A Machine-actionable Workflow for the Publication of Climate Impact Research Data from the ISIMIP Project

Jochen Klar, Matthias Mengel

Potsdam-Institut für Klimafolgenforschung (PIK)

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) aims to provide a quantitative and cross-sectoral synthesis of the different impacts of climate change. To this end, a simulation protocol defines a set of common experiments that are valid across sectors. Modelling groups around the world run their simulations following this protocol using climate and socio-economic forcing data provided by ISIMIP. The impact model output data is collected by the ISIMIP team at PIK and made publicly available via the ISIMIP repository. In total, more than 100 TB of up-to-date climate impact simulations are freely available. The data is broadly used by the international scientific community, but also by economic and civil society actors.

The workflow of data submission and publication was considerably improved by the the introduction of a machine-actionable protocol and the development of several interlinked software tools for data maintenance, quality control and data publication. While the specific implementation is tailored to ISIMIP, the general ideas should be transferable to other projects.

1 Introduction

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; Frieler et al. 2017) is a community-driven climate impact modeling initiative that aims to contribute to a quantitative and cross-sectoral synthesis of the various impacts of climate change, including associated uncertainties. It is designed as a continuous model intercomparison and improvement process for climate impact models and is supported by the international climate impact research community. ISIMIP is organized into simulation rounds, for which a simulation protocol specifies a set of common experiments. The protocol further describes a set of climate and direct human forcing data to be used as input data for all ISIMIP simulations. Based on this information, modelling groups from different sectors (e.g. agriculture, biomes, water) perform simulations using various climate impact mod-

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18077

els. After the simulations are performed, the data is collected by the ISIMIP data team, quality controlled and eventually published on the ISIMIP Repository. From there, it can be freely accessed for further research and analyses. The data is widely used within academia, but also by companies and civil society. ISIMIP was initiated by the Potsdam Institute for Climate Impact Research (PIK) and the International Institute for Applied Systems Analysis (IIASA).

In this paper we describe the data publication workflow from the modelling groups to the end user in detail (see Figure 1 for an overview). This workflow ensures a high data quality and follows the FAIR data principles (Wilkinson et al. 2016).

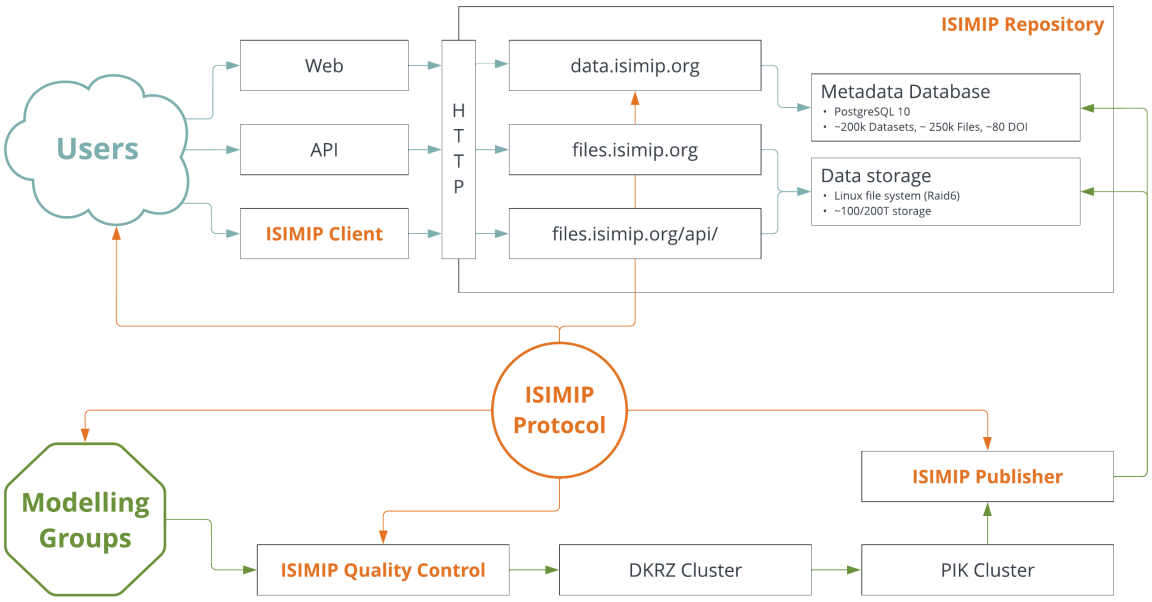

Figure 1: Overview of the ISIMIP publication workflow: After the simulations are performed by the modelling groups, the data is checked using the ISIMIP Quality Control tool. Only if the data passes all checks, it is made internally available to the modellers community at the ISIMIP space at the German Climate Computing Centre (DKRZ) and the PIK cluster. Finally, it is published on the ISIMIP Repository, using the ISIMIP Publisher. From there, the data can be retrieved by users all over the world, via the web page of the Repository, using it's API, or with the ISIMIP Client library. All components of the workflow are using information from the machine-actionable ISIMIP Protocol.

