ASCII text of the Fluctuations of Glaciers dataset, the abbreviations of the WGI dataset or Cyrillic and Chinese scripts. As a note, the RGI input sometimes shows different official variants which are used and accepted as a name, they are divided with a vertical slash, if a candidate is equivalent to either name, that can be labelled as correct.

Table 4.16: Manual glacier name validation check

Classified as false, manually changed to true:				
RGI Input	Reference output	Reference source	Validation before	Validation after
Ålfotbreen	AALFOTBREEN	fog	0	1
Djankuat Glacier Ледник Джанкуат	Dzhankuat	gi_russia	0	1
Leviy Aktru Glacier Ледник Левый Большой Актру Ледник Левый Актру	Levyi Aktru	gi_russia	0	1
Golubin Glacier	Голубина	osm_ply	0	1
Vodopadnyi Glacier	Водопадный	osm_ply	0	1
Langfjordjøkelen	LANGFJORDJOEKELEN	fog	0	1
Langfjordjøkelen	AV LANGF /LANGFD	wgi	0	1
Kleinfleißkees	KLEINFLEISS K.	fog	0	1
Langjökull (south)	LANGJOKULL	fog	0	1
	SOUTHERN DOME			
Tarfalaglaciären	Darfalaglaciären	gi_sweden	0	1
Nordenskiöldbreen	Nordenskiöldbreen	gi_svalbard	0	1
Barkrak Sredniy Glacier	Баркрак	osm_ply	0	1
Ningchan No. 1 Glacier	宁缠河 1 号冰川	gi_china	0	1
Kangwure Glacier	抗物热冰川	gi_china	0	1
Classified as true, manually changed to false:				
Parlung No. 94 Glacier Parlung Glacier No. 94	PARLUNG ZANGBO: NO. 12	fog	1	0
Vadret dal Corvatsch (south)	Vadret dal Corvatsch N	gi_switzerland	1	0

4.7 RGI regions methods

The same process as for the validation has been done for the 19 regions of the Randolph Glacier Inventory (RGI) dataset. In summary, the matching process was looped over, to compute the matching for each of the regions iteratively and finally give a CSV along with a FlatGeobuf as output. Thus, running the algorithm resulted in 19 CSVs and FlatGeobufs (each called G-01 to G-19). All 19 CSV files were concatenated and then assessed on their coverage, difference in point and polygon data per region, along with the scores of each query. The results are described in the following chapter (see 5.2). The application on all RGI regions requires the batch processing functions described in 4.5.4.

The parameters for the matching function, adapted to the `run_regions` loop, and applied to all unnamed RGI glaciers were chosen as follows:

```
code:
run_regions(regions=range(1, 20),
            datasets=datasets,
            max_dist=2000,
            point_max_dist=500.0,
            top_n_overall=5,
            min_combined_score=0.001)
```

4.8 Evaluation

To evaluate the performance of the matching function, different precision values were computed for each RGI query glacier. Precision is a standard measure in information retrieval (Purves et al., 2018). Per query, the candidate list is first sorted by decreasing combined score. The following metrics are computed with the Python package `pytrec_eval` (Van Gysel & De Rijke, 2018):

Precision@1 (P@1) indicates whether the top-ranked candidate is correct (1) or incorrect (0). The reported P@1 is the mean over all queries, corresponding to the fraction of queries where the best-ranked candidate is correct.

Precision@5 (P@5) indicates the strict, ranked precision of each query's top five results, which is then averaged over all queries. Average Precision (AP): For each query, AP summarises precision over the entire ranked list. Lastly, the mean of the ranked average precision (AP) per query over all queries is reported as MAP.

While the precision values can be computed for the validation dataset, the application on the RGI regions does not have ground truth. Therefore, the final results of the 19 RGI regions will be described based on coverage, point and polygon distribution, along with an assessment of the final scores (= `combined_score`).

5 Results

This chapter presents the empirical results of the proposed name-matching workflow. It addresses (i) how accurately the algorithm can identify correct glacier names for a sampled validation set, and (ii) to what extent the workflow can assign names to currently unnamed glaciers in the Randolph Glacier Inventory (RGI).

5.1 Validation results

Based on the procedure described in section 4.6, the validation of the matching algorithm resulted in an overall (unranked) precision of 0.774 over 253 queries. The mean average precision (MAP) was 0.915, while the ranked precision at one (P@1) reached 0.901 and precision at five (P@5) was 0.623.

On a query level, *Input glaciers* lists the number of input polygons per region, of which the number in *Correct name* have at least one correct candidate. Per query there are up to five candidates (as `top_n_overall = 5`), of which candidates can have a point or polygon geometry. This is listed respectively as *Correct points* and *Correct polygons*.

Table 5.1: Validation summary per RGI region

Region	Query level			Candidate level	
	Input glaciers	Correct name	Correct name [%]	Correct points	Correct polygons
G-01	12	12	100.0	30	13
G-02	21	21	100.0	44	43
G-05	6	5	83.3	8	2
G-06	12	12	100.0	24	14
G-07	15	15	100.0	37	28
G-08	31	31	100.0	75	27
G-09	1	1	100.0	4	0
G-10	3	3	100.0	3	5
G-11	56	55	98.2	104	130
G-12	3	3	100.0	5	7
G-13	30	29	96.7	39	17
G-14	15	15	100.0	21	8
G-15	14	14	100.0	17	6
G-16	9	9	100.0	12	5
G-17	17	17	100.0	25	10
G-18	3	3	100.0	8	5
G-19	5	5	100.0	12	0
Total	253	250	98.8	468	320

Table 5.2: Validation metrics over 253 queries (788 correct of 1,018 candidates)

Metric	Value
Overall precision P_{overall}	0.774
Precision at rank 1 $P@1$	0.901
Precision at rank 5 $P@5$	0.623
Mean average precision (MAP)	0.915

Out of the total 1,265 candidates which could have been proposed by the matching workflow, as a limit of five candidates per query was set, 1,018 possible candidates were found. In turn, from the 1,018 candidates, 788 were labelled correct, checked both with a string matching function and manually. At the query level, 250 of the 253 validation glaciers (98.8%) had at least one correct candidate. The average number of returned candidates per query glacier was 4.02. More precisely, the distribution of the number of candidates per query in the validation set is stated in table 5.3.

The high value of $P@1$ indicates that the matching algorithm typically succeeds in identifying the correct glacier name as the top-ranked candidate. The lower value of $P@5$ shows that, among the five best-ranked candidates per glacier, not all matches are correct, which is consistent with the manual validation. To further quantitatively analyse the matched validation dataset, the histograms below help at interpreting the match data.

Table 5.3: Number of returned candidates per query in the validation set

Candidates per query	Queries
1	26
2	18
3	32
4	25
5	152

From the ranked and scored validation set, a threshold for the next application to the RGI regions can be set. The chosen threshold logic is based directly on the validation statistics (5.1). As stated in the methods section, the scoring is experimentally designed and not probability-based. Therefore, its threshold cannot be set with statistical and probabilistic measures, but rather has to be based on the quantitative analysis of the validation. First, polygons and points should have a different threshold, due to the fact that the scoring system was designed so that: polygons have a range of 0-2, while points are handicapped: 0-0.50 (set point_handicap = 0.50). This means that a “good” point score cannot exceed 0.5 by construction and that a single global threshold would be wrong. For setting a threshold, the relevant results of the validation dataset are:

Table 5.4: Score distribution of polygon (left) and point (right) candidates

	Count	Mean	Med	Q1–Q3		Count	Mean	Med	Q1–Q3
Correct	320	1.45	1.59	1.22–1.73	Correct	468	0.49	0.50	0.50–0.50
Incorrect	87	0.53	0.68	0.28–0.79	Incorrect	143	0.37	0.50	0.22–0.50

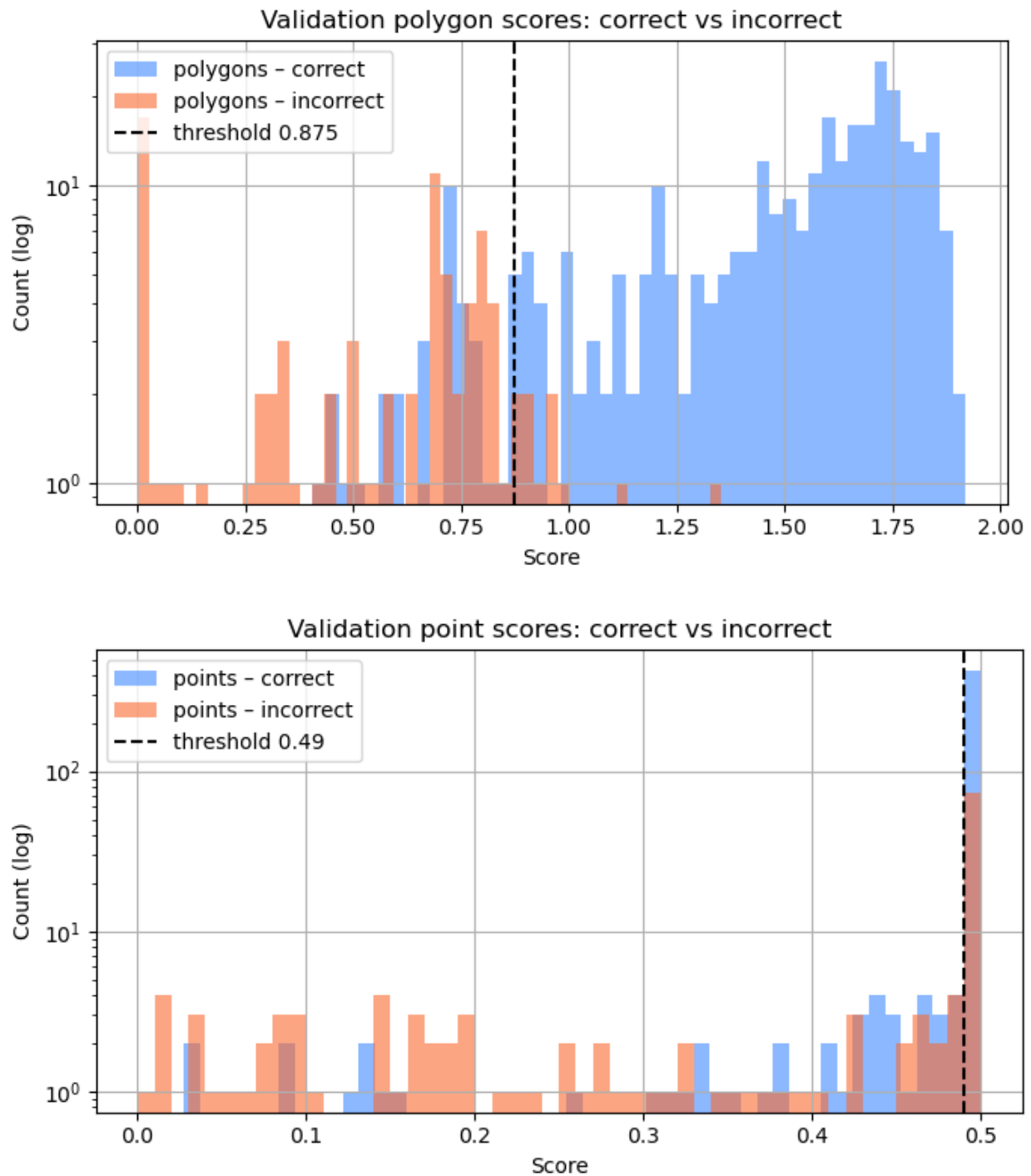


Figure 5.1: Score distributions in the validation dataset after matching, split by geometry type (polygons, points) and correctness (validation = 1 / 0). Dashed line indicates geometric confidence threshold set after validation.

Interpretation for setting a threshold

Correct polygon matches cluster strongly above 1.2, while incorrect ones cluster below 0.8. The optimal threshold for polygons should lie somewhere between 0.8 and 1.2. A threshold of 1 might lead to a maximal separation of distributions, in turn minimising false positives but while also excluding some true positives. To include more correct matches with lower scoring, such as the correct polygon cluster from 0.875 to 1, the lower threshold of 0.875 will be set. This on one hand excludes some true positives, but on the other hand excludes most true negatives, meaning incorrect polygon matches.

For points, the histogram shows less discriminative power, as the incorrect point scores range from 0-0.5. A meaningful separation which is rather conceptually, namely correct point inside glacier $\Rightarrow$ score = 0.5 and point further away $\Rightarrow$ score < 0.5. However, the histogram and the average of the correct point scores show that also slightly below 0.5 there are several correct point matches. Thus, in order to account for the "border cases" the threshold for points will be set at 0.49.

5.1.1 Validation: special cases

While the high precision values showed that the process succeeded overall, some cases left questions. This section lists and discusses some of the results, where the score would suggest a different validation. There was no match classified as correct, which had a score of zero. However, there are 10 queries classified as correct, which have a very low score (≤ 0.2 : 9), which is surprising. Also, manually control was repeated but the cases were left correct. Two of these matches will be discussed below.

Table 5.5: Examples of correct matches with low combined score (≤ 0.2)

query_id	query_name	name	dataset	combined_score	geom_type	validation
RGI2000-v7.0-G-02-14230	Sandalee Glacier	SANDALEE	wgi	0.0275	Point	1
RGI2000-v7.0-G-10-01583	Vodopadnyy Glacier	VODOPADNIY (NO.125)	fog	0.0323	Point	1
RGI2000-v7.0-G-02-14541	Yawning Glacier	Yawning Glacier	geonames	0.0872	Point	1
RGI2000-v7.0-G-13-56928	Ningchan No. 1 Glacier	1	wgi	0.0936	Point	1
RGI2000-v7.0-G-15-13192	Parlung No. 390 Glacier	PARLUNG NO. 390	fog	0.1243	Point	1
RGI2000-v7.0-G-18-02881	Rolleston Glacier	ROLLESTON	fog	0.1339	Point	1
RGI2000-v7.0-G-17-16255	Glaciar Martial Este	MARTIAL ESTE	fog	0.1389	Point	1
RGI2000-v7.0-G-11-03704	Wurtenkees	Wurtenkees	gns_world	0.1450	Point	1
RGI2000-v7.0-G-13-61748	Zhadang Glacier	ZHADANG	fog	0.1507	Point	1

Example 1: Wurtenkees

For the RGI glacier called Wurtenkees and located south of Salzburg, Austria with the ID "RGI2000-v7.0-G-11-03704", all five given candidates are correct. The first and best match is a polygon from the Austrian Glacier Inventory, which confirms that local specialised data is best. However, the last listed point, has a distance of 355m from the input polygon. This shows for one, that a buffer radius for points is needed, as not all are located within the query bounds. Even though the dataset from the US Geographic Names Server (National Geospatial-intelligence agency, 2025) is said to be updated regularly, the point is located at where the former bigger Wurtenkees might have had its terminus and shows that the data might be older. A research article by Schöner et al. (1997) confirms this.

