Icy Displacement: Integrating Glacier Fluctuation and Species Occurrence Data to Study Distribution Shifts Near Retreating Glaciers

Christian Marino Latuff (s3126153)

Supervisors:
Rita Pucci & Jens van Bijsterveld

Bachelor Thesis– Informatica

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl July 1, 2026

Abstract

Glacier retreat is reshaping mountain ecosystems by exposing new terrain, altering local environmental conditions, and potentially influencing the spatial and elevational distribution of nearby species. This thesis investigates to what extent glacier fluctuation records and GBIF species occurrence data can be integrated into a reproducible pipeline for identifying spatial and elevational patterns in plant species distributions near retreating glaciers. A geospatial analysis pipeline was developed using World Glacier Monitoring Service glacier records, GBIF occurrence data, and Copernicus GLO-30 digital elevation data. The pipeline performs occurrence cleaning, spatial filtering around glacier reference coordinates, DEM-based elevation completion, yearly metric construction, correlation analysis, and diagnostic checks for temporal dependence. The analysis centered on the Maladeta glacier in the Spanish Pyrenees and examined two biological cases: Orchidaceae as a broad family-level exploratory case, and *Salix pyrenaica* as a single-species case. The Orchidaceae analysis showed strong raw associations between centroid distance and glacier state variables, but these relationships were strongly structured by year and therefore interpreted as temporal co-movement rather than direct causal evidence. The *Salix pyrenaica* analysis produced more ecologically interpretable elevation-based patterns, with mean occurrence elevation showing negative correlations with glacier area and glacier length. However, time-controlled regressions again showed that the available data was not sufficient to isolate glacier variables as independent drivers. A third regional stress test in the Hengduan Mountains showed that the pipeline can be conceptually extended to multi-glacier settings, but also that data availability in FoG and GBIF can prevent usable analysis even in ecologically relevant regions. Overall, the thesis demonstrates that glacier and biodiversity datasets can be integrated into a reproducible exploratory framework, while also showing that stronger causal or predictive claims require better temporal coverage, systematic species observations, and more precise glacier geometry data.

Contents

1 Introduction

The rise in temperature levels, as a consequence of human-induced climate change, has resulted in the extensive retreat of glacier levels world wide. Of greatest importance to ecologists of our time has been the study of climate change's effect on the distribution, migration, and extinction levels of species on Earth. With 70 % percent of the earth's entire freshwater being stored in ice caps and glaciers, these alarming values have drastic results to the future of all the fresh-water dependent biodiversity, which includes the largest part of terrestrial animals (and consequently, us humans too). Thus, glacier retreat has been extensively documented as a key factor in the state of polar and subpolar ecosystems, as the resulting sea-level rise and remaining barren land has great and far-reaching species-dependent consequences. Nonetheless, even if a lot of the resulting ramifications from ice-loss surmised from the clear relation cryophile biota have to the surrounding ice have been verified, researchers still lack a clear understanding of how glacier retreat has directly affected species distribution. Analysis of the effects has been spatially and conditionally restricted and specific to a certain case of interest. The complexity of ecosystems as multivariate, resilient, and widespread systems containing both biotic and abiotic counterparts makes large-scale predictive analysis a complicated issue. With glaciers the complexity increases as glacier falls can have both land-loss, through increasing sea levels, and land-gaining, through appearance of moraines, effects.

The driving idea behind the study is the realization that, even if temperature rise is predictably a cause for changes in species' biogeography worldwide, for periglacial species (taxa situated on the edges of past or present glaciers) a more fundamental explanation, and consequently a stronger predictor, for dispersal might be glacier related metrics. To give an example relevant to the use cases that will be presented later on, plants around glaciated alps.

This because whatever land is being colonized by, to give an example, altitude climbing plants is previously glaciated land now barren from glacier retreat as a consequence of the decreasing mass-balance. Despite temperature rise being the primary general cause, which has caused hotter years which in turn made negative glacier mass balance years ordinary and so glacier retreat a certainty, I hypothesize that it is the final more geological step of glacier retreat (and its corresponding measures) that correctly determines the overall movement. To further justify the belief, numerous of the papers cited in the scientific background section to come mention the necessity of considering the possible implications of glacier related indicators as driving factors for dispersal.

Nonetheless, since no study has researched this eco-glaciological connection in a more data-based manner, this thesis project will be exploratory than causational in the sense that its main objective will be the creation of a reproducible pipeline to study the relation between glacier fluctuation and species occurrences publicly available data. Rather than claiming more causal relationships between the two variables. Correlational values and other relevant data acquired in the exploration of the data is thus both used as a scientifically informative result but more so as a validating mechanism for the implemented pipeline.

Why does this matter? Increase in predictive abilities is the natural objective of any modelling methodology and so the idea vindicates itself as being useful. Considering the drastic consequence biodiversity loss is predicted to incur upon us and the remaining species post the extinction sweeps we are to see from climate change, greater precision in our understanding of the effects is beneficial. Furthermore, the more we understand, the better (theoretically) our conservation efforts will be when trying to save species as no system we seek to fix benefits from ignorance.

1.1 Thesis overview

This bachelor thesis is conducted at the Leiden Institute of Advanced Computer Science (LIACS) under the supervision of Rita Pucci as the main supervisor. The thesis investigates whether glacier fluctuation records can be combined with species occurrence data to identify spatial and elevational patterns in plant species distributions near glaciers. The remainder of the thesis is structured as follows. Sections 1 and 3 introduce the necessary background on glacier retreat, species distribution shifts, the datasets used in this study, and the scientific foundation of the research. Section 4 briefly summarizes related previous analyses on snow associated parameters on implemented models. Section 5 describes the World Glacier Monitoring Service's (WGMS) Fluctuations of Glaciers database, GBIF species occurrence data, and DEM-derived elevation data. Section 6 explains the proposed geospatial analysis pipeline, including data cleaning, spatial filtering, distance calculation, yearly metric construction, and correlation analysis. Section 7 presents the case-study experiments on the Maladeta glacier and selected plant taxa, together with the third stress test, plus the resulting statistical and spatial analyses where applicable. Throughout it I discuss the interpretation, robustness, and limitations of the results gotten. Finally, Section 8 gives an overview of all the thesis report, and Section 9 outlines possible directions for future work.

2 Research Question

> To what extent can satellite-derived glacier fluctuation records and GBIF species occurrence data be integrated into a reproducible pipeline for identifying spatial and elevational patterns in plant species distributions near retreating glaciers?

This study will address the following sub-questions:

- Can a digital elevation model (DEM) and coordinate points be used to reduce loss of occurrence records missing elevation values?
- Which glacier fluctuation variables can be extracted and transformed into yearly metrics suitable for comparison with species occurrence data?
- Which species-side metrics, such as occurrence count, centroid distance, and elevation, show interpretable relationships with glacier fluctuation metrics?
- How sensitive are the observed relationships to spatial buffer size, sampling effort, and uncertainty?
- Are elevation values derived from a digital elevation models (DEM) good to reduce the loss of occurrence records?

3 Scientific Background

Ecological effects of glacier adjacent related causes are varied and non linear. Few effects have been able to be generalized as small geological or ecological changes can have great variation in the impact whichever effect is measured, both in magnitutde and direction. As a good example of this, one study has found that there are "significant differences between predictions from different models, with predicted changes in range size by 2030 differing in both magnitude and direction (e.g. from 92% loss to 322% gain) [15]. In this section I'll go through some scientific results that prove the complexity of the problem, yet still clarify some greater overarching patterns in ecological distributional shifts.

A meta-analysis of the current findings on glacial retreat effects on fjord habiting marine species found that glacial retreat causes the population loss of the previously habiting species. The uniquely specialized species to glacial environments receive the hardest hit, with the winning colonizing species being generalist taxa species from downstream who's resilience towards environmental conditions allows them to occupy the barren ecological niche that is the previously glaciated, eroded land [4]. Relative to plant colonization the consequences are uneven, with plant diversity initially increasing after glacier retreat but ultimately reaching a substantial decline with an extinction of a fifth of the plant species. Global warming causes a shift from facilitative, positive species interactions to competition, which results in the faulty perseverance of colonizing species [11]. For the bottom dwelling, benthic zoology, colonization and distribution shifts are highly dependent on feeding type, as filter and suspension feeders just require the presence of nutritionally rich water for consumption. Nonetheless, the intense hydrological dynamics resulting from glacier fall diminishes water food supply and so reduces speed and possibility of colonization. This further constrains the second settlement of benthic predators, scavengers, and deposit feeders since the presence of an already partially established biodiversity is necessary for survival [12].

The resulting dynamics of ecological communities in glacial retreats are defined primarily by the taxa of pioneering species as whatever food chain systems that prevails in the region will be founded on them. Aerial species are faster in colonizing compared to their land dispersing counterparts and the most common colonizers are bacteria and other abiotic feeding microscopic animalea, beetles (specially flying beetles), and spiders [7]. Large animals might forage and hunt in the area briefly, but actual settlement is predicted to be at the scale of thousands of years. The rate of colonization is not linear through time, with the rapid colonization of pioneer species taking place during the first 20-50 years post glacial fall. This long pioneering timespan is what results in the fragility of these ecosystems, as other environmental events such as the further retreat of adjacent glaciers leads to disturbances that can prove fatal to a large part of the population [7]. For alpine biodiversity in the Neotropics, the recession of glaciers, although creating new lakes with promising possibilities for biodiversity, also results in the eventual loss of surrounding wetland ecosystems and eventual loss of alpine biota [5]. At a contrast to low-elevation plant dynamics, biodiversity loss due to glacier retreat was found to be at a lower incidence than predicted. Nevertheless, the survival of species was found to be drastically dependent on the presence of habitable microrefugia at lower altitudes for high-altitude species, which is an inconsistent and non-dependable reality from alpine glacial falls [8]. Finally, the abiotic composition of the land and resulting regions is a globally significant defining characteristic of the established communities as soil composition, the most notable geological attribute, determines the pioneering bacterial biota which further defines the rest of the trophic levels [1].