2 The machine actionable Protocol

The simulation protocol is used by the modeling groups contributing to ISIMIP to set up their simulation runs. In the past, the simulation protocol was a text document of ca. 100 pages, which was collaboratively edited using Microsoft Word. The protocol was published as a PDF document and subsequently updated every few months. It took considerable

effort to manage the different changes in the protocol and it was not possible to update the protocol in an appropriate frequency to follow the development of the models in the different sectors. Since the chapters for the different sectors where edited independently, inconsistencies in the output variables occurred frequently.

To address these issues, the protocol for the current simulation rounds was implemented using a novel approach. Instead of one large document, the information is now stored in a set of smaller documents. The text part of the protocol, which contains mainly the introduction and notes on specific sectors, is written in markdown. Structured information is stored in JSON files. Both file formats are plain ASCII and are therefore perfectly suited to be version-controlled using Git. A set of Python scripts was developed to create a human readable, interactive web page[1] and a set of static sector specific, machine readable JSON files[2]. Using GitHub Actions we are able to run the build process every time a new commit is pushed to the GitHub repository. In addition to the JSON definitions, the protocol contains file name patterns for each sector. These patterns are regular expressions which convert the file names of the input and output files to a dictionary of specifiers (cp. Figure 2).

Figure 2: Metadata extraction using regular expressions: The file name consist of a set of specifiers which are defined in the ISIMIP Protocol. Using regular expressions we validate the file names and extract the specifiers into a dictionary.

1 ISIMIP protocol: `https://protocol.isimip.org`.
2 e.g. `https://protocol.isimip.org/definitions/ISIMIP3a/OutputData/water_global.json`.

3 Quality control

After the model groups finished their simulations, the data, which is stored in the NetCDF[3] format, is transferred to the ISIMIP data space at the German Climate Computing Centre (DKRZ), where the ISIMIP team performs different quality control checks on the data to ensure that the data products are in agreement with the specifications of the protocol. In the past, this was done using a number of different scripts which were hard to maintain. In order to connect the quality control with the new, actionable protocol, we developed a dedicated command-line tool written in the Python programming language. Its source code is available on GitHub[4] and it is published on the Python Packaging Index[5]. The tool can be used independently by the modelling groups to find and fix most errors *before* uploading terrabytes of data.

With only the *protocol path*, which specifies simulation round, data product, and sector[6] as argument, the tool is able to fetch the *current* sector specific machine readable protocol via the internet. The tool then scans the current directory and its subdirectories recursively. For each file individually, it checks its filename against the file pattern from the protocol. Only if a filename matches, the file is opened and a series of checks (NetCDF data model, compression, dimensions, grid, global attributes, variables, units) is performed. The tool is also able to perform limited checks on the actual data stored in the files. As part of the protocol, each variable has a minimum and maximum value assigned to it. Examples include negative water consumption or unnaturally high tree heights.

If one of the steps fails or any other information needs to be communicated to the user, the program gives a meaningful response, so the users can alter the files accordingly. Some of the less severe errors can be fixed by the tool itself (if the user so chooses).

4 Data publication

After the data is checked and an optional embargo period has passed, the data is published on the ISIMIP Repository. From a technical point of view, the repository contains three main components:

1. A file server, which can be accessed via the internet using a simple web server[7]. As of 2023, the server has a storage capacity of about 200 TB.

2. A relational database, which contains metadata entries for all files on the file server, as well as entries for *datasets*, in which files with the same variable and from the same experiment, but for different decades are combined.

3 Network Common Data Form (NetCDF):`https://www.unidata.ucar.edu/software/netcdf`.
4 ISIMIP Quality Control on GitHub: `https://github.com/ISI-MIP/isimip-qc`.
5 ISIMIP Quality Control on PyPI: `https://pypi.org/project/isimip-qc`.
6 e.g. `ISIMIP3a/OutputData/agriculture` for the agriculture sector in the ISIMIP3a simulation round.
7 ISIMIP files server: `https://files.isimip.org`.

3. The main repository web page[8], which allows for a convenient search of the metadata database and guides the user to the corresponding downloads from the file server.