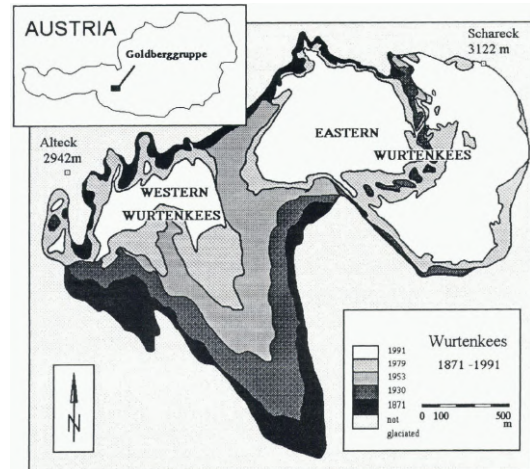


Figure 5.2: Wurtenkees map by Schöner et al. (1997)

Figure 5.3: Top candidate matches for "Wurtenkees"

Name	Dataset	Ed	IoU	overlap _q	overlap _r	Score	Geometry
Wurtenkees + Toteis	GI_AUSTRIA	0.00	0.50	0.61	0.74	1.20	Multipolygon
Wurtenkees	GEONAMES	0.00	0.00	1.00	0.00	0.50	Point
WURTEN	WGI	0.00	0.00	1.00	0.00	0.50	Point
OE. WURTEN K.	FOG	0.00	0.00	1.00	0.00	0.50	Point
Wurtenkees	GNS_WORLD	355.01	0.00	0.00	0.00	0.15	Point

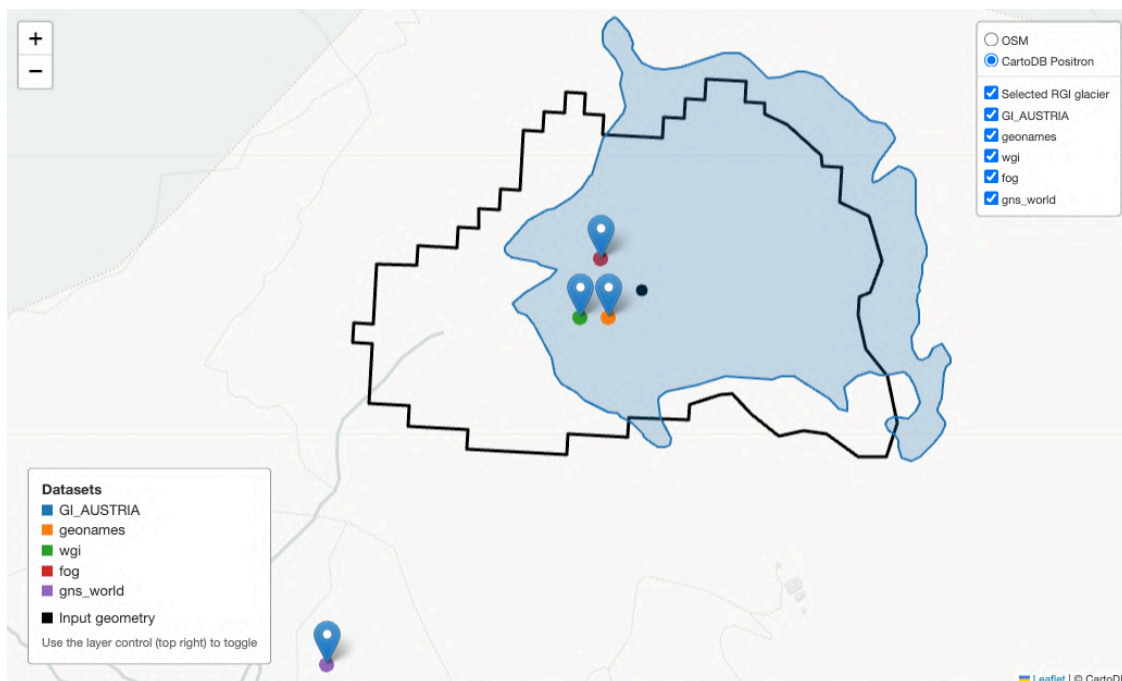


Figure 5.4: Matches for RGI Wurtenkees glacier from the validation dataset. Basemap: CartoDB

Example 2: Sandalee Glacier

The Sandalee glacier is located in the Northwest of the US, close to Vancouver. On the map one can see that the OpenStreetMap polygon (orange) covers five separate polygons, all belonging to the Sandalee Glacier. However, the RGI input has only the middle biggest polygon (black), named Sandalee. This is in tune with the US Glacier Inventory data (blue), which also only covers the middle largest polygon. The FoG and GeoNames points are also both within the RGI input, which makes them correct candidates. Lastly, the World Glacier Inventory (WGI) point is 472.5m away from the RGI bounds, but the name confirms its correctness. This discrepancy could not be fully resolved, as no other nearby glaciated areas are present that would justify this location.

Table 5.6: Top candidate matches for "Sandalee Glacier"

Name	Dataset	Ed	IoU	$overlap_q$	$overlap_r$	Score	Geometry
Sandalee Glacier	GI_USA	0.00	0.6871	0.7799	0.8523	1.5177	Multipolygon
Sandalee Glacier	OSM_PLY	0.00	0.3606	0.7665	0.4051	0.8741	Multipolygon
Sandalee Glacier	GEONAMES	0.00	0.0000	1.0000	0.0000	0.5000	Point
SANDALEE	FOG	0.00	0.0000	1.0000	0.0000	0.5000	Point
SANDALEE	WGI	472.53	0.0000	0.0000	0.0000	0.0275	Point



Figure 5.5: Matches for RGI Sandalee glacier from validation dataset. Basemap: OpenStreetMap contributors.

5.2 RGI regions results

In this subsection, the results of the final matching between all unnamed RGI glaciers and potential candidates are presented. While the methodology for the RGI regions is the same as for the validation, the limitation for the RGI is that there is no ground truth to assess whether the proposed matches are correct. However, as the validation of the algorithm resulted in a high precision, score thresholds could be set and the process can now be applied to inputs, where no actual ground truth is available. This showcases the application in the working world. Thus, in contrast to the validation, the application of the matching process to the RGI regions is a real-life example and need of the World Glacier Monitoring Service (WGMS) for such a matching framework.

Based on the procedure described in 4.7, running the algorithm resulted in 19 CSV and 19 geometry files (each called G-01 to G-19). The concatenated region CSV files were then assessed on their coverage, the difference in point and polygon data per region, along with the scores of each query. When running the matching function, an immediate output given in Python are the number of queries per region and how many inputs cannot be matched at all. These are listed on the go, as the function loops through all the RGI regions, in table 5.7 these non-matched queries are called *Empty*. This is a meaningful initial statistic, displaying the regions with a high data coverage and others with a lack thereof.

The thresholds were set based on validation dataset histograms and statistics after matching, the framing can be read at the bottom of section 5.1. As no ground truth is available to check the final matches, the word "correct" in this section has to be read with care. The thresholds were set at 0.875 for polygon and 0.49 for point geometries.

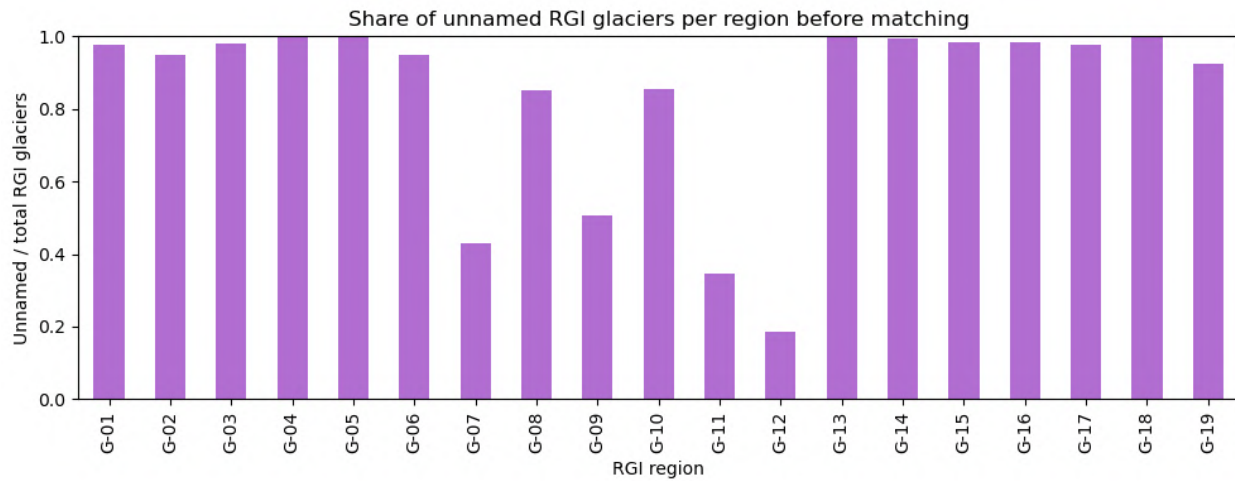


Figure 5.6: Fraction of unnamed glaciers out of total glaciers in each RGI region before application of matching framework

Figure 5.7: Unnamed RGI glacier counts per country before matching, equivalent to column *before matching* - *unnamed* in table 5.7. Country borders with Natural Earth (2009), map projection: Eckert IV (ESRI:54012).

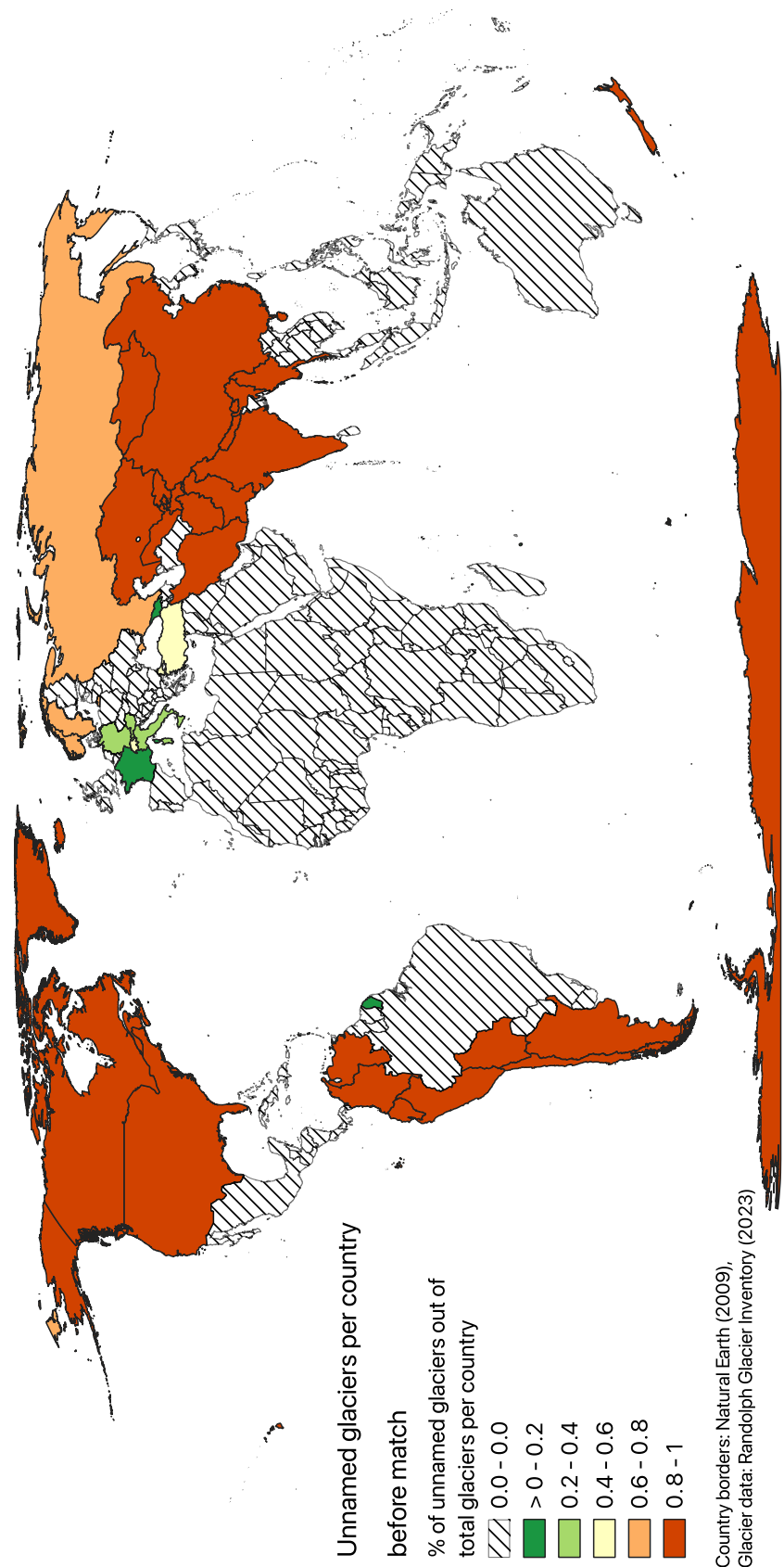


Table 5.7: Summary of match outcome in 19 regions of the Randolph Glacier Inventory. Before matching: total number of glaciers and number of unnamed glaciers. After matching: query glaciers which found no match are called "empty", where a match was found called "matched". A match with a confidence score above the thresholds (points: 0.49, polygons: 0.875) is considered "correct".

Region	Numbers before matching			Numbers after matching				
	Glaciers	Unnamed	Unnamed Glaciers [%]	Empty	Matched	Matched Unnamed [%]	Correct	Correct Unnamed [%]
G-01	27,509	26,857	97.6	23,661	3,196	11.9	477	1.8
G-02	18,730	17,753	94.8	16,173	1,580	8.9	268	1.5
G-03	5,216	5,114	98.0	3,840	1,274	24.9	80	1.6
G-04	11,009	11,000	99.9	10,694	306	2.8	95	0.9
G-05	19,994	19,994	100.0	18,454	1,540	7.7	434	2.2
G-06	568	538	94.7	7	531	98.7	322	59.9
G-07	1,666	717	43.0	439	278	38.8	104	14.5
G-08	3,410	2,905	85.2	831	2,074	71.4	1,259	43.3
G-09	1,069	542	50.7	90	452	83.4	267	49.3
G-10	7,155	6,115	85.5	4,533	1,582	25.9	315	5.2
G-11	4,079	1,413	34.6	149	1,264	89.5	575	40.7
G-12	2,275	420	18.5	260	160	38.1	38	9.0
G-13	75,613	75,363	99.7	53,621	21,742	28.8	5,944	7.9
G-14	37,562	37,392	99.5	35,400	1,992	5.3	457	1.2
G-15	18,587	18,291	98.4	16,432	1,859	10.2	407	2.2
G-16	3,695	3,632	98.3	2,245	1,387	38.2	343	9.4
G-17	30,634	29,878	97.5	26,805	3,073	10.3	358	1.2
G-18	3,018	3,016	99.9	1,607	1,409	46.7	387	12.8
G-19	2,742	2,539	92.6	1,770	769	30.3	287	11.3
Total	274,531	263,479	96.0	217,011	46,468	17.6	12,417	4.7

Before matching

The choropleth map in [5.7](#) serves as a visual aid, where the unnamed glaciers per country are mapped on a diverging scale. Many countries appear red, highlighting the global scale and distribution of the missing glacier names. Further, the table [5.7](#) shows the exact numbers per RGI region and in total. The region titles are listed in table [3.4](#). Before applying the matching workflow, the vast majority of RGI glaciers lacked an assigned name. Inspecting this regionally reveals that most RGI regions contain over 90% unnamed glaciers, and region 5 (Greenland Periphery) is fully unnamed. Several regions show almost complete coverage of unnamed glaciers, such as G-04, G-13, G-14, and G-18, all with $\geq 99.5\%$ unnamed features. Large regions like G-13 (Central Asia) and G-14 (South Asia West) contribute disproportionately to the global unnamed count because of their high glacier numbers, together accounting for over 110,000 unnamed glaciers. Only six regions show substantially lower proportions of unnamed glaciers (18-85%), with region 12 (Caucasus and Middle East) containing the least unnamed, namely 18.4%. In total, 263,479 out of 274,531 glaciers (96%) in the RGI lack a name before processing, highlighting both the scale of the problem and the need for systematic enrichment.

After matching

The first output after the match, *Empty*, reveals the number of queries without any candidates: 217,011 (82.4%) non-matched queries out of 263,479 unnamed glacier inputs. Spatially, this is not distributed evenly: in several regions (G-02, G-04, G-05, G-14) less than 10% of unnamed RGI glacier inputs had a nearby named candidate polygon or point, whereas in other regions up to 98% of inputs resulted in a match. Regions such as G-06 (Iceland), G-08 (Scandinavia), G-09 (Russian Arctic), and G-11 (Central Europe) show the highest success rates. The column *Correct* reflects the high quality and density of external name sources in these areas. In contrast, regions with extremely large glacier counts and sparse naming information, notably G-04, G-05, G-14, and G-15 show comparatively low matching yields (<10%). Finally, after running the matching workflow, a notable fraction of previously unnamed glaciers can now be linked to at least one candidate name: 46,468 unnamed glaciers (17.6%) receive a potential match, and 12,417 (4.7%) have at least one correct match according to the validation thresholds.

