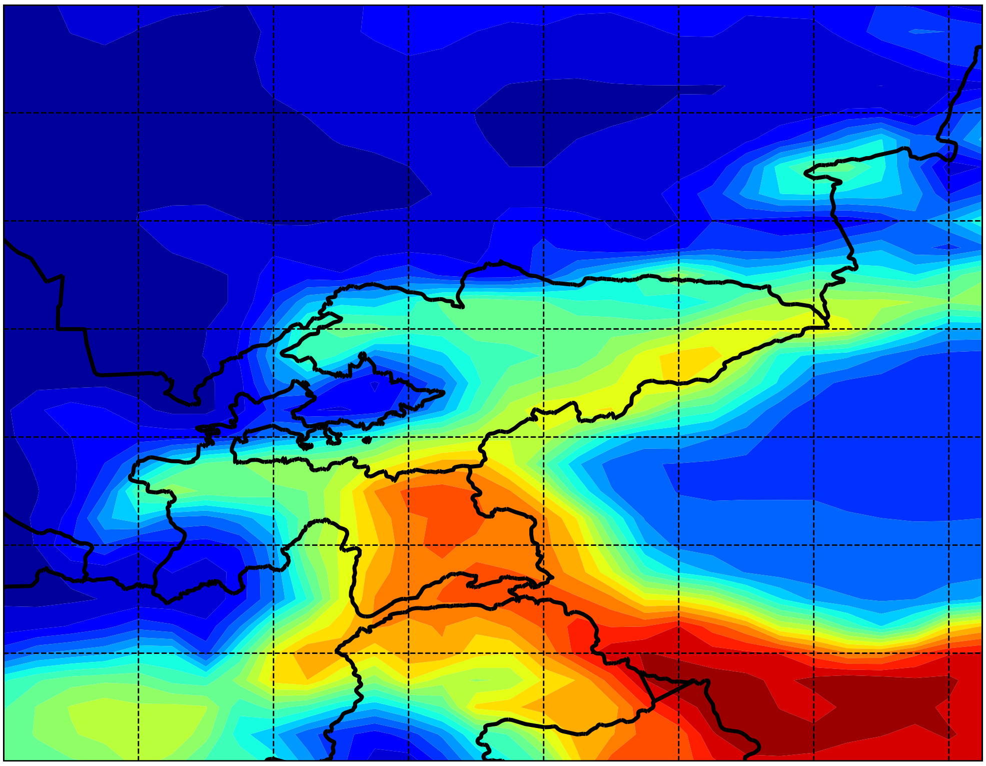
American University of Central Asia

Internship Report RS RAS



Place of Internship

Research Station

of the Russian Academy of Sciences, Bishkek

Program

Applied Mathematics and Informatics

Student

Moritz M. Konarski

Group

2017

November 11, 2020 – Bishkek

Contents

1 Introduction

This report covers the tasks and results of my internship at the Federal State Budgetary Institution of Science Research Station of the Russian Academy of Sciences in Bishkek (RS RAS). RS RAS is a subsidiary of the Ministry of Science and Higher Education of the Russia Federation and employs 137 people. Since 1978 it has been researching seismic processes and developing geodynamic models. The internship took place from the 7th of September 2020 to the 7th of November 2020 and was conducted remotely due to the continuing COVID-19 pandemic. My AUCA supervisor for this internship was Olga Zabinyakova, Scientific Secretary of RS RAS and my RS RAS supervisor was Sanzhar Imashev, Acting Head of the Laboratory for Integrated Research of Geodynamic Processes in Geophysical Fields.

The internship was split into an educational section and an industrial section. According to my internship dairy form, the aim of the educational internship was to acquire the knowledge necessary to understand the research carried out by RS RAS. During the industrial part of the internship I should participate in a certain part of their work or in work that is similar to theirs. To fulfill these requirements my RA RAS supervisor gave me the following tasks for my educational internship:

1. familiarize yourself with web resources providing access to NASA Earth Remote Sensing data;
2. familiarize yourself with the scientific data format netCDF (Network Common Data Form);
3. study libraries used to work with the netCDF format in various computing environments.

For the industrial internship I was tasked to:

1. register on the NASA Earthdata platform to access satellite data;
2. develop a library for working with netCDF files in the Python programming language (using satellite data as an example);
3. develop a computer application for data visualization and reanalysis of NASA MERRA2 satellite data.

These tasks are outlined in my internship diary form. In the next section of this report I will cover the educational part of the internship and talk about NASA Earth Remote Sensing data, the netCDF data format, and libraries used to work with netCDF files. Then, I will detail the industrial part of the internship and talk about registering on the NASA Earthdata platform, downloading NASA MERRA2 data, and the development of the Python application.

2 Educational Internship

NASA Remote Sensing data is available via the Earth Science Data Systems (ESDS) Program (see `https://earthdata.nasa.gov/esds`). The program covers the data acquisition, processing, and distribution with the goal to enable the widespread use of NASA mission data. The data is available for free and their software is publicly available as Open Source Software.

Part of ESDS is the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) which provides data on atmospheric composition, water and energy cycles, and climate variability. The GES DISC provides over 3.3 Petabytes of data, including the MERRA2 dataset. MERRA2 (the Modern-Era Retrospective analysis for Research and Applications version 2) focuses on historical climate reanalysis using satellite data. The dataset I worked with (M2I3NPASM) contains data from January 1st, 1980 until October 1st, 2020 (at the time of writing). It covers the whole globe with measurements taken every 3 hours. M2I3NPASM includes 14 measured variables in addition to latitude, longitude, time, and a pressure level. The measured variables include the surface pressure, specific humidity, eastward and northward wind, and temperature. The measurements are done on a cube sphere grid (add explanation) and later processed to fit the standard latitude and longitude grid. This processing generally involved a bilinear interpolation of data values.