4 Related Work

The main driving force for the selection of this topic is the lack of related methodologies doing the sort of work explored in this thesis. There is lots of field research investigating the snow-loss to species relation, but none apply the data-science approach I take, and so I mention them in the scientific background section above (Section 3). Therefore I'll give a short overview of the brief data-analysis research done on the effect of snow as a model parameter. A 2018 paper investigating the implementation of snow information on snow-adjacent Species Distribution Models (SDM's) finds that "it is only rarely used as a predictor", this being a consequence of "methodological difficulties limiting both the quality and the quantity of available snow information in SDM's" [13]. The paper finds that out of the 273 species gathered for the study from 1200 previous studies, "Snow was the most influential predictor for 36% of the species and had, on average, the second highest variable importance scores of all environmental variables considered" [13], justifying inclusion. Snow and ice cover are amongst the most important characteristics of cold regions [14], with high winds and regional isolation being equally important in alpine terrains, making their omission as modelling parameters a clear problem.

5 Data

5.1 Glacier Data

The main glacier database used in the creation of the pipeline was the World Glacier Monitoring Service's (WGMS) database. It is one of the most comprehensive publicly available glacier monitoring databases and contains reference identifiers to other glacier inventories and databases. The database compiles and standardizes glacier data worldwide, tracking glacier fluctuations and changes for public accessibility. Although the Global Land Ice Measurements from Space (GLIMS) database also includes glacier geometry, glacier area, snowlines, rock debris, and other glacial attributes of interest, WGMS structures its data in a way that is especially useful for temporal analysis. Specifically, the Fluctuations of Glaciers (FoG) database provides multi-temporal records on glacier state, front variation, elevation and volume change, and mass balance. The availability of these records is not uniform for all glaciers or all years, but the structure of the database makes it possible to extract year-based and interval-based glacier variables for selected glaciers.

All glaciological data used in this thesis was retrieved from the Fluctuations of Glaciers (FoG) database prepared by WGMS. The tables of primary interest for this analysis were **glacier**, **state**, **front_variation**, and **mass_balance**. The **glacier** table provides the general metadata needed to identify and locate each glacier, including the glacier identifier, country, name, geographical coordinates, GTN-G region, and references to external inventories such as GLIMS and the Randolph Glacier Inventory. This table was used to select the glacier case study and to retrieve the reference coordinate used for distance calculations.

The **state** table contains measurements of the physical state of a glacier at specific dates. These measurements include glacier area, length, lowest elevation, highest elevation, and mean elevation. In this thesis, these values were used to construct a yearly glacier table for the selected glacier. When multiple state measurements were available for the same year, they were aggregated to obtain a single yearly value. The resulting variables, such as glacier area and glacier length, were then used as glacier-side explanatory variables in the correlation analysis.

The **front_variation** table contains measurements of glacier front change, usually describing the change in glacier length between an earlier reference date and a later measurement date. This table is therefore not always a simple yearly table, but rather an interval-based record. For the exploratory correlation analysis, front variation measurements were assigned to the year of the later measurement date. Additionally, where possible, the front length change was annualized by dividing the measured length change by the duration of the interval in years. This allowed front variation records to be compared more easily with yearly species occurrence metrics.

The **mass_balance** table contains yearly glacier-wide mass balance information, including variables such as winter balance, summer balance, annual balance, equilibrium line altitude, accumulation area ratio, and the area considered in the balance calculation. These variables are particularly relevant to this study because mass balance reflects whether a glacier is gaining or losing mass in a given hydrological year. Negative mass balance years are an important glaciological expression of the climatic conditions that drive glacier retreat, and therefore provide a meaningful additional glacier fluctuation variable to compare against species occurrence patterns.

The glacier selected for the initial case study was the Maladeta glacier in Spain. This glacier was selected because it has usable WGMS records and is located in a mountain region where plant occurrence records from GBIF are available. The Maladeta glacier therefore provided a practical first case study for testing whether glacier fluctuation data and species occurrence data can be spatially and temporally integrated. The glacier variables extracted from WGMS were not treated as direct causal explanations by themselves, but as measurable indicators of glacier fluctuation that could be compared against distributional metrics derived from species occurrence data.

5.2 Species Occurrence Data

Species occurrence data was obtained from the Global Biodiversity Information Facility (GBIF), an international open-access infrastructure that aggregates biodiversity records from museums, herbaria, research institutions, monitoring programs, and citizen science platforms. GBIF occurrence records typically represent observations or collected specimens of a taxon at a particular location and time. For this thesis, the relevant information extracted from GBIF consisted mainly of the species or taxon name, geographic coordinates, observation year, and, when available, elevation.

The use of GBIF data is appropriate for this thesis because the research objective is not to estimate absolute population size, but rather to explore whether observed species distributions show spatial or elevational patterns that can be compared with glacier fluctuation records. GBIF records have been widely used in species distribution and biogeographical research because they provide broad spatial and temporal coverage that would be difficult to obtain from individual field campaigns alone. However, GBIF records are not collected through a single standardized sampling design. They are affected by observer effort, accessibility, research interest, digitization history, and the uneven spatial distribution of citizen science and institutional records. For this reason, occurrence counts cannot be interpreted directly as population abundance.

This limitation was treated as an important part of the analysis rather than as a problem to ignore. The pipeline therefore calculated several yearly species metrics instead of relying only on raw occurrence counts. These included the number of occurrence records, the number of unique species where relevant, mean and median distance to the glacier reference coordinate, centroid distance to the glacier reference coordinate, and mean and median elevation. For broad taxonomic groups such as Orchidaceae, unique species count was used as a simple diversity-related metric, although it remains sensitive to sampling effort. For the single-species case study of *Salix pyrenaica*, unique species count was not informative, and the analysis focused instead on occurrence count, elevation, and spatial distance metrics.

The species records were spatially filtered using circular buffers around the selected glacier reference coordinate. Three main buffer sizes were used: 25 km, 50 km, and 100 km. The 25 km buffer represents a stricter local study area, the 50 km buffer was used as the main analysis scale, and the 100 km buffer was used as a broader sensitivity check. The purpose of using multiple buffers was to avoid basing the analysis on a single arbitrary spatial scale and to evaluate whether detected relationships remained stable when the spatial extent of the study area changed.

Before analysis, GBIF occurrence records were cleaned by keeping only records with valid coordinates and usable temporal information. Coordinates were converted to numeric latitude and longitude values, impossible coordinate values were removed, and observation years were extracted either directly from the year column or from the event date when necessary. Records before the selected analysis period were removed. Where coordinate uncertainty information was available, it was used to filter out records with high spatial uncertainty. These cleaning steps were necessary to make the occurrence records comparable with the glacier data and to reduce the risk of including spatially unreliable points in the analysis.

5.3 DEM-derived Elevation

A substantial limitation of the GBIF occurrence data was that many records contained valid geographic coordinates but did not contain an elevation value. Since elevation is central to this thesis, especially for mountain plant species and potential upward distribution shifts, removing all records without original GBIF elevation would have resulted in unnecessary data loss. This would be especially problematic in alpine and cold-region contexts, where occurrence records are already less abundant than in more accessible and densely sampled regions.

To address this, missing elevation values were supplemented using a Digital Elevation Model (DEM). A DEM is a pixelated representation of the Earth's surface elevation, where each grid cell stores an elevation value for a specific location. In this thesis, the Copernicus GLO-30 elevation product was used to infer elevation values for species occurrence coordinates. For each occurrence record with valid latitude and longitude, the corresponding DEM cell was sampled and its elevation value was assigned to the occurrence point.

The use of DEM-derived elevation is reasonable in this context because the occurrence coordinates already define the spatial position of the observation. Similar to how coordinates can be used to infer spatial attributes such as country, region, or distance to a glacier, they can also be used to infer the approximate elevation of the location. However, DEM-derived elevation should not be interpreted as an originally measured field value. It is an inferred environmental attribute based on the recorded coordinates and the resolution and accuracy of the elevation model.

To preserve transparency, the analysis distinguished between original GBIF elevation and DEM-derived elevation. When an original GBIF elevation value was available, it was retained, when it was missing, the DEM-derived value was used instead. This produced a final elevation variable that allowed more occurrence records to be included in the yearly elevation metrics. The resulting elevation data was then used to compute mean and median occurrence elevation per year and per spatial buffer. These metrics were central to the species-side analysis because elevational shifts are one of the most ecologically meaningful forms of distributional change in mountain environments.

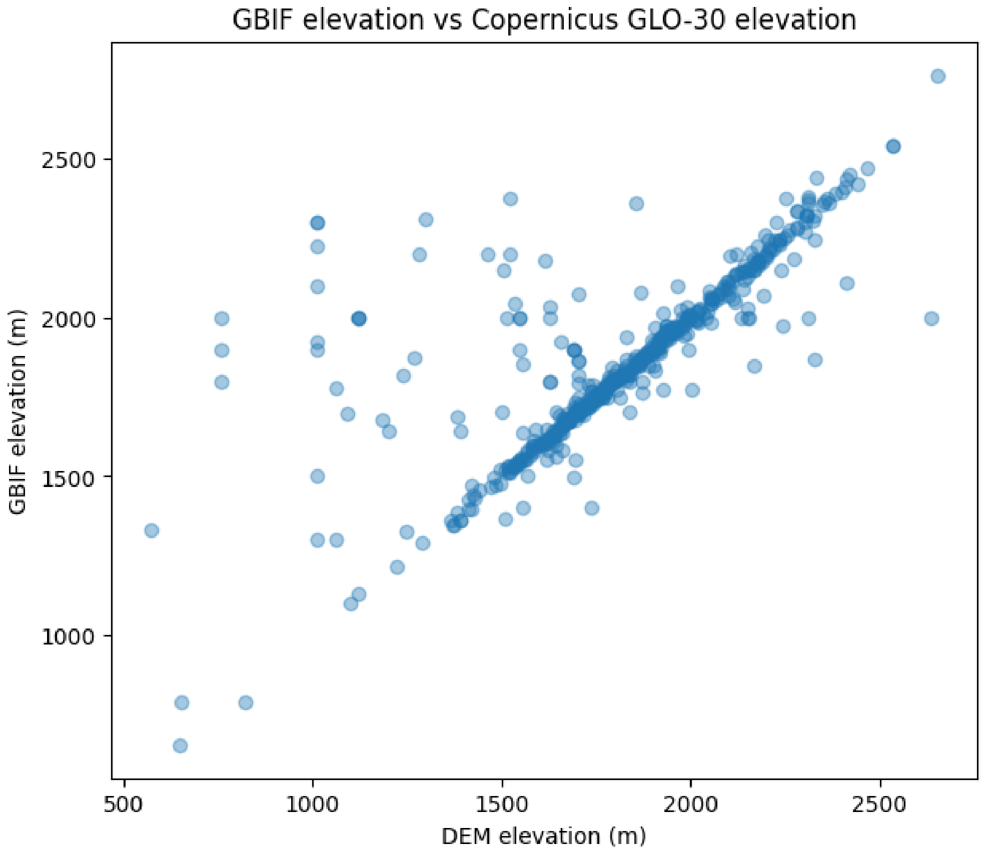


Figure 1: Comparison between original GBIF elevation values and Copernicus GLO-30 DEM-derived elevation values for occurrence records of *Salix Pyrenaica* where both elevation sources were available.