A visual analysis further confirms strong regional differences: G-06, G-08, G-09, G-11, and G-18 show relatively high shares of correctly matched glacier names, whereas G-02, G-04, and G-14 remain poorly resolved despite having many unnamed glaciers. This is shown in the two histograms in [5.8](#). The coverage histograms refine the picture by separating polygon and point contributions: point datasets provide a larger fraction of possibly correct matches in several regions (e.g. G-06 to G-09 and G-18–G-19), whereas polygons play a more important role in others (e.g. G-11 and G-12). For each unnamed RGI glacier, all candidate matches from all datasets are evaluated; a glacier is counted as matched if it has at least one polygon candidate above the polygon threshold or at least one point candidate above the point threshold, so each glacier contributes at most one polygon match or one point match. The side-by-side bar plot of correct point and polygon matches makes this contrast explicit: the best-performing regions reach combined coverages of roughly 50–60% of the total unnamed glaciers in the respective region with at least one good match, while many regions remain below 20%. The combined coverage histogram (using $N_{\text{correct}}/N_{\text{unnamed}}$) therefore directly answers the central question of this section: in each RGI region, which share of currently unnamed glaciers can be assigned at least one plausible name candidate by the present workflow?

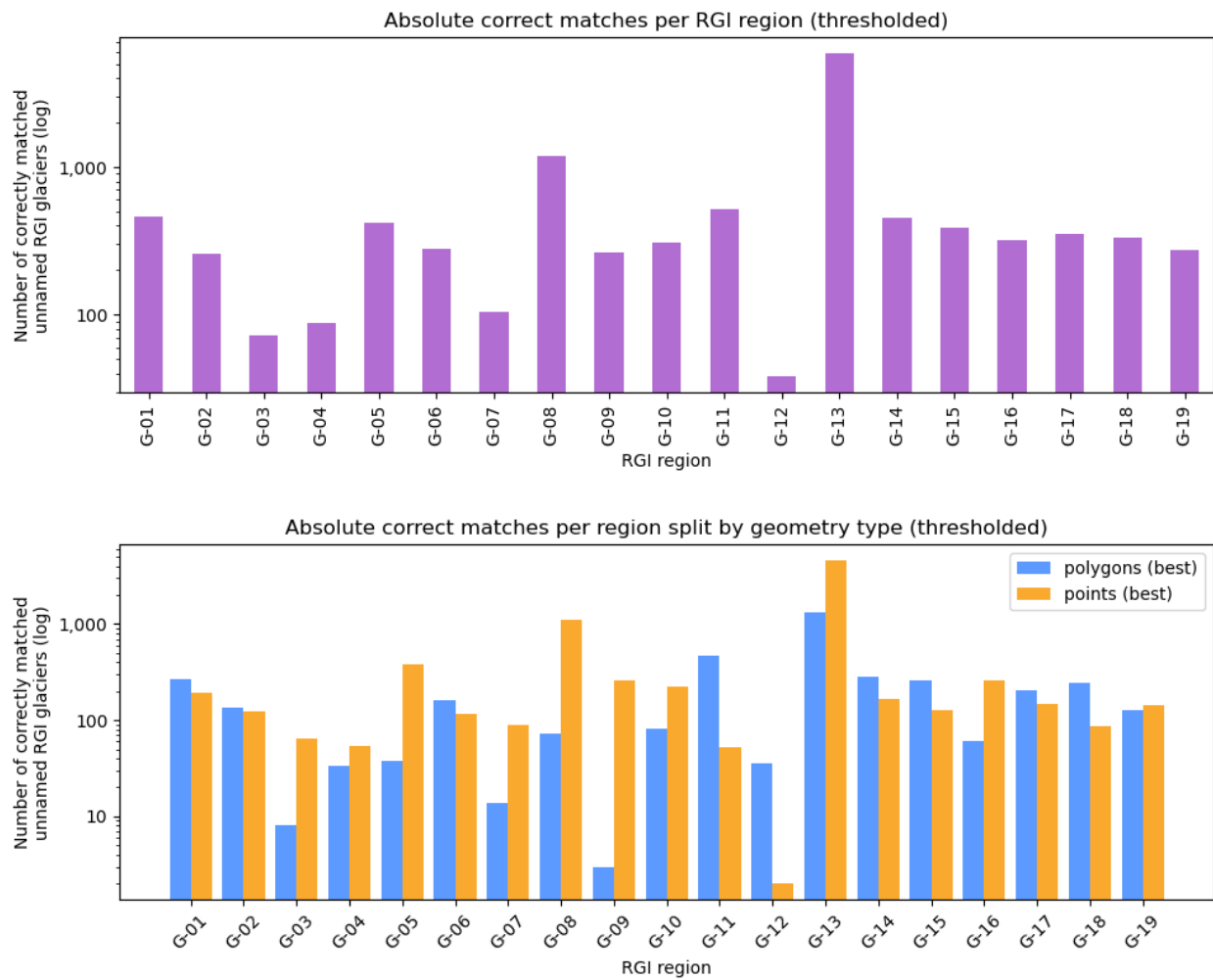


Figure 5.8: Number of unnamed RGI glaciers (in a logarithmic scale) with at least one correct candidate after matching shown combined per region and split by geometry type (polygons, points). Counts are equivalent to column *Correct (After matching)* in table [5.7](#).

Score distribution

To compare the distribution of confidence scores for polygon and point candidates, it was visualised as a histogram with a logarithmic y-axis to reveal the magnitude of occurrences. The score distribution clearly separates polygon and point candidates. While polygons span the scoring range of 0–2, with a cluster between 0.2–0.4 and a visible long tail towards higher values. Points, in contrast, are restricted to the score range 0–0.5, with a strong concentration near 0.5.

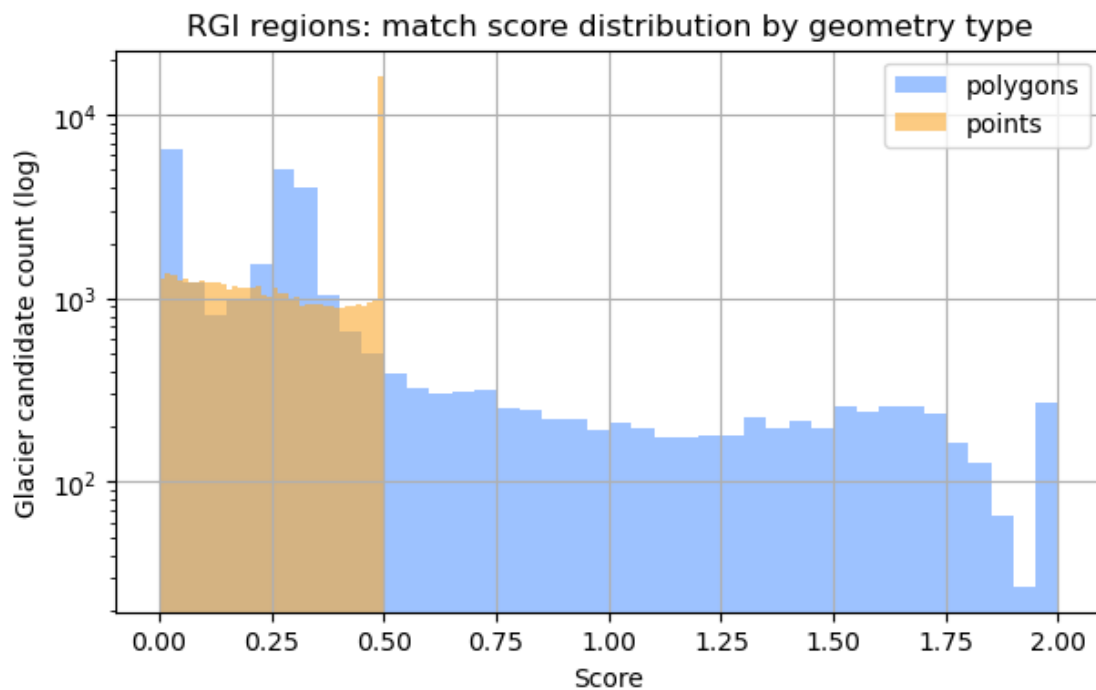
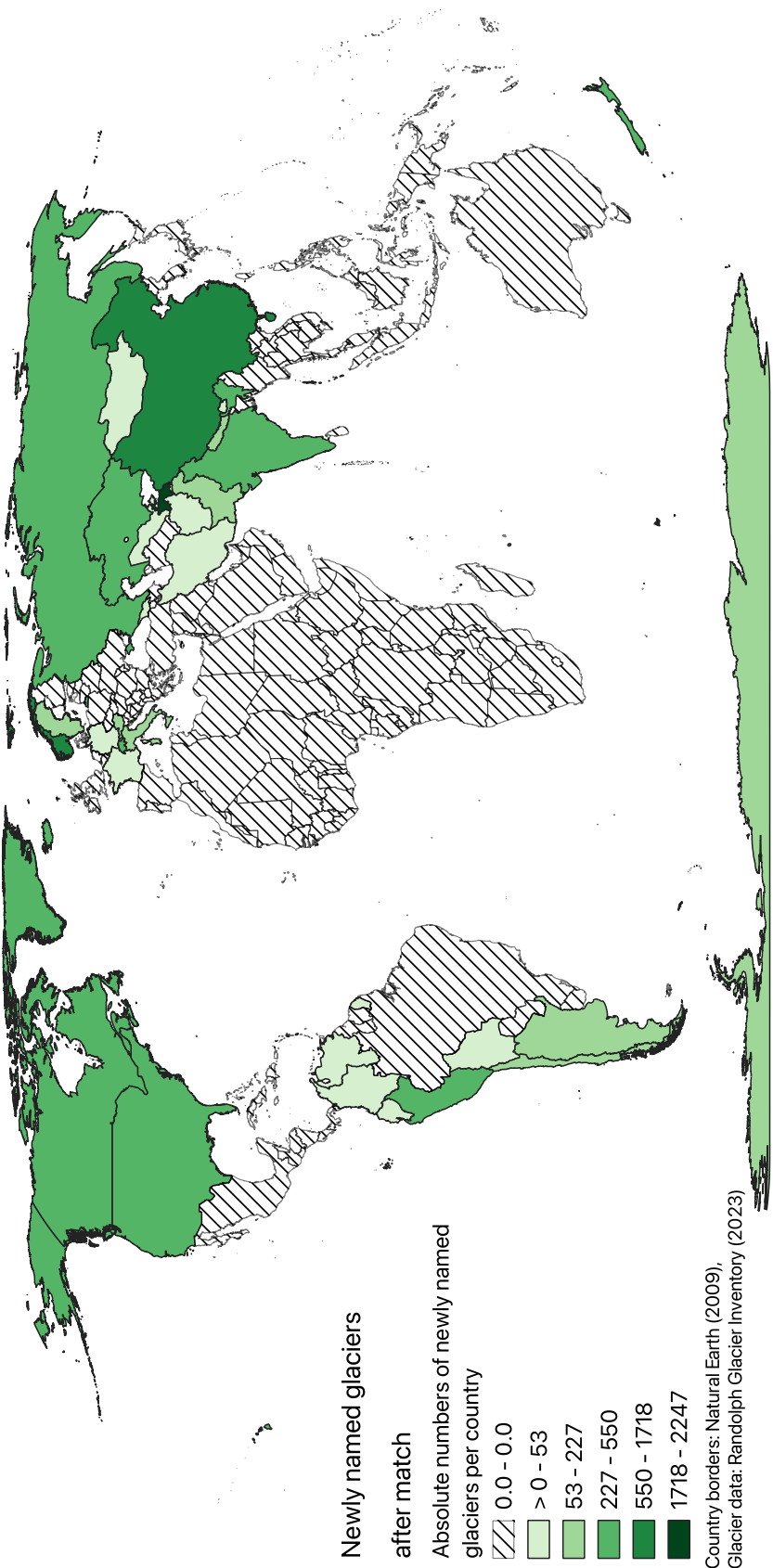


Figure 5.9: Score distribution of 46,468 matched polygons in the RGI dataset divided into points and polygons (y axis in a logarithmic scale). Polygons can achieve a score range of 0–2, points 0–0.5.

To provide a geographic overview of naming completeness, the following choropleth map display the proportion of unnamed glaciers per country. Both choropleth maps [5.7](#) and [5.10](#) were generated with Natural Earth 1:10m resolution country boundaries (Natural Earth, [2009](#)) and depicted in the Eckert IV World projection (ESRI:54012).

Figure 5.10: Absolute numbers of possibly correct names, newly associated in the matching workflow and equivalent to column *after matching - correct* in table 5.7. Country borders with Natural Earth (2009), map projection: Eckert IV (ESRI:54012).



5.3 Contribution of reference datasets

In order to facilitate further work with my results, for example for the World Glacier Monitoring Service (WGMS), the strongest reference data sets globally and per region are discussed here. And to further assess the relative importance of the reference datasets used in the matching workflow, an additional analysis was conducted on the dataset level. For each unnamed RGI glacier, only the top-ranked candidate was considered, ensuring consistency with the regional matching statistics presented above. Two cases were analysed: 1) the dataset of the top-ranked candidate regardless of score thresholds, and 2) the dataset of the top-ranked candidate passing the geometric validation thresholds (polygons: 0.875, points: 0.49).

5.3.1 Global dataset contribution

On a global scale, the World Glacier Inventory (WGI), a point dataset, contributed the largest share of both top-ranked candidates and possibly correctly matched glacier names, accounting for approximately half of all matches. Polygon-based datasets, in particular OpenStreetMap polygon features (OSM_PLY), formed the second most important source, while other global gazetteers and national inventories contributed smaller but regionally significant shares. The full tables on a global and national scale can be seen in the appendix [9.2.2](#).

Table 5.8: Datasets contributing the most top-ranked matches globally (datasets with > 1% shown)

Top-ranked matches (no threshold):				
Rank	Dataset	Geometry	Matches	% of queries
1	WGI	Point	23653	50.902
2	OSM_PLY	Polygon	9258	19.923
3	GLIMS_NAMES	Polygon	6820	14.677
4	GI_CHINA	Polygon	2145	4.616
5	FOG	Point	1596	3.435
6	GEONAMES	Point	899	1.935
Top-ranked correct matches (thresholded):				
Rank	Dataset	Geometry	Matches	% of correct queries
1	WGI	Point	6232	52.029
2	OSM_PLY	Polygon	2510	20.955
3	FOG	Point	875	7.305
4	GLIMS_NAMES	Polygon	724	6.044
5	GI_CHINA	Polygon	263	2.196
6	GEONAMES	Point	255	2.129
7	NMA_GREENLAND	Point	174	1.453
8	GNS_WORLD	Point	168	1.403
9	GI_SWITZERLAND	Polygon	166	1.386
10	OSM_PNT	Point	137	1.144

Across all unnamed RGI glaciers, a small number of reference datasets account for the majority of top-ranked matches. Without applying quality thresholds, the World Glacier Inventory (WGI) dominates the results, providing the top-ranked candidate for 50.9% of all queries. This is followed by OpenStreetMap polygon data (OSM_PLY, 19.9%) and GLIMS

names (14.7%). Together, these three datasets account for over 85% of all highest ranked matches. While point data of the WGI contributes the most matches, polygon-based sources are more prominent among the highest ranks than point data.

After applying geometry-specific quality thresholds, the overall ranking remains similar for the two highest positions. WGI continues to provide the largest share of correct matches (52.0%), followed by OSM_PLY (21.0%). However, the relative importance of lower-ranked datasets changes. FoG becomes the third most important contributor of correct matches (7.3%), surpassing GLIMS_NAMES (6.0%). In total, the top four datasets (WGI, OSM_PLY, FoG, and GLIMS_NAMES) account for more than 86% of all correctly matched glaciers. Beyond these, contributions are distributed across several regional and global gazetteers, each contributing between 1% and 2% of correct matches.

5.3.2 Regional dataset contribution

After computing the analysis per RGI region, the ranking looks more diversified. A strong regional dependence of dataset performance can be observed. In several regions, including Central Asia (G-13) and Arctic regions, WGI clearly dominates the set of correctly matched names, whereas in regions with dense mapping coverage, such as Western Canada and parts of Europe, OpenStreetMap polygon data and national glacier inventories provided the majority of correct matches. These results demonstrate that no single dataset is sufficient for global glacier name enrichment and highlight the importance of integrating heterogeneous sources within a spatially informed matching framework.

Important to note here is that two complementary perspectives are used in this evaluation. The regional results of table 5.7 quantify whether an RGI glacier received at least one candidate or correct match, while the dataset-based analyses in this section assign each glacier to a single top-ranked dataset only. Because this attribution is deliberately restrictive, the resulting totals in table 5.9 are lower than the overall matching counts reported in table 5.7.

