REPORT

Visual inspection of weather data and there effect on glacier retreat

Mario Viehböck (k12345678)

Supervisors
DI Dr. Andreas Hinterreiter
DI Dr. Christina Humer
Univ.-Prof. Dr. Marc Streit

Submitted on February 24, 2025
for the Practical Work in AI (BSc)
in the WS 2024/25

Abstract One fundamental part of an machine learning task is data selection. The more complex and larger the datasets are, the more important this is.

To select only necessary data and reject any redundant or unrelated ones, will potential increases the model accuracy and decreases complexity and the use of resources during training. The objective of this project is to determine witch weather-data are most influential on glaciers. In order to do so, a interactive tool is provided, to discover different kinds of historical weather-data.

Keywords Visual Analytics • Climate Change • Glacier • Weather • Artificial Intelligence • Machine Learning

1 Introduction and Motivation

Glaciers are often seen as thermometer or a scale[1] in terms of climate changes. They are not sensitive on short term weather variation but on long term changes. The understanding which parameters are responsible for the, at the time, fast shrinking of the ice masses are potentially important to understand the climate change better in general. To predict the behavior of this complex ice structures is difficult. To use the help of machine learning, might be a good idea. But you immediately come across a huge amount of data as the basis for such projects. This is where good data selection into play.

The goal of this project in *Practical Work in AI* is to implement a basic interactive visualization of sample weather data in combination with measured data from glaciers.

For this, the project is split into the following **sub tasks**:

- Find data sources2
- Restrict the area of interest3
- Determine there usability and check the data quality4
- Preprocess the data to make them usable in a visualization framework5
- Implement an interactive visualization6

The focus is first on finding and processing data of high quality and high resolution where needed. And second on the implementation of the user interface.

The tasks are tackled in an iterative approach. So each step gains the knowledge and requirements for the next one.

2 Finding Data Sources

Two datasets are needed:

- Weather data (temperature, wind, ...)
- Glacier data (change of the ice mass)

Weather Data:

The ECMWF (European Center for Medium-Range Weather Forecasts) provides different kinds of suitable dataset. After some research the "*ERA5 hourly data on single levels from 1940 to present*[2]". The ERA5 datasets are widely used in scientific projects. More about the quality of the data in Section 4 ERA5 covers the period from 1940 to present with an data point at every hour on a 31km grid over the whole planet. This with a variety of about 260 variables (not all of them cover the whole planet). The data were produced by the ECMWF Integrated Forecast System (IFS) and were the best I could find. It is also possible to use the ERA5 Land dataset for the Alps with a more dense grid of 9km if needed.

An account is needed to request data from this source. But the account and the requests are free of charge.

Glacier Data:

There exists one dedicated database for glacier measurement, the WGMS[3] (world glacier monitoring service). This organization collects global glacier measurements and provides a regularly updated dataset. Many information and measurements are included like: location, area, mass, and so on.

3 Area of interest

The dataset from WGMS consists >20,000 glaciers worldwide. ERA5 covers the whole world with a 31km grid. Therefore, a restriction is necessary to reduce the amount of data to process. I decided to consider only the Apls without the Massif des Écrins.

WGMS data:

This dataset contains 3909 glaciers. Some of them are subparts from bigger glacier structures but still independent. This fast and the high variaty of different variables (observations) needs further restriction. Goal 10-15 glaciers 1-2 variables.

Most glaciers in this database are very inconsistent in their frequency of observation. During the research I found, there are reference glaciers. They are perfect for this task. Data at least once a year and high quality of each observation. So I picked all 11 reference glaciers in the Alps.

ERA5 data:

The restriction of the weather data is basically the the area of the central alps. This are can be specified when requesting data from the system.

To be precise, this is the area of:

latitude	49° - 45°
longitude	5° - 17°

[1] https://en.wikipedia.org/wiki/Retreat_of_glaciers_since_1850

[2] https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=documentation

[3] https://wgms.ch/data_databaseversions/

4 Usability and quality of the data

The next question was which measurement is the best suitable to describe the glacier change the best. After exploring all possible sub-datasets, the variable *Mass Balance* is the best choice for this. Values like *Area* or *Thickness* are very specific to the landscape and make different shapes of glaciers incomparable to each other.

