

Web-based, user-driven climate impact data extraction

Jochen Klar and Matthias Mengel
Potsdam Institute for Climate Impact Research (PIK)

Date of submission: May 19, 2023
Required funding: 1 FTE for 6 months

Abstract

The ISIMIP Repository¹ of the ISIMIP project² holds the world's largest collection of global climate impact model data. However, both the format (NetCDF³) and file sizes represent a major barrier for many users. We here propose to build an innovative web-based service that allows users to extract, process and download subsets of the data. User-defined extraction, chained processing and data interaction through scripts and interactive Jupyter notebooks⁴ will largely widen the user base. Users can initiate processing tasks in the cloud and download the resulting files in different formats. The code will be released as open source software and, as the application is not tied to ISIMIP or the ISIMIP conventions, can be adopted for similar archives of NetCDF files.

I. Introduction

The Inter-sectoral Impact Model Intercomparison Project (ISIMIP) brings together worldwide climate impact modeling communities through a consistent modeling setup across sectors⁵, the provision of high-quality forcing data and the curation of the output data in the ISIMIP Repository. As climate impact data provides the link between climate change and society, there is a high demand for reuse of the data across many fields. Data is typically stored on a grid covering the global domain with typical file sizes of hundreds of GB and uses the NetCDF format, which is very common in climate science, but lacks distribution in other communities. The size and the format are strong barriers for many users. In the past, the ISIMIP data team needed to manually prepare special conversions of datasets for download, which was a cumbersome and time consuming task. Many users are only interested in data on particular countries, regions or individual places with much smaller potential sizes than the source files. To overcome these barriers, we initiated the development of a web-based service as part of the BMBF funded ISI-Access project: the ISIMIP Files API enables the user to interact with the storage cloud via the internet from anywhere in the world, run a predefined set of regularly requested processing tasks and download the results (e.g. as .CSV). As of now, the API is able to perform cut-outs of regions, masking of areas and countries and selecting time series for points, countries or areas. The service is already in use and enjoys great popularity. Based on

¹ ISIMIP Repository, <https://data.isimip.org>, online.

² ISIMIP Website, <https://www.isimip.org>, online.

³ NetCDF, <https://www.unidata.ucar.edu/software/netcdf/>, online.

⁴ Jupyter Notebooks, <https://jupyter.org>, online.

⁵ ISIMIP Protocol, <https://protocol.isimip.org>, online.

the feedback we received from our community, we would like to extend the functionality of the API through the proposed NFDI4Earth Incubator Project.

II. *Incubator Project description*

The ISIMIP Files API is developed as a web based application, is written in Python using the popular Flask framework⁶ and the RQ⁷ library and is already publicly available as Open Source Software on GitHub⁸. As input, it expects one or many file paths and additional arguments like a point, an area, or a country, and then, based on these arguments, it starts asynchronous tasks on the server. These tasks use standard NetCDF tools like cdo⁹ and nco¹⁰ to perform the desired operations on the files specified by the input path and create new files which are at first stored on the server and then downloaded by the user via a regular web server. The API makes no further assumptions on the files besides that they are globally gridded NetCDF files. In particular, no ISIMIP internal conventions are used. It can therefore be reused by other archives easily. Currently it is mainly used through the ISIMIP Repository.

The proposed project has two main tasks:

Task 1 (Month 1 to 4) Add key functionality to API

User-defined extraction, chained operators (1a): We implement the functionality to upload user-defined shapes, e.g. ocean regions or river catchments, to perform cutout and extraction. We plan to provide compatibility with GeoJSON and shapefile formats (with file specifications to be specified). The extracted shapes can be further processed, e.g. their data averaged or summed and then downloaded as time series. Inspired by the cdo feature of chained operators¹¹, we implement a similar feature in the ISIMIP Files API. A use case for this would be the extraction of the land area in a rectangular bounding box. Operators will be taken from a predefined list of nco and cdo commands, which ensures flexibility but also security.

Add script and Jupyter interaction and scripting capabilities (1b): We implement a user-friendly connection to the service via scripts and Jupyter notebooks. In addition to the already available integration into the ISIMIP Repository, this will enable the users to access all functionality and all newly developed functions in a programmatic way following current practice in data science. It will also allow the users to run larger batches of tasks in a reproducible way. The employed robust queuing system prevents overuse of the API and ensures availability for all users.

Task 2 (Month 4 to 6) User feedback and development towards reusable package

Towards the objective of wide usability, we will advertise the functionality through the ISIMIP mailing list and collect feedback after Task 1 is concluded, including additional operations on the files and required data formats. We will then revise the features accordingly and add missing functionality. At the end of month 6, we will release the package on GitHub and on PyPI¹² for a convenient installation using the “pip” tool. We will use automatic tests and continuous

⁶ Flask, <https://flask.palletsprojects.com/en/2.1.x/>, online.

⁷ RQ, <https://python-rq.org/>, online.

⁸ ISIMIP File API on GitHub, <https://github.com/ISI-MIP/isimip-files-api/>, online.

⁹ CDO, <https://code.mpimet.mpg.de/projects/cdo/>, online.

¹⁰ NCO, <http://nco.sourceforge.net>, online.

¹¹ Tutorial Combining Operators in CDO, <https://code.mpimet.mpg.de/projects/cdo/wiki/Tutorial>, online.

¹² Python Package Index (PyPI), <https://pypi.org>, online.



integration (using pytest and GitHub actions) to ensure the quality of the developed code and to prevent the introduction of bugs into the existing functionality. All new functionalities will be implemented in the existing code base.

The work on this Incubator Project requires both experience in software development (web based services) and data science in the earth and environmental sciences (scientific workflow, Jupyter notebooks, NetCDF processing). We therefore apply for 6 month full time equivalent PostDoc funding for Dr. Jochen Klar, who is the main developer of the ISIMIP Files API.

III. *Relevance for the NFDI4Earth*

Existing users of ISIMIP data in the earth and environmental sciences and related fields will benefit from the work, since they can use the extended functionalities to better customize their data downloads to fit their specific tools and workflows. We expand this user group to researchers who are not familiar with the NetCDF file format and are without access to supercomputers. Beyond that, transdisciplinary and non-scientific users will benefit from the simplification of the data access, in particular those with poorer infrastructure or internet access. Data curators and infrastructure providers can freely install and reuse the proposed package. This is attractive for all providers that host archives of NetCDF data, do not provide processing of it on the server so far and do not have funding to implement large cloud-based solutions like DKRZ's jupyterhub¹³. This applies to many of the listed NFDI4Earth repositories and infrastructures.

IV. *Deliverables*

Deliverable 1 (end of month 4)

New features implemented: User-defined shapes and chained commands, Jupyter notebooks and script interaction available on GitHub.

Deliverable 2 (end of month 6)

Revised version with documentation released on GitHub and PyPI. This will include example scripts and Jupyter notebooks.

All features developed as part of this proposal will be integrated into the ISIMIP Repository and will be usable by the community when downloading data from the Repository.

¹³ Jupyterhub@DKRZ, <https://docs.dkrz.de/doc/software%26services/jupyterhub/index.html>, online.