Table 5.9: Per RGI region: reference dataset contributing the most top-ranked matches (overall vs. thresholded correct). PT = point data, all others are polygon data.

Region	Region name	Top-ranked overall (no threshold)		Top-ranked correct (thresholded)	
		Dataset	Matches	Dataset	Correct matches
G-01	Alaska	GLIMS_NAMES	1604	GLIMS_NAMES	182
G-02	Western Canada and USA	OSM_PLY	854	OSM_PLY	111
G-03	Arctic Canada North	OSM_PLY	978	FOG (PT)	24
G-04	Arctic Canada South	OSM_PLY	232	OSM_PLY	34
G-05	Greenland Periphery	OSM_PLY	524	NMA_GREENLAND (PT)	174
G-06	Iceland	OSM_PLY	247	OSM_PLY	154
G-07	Svalbard and Jan Mayen	NMA_SVALBARD (PT)	81	NMA_SVALBARD (PT)	66
G-08	Scandinavia	WGI (PT)	1721	WGI (PT)	1026
G-09	Russian Arctic	WGI (PT)	405	WGI (PT)	257
G-10	North Asia	WGI (PT)	1214	WGI (PT)	220
G-11	Central Europe	GLIMS_NAMES	388	GI_SWITZERLAND	166
G-12	Caucasus and Middle East	GLIMS_NAMES	68	GLIMS_NAMES	16
G-13	Central Asia	WGI (PT)	18008	WGI (PT)	4174
G-14	South Asia West	GLIMS_NAMES	729	OSM_PLY	232
G-15	South Asia East	GI_CHINA	573	OSM_PLY	188
G-16	Low Latitudes	WGI (PT)	823	WGI (PT)	233
G-17	Southern Andes	GLIMS_NAMES	1973	GLIMS_NAMES	171
G-18	New Zealand	OSM_PLY	695	OSM_PLY	210
G-19	Subantarctic and Antarctic Islands	GLIMS_NAMES	305	GLIMS_NAMES	80

Across the 19 RGI regions, the dataset contributing the most top-ranked matches is often unchanged after thresholding (e.g., WGI in G-08 to G-10 and G-13). However, several regions show a different source between overall and correct matches, e.g., G-03 and G-05 (OSM polygons vs. FoG/NMA_GREENLAND) and G-11 (GLIMS_NAMES vs. Swiss Glacier Inventory), the implications are addressed in the discussion.

5.4 Case studies

In this section, three example applications for the matching workflow are given. This is done in order to show the various situations in which the matching workflow can be employed. The case studies are distinguished into the four situations: 1) Unnamed input with ambiguous candidates, 2) Well-known glacier with complex sub-glacier systems, and 3) Unnamed glacier with a clear match in order to represent the best-case scenario.

5.4.1 Case 1: Unnamed RGI polygon

The Sally / Hurd glacier appeared in the WGMS validation dataset, where its correct name is Hurd glacier. However, in the RGI database it has no name and the candidates showed conflicting names. As multiple datasets disagree, in this case, the workflow might be helpful to disambiguate. Thus, location and spatial matching can help to find out which inventories or gazetteers show the best results. The polygon input was taken from RGI. The unnamed glacier carries the GLIMS ID = G299601E62691S and RGI ID = RGI2000-v7.0-G-19-00059. With this, the polygon geometry was defined and given as input into the match function. The maximum distance was set to 2000 m or 2 km, along with the number of results, set to 10.

Table 5.10: Best 10 candidate matches for RGI Glacier RGI2000-v7.0-G-19-00059.

Name	Dataset	Geom	Ed	IoU	overlap _q	overlap _r	Score
Sally (Hurd) Glacier	GLIMS_NAMES	MultiPolygon	0.000	1.000	1.000	1.000	2.000
Hurd (Sally/Argentina/Las Palmas) Glacier	GLIMS_NAMES	MultiPolygon	0.000	0.354	0.795	0.390	0.866
Sally Gl	WGI	Point	0.000	0.000	1.000	0.000	0.500
HURD	FOG	Point	0.000	0.000	1.000	0.000	0.500
Sally Glacier	SCAR	Point	15.620	0.000	0.000	0.000	0.484
Sally Glacier	GEONAMES	Point	16.098	0.000	0.000	0.000	0.484
Sally Rocks, Lóbulo	SCAR	Point	76.329	0.000	0.000	0.000	0.424
Hurd, Glaciar	GEONAMES	Point	267.005	0.000	0.000	0.000	0.233
Argentina Glacier	SCAR	Point	280.564	0.000	0.000	0.000	0.219
Las Palmas Gl	WGI	Point	286.990	0.000	0.000	0.000	0.213

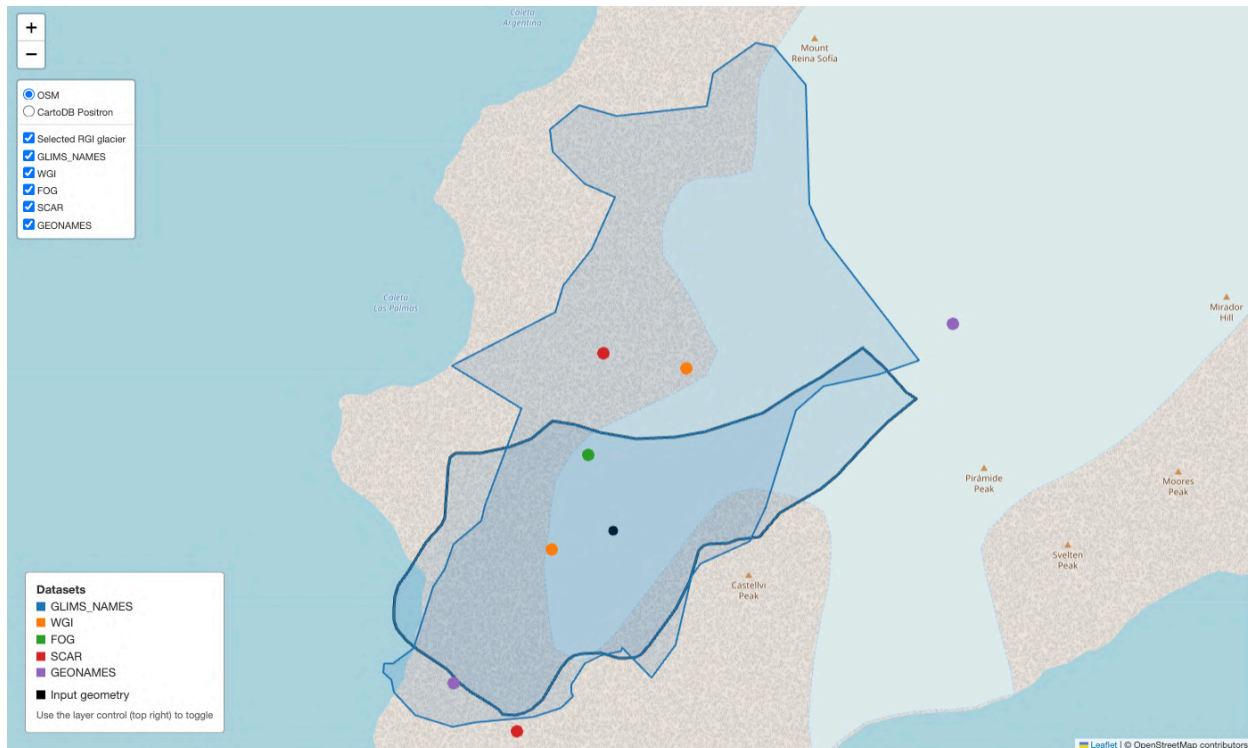


Figure 5.11: Map of potential matches for RGI glacier with ID RGI2000-v7.0-G-19-00059. Basemap: OpenStreetMap.

5.4.2 Case 2: Comparison for a well-known glacier

The Gornergletscher is named in RGI, therefore it does not represent the “unnamed glacier” workflow but is used to illustrate dataset disagreement, as well as many overlapping geometries and alternative names (Gornergletscher / Gorner Glacier / Grenzgletscher / etc.). Its RGI ID is: RGI2000-v7.0-G-11-01225.

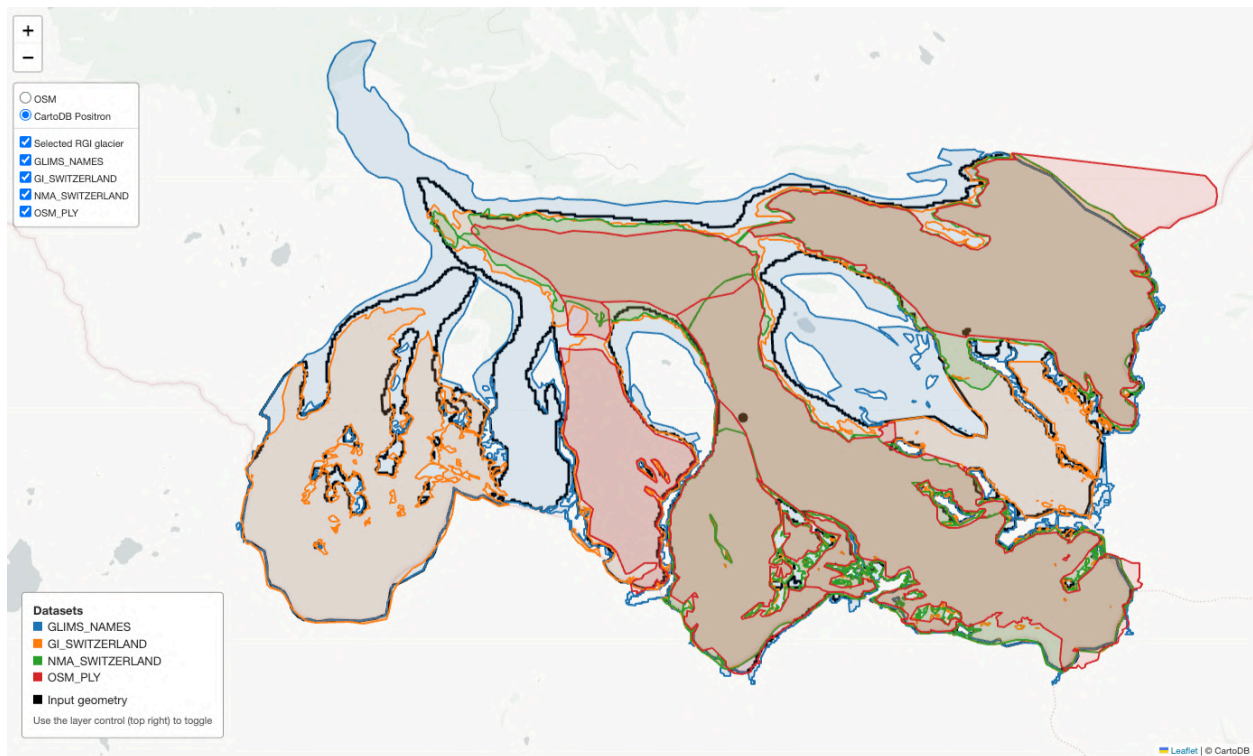
In the following table, 20 possible matches for the Gornergletscher polygon are shown, ranked by their confidence score. The candidates are all multi-polygons, which is why the edge distance (Ed) and normalised distance (Nd) are removed, as well as the geometry type (geom_type). Grenzgletscher is a tributary of the Gornergletscher (*The Swiss Glaciers 2007/08 and 2008/09, 2014*).

The geometry of Gornergletscher differs substantially across datasets. While GLIMS and RGI shows the full system including multiple tributaries, OpenStreetMap and national mapping agencies partition parts as “Grenzgletscher” GI Switzerland provides a modern 2016 outline.

This not only visually demonstrates the dataset along with name disagreement, but also that the matching workflow succeeds at collecting all relevant points and polygons that have an appropriate spatial relation to the RGI outline. The Gornergletscher example further shows why matching polygons is not always straight-forward but that intersection over union based scoring can help this problem.

Table 5.11: Best 20 candidate matches for “Gornergletscher”

Name	Dataset	IoU	overlap _q	overlap _r	Score
Gornergletscher	GI_SWITZERLAND	0.6972	0.7100	0.9748	1.5925
Gornergletscher	NMA_SWITZERLAND	0.2578	0.2592	0.9787	1.0206
Grenzgletscher	NMA_SWITZERLAND	0.1993	0.2002	0.9789	0.9446
Grenzgletscher	OSM_PLY	0.1887	0.1896	0.9753	0.9283
Gornergletscher	OSM_PLY	0.2224	0.2298	0.8745	0.9036
Unterer Theodulgletscher	GI_SWITZERLAND	0.1610	0.1623	0.9535	0.8771
Zwillingsgletscher	NMA_SWITZERLAND	0.0931	0.0933	0.9824	0.8088
Zwillingsgletscher	OSM_PLY	0.0813	0.0814	0.9908	0.7993
Schwarzegletscher	OSM_PLY	0.0667	0.0667	0.9877	0.7781
Schwarzegletscher	NMA_SWITZERLAND	0.0638	0.0639	0.9801	0.7690
Breithorngletscher	OSM_PLY	0.0375	0.0376	0.9802	0.7350
Breithorngletscher	NMA_SWITZERLAND	0.0300	0.0300	0.9820	0.7264
Unterer Theodulgletscher	NMA_SWITZERLAND	0.0192	0.0192	0.9995	0.7246
Klein Matterhorngletscher	NMA_SWITZERLAND	0.0421	0.0422	0.9568	0.7245
Theodulgletscher	OSM_PLY	0.0669	0.0673	0.9094	0.7237
Breithorngletscher (Zermatt)	GI_SWITZERLAND	0.0327	0.0327	0.9483	0.7063
Plateau Rosa	NMA_SWITZERLAND	0.0040	0.0040	1.0000	0.7052
Triftjiplateau	NMA_SWITZERLAND	0.0012	0.0012	0.9967	0.6993
Triftjigletscher	OSM_PLY	0.0122	0.0122	0.9731	0.6971
Scholle	NMA_SWITZERLAND	0.0008	0.0008	0.9821	0.6885

**Figure 5.12:** Map of potential matches for RGI Gornergletscher with ID RGI2000-v7.0-G-11-01225. Basemap: CartoDB.

5.4.3 Case 3: Unnamed polygon with clear match

In order to present an example of a succeeded final match, an unnamed RGI glacier was chosen, which has a match confidence score above the correctness threshold, 0.875 for polygons and 0.49 for points. This example also shows matches of both geometry types, which is useful for the illustration. The unnamed glacier is in the RGI region 6, the Iceland. Its RGI ID is RGI2000-v7.0-G-06-00095.