The data is provided on the GES DISC website (see `https://disc.gsfc.nasa.gov/datasets/M2I3NPASM_5.12.4/summary`) and can be accessed from there. The website gives the option to download only a subset of the data by selecting a certain time range, latitude and longitude range, and group of variables. This is advantageous because a full file for 1 day is about 1.1 GB in size. For my internship I worked with a subset of the data that includes all variables but is restricted to a latitude of 34°N to 48°N and a longitude of 65°E to 83°E. This restricts the area to Kyrgyzstan and sections of all surrounding countries. Additionally, the file size is reduced to a manageable 6 MB per file so that a whole year of data only takes up 2.2 GB (in 365 files), the same space two files of the complete data would take up.

The M2I3NPASM data is provided in a file format called the Network Common Data Form (netCDF). NetCDF is made up of a data format and libraries that can read and write its data. NetCDF is developed and maintained by Unidata, a community of research institutions, with the goal of sharing geoscience data and the tools to use and visualize it. Unidata is funded by the National Science Foundation, a US government agency that promotes and supports science and research. Unidata also maintains libraries (programming interfaces) for C, Java, and Fortran. Based on these, multiple other interfaces are available, including one for Python.

The netCDF data format is specifically designed to hold scientific data. According to the Unidata website, the netCDF data format has the following features:

- **Self-Describing**. A netCDF file includes information about the data it contains.
- **Portable**. A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- **Scalable**. Small subsets of large datasets in various formats may be accessed efficiently through netCDF interfaces, even from remote servers.
- **Appendable**. Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- **Sharable**. One writer and multiple readers may simultaneously access the same netCDF file.
- **Archivable**. Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

The NASA M2I3NPASM data is available in ”classic” netCDF–4 format, meaning that it is backwards compatible. These files have 4 dimensions:

1. longitude in degrees east (meaning west is represented as negative),
2. latitude in degrees north (making south negative),
3. pressure in hPa,
4. time in minutes since the first time point in a file.

To be self-describing, the NASA M2I3NPASM netCDF files contain information about, among others, the intitution that created the file, date and time of the beginning and end of the dataset, and the minimum and maximum latitude and longitude values. The files furthermore contain metadata for each of the measured variables. The most important of these are the fill values that identify missing data, the long name, a full version of the short variable name abbreviation, and the units of the variable.

Because I am working with the Python programming language for my internship I require a netCDF library that works in that programming language. Unidata provides an interface between the netCDF library for the C programming language and Python. This interface is called netCDF4. It has most of the features of the C library and enables the creation and reading of netCDF files using Python. This interface is used in my industrial internship work to work with the netCDF files downloaded from GES DISC.

3 Industrial Internship

To access data on the NASA Earthdata platform (which includes GES DISC), a registered account is required. Then, that account needs to be linked to a GES DISC account.

4 Data Processing Library

What does it do? How does it do it? What are the implementation details?

5 Plotting Library

What does it do? How does it do it? What are the implementation details?

6 GUI

What does it do? How does it do it? What are the implementation details?

7 Conclusion

One of the desired outcomes of educational and industrial practitioner is the formation of research directions and topics student's graduation bachelor's work.

- what did I learn during the internship
- which skills did I receive

References

[1] "NumPy reference," 06.2020. Release 1.19.0. Retrieved from: https://numpy.org/doc/1.19/numpy-ref.pdf.

[2] W. KcKinney and Pandas Development Team, "Pandas: powerful python data analysis toolkit," 08.2020. Release 1.1.3. Retrieved from: https://pandas.pydata.org/docs/pandas.pdf.

[3] "The netcdf user's guide," Accessed 23.10.2020. URL: https://www.unidata.ucar.edu/software/netcdf/docs/user_guide.html.

[4] "Netcdf4 module documentation," Release 1.5.4. Accessed 23.10.2020. URL: https://unidata.github.io/netcdf4-python/netCDF4/index.html.

[5] J. Hunter, D. Dale, E. Firing, M. Droettboom, and the matplotlib development team, "Matplotlib," 09.2020. Release 3.3.2. Retrieved from: https://matplotlib.org/3.3.2/Matplotlib.pdf.

[6] "Cartopy documentation," Release 0.18.0. Accessed 23.10.2020. URL: https://scitools.org.uk/cartopy/docs/latest/.

[7] "Python documentation: json – json encoder and decoder," Release 3.9.0. Accessed 23.10.2020. URL: https://docs.python.org/3/library/json.html.

[8] "Python documentation: Built-in types – dict," Release 3.9.0. Accessed 23.10.2020. URL: https://docs.python.org/3/library/stdtypes.html#typesmapping.

[9] "Python documentation: datetime – basic date and time types," Release 3.9.0. Accessed 23.10.2020. URL: https://docs.python.org/3/library/datetime.html.

[10] "Ipython documentation," Release 7.18.1. Accessed 23.10.2020. URL: https://ipython.readthedocs.io/en/stable/.

Appendices

A Code Listings

What does it do? How does it do it? What are the implementation details?

B Supplemental Graphs

put graphs here that don't fit into the main body

C Tables

tables that contain data values for illustration maybe