Metric	**Value**
Number of records	479
Pearson correlation (r)	0.740
Pearson p-value	< 0.001
Spearman correlation (ρ)	0.795
Spearman p-value	< 0.001
Mean absolute error (m)	87.92
Median absolute error (m)	12.29
Root mean squared error (m)	229.64
Mean error bias (m)	55.71

Table 1: Validation statistics comparing GBIF elevation values with Copernicus GLO-30 DEM-derived elevation values.

The use of DEM-derived elevation therefore served two main purposes. First, it reduced the loss of occurrence records caused by missing elevation values. Second, it made it possible to examine whether species occurrence elevations changed over time in relation to glacier fluctuation metrics. This was especially important for the *Salix pyrenaica* case study, where elevation-based metrics provided a more direct species-level distribution signal than raw occurrence counts.

To evaluate whether DEM-derived elevation could be used to supplement missing GBIF elevation values, I compared the original GBIF elevation values with the corresponding Copernicus GLO-30 elevation values for records where both were available. The comparison is shown in Figure 1, and the summary statistics are presented in Table 1. Overall, as shown in Figure 1, the two elevation sources showed a strong positive relationship, with a Pearson correlation of $r = 0.740$ and a Spearman correlation of $\rho = 0.795$ across $n = 479$ records. Both correlations were statistically significant ($p < 0.001$). This indicates that the DEM-derived values preserve the general elevational ordering of occurrence records and are therefore suitable for estimating missing elevation values in the context of this analysis.

The error statistics show that the mean absolute error between GBIF and DEM elevation was approximately 87.92 m, while the median absolute error was 12.29 m. The difference between these two values indicates that most records were closely aligned, but that some outliers produced larger deviations. This is also visible in Figure 1, where most points follow the expected positive one-to-one trend, while a smaller number of records deviate substantially. The root mean squared error was 229.64 m, confirming that large deviations exist for a limited number of points. The mean error bias, calculated as GBIF elevation minus DEM elevation, was 55.71 m, suggesting that GBIF elevation values were on average slightly higher than the corresponding DEM-derived values.

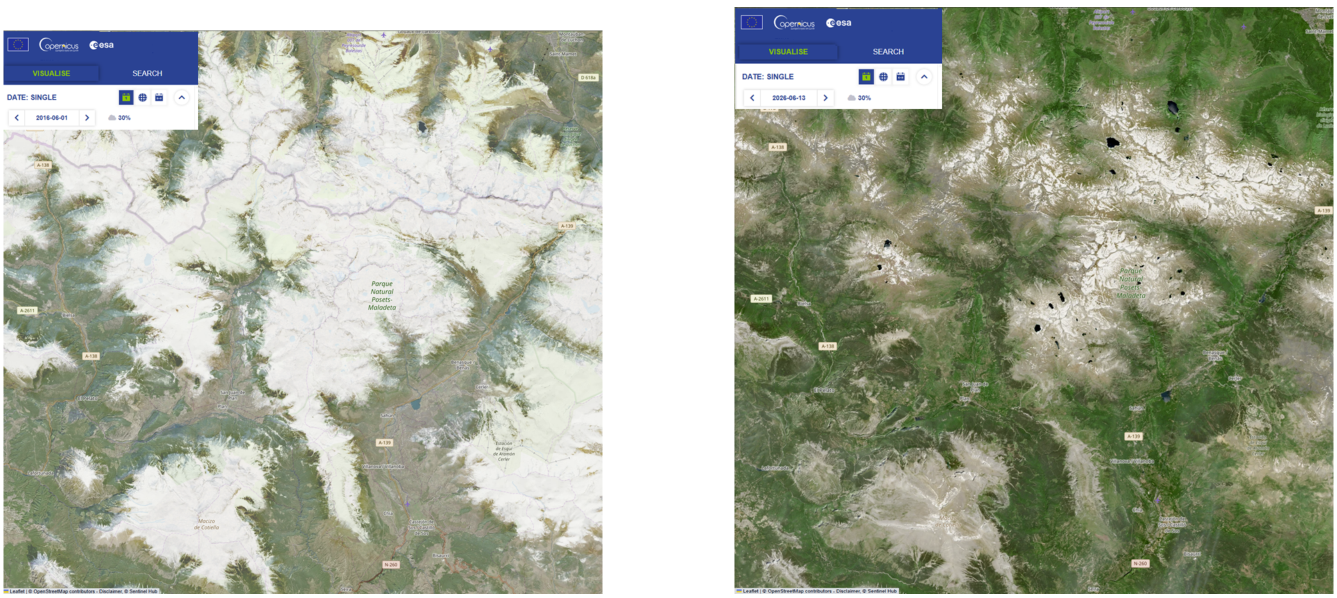
Based on this validation, DEM-derived elevation was considered sufficiently reliable for filling missing elevation values, especially because excluding all records without original GBIF elevation would have considerably reduced the number of usable occurrence records. Nevertheless, DEM-derived elevation was treated as an inferred environmental attribute rather than a directly observed measurement. For this reason, the analysis distinguishes between original GBIF elevation and DEM-completed elevation, and the potential uncertainty introduced by this step is considered in the interpretation of the results.

6 Methodology

6.1 Glacier and species selection

Glacier selection was grounded on a filtering process based on number of annual observations, rather than observations per year, as my analysis would be interannual. Considering that GBIF's main source of occurrences are images taken of an individual or group of a species, occurrence count was taken from the year 2000 till present. This filtering period was chosen because the advent of the mobile phone is the biggest cause for a spike in occurrence data in GBIF's citizen science tracking, but also because it approximated a nice 25 year study period till current times. After ordering based on number of yearly values, the Maladeta glacier ranked first with data for the past 24 years. Located in the Spanish province of Huesca (Aragón), near the highest peak in the Pyrenees (Pico de Aneto), it's total glacierized area has shrunk by 21.4% - 23.2% ([9] and [18], respectively) since the 1950's with an average loss of total thickness of 6.3 m [18]. These changes are clearly visible as shown in the Maladeta satellite imagery of Figure 2. Selection of the Maladeta glacier as the foundational glacier to build the pipeline around is further justified by the fact that it has been strongly affected by climate change as strongly affected glaciers with high extinction probabilities are of greatest interest for studies. Both because they have the greatest consequence for future species distribution predictions and because their elevated mass balance loss means makes them a disappearing study case as glacier extinction is imminent.

As such, the selection of the Maladeta glacier as the foundational glacier to build the pipeline around is further justified for being one that has been strongly affected by the changing climate, making it an important candidate as strongly afflicted glaciers with high extinction probabilities are of greatest interest for future species distribution predictions.



(a) 1st of June, 2016 (b) 13th of June, 2026

Figure 2: Satellite images of the Maladeta glacier 10 years apart (01/06/2016 and 13/06/2026), taken with the European Space Agency's Coopernicus online browser.

The first biological group selected for analysis was Orchidaceae. This group was chosen as an initial exploratory case because orchids are plant species with ecological relevance in mountain environments and sufficient GBIF occurrence availability in the Pyrenean region surrounding Maladeta. Their unique aesthetics and endemic nature make them more likely to be recognized and so photographed and reported in citizen science datasets such as GBIF's, compared with less visually distinctive plant groups. This was important for the first implementation of the pipeline because the objective was not only to test a biological hypothesis, but also to verify whether glacier and species occurrence data could be joined, filtered, aggregated, and statistically compared in a reproducible way. Orchidaceae therefore provided a useful first study group because it contained enough occurrence records to test the technical and methodological structure of the pipeline.

Plants were also appropriate study organisms for this thesis given that their static nature is easier to interpret in as factual spatial distribution in relation to local environmental conditions compared to other highly motile animals. Birds, mammals, and other mobile species can move across large distances and may be observed far away from the environmental conditions most relevant to their long-term survival [7]. In contrast, plants are fixed after establishment, making their recorded locations a more direct connection to the habitat conditions of the observation site. Compared with trees, many herbaceous and alpine plants also respond over shorter ecological timescales. For this reason, plant occurrence records are suitable for exploring whether distributional metrics such as elevation, distance to glacier reference point, and centroid movement show relationships with glacier fluctuation variables.

The second biological case selected was *Salix pyrenaica*. This species was used as a more specific validation case after the broader Orchidaceae analysis as a plant expansion studies in Maladeta region finds that "Salix pyrenaica, which occupies a broad range of habitats, shows a high level of range filling, a blurred genetic structure and an admixture cline between the two main genetic groups, congruent with rapid postglacial expansion" [3]. Unlike Orchidaceae, which represents a family-level community analysis, *Salix pyrenaica* allows the pipeline to be tested on a single species. This distinction is important because family-level metrics such as unique species count can be strongly influenced by sampling effort, whereas a single-species analysis focuses more directly on the spatial and elevational distribution of one taxon. *Salix pyrenaica* was therefore selected to evaluate whether the pipeline could produce a more species-specific ecological signal, especially in relation to elevation-based distribution metrics. Its occurrence in Pyrenean mountain environments and the availability of GBIF records near the Maladeta study area made it an appropriate second case study for testing the robustness and interpretability of the method.

6.2 Spatial Filtering

Spatial filtering was necessary because GBIF occurrence records are distributed across large geographic regions, while the objective of this thesis is to study species records in relation to a specific glacier. Without spatial filtering, the analysis would mix occurrence records from areas that are unlikely to be ecologically related to the Maladeta glacier. Therefore, occurrence records were filtered using circular distance buffers around the glacier reference coordinate provided in the WGMS glacier table. Distances were calculated as geodesic distances between each occurrence coordinate and the Maladeta reference coordinate. Importantly, this distance does not represent distance to the glacier front or glacier outline, but rather distance to the glacier reference point used in the FoG database. The glacier front (or terminus) is the lowest or outermost edge of the glacier, marking the physical end of a glacier at any point in time. On the other hand, a glaciers outline is WGMS's approximated edge margin of the glacier. Although these two metrics might have been of significance in my application, their greater uncertainty and thus less real factual veracity warranted their exclusion from this initial, more exploratory application.