Table 5.12: Best candidate matches for unnamed RGI glacier RGI2000-v7.0-G-06-00095

Name	Dataset	Geom	Ed	IoU	overlap _q	overlap _r	Score
Illagilsdalsjökull	OSM_PLY	MultiPolygon	0	0.836	0.890	0.933	1.756
Illagilsdalsjökull	GI_ICELAND	Point	0	0.000	1.000	0.000	0.500
Illagilsdalsjökull 2	GI_ICELAND	Point	179.121	0.000	0.000	0.000	0.321
Dýjafjallsjökull	GI_ICELAND	Point	432.463	0.000	0.000	0.000	0.068

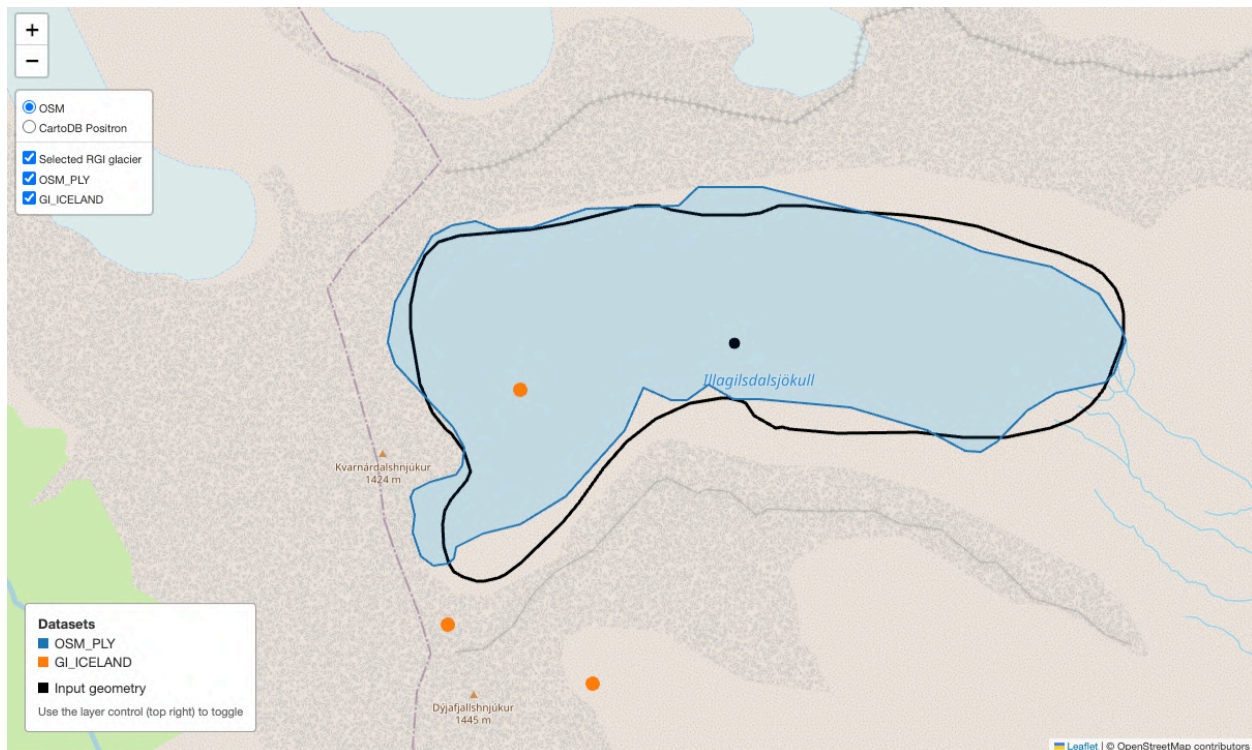


Figure 5.13: Map of potential matches for unnamed RGI glacier with ID RGI2000-v7.0-G-06-00095. Basemap: OpenStreetMap contributors.

The Python generated map shows the best fitting match, the first being a polygon from OpenStreetMap (OpenStreetMap contributors, 2025), also visible on the base layer. The confidence score is very high, namely 1.76, indicating a very good fit of the two geometries. The second best is a point from the Icelandic Glacier Inventory (Icelandic Meteorological Office, 2023) and lies within the polygon input. Another point by the same source indicates that either there are two glacier termini or two data entries for the same point, distinguished with a "2". When consulting the Icelandic Glacier report by Sigurðsson and Williams (2008), this confirms that there is only one glacier called Illagilsdalsjökull.

6 Discussion

This thesis addresses the overarching research question: How can glacier geometries in the Randolph Glacier Inventory be automatically matched with their correct glacier names using external, openly available sources?

In response to this challenge, a geometry-aware matching workflow was generated which identifies candidate names from fourteen polygon and ten point datasets, evaluates spatial similarity using intersection based metrics and distance measures, and presents ranked name suggestions for unnamed glaciers in the Randolph Glacier Inventory. The workflow was implemented in Python and PostGIS, validated on a manually annotated sample, and initially tested in Switzerland before being generalised to global data.

To operationalise this broad question and turn it into a framework, three more specific research questions were formulated, which will be answered accordingly in the following discussion.

1. How can glacier geometries in the Randolph Glacier Inventory (RGI) be linked to glacier toponyms using an automated, spatially informed matching workflow?
2. To what extent can gazetteers and glacier inventories contribute additional glacier name information to the RGI, and how effectively can these heterogeneous sources be integrated?
3. How accurately can spatial entity-linking techniques based on overlap, containment, and distance metrics identify correct name-geometry pairs in a validation setting?

While the overarching topic of this thesis is interdisciplinary between geographic information retrieval (GIR) and glaciology, this implementation is a GIR contribution and its values lie in geometry-aware matching, multi-source integration and name enrichment under a situation of uncertainty. On the glaciology side, the World Glacier Monitoring Service (WGMS) receives clear starting points for glacier name integration from external sources, along with a quantitative analysis of the added value of the name enrichment. However, this work matters beyond glaciers. While the matching workflow has been generated for and applied to polygon and point geometries of glaciers, this process, when adapting source data and function parameters, can also be applied to other natural features, which as discussed in the literature review, remain under-represented in computational studies (Acheson, De Sabbata & Purves, 2017).

6.1 Research question 1: methodology

RQ1: How can glacier geometries in the Randolph Glacier Inventory (RGI) be linked to glacier toponyms using an automated, spatially informed matching workflow?

This question highlights the topic of entity linking and how location and geometric information aid the aim of finding fitting matching pairs. It can be answered by inspecting the methodological foundation and the development of the matching workflow.

Before developing the matching algorithm, an exploratory analysis was conducted to understand the characteristics, completeness, and spatial behaviour of the available glacier datasets. This exploration informed the selection of suitable metrics and highlighted the strengths and limitations of each dataset for name attribution. To investigate within Switzerland which datasets bring additional value and to better understand the distances between glaciers of interest, spatial intersections and distance calculations have been computed and mapped as histograms and discussed briefly within the methods section 4.3.

Additionally, literature showed that several metrics, most coming from object-detection techniques can be used to match geometry A to geometry B. Walter and Fritsch (1999) provided the methodological basis for this thesis in form of a five step strategy for developing a matching workflow. Step two, including a geometric buffer to reduce the search window and thus computation time is of key importance. Furthermore, several independent authors listed the intersection over union (IoU) metric, also called the Jaccard Index, as one of the most important in similarity comparisons between objects (Houde et al., 2025; Rezatofighi et al., 2019; Zheng et al., 2021). Several adaptations of the metric have been proposed as the basic IoU metric can have shortcomings in non-overlapping cases. However, for this project, only polygons are taken into consideration which actually do overlap with the input geometry, which is why the original IoU metric can be used. What is more, not only the intersection over union but also its component parts are taken into the metric framework to compare two geometries. Specifically, the area of intersection compared to the input area (called $overlap_{query}$), as well as compared to the reference area (called $overlap_{reference}$). Analysing this can help to pinpoint containment cases such as the two case studies below.

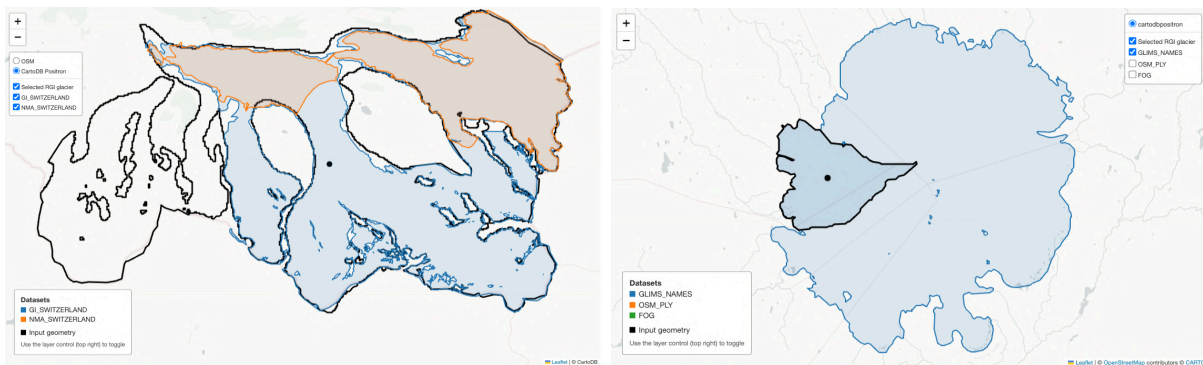


Figure 6.1: Illustrative examples of overlap/containment cases. Left: Gornergletscher in Switzerland, right: Hofsjökull in Iceland, basemaps: CartoDB.

Not only have the IoU, along with the two overlap measures been integrated into the matching framework, but also integrated accordingly into the scoring of a match. Most importantly, a so-called boost has been added for fine-tuning the confidence score. This boost can attain at most 0.5 scoring points additionally for polygons only. This focus on confinement for polygon geometries has shown to produce accurate geometric matches between input and reference and can be seen as a strong point of the metrics and scoring. Points on the other hand are handicapped by 0.5 and scored solely on their normalised distance, which can at best reach a score of one, and thus points attain an overall confidence score of 0.5 if they are within the input glacier polygon. Specifically, the scoring matrix was generated using concrete glacier cases. Two of them being Gorner Glacier in Switzerland and Hofsjökull ice cap in Iceland, visible in figure 6.1. These two situations represented the input polygon as either parent glacier or sub-glacier, respectively comprising other, smaller glacier polygons or being part of an ice dome.

Overall, the developed matching workflow performs well across a wide range of glacier configurations, from small Alpine cirques to extensive ice caps and glacier complexes. The geometric pre-filtering ensures that only spatially plausible candidates are evaluated, and the scoring function systematically favours true polygon matches over point-based candidates.

Another strong point of this thesis is also the generated database. While, the database leaves room for improvement (discussed in section 6.4), it can be seen as a product similar to a glacier gazetteer. It contains 14 polygon and 11 point sources, some national, while others are global. Which is also why the decision was taken to use a standard projection for the whole computation. This resulted in a fast process and proved to be effective when applied on the 263,479 unnamed glaciers in the RGI. That sufficient geometric confidence was maintained could be shown with the high precision for the validation dataset. Therefore, the data sources are varied - from glacier inventories, to national mapping agencies and global gazetteers, all named glacier entries were imported into a harmonised schema (*core*), while still keeping all source layers as they are in another schema (*raw*). This also means that for one glacier, there are sometimes several names and name variation, each coming from a different dataset. The outcome of a query can thus also be seen as a direct comparison between those reference dataset, not only in their data availability but also the geometric accuracy and name quality.

6.2 Research question 2: results

RQ2: To what extent can gazetteers and glacier inventories contribute additional glacier name information to the RGI, and how effectively can these heterogeneous sources be integrated?

Research question two has a clear focus on the enrichment goal of the matching workflow. It further highlights the heterogeneity of the sources, which can be observed on an inter- and intra-dataset level. RQ2 can be answered by inspecting the results of the matching workflow. For how many unnamed RGI glaciers has there been found a name, which datasets have performed best, and are there regional differences? This will be deduced by analysing the results of the application of the matching framework to the 263,479 unnamed glacier features in the Randolph Glacier Inventory. Finally, after running the matching workflow, a notable fraction of previously unnamed glaciers could be linked to at least one candidate name. Out of the 263,479 unnamed glacier features in the RGI, 46,468 (17.6%) receive a potential name match, of which 12,417 (4.7%) are very likely to be correct due to passing the geometric confidence threshold. Overall, the table 5.7 demonstrates that while the automated workflow cannot fully compensate for global data gaps, it substantially enriches the naming completeness in well-documented regions and provides a scalable approach for future improvements.

The regional analysis highlights that matching performance varies strongly with data coverage. In several regions, more than 90% of unnamed RGI glaciers had no named polygon or point in their vicinity, indicating that the limiting factor is the absence of reference data. Where reference datasets do contain names, such as Iceland, Scandinavia, the Russian Arctic, or the European Alps, the workflow performs consistently and yields spatially coherent matches. However, the regions where matching showed a higher success rate, also are the smaller regions, thus with less glacial ice. In absolute terms, Scandinavia and Central Asia (G-08 and G-13) yielded the highest count of correct name matches. A key conclusion from this analysis is that the workflow's performance is primarily bounded by reference data availability rather than algorithmic limitations.

Reference dataset contribution

An analysis on the dataset level adds conceptual value for further work with the results of this thesis. It moreover answers four distinct questions on the global and regional level.

Global contribution:

- Which datasets dominate glacier name attribution overall?
- Which datasets dominate correct name attribution (thresholded)?

Regional specialisation:

- Which datasets dominate in specific RGI regions?
- Which datasets dominate correct name attribution (thresholded) in specific RGI regions?

The global analysis of reference datasets (see table 5.8) supports a further conclusion: a small number of datasets accounts for the majority of successful name matches. Both the absolute and the correct only count show the World Glacier Inventory as the best source to find names for unnamed RGI glaciers. In second place comes OpenStreetMap, and together, the WGI and OSM cover 70% of all matches (WGI 50%, OSM 20%). The strong dominance of WGI in both the unfiltered and thresholded results highlights the continued relevance of curated glacier inventories for global name coverage. Despite being point-based and often spatially coarse, WGI appears well aligned with RGI glacier identities at a global scale, allowing it to outperform many polygon-based datasets. The consistently high performance of OSM suggests that volunteered geographic information, when represented as polygons, can provide robust spatial correspondence with RGI outlines. Its stable second rank across both unfiltered and correct matches indicates that OSM polygon geometries often coincide well with RGI glaciers, even when name quality varies and its source is often unsure (Acheson, De Sabbata & Purves, 2017). The other global gazetteer, GeoNames shows no regional matching dominance and also globally is only ranked sixth (with 1.9% of matched names, and 2.1% correctly matched names). This supports the fact that natural feature types such as glaciers are not represented well enough in global gazetteers (Acheson, De Sabbata & Purves, 2017). The shift in third place from GLIMS_NAMES (unfiltered) to FoG (thresholded) is particularly revealing. While GLIMS_NAMES produces a relatively large number of high-scoring candidates overall, a smaller proportion of these candidates satisfy the stricter correctness criteria. In contrast, FoG contributes fewer candidates in total, but a higher fraction of these pass the quality thresholds. This suggests that FoG, as a curated dataset focused on glacier monitoring, provides fewer but more reliable name–geometry associations.

Overall, these results indicate that datasets differ not only in their coverage but also in the reliability of their matches. Polygon-based datasets tend to dominate the unfiltered top ranks, whereas curated point-based inventories play an increasingly important role once correctness constraints are applied. This distinction underscores the importance of separating candidate abundance from match quality when evaluating reference datasets for automated glacier name attribution. The regional specification of some gazetteers, visible in table 5.9, undermines the assumption that no single gazetteer or inventory can cover all lacking RGI data. This is further stated by Smart et al. (2010), who for this reason developed a meta-gazetteer out of several different gazetteers. Even the worldwide glacier inventories such as World Glacier Inventory (WGI) and Fluctuations of Glaciers (FoG) showed regional variation in data coverage. Specifically, the WGI shows the most overall and correct matches in Scandinavia, the Russian Arctic, North Asia, Central Asia, and the Low Latitudes. Whereas the FoG only shows the highest count of correct matches in the northern Arctic Canada (RGI region 3). Otherwise, the FoG does not dominate in any other RGI region. However, on a global scale, the FoG shows the fifth highest ranking of glacier name contribution, and when analysing the correct matches only, even the third best.

The pronounced regional dependence observed in the dataset contributions reflects differences in data provenance, mapping traditions, and glacier monitoring practices across the world. The dominance of WGI in large and remote regions, such as Central Asia and the Arctic, suggests that historical and globally compiled inventories remain particularly valuable where contemporary, high-resolution mapping is sparse or uneven.

In contrast, regions with dense mapping coverage or strong national data infrastructures benefit more from region-specific or polygon-based datasets. The strong performance of OpenStreetMap polygons in Western Canada, Iceland, New Zealand, and parts of South Asia indicates that volunteered geographic information can provide high-quality spatial representations where community mapping efforts are active. Likewise, the emergence of national inventories, such as NMA_GREENLAND and GI_SWITZERLAND, as leading contributors of correct matches underscores the importance of locally curated datasets for reliable name attribution.

The observed discrepancies between overall and correct top-ranked datasets further emphasize that candidate frequency alone is insufficient as a measure of dataset quality. Some datasets generate many plausible candidates but fewer correct matches, while others contribute fewer candidates with higher reliability. These findings reinforce the need for an integrated, multi-source matching approach that accounts for regional context and explicitly separates candidate generation from quality-controlled validation.

What the global and regional contributions of different reference datasets used for matching show is that the enrichment of glacier names in the Randolph Glacier Inventory can only be as substantial as the underlying datasets' coverage. This is also suggested by De Sabbata and Acheson (2016), who state that the success of an application is strongly related to the content of a gazetteer used as a base.

6.3 Research question 3: validation

RQ3: How accurately can spatial entity-linking techniques based on overlap, containment, and distance metrics identify correct name-geometry pairs in a validation setting?