The publication process contains a number of processing steps, which are performed using the specifically developed command-line-tool ISIMIP Publisher[9]. First, the files to be published are copied to a working directory on the file server. There, the file name patterns (cp. Section 2) are used to extract the metadata from the file names. Files are then logically combined to datasets, so that all files for the same experiment and variable, but different years are contained in one dataset. Each dataset and file is assigned a UUID as unique persistent internal identifier (called `isimip_id`) and a checksum is computed for each file. For each dataset and file, an entry is created in the database, containing the described metadata, a version string based on the date (e.g. `20230101`), and the licence under which the dataset is published. In the final step of the publication process, the files are moved to the public directory on the server, and the entries in the database are included in searches on the repository website.

The Repository assigns Digital Object Identifiers (DOI) to all datasets corresponding to a specific sector and simulation round (e.g., Marcé et al. 2022). The DOI are registered with DataCite (PIK is a member of the German DataCite consortium). The metadata according to the DataCite Metadata Schema 4.4 is manually prepared and stored in the database with a link to all datasets that are referenced by the DOI. This rigid link between the DOI and the specific version of the data ensures traceability and reproducibility.

When datasets need to be replaced because problems were discovered or improved data is available, the old datasets are archived. The metadata for archived datasets remains in the database, a corresponding page remains online, and the datasets are available on request. When the new dataset is published, the repository links the new and the old dataset together. When files are replaced, a new DOI is created for the sector. The old DOI page stays online and a reference to the new DOI is added. A caveat system[10] provides the users with information about changes to the data.

To further improve data access, a *Configure Download* option can be used to perform operations on the server, e.g. a cut-out of a specific country or region or the extraction of a time series for a point as CSV. This functionality is provided by a dedicated web service[11]. The Repository can also be accessed by it's API and a dedicated Python client library[12]. This programmatic access can be used in scripts and Jupyter notebooks to search and download large sets of data directly.

8 ISIMIP Repository: `https://data.isimip.org`.
9 ISIMIP Publisher on GitHub: `https://github.com/ISI-MIP/isimip-publisher`.
10 ISIMIP Caveats and Updates: `https://data.isimip.org/caveats`.
11 ISIMIP Files API on GitHub: `https://github.com/ISI-MIP/isimip-files-api`.
12 ISIMIP Client on GitHub: `https://github.com/ISI-MIP/isimip-client`.

5 Discussion and Outlook

As the world's largest data archive of model-based climate impact data, ISIMIP output data is used by a diverse audience inside and outside of academia for research and analyses. It is therefore crucial to ensure a high-quality of the published data, both formally and content-wise. Using the machine-readable protocol as *single source of truth* for quality control, metadata extraction and data publication has greatly improved our workflow. Using a git-based workflow allows us to adjust and track the changes suggested by the impact modeling sectors. Changes only need to be made in one place, and the history of the protocol is available in a transparent and open way.

The presented workflow has proven useful to ensure the conformance of the provided data with the simulation protocol throughout the whole data publication chain. In particular the possibility for modellers to check their output files prior to the upload to the ISIMIP data space, using the same tools as in the final quality control step, has been productively adopted by several modelling groups.

Although the work presented is tailored to the ISIMIP project, we believe it can be seen as a best practice example for similar collaborations that have a common set of simulations experiments and a data curation process. The combination of a machine-readable protocol, automatic deployment on the Internet using Continuous Integration (CI), and subsequent use by tools and services can be useful for many data intensive research areas. All the components presented are available as open source software and can be adopted for other contexts.

In the future, we will extend the quality control process, which currently only checks for formal deviations from the ISIMIP protocol and some minimum or maximum violations, to include a quality assessment of the data content itself. We are working on a new tool[13], which is able to automatically check model outputs against an ensemble of already published models. This will allow us to identify and correct not only formal errors to formats and naming conventions, but errors in the data itself.

Acknowledgements

This research has received funding from the German Federal Ministry Ministry of Education and Research (BMBF) under the research project ISIAccess (16QK05) and from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821010.

References

Frieler, Katja, Stefan Lange, Franziska Piontek, Christopher P. O. Reyer, Jacob Schewe, Lila Warszawski, Fang Zhao, et al. 2017. "Assessing the impacts of 1.5°C global

13 ISIMIP Quality Assessment on GitHub: `https://github.com/ISI-MIP/isimip-qa`.

warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b)". *Geoscientific Model Development* 10 (12): 4321–4345. DOI: https://doi.org/10.5194/gmd-10-4321-2017.

Marcé, Rafael, Donald Pierson, Daniel Mercado-Bettin, Wim Thiery, Sebastiano Piccolroaz, Bronwyn Woodward, Richard Iestyn Woolway, et al. 2022. *ISIMIP2b Simulation Data from the Local Lakes Sector.* Version 1.0. DOI: https://doi.org/10.48364/ISIMIP.563533.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.