Several buffer sizes were tested in order to evaluate the trade-off between ecological locality and data availability. The tested buffers were 5 km, 10 km, 25 km, 50 km, 100 km, 200 km, 300 km, and 400 km. Figure 3 shows the number of Orchidaceae occurrence records around Maladeta for each buffer through time. The smallest buffers, 5 km and 10 km, contained very few yearly observations and several years with zero or near-zero records. Although these buffers are the most local and therefore the most directly connected to the glacier surroundings, their low data availability makes yearly correlation analysis unstable and overly dependent on individual records.

The 50 km buffer provided a better balance between ecological locality and sufficient temporal coverage, and was therefore selected as the main radius for the analysis. This radius remained close enough to the Maladeta region to preserve glacier-specific interpretation while containing

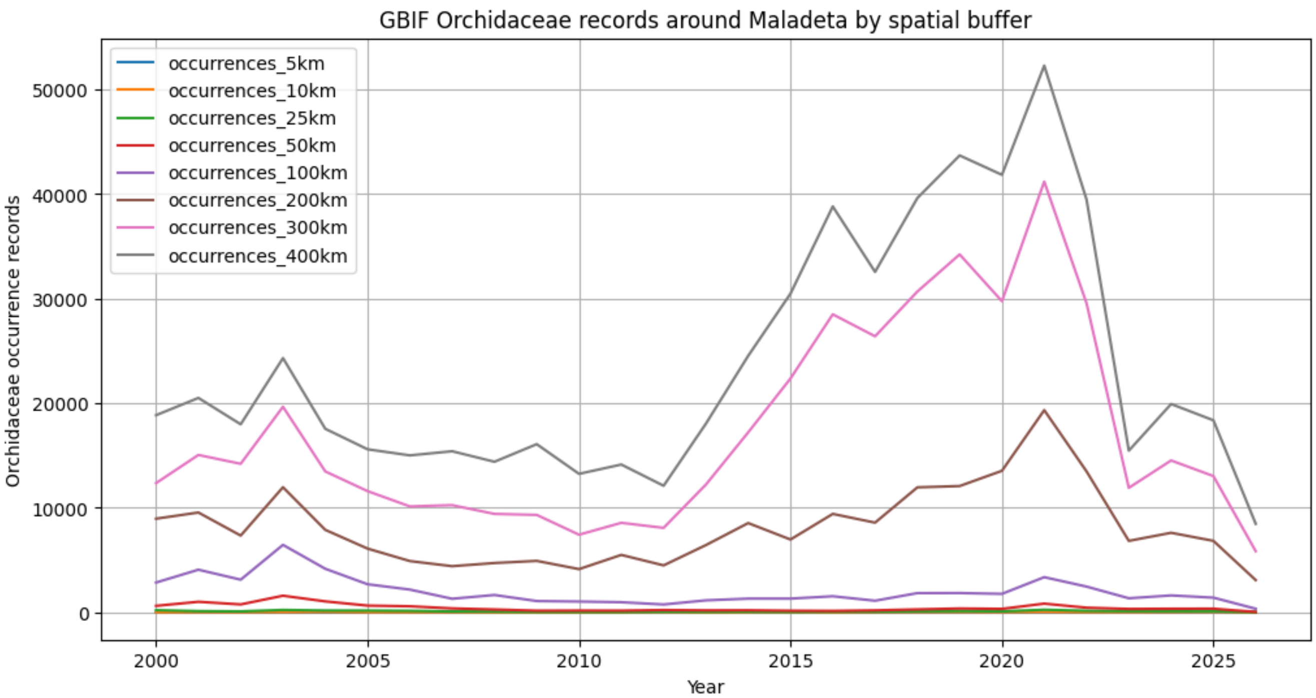


Figure 3: Yearly GBIF Orchidaceae occurrence records around the Maladeta glacier reference coordinate for spatial buffers ranging from 5 km to 400 km. Smaller buffers preserve local glacier relevance but contain fewer records, while larger buffers increase data availability but increasingly represent broader regional sampling patterns.

enough yearly occurrence records to compute distribution metrics such as mean elevation, median elevation, distance to glacier reference point, and centroid distance.

The larger buffer ranges of 100 km and up predictably captured substantially more occurrence records. However, this expansion increasingly captures a wider Pyrenean and regional biodiversity signal rather than a local Maladeta signal. This is visible in Figure 3, where the 200 km, 300 km, and 400 km buffers show much larger yearly occurrence counts and stronger temporal peaks. These large increases improve statistical sample size but weaken the ecological interpretation of the records as being specifically related to the Maladeta glacier. The strongest peaks at larger radii are also likely to reflect regional sampling intensity in GBIF rather than only biological change. At what point the buffer size becomes too big is taxon dependent. Since both analysis cases implemented in this thesis were of plant taxons that disperse through wind, buffer values of interest were maintained the same.

The spatial filtering results therefore supported the use of Maladeta and Orchidaceae as suitable initial subjects for building and testing the pipeline. The occurrence counts showed that the area surrounding Maladeta contains enough Orchidaceae records for analysis once the buffer is expanded beyond the very local scales.

6.3 Glacier Metrics

The glacier-side variables used in the analysis were extracted from the WGMS FoG tables for the selected glacier. The main source was the **state** table, which contains glacier state measurements at specific dates. Specifically, the **state** variables were grouped by year and then yearly average glacier values were computed: glacier area, glacier length, mean elevation, lowest elevation, and highest elevation. These variables describe the physical state of the glacier at a given observation date and are therefore the most direct glacier metrics for yearly comparison with species occurrence data.

As the analysis scope is centered around year comparisons, interannual changes between the area, glacier length, mean elevation, lowest elevation, and highest elevation parameters were computed. The generalized formula was

$$\Delta P_y = P_y - P_{y-1} \tag{1}$$

where P_y stands for the value of the parameter P in year y. Percentage change variables, the relative change compared with the previous available year and expressed as a percentage, were computed with

$$\Delta A_y^{\%} = \frac{A_y - A_{y-1}}{A_{y-1}} \times 100 \tag{2}$$

These percentage change values are useful because they not only express the state of the glacier in a given year, but also the direction and magnitude of change between consecutive observations, perceiving hidden overarching trends. The code calculates these annual change values after sorting records by year. Thus, although the previous row usually aligns with the previous year, technically it is actually calculating the values of P_y with the previous *available* glacier observation year.

In addition to the state table, the **change** table was considered as a source of geodetic glacier change information. This table is not a yearly table, but an interval-based table. Each record describes glacier-wide elevation or volume change between an earlier reference measurement, begin_date, and a later measurement, end_date. When used, the duration of the interval was calculated as $T = \frac{\text{end\_date} - \text{begin\_date}}{365.25}$, where T is expressed in years. Elevation and volume change could then be annualized as annualized elevation change $= \frac{\text{elevation change}}{T}$ and annualized volume change $= \frac{\text{volume change}}{T}$. This distinction is important because the change table does not represent a single observation year, but a measured change across a time interval.

The **front_variation** table was also used to describe changes in glacier front position, which specifically refers to the advance or retreat of a glaciers terminus [19]. As with the change table, front variation is interval-based rather than purely annual and so for the yearly exploratory analysis, each front variation record was assigned to the year of its end_date. The interval duration was calculated using the difference between end_date and begin_date, divided by 365.25. An annualized front length change was then calculated as annualized front length change $= \frac{\text{front length change}}{T}$. This allowed front variation records with different time intervals to be compared more directly with yearly species metrics.

Finally, the **mass_balance** table, which simply refers to the resulting gain and loss of ice for the glacier throughout the year, was used as a complementary glacier fluctuation source [6]. From this table I used annual balance, winter balance, summer balance, equilibrium line altitude, accumulation area ratio, and the area considered in the balance calculation when these values were available. These variables are relevant because they more explicitly conceptualize yearly changes in mass balance. In this sense, mass balance provides a process-based glaciological variable, while the state and front_variation tables provide more geometric descriptions of glacier size, elevation structure, and retreat.

6.4 Observation Metrics

After cleaning the GBIF occurrence records and extracting the glacier records, both datasets were transformed into yearly metric tables. This was necessary because the analysis is interannual: the objective is to compare yearly glacier fluctuation variables with yearly species distribution variables. The pipeline therefore first cleaned the occurrence records, then calculated distances to the selected glacier, filtered the records by spatial buffer, computed yearly species metrics, and finally merged the species and glacier tables by year.

In the main yearly analysis, I mainly used glacier state variables because they can be assigned directly to calendar years. Interval-based variables were treated separately and interpreted more cautiously. When interval-based glacier variables were used, they were annualized and assigned to the end year of the interval, but they do not represent a true yearly observation in the same way as GBIF occurrence year. These state variables were therefore the most direct glacier-side variables for merging with yearly species metrics. Therefore, interval-based variables were interpreted as approximate rates of change associated with the end of a measurement interval, not as true single-year observations. For this reason, interval-based glacier variables were treated more cautiously than state-table variables in the analysis.

Distance from the glacier was calculated using the geographic coordinates of each GBIF occurrence record and the reference coordinate of the selected glacier from the WGMS glacier table. Since these coordinates are latitude and longitude values, a simple Euclidean distance would not be appropriate as they are not in a flat x/y axis plane. Instead, the pipeline calculated geodesic distance on the WGS84 ellipsoid, the standard geographic coordinate system used by GPS systems

[10]. For an occurrence point i, with latitude and longitude (ϕ_i, λ_i), and the glacier reference point (ϕ_g, λ_g), the distance was defined as

$$d_i = Geod_{WGS84}((\phi_i, \lambda_i), (\phi_g, \lambda_g)). \tag{3}$$

The result was returned in meters and converted to kilometers. This value was stored as **distance_to_glacier_km**. It is important to clarify that this is the distance to the glacier reference coordinate, not distance to the glacier front, glacier boundary, or glacier outline. Therefore, this variable is interpreted as a proximity metric to the glacier reference point. These individual distance values were then used for the spatial filtering described previously. For each radius, the pipeline returned only the records whose geodesic distance to the glacier reference point was less than or equal to the selected radius.