In this thesis, accuracy is understood as the ability of the matching workflow to assign a high geometric confidence score to the correct name–geometry pair, rather than as a classification accuracy measure commonly used in geographic information retrieval. Accordingly, RQ3 is addressed through a validation experiment that evaluates the ranking behaviour of the geometric confidence score, which integrates overlap, containment, intersection over union, and edge distance metrics.

The validation against a WGMS ground truth dataset (provided by Michael Neuwirth) resulted in a high mean average precision (MAP) of 0.915 across 253 glaciers. Precision at rank one reached 0.901, indicating that the correct name was most often ranked as the top candidate. Precision at rank five (P@5) was lower (0.623), which can largely be explained by the fact that fewer than half of the validation glaciers had five plausible candidates in the first place; the average number of candidates per query was four. Overall, these results show that when sufficient reference data are available, the matching workflow reliably identifies the most plausible spatial match.

A more detailed examination reveals systematic differences between polygon- and point-based candidates. In absolute terms, both geometry types performed similarly: 78.6% of polygon candidates (320 out of 407) and 76.6% of point candidates (468 out of 611) were classified as correct. However, their score distributions differed substantially. Correct polygon candidates exhibited clearly separable confidence scores from incorrect ones, with interquartile ranges of 1.22–1.73 for correct matches compared to 0.28–0.79 for incorrect candidates. This separation supports the use of a robust score threshold for polygon-based matching. In contrast, point-based candidates showed a much narrower and overlapping score range. Correct point matches clustered tightly around a score of 0.50, while incorrect candidates frequently occupied a similar range. As a result, geometric confidence scores derived from point data have lower discriminative power, making it considerably harder to define a reliable threshold that separates correct from incorrect matches. Based on these distributions, geometry-specific score thresholds could be derived from the validation experiment. While such thresholds perform well for polygon-based candidates, they must be applied more cautiously to point-based datasets, where ambiguity remains higher even for spatially close matches. This behaviour reflects an important aspect: the limited geometric information encoded in point representations in comparison to polygons (Goldberg et al., 2008; Hill et al., 1999; Machado et al., 2011) and confirms the conceptual assumptions underlying the scoring design. Polygons allow the distinction between, for example, a named ice stream contained within a larger RGI outline or two neighbouring glaciers sharing a boundary. Points, in contrast, remain ambiguous because they do not describe the extent of a glacier but only provide a single location. Finally, the observed score distributions empirically confirm the conceptual assumptions introduced in the scoring design.

It is important to note that the 253 WGMS glaciers used for validation represent a set of well-documented and frequently monitored glaciers that are more likely to appear in multiple reference datasets. The validation therefore reflects a favourable scenario with comparatively high data availability per glacier. Consequently, while the validation confirms the effectiveness of the matching approach under controlled conditions, similar performance cannot be assumed for the global application to all unnamed RGI glaciers, particularly in regions with sparse or inconsistent reference data. What stood out after having completed and analysed the validation of the matching algorithm, is that polygons usually succeed at being matched, while points were less simple to classify as correct or false. Confidence scoring of the latter reflects this. Polygons are more complex, shown in the scoring matrix, while points were graded on their distance alone. Therefore, future work would need to comprise more requirements and maybe other metrics for points to be ranked more differentiated. Furthermore, when applying the matching workflow on the many unnamed RGI glaciers globally, there is no ground truth available to say with confidence that a match was indeed correct. For this reason, geometrically informed thresholds were set, which retrospectively were set rather low. Therefore, the term *correct* in the RGI regions context must be read with care. This thesis successfully demonstrated how glacier geometries in the Randolph Glacier Inventory can be matched to external, openly available name sources in order to improve its information value for organisations such as the World Glacier Monitoring Service. Taken together, the methodological design (RQ1), the dataset-dependent enrichment potential (RQ2), and the validation results (RQ3) serve as an important pilot study and answer to the interdisciplinary problem of missing data. It further shows the strength and added value resulting from combining knowledge from seemingly different study fields, precisely, geographic information retrieval and glaciology. At the start of this thesis the first law of geography by Tobler was addressed, in essence meaning that close things are more related than distant things. However, is that true for glacier geometries? What can be said after generating a matching framework and applying it on the Randolph Glacier Inventory, is that the location is indeed the most important aspect of matching. Nevertheless, the distance alone from one entity to another is not sufficient to express if a match is plausible or not, but rather the geometric agreement between the two. This fact is easier and more reliably assessed for polygon than point geometries. Therefore, Tobler's first law can be said to apply to the matching approach in general, as all matched polygons have a distance of zero, and points in the best cases as well. While necessary, it is not sufficient. The geometry and geometric similarity is missing in Tobler's view, which is where the generated framework provides a combination of metrics, justifying an overall spatially informed match.

6.4 Limitations

A fundamental limitation of this thesis is data availability. The matching workflow relies exclusively on existing, openly available glacier inventories and gazetteers. While these sources serve an essential name basis, they represent only a subset of the glacier names that are known and used in practice. Many glacier toponyms appear in written sources such as scientific publications, historical documents, maps, or local narratives, but are not formalised in structured datasets (Rignot & Mouginot, 2012; Sigurðsson & Williams, 2008; Weidick, 1995). As a result, the matching approach presented in this thesis cannot recover names that are absent from digital inventories.

Closely related to this is the limited transparency of name provenance and quality in many reference datasets. Across several inventories and gazetteers, glacier names are provided without metadata on their origin, temporal validity, or naming authority. Depending on the dataset, the primary focus lies on remote sensing and geometry extraction (e.g. GLIMS) or long-term mass balance observations (FoG) with glacier naming treated as a secondary attribute. Even gazetteers, which are explicitly name-focused, often lack clear documentation on how names were collected, standardised, or updated. Both, GeoNames and OpenStreetMap further include participative or volunteered data (Acheson, De Sabbata & Purves, 2017), which further reduces the transparency of name sources, as this is not available in the freely downloadable data (Acheson, Villette et al., 2017). This reduces the interpretability of matching results and complicates any assessment of name reliability beyond geometric confidence.

Another limitation arises from the uneven regional distribution of reference data. National glacier inventories and national mapping agencies play a central role in documenting glacier names. Therefore, glacier name coverage and naming conventions can vary between countries. Furthermore, not all national mapping agencies provide their data freely. This results in strong regional differences in candidate availability and hence matching performance, a pattern that has also been observed by Walter and Fritsch (Walter & Fritsch, 1999), whose matching approach has been instrumental for this work. This limitation has also been noted for global gazetteers more generally (Acheson, De Sabbata & Purves, 2017). In some regions, glacier names follow descriptive or historical conventions (Hill, 2006; Sigurðsson & Williams, 2008), while in others numerical identifiers are common. Such heterogeneity limits the transferability of matching thresholds and confidence scores across regions.

From a geometric perspective, point-based reference data constitute an additional source of uncertainty. For several datasets, including OpenStreetMap (point dataset), GeoNames, Fluctuations of Glaciers, and the World Glacier Inventory, it is often unclear how glacier points were placed. Whether they represent centroids, glacier termini or approximate locations. This information lacks in literature and data documentations. In regions with dense glacier coverage, this ambiguity can lead to misleading distance relationships, particularly where multiple glaciers lie spatially close. While the matching algorithm accounts for this by treating point-based candidates differently from polygons, the reduced geometric information content of points constrains their discriminative power.

Lastly, while the scoring schema of the matching framework has been based on concrete glacier cases such as Gornergletscher and Hofsjökull and user requirements, such as Ethan Welty's (database manager at WGMS) its base is still rather experimental. Thus, in further work, a quantitative analysis on the information value of points and polygons in a glacier setting might help to set weights and handicap values more empirically. This can help to refine scoring and ranking outcomes.

6.5 Outlook

There are several directions which future work could take up this thesis as groundwork. One being generation of a graphical user interface (GUI), for the matching workflow, for example as a web-based application. Such an interface would allow users to upload glacier geometries in the form of points or polygons and receive ranked name suggestions without requiring programming expertise. The interface could support exploratory analysis by visualising candidate geometries, confidence scores, and alternative matches. Ultimately, the application could potentially be used in field to ascertain or check the name of the glacier of interest.

To address the lack of glacier name integration from written documents, traditional geographic information retrieval tasks could be of use. Therefore, one branch of future work could go into the direction of retrieving glacier names from historic documents, maps and glacier reports (such as Rignot and Mouginot, 2012; Sigurðsson and Williams, 2008; Weidick, 1995). What is more, while the topic of disambiguation has been touched on in the literature review, it has not been executed within this project. Thus, the available names in the Randolph Glacier Inventory were not modified or examined in detail. Nevertheless, there is potential in analysing the name quality of glacier names within the RGI and augment this information through disambiguation.

While the information discrepancy between points and polygons has been discussed thoroughly in this work, it has not been quantitatively analysed. For example, where a toponym point is located could be geometrically analysed and compared to toponym polygons, in turn substantially helping the refinement of the matching and scoring framework.

At the data level, another continuation would be the construction of a dedicated glacier gazetteer. While this thesis already integrates multiple inventories and gazetteers within a unified database, the glacier gazetteer would require clearer standards for name origin, temporal uniformity and multilingual variants. What is more, a differentiation or link between historical and contemporary names is also missing in my database. This specified glacier gazetteer would provide a more stable reference framework for automated name matching.

More broadly, the matching workflow developed here could be extended beyond glacier naming to other geographic feature types that are underrepresented in computational work, such as lakes, streams, hills and mountains (Acheson, De Sabbata & Purves, 2017). In this sense, the thesis contributes not only a glacier specific solution, but also a transferable framework for spatial entity linking in contexts where names, geometries, and data quality vary across sources.

7 Conclusion

The Randolph Glacier Inventory (RGI) is a dataset which contains over 96% unnamed glacier features. While regional centres are responsible for delivering data for integration into the inventory, the origin and completeness of name information remains vague. Furthermore, the RGI was generated in preparation for the fifth IPCC report, with the focus of global completeness of glacier coverage but with limited time, shortcomings in the data were accepted (RGI Consortium, 2017a). Nonetheless, glacier names are focal for scientific communication, such as by the World Glacier Monitoring Service (WGMS). Furthermore, missing names in geographical data also risks the loss of valuable cultural and historical knowledge.

Against this background, this thesis set out to address the question of how glacier geometries in the RGI can be spatially linked to their correct glacier names using external, openly available sources. For this, a geometry-aware matching framework was developed that integrates heterogeneous polygon and point datasets, evaluates spatial similarity using overlap and distance based metrics, and produces ranked candidate name suggestions for unnamed RGI glaciers.

The matching framework has been applied to i) a validation dataset consisting of 254 reference glaciers of the World Glacier Monitoring Service (WGMS) and ii) 263,479 unnamed glacier features in the Randolph Glacier Inventory (RGI). The first application led to a mean average precision of 0.915, along with a precision at rank one of 0.901 and a precision at five of 0.623. The second application led to the matching of 46,468 names to glacier geometries, of which a quarter (12,417) can be stated to be a match with very high geometric accuracy and thus highly probably a correct match. In a further analysis on global and regional name contribution by different sources, the World Glacier Inventory, a point dataset, is shown to contribute the most names which are not yet in the Randolph Glacier Inventory. Following, OpenStreetMap, a polygon dataset, ranks second in the name contribution count.

The results demonstrate that glacier geometries can indeed be linked to potential toponyms through an automated, spatially informed workflow. By combining geometric pre-filtering, standardised spatial metrics, and a unified scoring scheme, the workflow enables a systematic candidate generation and ranking without (much) manual intervention. The approach proved robust across a wide range of glacier sizes, two geometry types and global application. Furthermore, the results show that external glacier inventories and gazetteers can substantially enrich the RGI with additional name information, however, with strong regional dependency. Datasets such as the World Glacier Inventory (WGI) and national glacier inventories contribute a high number of names that are not yet present in the RGI, particularly in regions with a long history of glaciological observation. In contrast, in many parts of the world more than 90% of unnamed RGI glaciers had no named polygon or point within the defined search radius, indicating that the main limitation is the absence of reference data rather than shortcomings of the matching method itself. Additionally, the validation experiment using a WGMS ground-truth dataset confirmed that the proposed spatial entity-linking approach can indeed identify correct name-geometry pairs with high accuracy, where data is available. The statistics generated from the validation could be used to set geometric thresholds, which discriminate correct from false matches with a high confid-

ence. The findings empirically support the conceptual assumptions underlying the scoring design and highlight the importance of geometry type in spatial matching tasks.

This research, however, is subject to several limitations. Firstly, the main limiting factor to successful name to geometry matching was the absence or strong regional dependency of reference data. Furthermore, the data available sometimes lacks transparency of its name origin and quality. Secondly, while the scoring schema of the matching framework has been based on concrete glacier cases and user requirements, its base is rather experimental. Thus, in further work, a quantitative analysis on the information value of points and polygons in a glacier setting might lead to a more precise reflection of this information value in the scoring and ranking outcome.

Further work could entail creating a graphical user interface for easier data import and process handling as well as immediate visualisations. This would also cater to users who are not familiar with geoinformation. Another approach that could be beneficial is to employ geographic information retrieval from text such as historic documents and maps in order to include even more glacier names. Beyond the specific case of glaciers, this thesis contributes to the broader fields of geographic information retrieval by demonstrating spatial matching on natural feature geometries. By specifically addressing polygon- and point-based representations, and by focusing on a globally distributed natural feature class, the study helps bridge an existing gap between methodological research on spatial entity linking and real-world applications involving heterogeneous, incomplete, and unevenly distributed data.

Finally, the World Glacier Monitoring Service's need for an appropriate method to automatically manage the topic of missing names in the Randolph Glacier Inventory can be answered. Namely by applying the matching framework generated in this thesis and by matching them to correct glacier names using external, openly available sources. Not only has this project provided the according framework, but also compiled a vast database with all found and openly available glacier name sources. This workflow significantly reduces manual effort for finding and comparing available glacier names. The outcomes also showed the data sources with the most potential to find such glacier names, providing a tangible starting point for data conflation.

8 Data and code availability

8.1 Code and environment

The Jupyter Lab scripts for data processing, matching, and analysis are available on GitHub:

https://github.com/laurui/Masterthesis_25/tree/main/mscthesis. The exact software environment is documented in `requirements.txt` and `environment.yml` in the repository.

8.2 Database & data

A dump of the SQL database schema is provided without data in the repository: (`database/glaciers_schema.sql`).

The generated database of harmonised glacier toponyms can be shared upon request. Furthermore, a results folder is provided, containing the output CSVs from the matching application to RGI regions and the validation.

Most reference datasets are not redistributed due to licensing, all sources are cited in this document. For a national test case, users may obtain the Swiss Glacier Inventory 2016 (GLAMOS, CC BY 4.0), the setup instructions are in `data_work/README_DATA.md` and use the test geometry provided alongside `test_glaciers/test_polygon_ch_moll.fgb`.

