EDI Internship: MetRecord Software Developer

Lead Supervisor: Hua Lu

Project description:

The Antarctic Peninsula (AP), the northernmost part of the Antarctic continent, is located to the east of the '*pole of variability*', a region in the South Pacific where the mean sea-level pressure is more variable than any other places in the southern hemisphere. The region has experienced the largest variability in annual mean temperature and sea ice concentration and has been projected to become significantly warmer with sustained high temperature events above freezing during summer months. This indicates significantly enhanced surface melt by the end of the century. Understanding climate variability at the AP is thus very important for Antarctic glaciological, oceanographic, chemical, and biological processes. However, due to large natural variability of the region, determining long-term climate impact remains challenging because there are very few in-situ observations before 1979.

Since 1944, some continuous or intermittent surface weather observations were taken at fifteen stations near the AP by British Antarctic Survey (BAS) and/or its antecedents. Among these fifteen stations, only four stations are still in operation now-a-days. But paper-based historical meteorological records taken from those closed stations are archived at British Antarctic Survey and/or Met Office. Those records cover a range of subperiods between 1944 and 1967. Up to date, these pre-satellite era weather observations have been severely under-harnessed due to labour-intensive, tedious manual digitization.

With recent advances in artificial intelligence and machine learning for text transformation and recognition, it becomes possible to extract those valuable historical records both rapidly and accurately. A BAS team has recently developed a trial version of Optical Character Recognition (OCR) software that converts scanned images with handwritten, historical weather data into a fully digitized format. The software or BASOCR reduces the extensive manual data entry as well as setting up templates for various form layouts. This project aims to pursue excellence in digital technology, environmental science, and interdisciplinary research through advancement of diversity, equity, and inclusion, with a specific focus on improving the accuracy and the speed of BASOCR. By working collaboratively with the existing team, we aim to lift BASOCR to a new level so that it can digitise over 5000 pages of handwritten records on old paper-based weather observation registry covering the AP region rapidly so that those valuable data sets can be more widely used for climate research.

Objectives and Methodology: the purpose-built digitization software or BASOCR has an OCR engine that converts the handwritten numbers into machine-encoded digits. An accurate conversion of BASOCR highly depends on the quality of input images. The EDI intern will work collaboratively with existing BASOCR team to develop a customized image pre-processor to improve the quality of input images so that the BASOCR can attain its best outcome. The intern will also work with the software engineer to make BASOCR run parallel with BAS's HPC clusters.

Job description:

The student will be based at the British Antarctic Survey in Cambridge and work collaboratively with the existing team members to further develop BASOCR. Under the supervision of Dr Hua Lu, the student will develop a python-based pre-processor and to refine the existing code to speed up BASOCR via parallel computing.

The first part of the work involves developing a customized pre-processor for BASOCR by taking full advantage of the state-of-art imaging processing. The newly developed pre-processor will not only improve the quality of the scanned input images but also enable an automated procedure that recognizes different meteorological forms and routes different record images to the most appropriate, pre-trained BASOCR models. The second part of the work is to optimize BASOCR to improve both its accuracy and speed via parallel computing and other digital technologies.

By working with a core team that is gender-balanced and multi-ethnical, this project will provide science inspiration, research training and hand-on experience to a junior researcher from an underrepresented background, in the areas of advanced digital technology and polar meteorology. The student will have the opportunity to interact with other scientists specialising in atmospheric dynamics, polar ocean, sea ice, climate modelling, and analysis of big data. Also, the student will have the opportunity to present the results at the BAS Atmosphere-Ice-Climate fortnightly meetings. This project would suit applicants with previous programming experiences or recent graduates studied in computing or IT related fields. Working knowledge in Python and image processing will be advantageous.

Project work:

Depending on possible changes of the government Covid rules, the project will be carried out in a hybrid mood with remote and in-person work.

All the input images are accessible via Teams, which can be downloaded remotely. The student will need the access to BAS HPC and other computing facilities because BASOCR is currently running on BAS HPC. The main supervisor will set-up a rota with weekly meetings and item-by-item plan to address the issues that the intern is most likely to face. The intern will be introduced to the team and attend the PCAP weekly meeting and the AIC monthly meeting.

The supervisory team will meet the student in person at BAS office at least once a week and more often in the first couple of weeks to get the student up to speed with the project. Additional supervision will be provided with mixed Zoom/Teams meetings and email exchanges.