For each year and buffer, several species-side metrics were computed. The first was **occurrence_count**, which is the number of GBIF records available in that year and radius. For broader taxonomic analyses, such as Orchidaceae, the pipeline also calculated **unique_species**, which is the number of distinct species recorded in that year. Since the number of unique species can increase simply because more records were collected, a normalized richness metric was also computed:

$$mathrmunique\ species\ per\ 100\ records = \frac{\text{unique species}}{\text{occurrence count}} \times 100. \tag{4}$$

This metric does not fully remove sampling bias but it does provide a simple way to compare species richness relative to the number of records collected. To clarify its effect, even if a specific year has more unique species in absolute terms, it can still have a lower value because it had many more total records, reducing the effect of years with much higher sampling effort.

Further, yearly distance metrics are also computed, with the mean distance to the glacier reference point calculated as

$$\bar{d}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} d_i, \tag{5}$$

where n_y is the number of occurrence records in year y, d_i being the geodesic distance calculated in Equation 3. Same process is applied for the median distance. Since median values are less sensitive to extreme records than means I decided to include both. Together, these two metrics describe whether the occurrence records in a given year were, on average, closer to or farther from the glacier reference coordinate.

Most important to my analysis was the centroid for that years occurrences, used as a compact representation of the central spatial position of the occurrence records in a given year. However, it was interpreted carefully because a centroid can hide the internal spatial spread of records. For each year, the centroid latitude and longitude were calculated as the arithmetic mean of the occurrence coordinates:

$$\bar{\phi}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} \phi_i \tag{6}$$

and

$$\bar{\lambda}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} \lambda_i \tag{7}$$

The distance between this yearly centroid and the glacier reference coordinate was then found using the same WGS84 geodesic distance method. This value was stored as **centroid_distance_km**.

Other elevation metrics were defined in parallel, with mean and median occurrence elvations calculated. Where DEM data completion was used, these values were based on the completed final elevation variables.

The final step was to merge the yearly species metrics with the yearly glacier metrics. The merge was performed by year, producing one analysis table per spatial buffer. Each row represented one year and contained both glacier-side variables and species-side variables. This merged table was then used for correlation analysis, diagnostic checks, and visualization. Repeating the same process across several buffer radii made it possible to test whether observed relationships were robust to the choice of spatial scale.

6.5 Correlational Analysis

As the objective of the thesis is more exploratory than causal, correlational analysis was used as the main statistical method. The aim to prove that glacier retreat directly causes species redistribution, but to test whether glacier fluctuation variables and species occurrence metrics show consistent temporal associations. This is appropriate because the datasets used in the study are observational, spatially heterogeneous, and not produced under a controlled experimental design. Therefore, stronger causal claims would not be justified from the available data alone. Correlation provides a suitable first method for identifying whether variables such as glacier area, glacier length, mass balance, occurrence elevation, and distance to glacier reference point move together through time. In this context, the correlation values are used both as scientific results and as a validation mechanism for the proposed pipeline: if the pipeline is able to detect interpretable relationships, or show where apparent relationships are weak or sampling-dependent, then it provides a useful framework for future glacier–species analysis. Spearman and Pearson correlation measure related but different forms of association between two variables. Pearson correlation measures the strength of a linear relationship, meaning that it evaluates whether one variable tends to increase or decrease at a constant rate with another variable. Spearman correlation, in contrast, is based on the ranked order of the values rather than the raw values themselves. It therefore measures whether two variables have a monotonic relationship: whether one variable generally increases or decreases as the other increases, even if the relationship is not perfectly linear. Correlation values range from -1 to 1, where values close to 1 indicate a strong positive association, values close to -1 indicate a strong negative association, and values close to 0 indicate little or no consistent association. Thus, Spearman correlation was used as the primary measure since the relationships between glacier and species variables are not necessarily linear and the yearly sample size is limited. Pearson correlation, on the other hand, was used only as a secondary comparison when useful. For the scope of this bachelor thesis, this level of analysis is sufficient because the main contribution is the integration, cleaning, spatial filtering, metric construction, and exploratory comparison of glacier and species datasets. Predictive modelling would require stronger temporal coverage, more glacier case studies, and better control of sampling effort than was available in the present work.

7 Results and Discussion

7.1 Analysis 1: Maladeta and Orchids

The first analysis tested the basic implementation of the pipeline using the Maladeta glacier and Orchidaceae occurrence records. The analysis was intentionally limited to the core pipeline: GBIF occurrence cleaning, spatial filtering, yearly metric construction, and comparison with glacier state variables. DEM-derived elevation, front variation, and mass balance variables were not implemented for this first case. This was done because Orchidaceae was used primarily as the initial test of the basic glacier-species integration workflow. The more extended version of the pipeline, including DEM-completed elevation and additional glacier fluctuation variables, was then applied in the second case study with *Salix pyrenaica*.

Species variable	**Glacier variable**	**Correlation**
Centroid distance (km)	Glacier area	0.922
Centroid distance (km)	Glacier length	0.886
Centroid distance (km)	Glacier lowest elevation	-0.809
Mean distance (km)	Glacier area	0.806
Median distance (km)	Glacier area	0.783
Median distance (km)	Glacier mean elevation change	-0.599
Mean distance (km)	Glacier mean elevation change	-0.541

Table 2: Strongest Spearman r correlation values between yearly Orchidaceae occurrence metrics and Maladeta glacier state variables at the main 50 km buffer.

From Table 2 we can see that despite the experimentational nature of this first implementation, there were strong relationships found. The strongest relationships in the Orchidaceae case were mostly distance-based, highest of those found between centroid distance and glacier area, followed by centroid distance and glacier length. This means that the yearly central position of Orchidaceae records around Maladeta changed in a way that was strongly aligned with the decreasing size and length of the glacier. Mean and median distance to the glacier reference point were also associated with glacier area, although less strongly than centroid distance. These results show that the pipeline was able to identify a clear spatial signal in the occurrence data.

Although strong correlations were expected, such high correlational values with such small p values caused a bit of alarm (absolute p-value values range from 0.922 to 0.783 for the first five strongest correlational values). Yet, the code and all calculations were verified and everything looked correct. Also, the scatter plots of Figure 4 for the first four highest correlations does show strong linear trends (specifically in the first two). The strongest relationships are orderly and consistent, especially for centroid distance against glacier area and glacier length. Upon looking at the labelled years in the plots and identifying clear patterns the cause of such statistical strength made sense. The strongest relationships are between glacier variables that are themselves changing strongly over time and species metrics that also show a strong temporal pattern. Therefore, the next diagnostic step was to check which variables were also tracking year.

Table 3 confirms that several of the variables are strongly associated with year. Glacier area, glacier length, and centroid distance all show very strong negative Spearman correlation values with year (-0.993, -0.970, and -0.904). We can understand this as year values increase (so going from

Variable	Correlation with year
Year	1.000
Glacier area	-0.993
Centroid latitude	-0.974
Glacier length	-0.970
Centroid distance (km)	-0.904
Glacier lowest elevation	0.894
Mean distance (km)	-0.577
Median distance (km)	-0.559

Table 3: Glacier parameters that have the highest Spearman correlations with year, indicating that part of the glacier-species relationship may be driven by shared temporal structure.

the year 2000 up to present years), glacier area, glacier length, and centroid distance all strongly decline. In other words,

both decrease strongly through time, with Spearman values of -0.993 and -0.970, respectively. On the other hand, centroid distance also decreases through time. This is important because it means that the raw correlations in Table 2 are not independent of temporal structure. In other words, the data shows strong temporal co-movement between glacier shrinkage and Orchidaceae spatial metrics, but it does not by itself prove that glacier area or length directly caused the observed centroid movement.

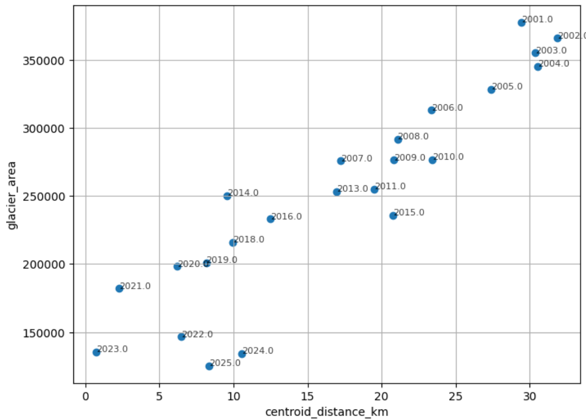
The scatterplots in Figure 4 make this temporal structure visible. The strongest relationships are orderly and consistent, especially for centroid distance against glacier area and glacier length. At the same time, the labelled years show that the points are not randomly distributed through time, but follow a directional temporal sequence. This supports the interpretation that the Orchidaceae case provides strong evidence of spatial-temporal association, but not enough evidence to isolate a direct glacier-area effect.

Species variable	Glacier variable	Spearman		Pearson	
		r	p-value	r	p-value
Centroid distance (km)	Glacier area	0.922	4.15×10^{-10}	0.919	6.29×10^{-10}
Centroid distance (km)	Glacier length	0.886	1.95×10^{-8}	0.814	2.27×10^{-6}
Centroid distance (km)	Glacier lowest elevation	-0.809	3.03×10^{-6}	-0.780	1.12×10^{-5}
Mean distance (km)	Glacier area	0.806	3.39×10^{-6}	0.747	4.24×10^{-5}

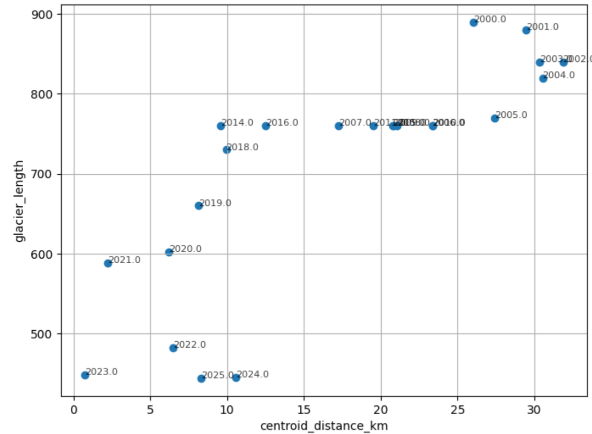
Table 4: Spearman and Pearson correlation tests for the strongest Orchidaceae-glacier relationships at the 50 km buffer.

Once again, the statistical tests in Table 4 show that these relationships are not only visually apparent, but also statistically strong in the raw correlation analysis. Both Spearman and Pearson correlations are high for the main relationships, and their p-values are very small. This confirms that the association is not random in the observed data. Nevertheless, because the same variables also track year strongly, the correlations need to be interpreted as exploratory indicators of co-movement rather than as causal explanations.