References

- Acheson, E., De Sabbata, S., & Purves, R. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64, 309–320. <https://doi.org/10.1016/j.compenvurbsys.2017.03.007>
- Acheson, E., Villette, J., Volpi, M., & Purves, R. S. (2017). Gazetteer matching for natural features in switzerland. *Proceedings of the 11th Workshop on Geographic Information Retrieval*, 1–2. <https://doi.org/10.1145/3155902.3155913>
- Ajayakumar, J., & Ghazinour, K. (2017). I am at home: Spatial privacy concerns with social media check-ins. *Procedia Computer Science*, 113, 551–558. <https://doi.org/10.1016/j.procs.2017.08.278>
- Ardanuy, M. C., Hosseini, K., McDonough, K., Krause, A., Strien, D. v., & Nanni, F. (2020, September 22). A deep learning approach to geographical candidate selection through toponym matching. <https://doi.org/10.48550/arXiv.2009.08114>
- Bjørk, A., Lisathe Kruse & Michaelsen, P. (2015). Greenland glacier names (GGNv01) [Artwork Size: 0 Bytes Pages: 0 Bytes]. <https://doi.org/10.6084/M9.FIGSHARE.1449148.V2>
- Buckel, J., & Otto, J.-C. (2018). The austrian glacier inventory GI 4 (2015) in ArcGis (shapefile) format. <https://doi.org/10.1594/PANGAEA.887415>
- Bundesamt für Landestopografie swisstopo. (2025, May). swissNAMES3d. Retrieved May 20, 2025, from <https://www.swisstopo.admin.ch/de/landschaftsmodell-swissnames3d#Weiterf%C3%BChrende-Informationen>
- Cho, Y.-J. (2024). Weighted intersection over union (wIoU) for evaluating image segmentation. *Pattern Recognition Letters*, 185, 101–107. <https://doi.org/10.1016/j.patrec.2024.07.011>
- Cogley, J., Hock, R., Rasmussen, L., Arendt, A., Bauder, A., Braithwaite, R., Jansson, P., Kaser, G., Möller, M., Nicholson, L., & Zemp, M. (2011). *Glossary of glacier mass balance and related terms* (IACS Contribution No. 2 No. 86). UNESCO-IHP. Paris.
- De Sabbata, S., & Acheson, E. (2016). Geographies of gazetteers in great britain. *De Sabbata, Stefano; Acheson, Elise (2016). Geographies of gazetteers in Great Britain. In: 24th GIS Research UK (GISRUK 2016) conference, Greenwich, 30 March 2016 - 1 April 2016, University of Leicester*. <https://doi.org/10.5167/uzh-150584>
- Fountain, A., Glenn, B., & McNeil, C. (2022). Inventory of glaciers and perennial snowfields of the coterminous USA. <https://doi.org/10.15760/geology-data.03>
- Gärtner-Roer, I., Nussbaumer, S. U., Raup, B., Paul, F., Welty, E., Windnagel, A. K., Fetterer, F., & Zemp, M. (2022). Democratizing glacier data –maturity of worldwide datasets and future ambitions [Publisher: Frontiers Research Foundation]. *Frontiers in Climate*, 4, 4:841103. <https://doi.org/10.3389/fclim.2022.841103>
- GeoNames. (2025). *GeoNames*. Retrieved March 4, 2025, from <https://www.geonames.org/>
- GLIMS Consortium. (2005). GLIMS glacier database [NASA National Snow and Ice Data Center Distributed Active Archive Center]. <https://doi.org/10.7265/N5V98602>
- GNBC Secretariat, Canada Centre for Mapping and Earth Observation. (2011, June 1). Canadian geographical names database [Last Modified: 2025-06-13]. Retrieved October 28, 2025, from <https://natural-resources.canada.ca/maps-tools-publications/maps/geographical-names-canada/download-geographical-names-data>
- Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2008). Exploring the use of gazetteers and geocoders for the analysis and interpretation of a dynamically changing world [Num

- Pages: 24]. In *Understanding dynamics of geographic domains* (1st ed., pp. 51–70). CRC Press.
- Google Maps. (n.d.). *General map* [Google Maps]. Retrieved February 28, 2025, from https://www.google.com/maps/@39.9389436,116.0678267,121180m/data=!3m1!1e3?entry=tту&g_ep=EgoyMDI1MDIyNS4wIKXMDSoASAFQAw%3D%3D
- Graham, M., & De Sabbata, S. (2015). Mapping information wealth and poverty: The geography of gazetteers [Publisher: SAGE Publications Ltd]. *Environment and Planning A*, 47(6), 1254–1264. <https://doi.org/10.1177/0308518X15594899>
- Hastings, J. T. (2008). Automated conflation of digital gazetteer data [Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/13658810701851453>]. *International Journal of Geographical Information Science*, 22(10), 1109–1127. <https://doi.org/10.1080/13658810701851453>
- Hill, L. (2006). *Georeferencing: The geographic associations of information*. MIT Press.
- Hill, L., Frew, J., & Zheng, Q. (1999). Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1). <https://doi.org/10.1045/january99-hill>
- Hotaling, S., Hood, E., & Hamilton, T. L. (2017). Microbial ecology of mountain glacier ecosystems: Biodiversity, ecological connections and implications of a warming climate [_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.13766>]. *Environmental Microbiology*, 19(8), 2935–2948. <https://doi.org/10.1111/1462-2920.13766>
- Hou, F., Lei, W., Li, S., Xi, J., Xu, M., & Luo, J. (2021). Improved mask r-CNN with distance guided intersection over union for GPR signature detection and segmentation. *Automation in Construction*, 121, 103414. <https://doi.org/10.1016/j.autcon.2020.103414>
- Houde, K. V., Kamble, P. M., & Hegadi, R. S. (2025). A comparative analysis of IoU-based losses with YOLOv8: Fruit tree detection from UAV images. *Procedia Computer Science*, 260, 898–905. <https://doi.org/10.1016/j.procs.2025.03.272>
- Icelandic Meteorological Office. (2023). Glaciers. Retrieved October 28, 2025, from <https://en.vedur.is/hydrology/glaciers/>
- ICIMOD | RDS. (2020). Glaciers of central karakoram national park (CKNP), pakistan. Retrieved October 28, 2025, from <https://rds.icimod.org/Home/DataDetail?metadataId=9033&searchlist=True>
- Institute of Geography of the Russian Academy of Sciences. (2021, April 23). Glaciers of russia | каталог ледников россии. Retrieved September 18, 2025, from <https://www.glacru.ru/>
- Kilinc, D. (2016). An accurate toponym-matching measure based on approximate string matching [Publisher: SAGE Publications Ltd]. *Journal of Information Science*, 42(2), 138–149. <https://doi.org/10.1177/0165551515590097>
- Kohler, J., Lith, A., Moholdt, G., & Kohler, J. (2021). Svalbard glacier inventory based on sentinel-2 imagery from summer 2020. <https://doi.org/10.21334/NPOLAR.2021.1B8631BF>
- Kounadi, O., Lampoltshammer, T. J., Leitner, M., & Heistracher, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services [Publisher: Taylor & Francis]. *Cartography and Geographic Information Science*, 40(2), 140–153. <https://doi.org/10.1080/15230406.2013.777138>
- Land Information New Zealand. (2015, August 19). NZ place names - polygons (NZGB) | LINZ data service. Retrieved October 28, 2025, from <https://data.linz.govt.nz/layer/52424-nz-place-names-polygons-nzgb/>

- Li, D., Cova, T. J., & Dennison, P. E. (2017). Using reverse geocoding to identify prominent wildfire evacuation trigger points. *Applied Geography*, 87, 14–27. <https://doi.org/10.1016/j.apgeog.2017.05.008>
- Linsbauer, A., Huss, M., Hodel, E., Bauder, A., Fischer, M., Weidmann, Y., Bärtschi, H., & Schmassmann, E. (2021). The new swiss glacier inventory SGI2016: From a topographical to a glaciological dataset [Publisher: Frontiers]. *Frontiers in Earth Science*, 9. <https://doi.org/10.3389/feart.2021.704189>
- Liu, S., Guo, W., & Xu, J. (2012). The second glacier inventory dataset of china [National Tibetan Plateau / Third Pole Environment Data Center.]. <https://doi.org/https://doi.org/10.3972/glacier.001.2013.db>
- Ma, M., Zhong, Z., Guo, N., Jing, N., & Xiong, W. (2016). An efficient reverse geocoding method based on global subdivision model [ISSN: 2161-0258]. *2016 24th International Conference on Geoinformatics*, 1–9. <https://doi.org/10.1109/GEOINFORMATICS.2016.7578966>
- Machado, I. M. R., De Alencar, R. O., Campos, R. D. O., & Davis, C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17(4), 267–279. <https://doi.org/10.1007/s13173-011-0044-4>
- McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2), 125–137. <https://doi.org/10.1080/15230406.2014.880327>
- Ministry of Rural Affairs and Infrastructure: Lantmateriet. (2025). Hydrografi Nedladdning. Retrieved October 28, 2025, from <https://www.lantmateriet.se/sv/geodata/vara-produkter/produktlista/hydrografi-nedladdning/>
- National Geospatial-intelligence agency. (2025). Geographic names server: Hydrographic. Retrieved October 28, 2025, from <https://geonames.nga.mil/geonames/GNSData/>
- National Snow and Ice Data Center. (n.d.). *GLIMS: Global land ice measurements from space*. Retrieved March 12, 2025, from <https://www.glims.org/>
- National Snow and Ice Data Center. (2023). *World glacier inventory*. Retrieved January 20, 2025, from https://nsidc.org/data/glacier_inventory/index.html
- National Snow and Ice Data Center. (2024, June 29). *GLIMS glacier database: Summary statistics*. Retrieved April 2, 2025, from http://glims.colorado.edu/glacierdata/db_summary_stats.php
- National Snow and Ice Data Center. (2025). *Glacier quick facts* [National snow and ice data center]. Retrieved January 22, 2025, from <https://nsidc.org/learn/parts-cryosphere/glaciers/glacier-quick-facts>
- Natural Earth. (2009). Admin 0 –countries. vector at 1:10m scale. Retrieved December 3, 2025, from <https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/>
- Norwegian Polar Institute. (2011). Registry of place names in norwegian polar areas. <https://doi.org/10.21334/NPOLAR.2011.A2813EB6>
- NSIDC National Snow & Ice Data Center. (2012). World glacier inventory (WGI). Retrieved October 28, 2025, from https://nsidc.org/data/glacier_inventory/
- OpenStreetMap contributors. (2025). *OpenStreetMap* [OpenStreetMap]. Retrieved January 21, 2025, from <https://www.openstreetmap.org/>
- Paul, F., Barry, R., Cogley, J., Frey, H., Haeberli, W., Ohmura, A., Ommanney, C., Raup, B., Rivera, A., & Zemp, M. (2009). Recommendations for the compilation of glacier in-

- ventory data from digital sources. *Annals of Glaciology*, 50(53), 119–126. <https://doi.org/10.3189/172756410790595778>
- Penko Seidl, N. (2019). Engraved in the landscape: The study of spatial and temporal characteristics of field names in the changing landscape [Publisher: Routledge _eprint: <https://doi.org/10.1080/00277738.2017.1415539>]. *Names*, 67(1), 16–29. <https://doi.org/10.1080/00277738.2017.1415539>
- Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., Hagen, J.-O., Hock, R., Kaser, G., Kienholz, C., Miles, E. S., Moholdt, G., Mölg, N., Paul, F., Radić, V., Rastner, P., Raup, B. H., Rich, J., Sharp, M. J., & Consortium, T. R. (2014). The randolph glacier inventory: A globally complete inventory of glaciers. *Journal of Glaciology*, 60(221), 537–552. <https://doi.org/10.3189/2014JoG13J176>
- Purves, R., Clough, P., Jones, C., Hall, M., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, 12(2), 164–318. <https://doi.org/10.1561/15000000034>
- Qiu, Q., Zheng, S., Tian, M., Li, J., Ma, K., Tao, L., & Xie, Z. (2024). A deep neural network model for chinese toponym matching with geographic pre-training model [Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/17538947.2024.2353111>]. *International Journal of Digital Earth*, 17(1), 2353111. <https://doi.org/10.1080/17538947.2024.2353111>
- Racoviteanu, A. E., Paul, F., Raup, B., Khalsa, S. J. S., & Armstrong, R. (2009). Challenges and recommendations in mapping of glacier parameters from space: Results of the 2008 global land ice measurements from space (GLIMS) workshop, boulder, colorado, USA. *Annals of Glaciology*, 50(53), 53–69. <https://doi.org/10.3189/172756410790595804>
- Raup, B., Racoviteanu, A., Khalsa, S. J. S., Helm, C., Armstrong, R., & Arnaud, Y. (2007). The GLIMS geospatial glacier database: A new tool for studying glacier change. *Global and Planetary Change*, 56(1), 101–110. <https://doi.org/10.1016/j.gloplacha.2006.07.018>
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression [ISSN: 2575-7075]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666. <https://doi.org/10.1109/CVPR.2019.00075>
- RGI Consortium. (2017a). *Randolph glacier inventory (RGI) - a dataset of global glacier outlines: Version 6.0* (Technical Report). Global Land Ice Measurements from Space. Boulder, Colorado, USA. <https://doi.org/10.7265/N5-RGI-60>
- RGI Consortium. (2017b). *Randolph glacier inventory 6.0*. <https://doi.org/10.7265/N5-RGI-60>
- RGI Consortium. (2023a). *Global statistics –the randolph glacier inventory version 7*. Retrieved April 2, 2025, from https://www.glims.org/rgi_user_guide/06_dataset_summary.html
- RGI Consortium. (2023b). *Randolph glacier inventory - a dataset of global glacier outlines, version 7*. <https://doi.org/10.5067/F6JMOVY5NAVZ>
- Rignot, E., & Mouginot, J. (2012). Ice flow in greenland for the international polar year 2008–2009 [_eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL051634>]. *Geophysical Research Letters*, 39(11). <https://doi.org/10.1029/2012GL051634>
- Roth, J. (2015). Generating meaningful location descriptions. *2015 15th International Conference on Innovations for Community Services (I4CS)*, 1–8. <https://doi.org/10.1109/I4CS.2015.7294482>
- Ruiz-Lendínez, J. J., Ureña-Cámara, M. A., & Ariza-López, F. J. (2017). A polygon and point-based approach to matching geospatial features [Publisher: Multidisciplinary Digital Publishing Institute]. *ISPRS International Journal of Geo-Information*, 6(12), 399. <https://doi.org/10.3390/ijgi6120399>

- Salvatore, M., Zanoner, T., Baroni, C., Carton, A., Banchieri, F., Viani, C., Giardino, M., & Perotti, L. (2015). The state of italian glaciers: A snapshot of the 2006-2007 hydrological period. <https://repo2.igg.cnr.it/ghiacciaiCGI/data/datagralp.json>
- Schöner, W., Auer, I., Böhm, R., Hammer, N., & Wiesinger, T. (1997). Retreat of wurtenkees, european east alps, since 1850. *Annals of Glaciology*, 24, 102–105. <https://doi.org/10.3189/S0260305500012003>
- Scientific Committee on Antarctic Research. (2014). SCAR composite gazetteer of antarctica. Retrieved October 28, 2025, from <https://data.aad.gov.au/aadc/gaz/scar/>
- SIG INAIGEM. (2023, November 22). Inventario nacional de glaciares y lagunas de origen glaciar actualizado al año 2020. Retrieved October 28, 2025, from https://services6.arcgis.com/c9GFTG11debu0wsH/arcgis/rest/services/Inventario_Nacional_de_Glaciares_y_Lagunas_de_Origen_Glaciar_Actualizado_al_a%C3%B1o_2020/FeatureServer
- Sigurðsson, O., & Williams, R. S. J. (2008). *Geographic names of iceland's glaciers: Historic and modern* (Professional Paper No. 1746) (Series: Professional Paper).
- Smart, P. D., Jones, C. B., & Twaroch, F. A. (2010). Multi-source toponym data integration and mediation for a meta-gazetteer service [Series Title: Lecture Notes in Computer Science]. In S. I. Fabrikant, T. Reichenbacher, M. Van Kreveld & C. Schlieder (Eds.), *Geographic information science* (pp. 234–248, Vol. 6292). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15300-6_17
- The swiss glaciers 2007/08 and 2008/09* (Glaciological Report No. 129/130). (2014) (ISSN 1424-2222). Laboratory of Hydraulics, Hydrology and Glaciology (VAW), ETH Zurich. Retrieved December 4, 2025, from https://doi.glamos.ch/pubs/glrep/glrep_129-130.pdf
- The MathWorks Inc. (n.d.). *What is unit testing?* [MathWorks MATLAB]. Retrieved December 9, 2025, from <https://ch.mathworks.com/discovery/unit-testing.html>
- The PostgreSQL Global Development Group. (2025). 39.3. *materialized views* [PostgreSQL documentation]. Retrieved December 18, 2025, from <https://www.postgresql.org/docs/18/rules-materializedviews.html>
- Unxos GmbH. (2025). GeoNames. Retrieved October 28, 2025, from <https://download.geonames.org/export/dump/>
- Van Gysel, C., & De Rijke, M. (2018). Pytrec_eval: An extremely fast python interface to trec_eval. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 873–876. <https://doi.org/10.1145/3209978.3210065>
- Walter, V., & Fritsch, D. (1999). Matching spatial data sets: A statistical approach. *International Journal of Geographical Information Science*, 13(5), 445–473. <https://doi.org/10.1080/136588199241157>
- Waters, N. (2017). Tobler's first law of geography [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118786352.wbieg1011>]. In *International encyclopedia of geography* (pp. 1–13). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118786352.wbieg1011>
- Weidick, A. (1995). *Satellite image atlas of glaciers of the world - greenland*.
- Welty, E., & WGMS. (2025). GLIMS names.
- WGMS. (2012). Fluctuations of glaciers database. <https://doi.org/10.5904/WGMS-FOG-2012-11>
- WGMS. (2021). Global glacier change bulletin no. 4 (2018-2019) (M. Zemp, S. U. Nussbaumer, I. Gärtner-Roer, J. Bannwart, F. Paul & M. Hoelzle, Eds.). 4, 278. <https://doi.org/10.5167/uzh-209777>
- WGMS. (2022a). Fluctuations of glaciers database. <https://doi.org/10.5904/wgms-fog-2022-09>