5 Preprocess the data

The preprocessing of the data to use them in Plotly was an iterative process. I will here state only the last and working approach.

Preprocessing of the WGMS data:
The whole dataset contains 11 individual .csv files with different content. The idea is to get a single .csv file of the following shape:

Table 1: Example of a processed glacier dataset

Date	Name	Annual Mass Balance	Latitude	Longitude
1990	Silvretta	-100.0	46.0	10.0
...	...	...	...	...

Two things have to be done here:

1. **Calculate the annual balance.** The measure of (in millimeters of water equivalent[4]) the mass balance is sometimes done more than once a year. To have one representative value for each glacier every year, the arithmetic mean is calculated whenever more than one value per year is in the list.

2. **Combining location and annual balance.** Hence the values for latitude and longitude are not in the same file as the balance values, they have to be extracted from another file. This is done according to the glacier names.

Preprocessing of the ERA5 data:
The data are provident as a GRIB[5] file. the library pygrib[6] is used to handle this file type.

- The **temperature** values are in Kelvin and have to be converted into °C. Then stored in a .csv file

- **Wind** values are given as a vectors with an U and an V value. Both values are in separate lines in the GRIB file because they are basically two datasets fetched at once. So here the length of and angle of the vector have to be calculated which gives wind speed and direction. And also stored in once .csv file.

6 Interactive visualization

The implementation of the interactive user interface, done in Python, was again an iterative process. Here are only the most recent and important facts stated.

The **framework** of choice is Plotly + Dash[7]. This is because of the high adaptable behavior trough a good low level code implementation. Dash gives a point-and-click web interface. The visualization of the first data shows, Plotly express is to restricted in the function and the low level version of graph objects is better suitable.

Idea: As the reference, glacier data are always visible on the map and line plots. The user select the data point of interest on the map and the corresponding time-frame on the sliders an check boxes on the bottom of the interface.

Structure of the web interface:

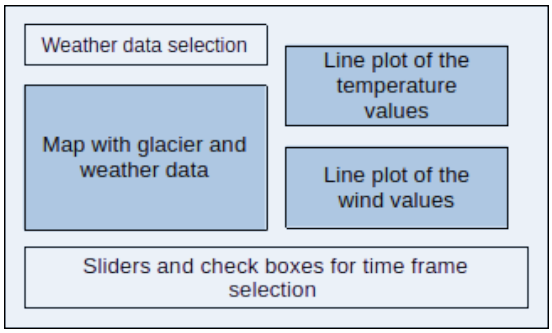

Fig. 1: This is an example of an SVG image.

This is done by 3 subplots in one figure. The object type of the map is a *scattergeo* and *scatter* for the line plots. Each as independent functions, so they can be called according to the selected variable.

So the user can visually observe the progression of the glacier ice mass related to the selected weather variable and time frame. If there is a meaningful corelation, this variable might be a good candidate for a future machine learning project.

7 State of the project and outlook

The user interface at the current state can be seen as a starting point for a more advanced project.

Plotly, especially *scattergeo*, can not handle the high amount of data points when working with datasets like ERA5. Also the approach with subplots in plotly is to restrictive to be used with various weather data. For the further development of the project, a new framework is needed and a more advanced handling of the data itself, like in a proper database. Link to the GitHub project:
https://github.com/viehbo/Thesis_Project_Glacier.git

8 Use of generative AI and web search

The main resource was the manual of plotly and the provided example code.
ChatGPT was mainly used to customize arrows in the scatter objects. But this was ultimately a dead end. Wherever parts of the code or the idea from ChatGPT is used, it is stated as comments in the code. Other sources for from some web searches are: stackoverflow[8] , geeksforgeeks[9]

9 Acknowledgment

The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

[4] https://www.antarcticglaciers.org/glaciers-and-climate/estimating-glacier-contribution-to-sea-level-rise/
[5] https://confluence.ecmwf.int/display/CKB/What+are+GRIB+files+and+how+can+I+read+them
[6] https://jswhit.github.io/pygrib/
[7] https://dash.plotly.com/
[8] https://stackoverflow.com/questions
[9] https://www.geeksforgeeks.org/python-plotly-tutorial/