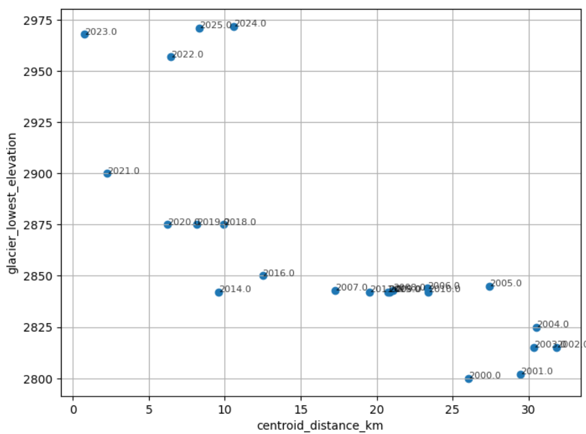
The time-controlled regression results in Table 5 show a more cautious picture. The R^2 value,



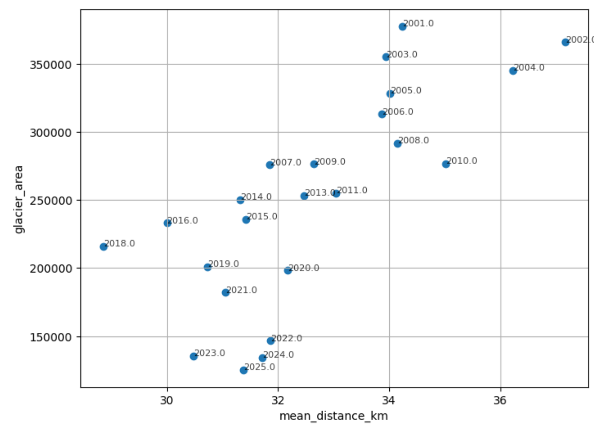
(a) Centroid distance vs glacier area



(b) Centroid distance vs glacier length



(c) Centroid distance vs glacier lowest elevation



(d) Mean distance vs glacier area

Figure 4: Scatterplots of the strongest Orchidaceae occurrence metrics against Maladeta glacier state variables at the 50 km buffer.

Response variable	**Predictor variable**	R^2	**Predictor *p*-value**	**Year *p*-value**
Centroid distance (km)	Glacier area	0.859	0.7652	0.1771
Centroid distance (km)	Glacier length	0.852	0.7051	6.00×10^{-4}
Centroid distance (km)	Glacier lowest elevation	0.855	0.4680	1.00×10^{-4}
Mean distance (km)	Glacier area	0.601	0.6424	0.2307

Table 5: Time-controlled regression results for selected Orchidaceae-glacier relationships. Models include year and occurrence count as controls.

also called the coefficient of determination, describes how much of the variation in the response variable is explained by the full regression model. In this case, the model includes the glacier variable, year, and occurrence count. A relatively high R^2 therefore means that these variables together explain a substantial part of the variation in the species metric being tested. However, the glacier predictor p-values are not significant after accounting for year and occurrence count. This means that, once temporal trend and sampling effort are included in the model, the glacier variable does not explain enough additional variation to be interpreted as an independent predictor of the species occurrence metric. Aditionally, in the case of centroid distance against glacier length and glacier lowest elevation, the year p-values remain significant, further supporting the interpretation that time is an important confounding structure in the Orchidaceae case.

This does not make the result meaningless. Rather, it shows that the pipeline is able to detect strong raw associations and then diagnose whether those associations remain interpretable after temporal controls. For a family-level group such as Orchidaceae, where occurrence records are strongly affected by sampling intensity and broad spatial coverage, this is an important result. The case validates the pipeline as an exploratory tool, while also showing the limitations of interpreting family-level GBIF occurrence patterns as direct ecological displacement.

Species variable	**Glacier variable**	**25 km**	**50 km**	**100 km**
Centroid distance (km)	Glacier area	0.759	0.922	0.902
Centroid distance (km)	Glacier length	0.715	0.886	0.869
Centroid distance (km)	Glacier lowest elevation	-0.560	-0.809	-0.816
Mean distance (km)	Glacier area	0.429	0.806	-0.363

Table 6: Buffer sensitivity comparison for selected Orchidaceae-glacier correlations across 25 km, 50 km, and 100 km spatial filters.

Finally, the predefined ranges were looked into as buffer sensitivity results and are presented in Table 6, the results showing that centroid-based relationships are more stable than mean-distance relationships. Centroid distance remains strongly associated with glacier area and glacier length at both 50 km and 100 km, while the 25 km values are weaker but still positive. In contrast, mean distance against glacier area changes substantially across buffers and even becomes negative at 100 km. This confirms that spatial scale affects the interpretation of occurrence-based metrics. For this reason, the 50 km buffer was retained as the main analysis radius for the next case, whilst the other two more extreme cases stayed as sensitivity checks.

7.2 Analysis 2: Maladeta and Pyrenaica

The second analysis applied the pipeline to *Salix pyrenaica*, a single species occurring in Pyrenean mountain environments. This case was used as a more specific test after the broader Orchidaceae analysis. Since *Salix pyrenaica* is a single species, the analysis focused less on diversity-related metrics and more on elevation, distance, and occurrence count. I also implemented the extended version of the pipeline, including DEM-derived elevation completion and additional glacier fluctuation variables from mass balance and front variation records.

The DEM step was especially important for this analysis because elevation is one of the most meaningful distributional variables for a mountain plant species. By using DEM-derived values to supplement missing GBIF elevation records, more observations could be retained for the yearly elevation metrics. Therefore, the *Salix pyrenaica* case represents a more complete version of the pipeline than the initial Orchidaceae test.

Species variable	**Glacier variable**	**Spearman**		**Pearson**	
		r	p**-value**	r	p**-value**
Occurrence count	Glacier area	0.689	2.76×10^{-4}	0.578	3.87×10^{-3}
Mean elevation (m)	Glacier area	-0.642	9.51×10^{-4}	-0.506	1.38×10^{-2}
Mean elevation (m)	Glacier length	-0.556	5.88×10^{-3}	-0.495	1.64×10^{-2}
Median elevation (m)	Glacier highest elevation	-0.177	0.420	-0.109	0.621

Table 7: Spearman and Pearson correlation tests between *Salix pyrenaica* occurrence metrics and Maladeta glacier state variables.

Table 7 shows that the strongest *Salix pyrenaica* relationships involved occurrence count and mean elevation. Occurrence count was positively correlated with glacier area, while mean elevation was negatively correlated with both glacier area and glacier length. The negative elevation relationships are the most ecologically interesting part of this case: years with smaller glacier area and shorter glacier length were associated with higher mean occurrence elevation. This is consistent with the general expectation of upward distributional movement in mountain species, although it cannot be interpreted as proof of glacier-driven migration.

The median elevation relationship with glacier highest elevation was weak and not statistically significant, suggesting that not all elevation metrics responded in the same way. Mean elevation showed clearer relationships than median elevation in this case, possibly because the yearly distribution of records includes upper or lower values that influence the mean more strongly.

Response variable	**Predictor variable**	R^2	**Predictor** p**-value**	**Year** p**-value**
Mean elevation (m)	Glacier area	0.354	0.2032	0.2991
Mean elevation (m)	Glacier length	0.315	0.4839	0.8910
Median elevation (m)	Glacier highest elevation	0.303	0.1620	0.0681

Table 8: Time-controlled regression results for selected *Salix pyrenaica*-glacier relationships. Models include year and occurrence count as controls.

The time-controlled regressions displayed in Table 8 again require caution. The glacier predictor p-values are not significant after controlling for year and occurrence count. This means that the raw correlations between *Salix pyrenaica* elevation and glacier area or length should not be interpreted as independent glacier effects.

Compared with the Orchidaceae case, the *Salix pyrenaica* analysis is more directly interpretable because it concerns one species and uses DEM-completed elevation values. At the same time, the regression results show that the dataset is still not strong enough to separate glacier fluctuation from broader temporal change. The relationship is therefore best understood as an exploratory signal: *Salix pyrenaica* occurrence elevation tends to be higher in years where Maladeta is smaller, but the analysis cannot prove that glacier shrinkage independently drives this pattern.

Table 9 presents the additional exploratory variables included in the second analysis. The strongest values are again elevation-based, with mean and median *Salix pyrenaica* elevation

Species variable	Glacier variable	Correlation
Mean elevation (m)	Mass balance area	-0.651
Median elevation (m)	Mass balance area	-0.507
Mean distance (km)	Winter balance	0.351
Mean distance (km)	Front interval years	-0.322

Table 9: Exploratory correlations between *Salix pyrenaica* occurrence metrics and additional Maladeta glacier variables from mass balance and front variation records.

negatively associated with mass balance area. The relationships with winter balance and front interval years are weaker and should be interpreted with caution, especially because front variation records are interval-based and mass balance data availability is not necessarily uniform across all years. The plots in Figure 5 support the same general conclusion as the tables: elevation metrics are the most interpretable species-side variables, while occurrence count and distance-based metrics are more difficult to interpret because they may reflect sampling effort or spatial recording patterns.

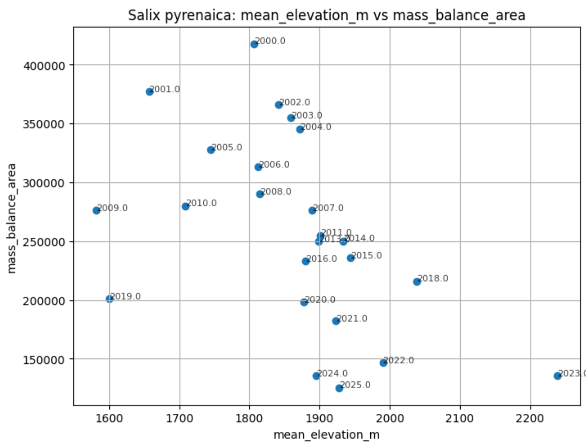
These results are useful mainly as a pipeline extension. They show that additional glacier fluctuation tables can be included in the same workflow, but they also demonstrate that not every glacier variable produces equally interpretable ecological relationships. In particular, state variables such as glacier area and length were easier to compare with yearly species metrics than interval-based front variation variables.