- WGMS. (2022b, December). *About gtn-g –global terrestrial network for glaciers*. Retrieved March 12, 2025, from https://www.gtn-g.ch/about_gtng/
- WGMS. (2025a). *Fluctuations of glaciers browser*. Retrieved December 15, 2025, from <https://experience.arcgis.com/experience/836c66d14c8b410f940355056ddb1bf8>
- WGMS. (2025b, February 20). *WGMS –world glacier monitoring service*. Retrieved March 13, 2025, from <https://wgms.ch/>
- World Glacier Monitoring Service (WGMS). (2025). *Fluctuations of glaciers (FoG) database* [Artwork Size: 857 MB Pages: 857 MB]. <https://doi.org/10.5904/WGMS-FOG-2025-02B>
- Wu, Y., & Winter, S. (2009). Inferring relevant gazetteer instances to a placename. *Proceedings of the 10th International Conference on GeoComputation*.
- Zemp, M. (2011). *The monitoring of glaciers at local, mountain, and global scale. parts a and b* [Habilitation]. University of Zurich. <https://doi.org/10.5167/uzh-53505>
- Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2019, November 19). Distance-IoU loss: Faster and better learning for bounding box regression. <https://doi.org/10.48550/arXiv.1911.08287>
- Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., & Zuo, W. (2021, July 5). Enhancing geometric factors in model learning and inference for object detection and instance segmentation. <https://doi.org/10.48550/arXiv.2005.03572>

9 Appendix

9.1 Tables of unit test

Table 9.1: Location matching of RGI / swissNAMES3D polygon input

Metric	min	max	mean	median	P@1
AI	0.00 km ²	4.46 km ²	0.39 km ²	0.06 km ²	0.60
RI	0.00	0.95	0.46	0.58	0.60
IoU	0.00	0.82	0.29	0.18	0.57
CD	37.22 m	1974.31 m	497.65 m	404.97 m	0.57
ED	0.00 m	389.15 m	52.55 m	0.00 m	0.53

Table 9.2: Location matching of RGI / SGI2016 polygon input

Metric	min	max	mean	median	P@1
AI	0.01 km ²	4.83 km ²	0.46 km ²	0.08 km ²	0.90
RI	0.09	0.95	0.68	0.76	0.90
IoU	0.01	0.82	0.53	0.58	0.90
CD	8.67 m	1169.2 m	151.06 m	53.24 m	0.87
ED	0.00 m	0.00 m	0.00 m	0.00 m	0.80

Table 9.3: Location matching of RGI / OSM polygon input

Metric	min	max	mean	median	P@1
AI	0.00 km ²	4.40 km ²	0.32 km ²	0.00 km ²	0.30
RI	0.00	0.91	0.21	0.00	0.30
IoU	0.00	0.76	0.18	0.00	0.30
CD	20.42 m	5226.93 m	1123.97 m	993.52 m	0.27
ED	0.00 m	2065.5 m	270.47 m	60.00 m	0.33

For the Fluctuations of Glaciers (FoG) and GeoNames datasets, only the centroid and edge distance could be evaluated in the precision test, as they are both point layers.

Table 9.5: Location matching of RGI / FoG point input

Metric	min	max	mean	median	P@1
CD	19.91 m	1074.6 m	209.24 m	108.47 m	0.90
ED	0.00 m	416.21 m	51.75 m	0.00 m	0.90

Table 9.6: Location matching of RGI / GeoNames point input

Metric	min	max	mean	median	P@1
CD	25.14 m	1849.71 m	517.41 m	346.77 m	0.57
ED	0.00 m	1612.17 m	319.85 m	26.31 m	0.57

Table 9.4: Location matching of RGI / GLIMS polygon input

Metric	min	max	mean	median	P@1
AI	0.02 km ²	5.67 km ²	0.58 km ²	0.15 km ²	1.00
RI	1.00	1.00	1.00	1.00	1.00
IoU	0.13	0.92	0.46	0.47	1.00
CD	12.51 m	706.6 m	220.43 m	204.13 m	1.00
ED	0.00 m	0.00 m	0.00 m	0.00 m	1.00

9.2 Results

9.2.1 RGI region results

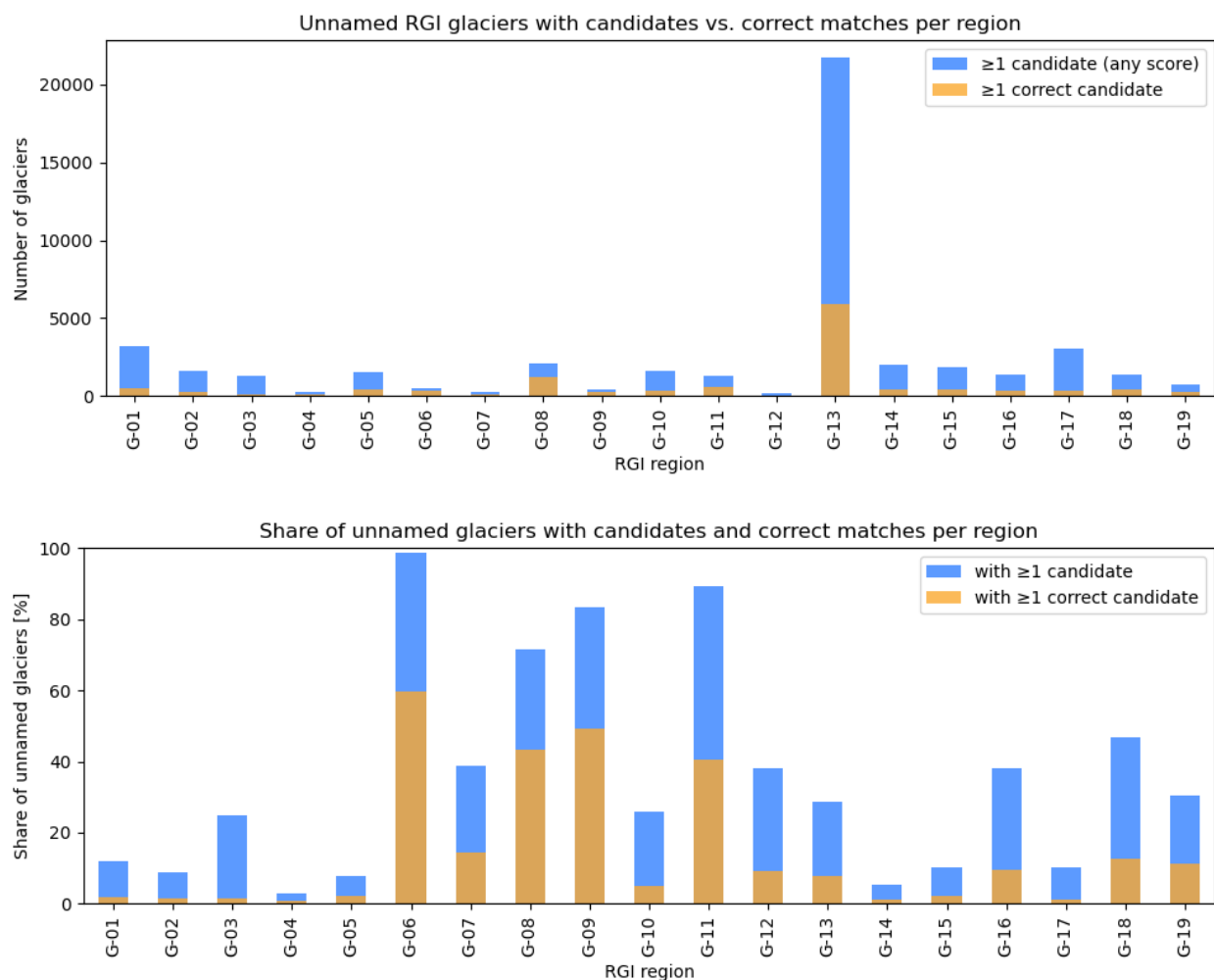


Figure 9.1: Number and percent of unnamed RGI glaciers with at least one candidate after matching and one correct candidate shown per region and split by geometric confidence threshold (any matched vs above threshold). Counts are equivalent to column *Matched*: 46,468 and *Correct*: 12,417 in table 5.7.

9.2.2 Global contribution of reference datasets

This is the full table as mentioned in [5.3](#).

Table 9.7: Datasets contributing the most top-ranked matches globally (not thresholded)

Rank	Dataset	Matched	% of queries
1	WGI	23653	50.902
2	OSM_PLY	9258	19.923
3	GLIMS_NAMES	6820	14.677
4	GI_CHINA	2145	4.616
5	FOG	1596	3.435
6	GEONAMES	899	1.935
7	GNS_WORLD	409	0.880
8	GI_SWITZERLAND	255	0.549
9	OSM_PNT	218	0.469
10	NMA_GREENLAND	213	0.458
11	NMA_CANADA	164	0.353
12	GI_RUSSIA	137	0.295
13	GI_ICELAND	120	0.258
14	SCAR	98	0.211
15	GI_ITALY	94	0.202
16	NMA_NEWZEALAND	91	0.196
17	NMA_SVALBARD	81	0.174
18	GI_SVALBARD	61	0.131
19	GI_PAKISTAN_KARAKORAM	59	0.127
20	GI_AUSTRIA	45	0.097
21	NMA_SWITZERLAND	42	0.090
22	GI_SWEDEN	5	0.011
23	GI_USA	3	0.006
24	GI_PERU	2	0.004

Table 9.8: Datasets contributing the most correct matches globally. Correct means match passed thresholds of 0.875 for polygons, 0.49 for points.

Rank	Dataset	Matched	$\frac{Matches}{Queries}$ [%]
1	WGI	6232	52.029
2	OSM_PLY	2510	20.955
3	FOG	875	7.305
4	GLIMS_NAMES	724	6.044
5	GI_CHINA	263	2.196
6	GEONAMES	255	2.129
7	NMA_GREENLAND	174	1.453
8	GNS_WORLD	168	1.403
9	GI_SWITZERLAND	166	1.386
10	OSM_PNT	137	1.144
11	GI_ICELAND	108	0.902
12	SCAR	72	0.601
13	NMA_SVALBARD	66	0.551
14	NMA_CANADA	64	0.534
15	GI_ITALY	42	0.351
16	NMA_NEWZEALAND	37	0.309
17	GI_RUSSIA	35	0.292
18	GI_AUSTRIA	21	0.175
19	NMA_SWITZERLAND	19	0.159
20	GI_SVALBARD	7	0.058
21	GI_SWEDEN	2	0.017
22	GI_PAKISTAN_KARAKORAM	1	0.008

9.2.3 National contribution of reference datasets

Table 9.9: Country-level leading dataset by match count (not thresholded).

Country	Dataset	Matched	Total	Share in country
Afghanistan	WGI	30	44	0.6818
Argentina	GLIMS_NAMES	564	922	0.6117
Antarctica	GLIMS_NAMES	199	385	0.5169
French Southern and Antarctic Lands	OSM_PLY	87	218	0.3991
Austria	GLIMS_NAMES	127	264	0.4811
Bolivia	OSM_PLY	202	221	0.9140
Bhutan	OSM_PLY	70	148	0.4730
Canada	OSM_PLY	2242	3286	0.6823
Switzerland	GI_SWITZERLAND	246	644	0.3820
Chile	GLIMS_NAMES	1370	2096	0.6536
China	WGI	5289	8110	0.6522
Colombia	OSM_PLY	41	51	0.8039
Germany	GLIMS_NAMES	1	2	0.5000
Ecuador	OSM_PLY	23	28	0.8214
France	GLIMS_NAMES	44	73	0.6027
Georgia	WGI	23	53	0.4340
Greenland	GLIMS_NAMES	510	1515	0.3366
Heard Island and McDonald Islands	OSM_PLY	24	30	0.8000
India	OSM_PLY	508	995	0.5106
Iran	OSM_PLY	10	12	0.8333
Iceland	OSM_PLY	247	531	0.4652
Italy	GI_ITALY	83	281	0.2954
Siachen Glacier	GLIMS_NAMES	38	41	0.9268
Kazakhstan	WGI	1062	1410	0.7532
Kyrgyzstan	WGI	6199	7133	0.8691
Mongolia	WGI	11	16	0.6875
Norway	WGI	1626	2179	0.7462
Nepal	GLIMS_NAMES	184	339	0.5428
New Zealand	OSM_PLY	695	1409	0.4933
Pakistan	GLIMS_NAMES	410	983	0.4171
Peru	WGI	806	1084	0.7435
Russia	WGI	1364	1814	0.7519
South Georgia and the Islands	OSM_PLY	49	80	0.6125
Southern Patagonian Ice Field	GLIMS_NAMES	31	34	0.9118
Sweden	WGI	140	172	0.8140
Tajikistan	WGI	5547	6412	0.8651
United States of America	GLIMS_NAMES	1439	3069	0.4689
Uzbekistan	WGI	211	226	0.9336
Venezuela	OSM_PLY	2	3	0.6667

Table 9.10: Country-level leading dataset by *correct* matches (thresholded).

Country	Dataset	Matched (correct)	Total (correct)	Share in country (correct)
Afghanistan	WGI	5	9	0.5556
Argentina	GLIMS_NAMES	56	129	0.4341
Antarctica	GLIMS_NAMES	75	188	0.3989
French Southern and Antarctic Lands	WGI	15	38	0.3947
Austria	GLIMS_NAMES	58	127	0.4567
Bolivia	OSM_PLY	9	15	0.6000
Bhutan	OSM_PLY	20	38	0.5263
Canada	OSM_PLY	134	479	0.2797
Switzerland	GI_SWITZERLAND	159	271	0.5867
Chile	GLIMS_NAMES	115	222	0.5180
China	WGI	1248	1658	0.7527
Colombia	OSM_PLY	23	24	0.9583
Germany	GLIMS_NAMES	1	2	0.5000
Ecuador	OSM_PLY	5	9	0.5556
France	GLIMS_NAMES	28	34	0.8235
Georgia	GLIMS_NAMES	5	8	0.6250
Greenland	NMA_GREENLAND	172	416	0.4135
Heard Island and McDonald Islands	OSM_PLY	17	19	0.8947
India	OSM_PLY	216	312	0.6923
Iran	OSM_PLY	8	9	0.8889
Iceland	OSM_PLY	154	277	0.5560
Italy	GI_ITALY	37	81	0.4568
Siachen Glacier	OSM_PLY	1	1	1.0000
Kazakhstan	OSM_PLY	239	480	0.4979
Kyrgyzstan	WGI	1571	2239	0.7017
Mongolia	GEONAMES	2	3	0.6667
Norway	WGI	953	1172	0.8131
Nepal	OSM_PLY	34	71	0.4789
New Zealand	OSM_PLY	210	334	0.6287
Pakistan	OSM_PLY	82	201	0.4080
Peru	WGI	230	267	0.8614
Russia	WGI	401	509	0.7878
South Georgia and the Islands	OSM_PLY	12	14	0.8571
Southern Patagonian Ice Field	WGI	1	1	1.0000
Sweden	WGI	86	112	0.7679
Tajikistan	WGI	1143	1704	0.6708
United States of America	GLIMS_NAMES	194	405	0.4790
Uzbekistan	WGI	41	46	0.8913
Venezuela	GLIMS_NAMES	1	2	0.5000

Declaration of Originality

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis. I used artificial intelligence as follows: ChatGPT was used with refining Python code and troubleshooting LaTeX documents. DeepL was used to ensure a fluent reading style and correct translations. The tools mentioned were used only to provide technical and linguistic support and did not influence the scientific and substantive core of my work.

Laura Bozzi, Winterthur, 19.12.2025

A handwritten signature in black ink, consisting of stylized initials 'LB' followed by a horizontal line.