Important to note is that, as shown in Figure 6, the centroid distance species metric does not correlate as initially thought it would with mean glacier elevation. This implies that within our selected buffer, horizontal distance from the Maladeta reference point is essentially not related to occurrence elevation. The finding makes sense as it only implies that elevation does not systematically increase or decrease as *Salix pyrenaica* records get farther from the glacier reference point. Distance to the glacier reference point is a horizontal proximity metric, while elevation is a vertical environmental metric. In mountain terrain, points farther from the glacier reference coordinate are not automatically higher or lower in elevation, because valleys, ridges, slope orientation, and local topography can produce complex spatial patterns. In other words, as should be expected once the geography of Maladeta is taken into consideration, the alpine terrain aroud Maladeta determines that distance to glacier does not correlate so clearly with elevation.

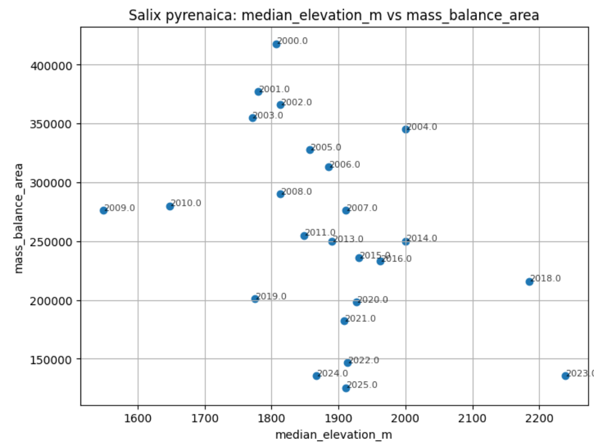
For this reason, centroid distance and elevation should not be treated as interchangeable measures of distributional shift. A species may show an elevation-related pattern without showing a clear horizontal movement away from or toward the glacier reference point. This distinction is important for the interpretation of the results: the strongest ecological signal in the *Salix pyrenaica* case is the elevation-based relationship with glacier area and length, while the horizontal distance metrics are better understood as complementary spatial descriptors.

7.3 Hengduan Timberline Case: Regional Pipeline Stress Test

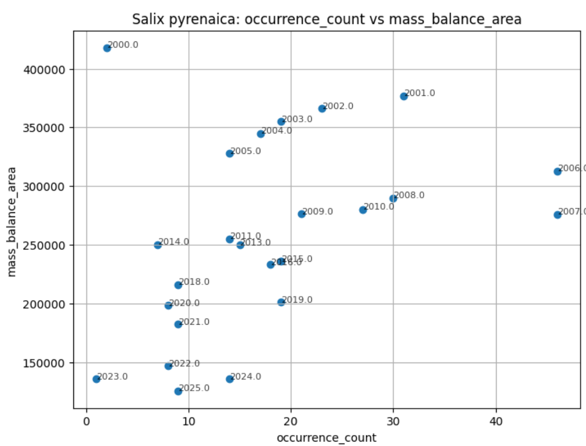
The planned third case study was the Hengduan mountain region, where upward movement of timberline and treeline systems has been reported in previous ecological studies [17]. This case was selected because it offered a different type of validation problem. Instead of testing one glacier and one taxon, the objective was to check whether the pipeline could be extended to a multi-glacier regional setting and compared with a broader elevation-related vegetation signal. Not only that,



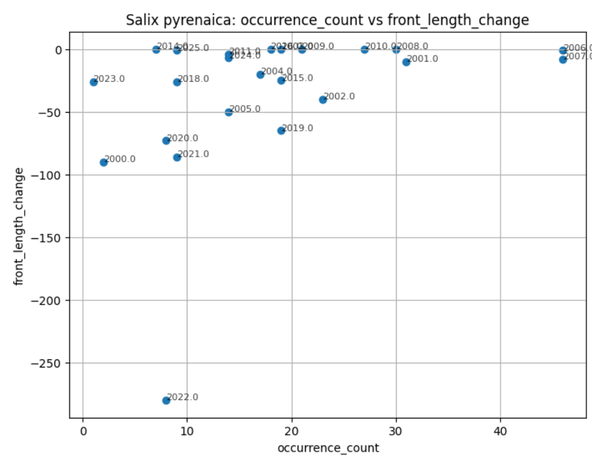
(a) Mean elevation vs mass balance area



(b) Median elevation vs mass balance area



(c) Occurrence count vs mass balance variable



(d) Occurrence count vs front variation variable

Figure 5: Exploratory *Salix pyrenaica* relationships with Maladeta mass balance and front variation variables.

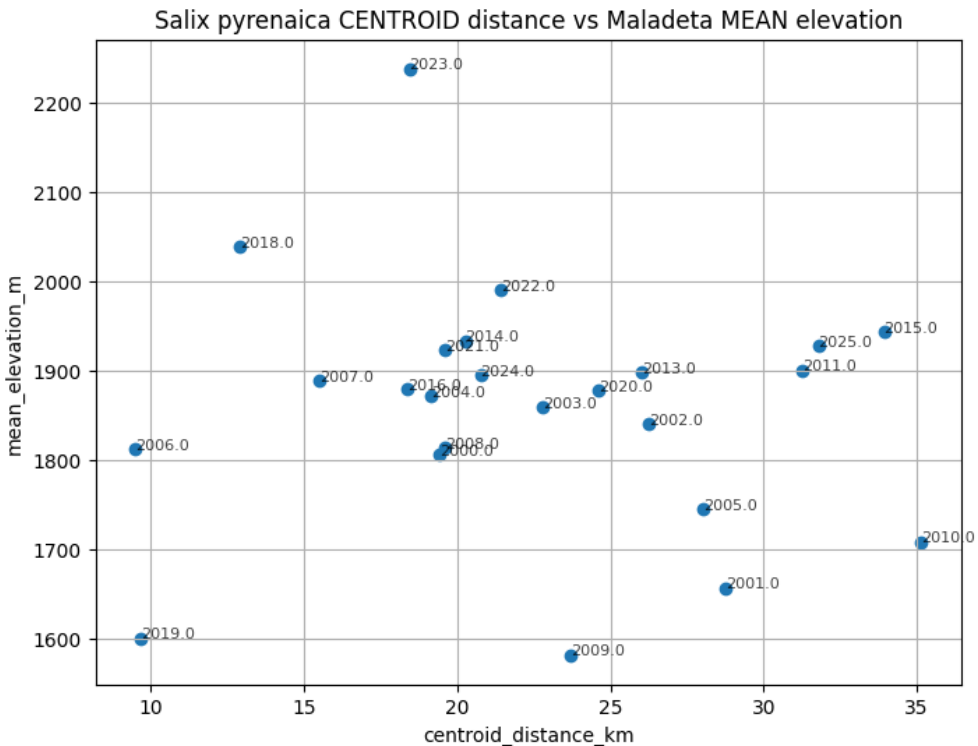


Figure 6: Relationship between centroid distance and elevation for *Salix pyrenaica*. The yearly relationship showed a Spearman correlation of $r = 0.500$ with $p = 0.0127$, and a Pearson correlation of $r = 0.386$ with $p = 0.062$. The points are labelled with the year they belong to.

but Hengduan's character as being one of the most biodiverse hotspots in the world, being one of the Earth's 34 identified hotspots, it could test the pipelines strength with dealing in such a scientifically interesting case.

The Hengduan case therefore required several changes to the pipeline. Glaciers had to be selected manually by coordinate bounds, because the Hengduan region does not correspond directly to one WGMS GTN-G region. Species selection was also more difficult than in the Orchidaceae and *Salix pyrenaica* cases, because timberline composition is not defined by one single species or one clean taxonomic group. Tree-line and timberline communities are made up of several taxa, trees themselves being polyphyletic (group of organisms sharing the same characteristics but not part of one individual taxonomic group), and the GBIF records available for those taxa are unevenly distributed. The pipeline was therefore extended to allow a multi-glacier structure, where occurrence records could be assigned to the nearest selected glacier and then summarized either at the glacier-year level or at the regional level.

However, the Hengduan implementation did not produce a usable final analysis. The main limitation was data availability. Although many glaciers could be selected geographically in the region, the number of usable post-2000 glacier-year observations in the FoG state data was too low to support the same type of yearly comparison used in the Maladeta case studies. In addition, after selecting GBIF species occurrences based after reviewing the literature on the tree composition ([17], [16]and [20]) the timberline-related species records were sparse after filtering and did not provide a strong enough yearly elevation signal for a regional glacier-species comparison.

Although a failed attempt, the result is still important for the conclusion of this thesis as it properly tested the scalability of the pipeline. The failed Hengduan implementation shows that the pipeline is not limited only by code structure, but also importantly by the temporal and spatial completeness of the underlying datasets. A region may be ecologically interesting and well studied in the literature, but still not be suitable for this specific GBIF-WGMS pipeline if either glacier observations or species occurrence records are too sparse.

Therefore, the Hengduan case was treated as a regional stress test rather than as a full third result. It demonstrated that the implemented approach can be conceptually extended to multi-glacier analyses, but also that future regional applications may require additional data sources. For example, reproducing timberline movement findings would likely require forest-line GIS layers, remote-sensing vegetation products, or systematic ecological survey data, rather than relying only on GBIF occurrence points. Timberline research uses satellite imagery for identification of its species relevant metrics such as density and highest elevations. This correctly identifies a natural extention for the pipeline: the usage of satellite images like those taken by Landsat for the identification of parameters not clearly verifiable only with species occurrences. This limitation helps define the practical scope of the pipeline: it is useful for exploratory glacier-species analyses where both glacier records and occurrence records have sufficient temporal overlap, but it cannot overcome missing data in either source.

7.4 Limitations

Several limitations should be considered when interpreting the results of this thesis, both from the pipeline or interpretative side. First, GBIF occurrence data does not necessarily represent systematic ecological sampling as records can come from preserved specimens, citizen science observations, herbarium collections, or institutional datasets, and these different basis types may not all imply the same ecological meaning. A preserved specimen collected decades ago is not equivalent to a recent field observation, and occurrence records in general should not be interpreted as population abundance. Thus, even though these checks were applied in the two analysis cases I implemented, a complete pipeline would have to specifically filter occurrence values based on record types as large number of occurrences does not imply correct ones. Further, not all GBIF regions have the same data types specified. For the first two cases GBIF's simple parquet datatype was used, but turned out not to be available for Hengduan's timberline species, so I used Darwin Archive Core. Thus, a more complete pipeline implementation would require either a clearer specification of occurrence datatypes or more robust verification methods.

To more definitely validate the pipeline and its results, comparison with direct ecological survey data would be required as only then could its results be said to be scientifically valid and so causational. Further, the Hengduan timberline case also pinpoint possible metrics that were not implemented but could've been of interest for the previous two cases and further complete the pipeline. Among them is annual rain fall, slope angle (steepness), and carindal positioning (in terms of which cardinal direction it is facing) as these were seen to have strong impact in timberline composition [2]and so consequently must be at least investigated in any other alpine biota researched. Aditionally, distances were calculated to the WGMS glacier reference coordinate, not to the glacier boundary, outline, or front. Therefore, distance-based metrics should be interpreted as proximity to a glacier reference point, not as exact distance to the ice margin. This further hinders the extrapolation of the correlational values and the completeness of the pipeline.

Although DEM-derived elevation helped reduce data loss, it remains an inferred attribute and introduces uncertainty from coordinate precision, terrain complexity, and DEM resolution. The high correlation found between DEM derived and GBIF elevation values argues that there is a strong relation and it can be trusted, but this was only investigated for elevation values of one species (*Salix pyrenaica*) and around one region (Maladeta). It is not know whether this relation would continue to be true for other cases and so a more in depth investigation into DEM derived values would need to be completed before it can be trusted entirely. The pipeline is therefore dependent on the temporal completeness of both glacier and species datasets.

8 Conclusion

This thesis set out to examine whether glacier fluctuation records and GBIF occurrence data can be integrated into a reproducible pipeline for identifying spatial and elevational patterns in plant species distributions near retreating glaciers. The answer is cautiously positive. The implemented pipeline successfully combined WGMS glacier records, GBIF occurrence data, DEM-derived elevation, geodesic spatial filtering, yearly metric construction, and exploratory statistical analysis into a coherent workflow. In the Maladeta case studies, the pipeline was able to detect strong raw associations between glacier variables and species occurrence metrics, especially between Orchidaceae centroid distance and glacier area, and between *Salix pyrenaica* elevation and glacier area or length. These results show that the approach is useful for exploratory glacier–species analysis and for identifying candidate relationships that may deserve further ecological investigation.

At the same time, the results also show the limits of what can be concluded from the available data. The strongest correlations were often strongly related to year, meaning that glacier shrinkage and species metrics may be moving together through time without the analysis being able to prove that glacier variables independently drive species redistribution. GBIF's occurrence sampling bias, the missing elevation values, the use of glacier reference points rather than glacier outlines, and incomplete glacier records all constrain the causational conclusion . Therefore, the main contribution of this thesis is not a causal proof of glacier-driven species displacement, but the development and testing of a reproducible computational framework that makes such relationships measurable, comparable, and critically diagnosable. The pipeline provides a foundation for future work in which richer glacier geometry, systematic ecological surveys, and broader regional datasets could support stronger predictive modelling.

9 Future Work

As mentioned in the limitations section (Section 7.4), future work should first improve the ecological and spatial precision of the pipeline. The distance metric used in this thesis is based on the WGMS glacier reference coordinate, but a more accurate implementation would calculate distance to glacier outlines, glacier fronts, or recently deglaciated terrain. This would make the spatial interpretation stronger, because proximity would then refer to the actual ice margin rather than to a representative glacier point. Future versions should also integrate glacier outline datasets such as GLIMS or Randolph Glacier Inventory geometries, and where possible use yearly or multi-temporal glacier outlines to measure exposed terrain directly.

A second direction is to improve the biological validation of the method. GBIF occurrence data is useful for exploratory work, but it is affected by sampling bias and mixed occurrence types. Future studies should compare the pipeline outputs with systematic ecological surveys, long-term monitoring plots, or species known from previous field research to be responding to glacier retreat. For plant species, upper-elevation boundary metrics such as the 90th or 95th percentile elevation could also be developed further, since these may better represent upward distributional shifts than mean elevation alone.

More importantly, the work I initiated with the Hengduan's mountainous region's rising timberline should be completed as it would correctly test the pipeline on a multi-taxon, multi-glacier analysis. My mistake was incorrect dependence on a well studied region where I considered occurrence data would be present. A literature review either on rising timberlines or any other species-dependent geographical changes could lead high regional occurrence values for correct usage of the pipeline.

Moreover, extending its application to non-plant species who, even if their dispersal methods makes GBIF's occurrence citizen science data acquisition less applicable, it could still lead interesting results. Clearly beyond the scope of this thesis is the usage of satellite imagery also on the species side for population, range, and distribution values but it seems a natural extension when considering lots of issues came from lack of data. This implementation would also remove the post 2000 year limit used in my applications.

References

[1] Ian S. Acuña-Rodríguez, Kevin K. Newsham, Peter Convey, Elisabeth M. Biersma, Gabriel I. Ballesteros, Cristian Torres-Díaz, William P. Goodall-Copestake, and Marco A. Molina-Montenegro. The role of the soil microbiome in the colonisation of glacier forefields by

Antarctic pearlwort (Colobanthus quitensis) under current and future climate change scenarios. *Soil Biology and Biochemistry*, 188:109249, 2024.

[2] T Eren Bilir, Inez Fung, and Todd E Dawson. Slope-aspect induced climate differences influence how water is exchanged between the land and atmosphere. *Journal of Geophysical Research: Biogeosciences*, 126(5):e2020JG006027, 2021.

[3] Pau Carnicero, Johannes Wessely, Dietmar Moser, Xavier Font, Stefan Dullinger, and Peter Schönswetter. Postglacial range expansion of high-elevation plants is restricted by dispersal ability and habitat specialization. *Journal of Biogeography*, 49(10):1739–1752, 2022.

[4] Sophie Cauvy-Fraunié and Olivier Dangles. A global synthesis of biodiversity responses to glacier retreat. *Nature Ecology & Evolution*, 3(12):1675–1685, 2019.

[5] Francisco Cuesta, Luis D Llambí, Christian Huggel, Fabian Drenkhan, William D Gosling, Priscilla Muriel, Ricardo Jaramillo, and Carolina Tovar. New land in the neotropics: a review of biotic community, ecosystem, and landscape transformations in the face of climate and glacier change. *Regional Environmental Change*, 19(6):1623–1642, 2019.

[6] Bethan Davies. An introduction to glacier mass balance. https://www.antarcticglaciers.org/glacier-processes/mass-balance/introduction-glacier-mass-balance/, 2020. Last updated 30 December 2020.

[7] Gentile Francesco Ficetola, Silvio Marta, Alessia Guerrieri, Mauro Gobbi, Roberto Ambrosini, Diego Fontaneto, Andrea Zerboni, Jerome Poulenard, Marco Caccianiga, and Wilfried Thuiller. Dynamics of ecological communities following current retreat of glaciers. *Annual Review of Ecology, Evolution, and Systematics*, 52(1):405–426, 2021.

[8] Rodolfo Gentili, Carlo Baroni, Cinzia Panigada, Micol Rossini, Giulia Tagliabue, Stefano Armiraglio, Sandra Citterio, Alberto Carton, and Maria Cristina Salvatore. Glacier shrinkage and slope processes create habitat at high elevation and microrefugia across treeline for alpine plants during warm stages. *Catena*, 193:104626, 2020.

[9] Eñaut Izagirre Ánchel Belmonte-Ribas Luis Carcavilla Juan Ignacio López-Moreno Ixeia Vidaller, Ana Moreno. Geomorphology of the maladeta massif (central pyrenees): the traces of the last remaining glaciers. *Journal of Maps*, 20, 2023.

[10] Land Information New Zealand. World geodetic system 1984 (WGS84). https://www.linz.govt.nz/guidance/geodetic-system/coordinate-systems-used-new-zealand/geodetic-datums/world-geodetic-system-1984-wgs84, 2026. Last updated 18 February 2026.

[11] Gianalberto Losapio, Bruno EL Cerabolini, Chiara Maffioletti, Duccio Tampucci, Mauro Gobbi, and Marco Caccianiga. The consequences of glacier retreat are uneven between plant species. *Frontiers in Ecology and Evolution*, 8:616562, 2021.

[12] Camila Neder, Kerstin Jerosch, Hendrik Pehlke, Luciana Torre, and Ricardo Sahade. Antarctic benthic species distribution models and compositional analysis in a coastal ecosystem under glacier retreat. *Marine Ecology Progress Series*, 750:1–18, 2024.

[13] Pekka Niittynen, Risto K Heikkinen, and Miska Luoto. Snow cover is a neglected driver of arctic biodiversity loss. *Nature Climate Change*, 8(11):997–1001, 2018.

[14] Pekka Niittynen and Miska Luoto. The importance of snow in species distribution models of Arctic vegetation. *Ecography*, 41(6):1024–1037, 2018.

[15] Richard G. Pearson, Wilfried Thuiller, Miguel B. Araújo, Enrique Martinez-Meyer, Lluis Brotons, Colin McClean, Lera Miles, Pedro Segurado, Terence P. Dawson, and David C. Lees. Model-based uncertainty in species range prediction. *Journal of Biogeography*, 33(10):1704–1711, 2006.

[16] Shi Peili, Wu Ning, and Gopal S Rawat. The distribution patterns of timberline and its response to climate change in the himalayas. *Journal of Resources and Ecology*, 11(4):342–348, 2020.

[17] Lei Tian, Wenxue Fu, Yu Tao, Mingyang Li, and Lei Wang. Dynamics of the alpine timberline and its response to climate change in the hengduan mountains over the period 1985–2015. *Ecological Indicators*, 135:108589, 2022.

[18] Ixeia Vidaller, Jesús Revuelto, Estibaritz Izagirre, Francisco Rojas-Heredia, Esteban Alonso-González, Simon Gascoin, Pierre René, Etienne Berthier, Iñaki Rico, Ana Moreno, Enrique Serrano, Alfonso Serreta, and Juan Ignacio López-Moreno. Toward an ice-free mountain range: Demise of pyrenean glaciers during 2011–2020. *Geophysical Research Letters*, 48(18):e2021GL094339, 2021.

[19] World Glacier Monitoring Service. Glacier front variation. https://wgms.ch/glacier_front_variation/, 2024. Last changed 30 August 2024.

[20] Bo Xu, Zhi-Min LI, and Hang Sun. Plant diversity and floristic characters of the alpine subnival belt flora in the hengduan mountains, sw china. *Journal of Systematics and Evolution*, 52(3):271–279, 2014.