Implementation of an InfraStructure for dAta-BasEd Learning in environmental sciences (ISABEL)

Marcus Strobl[1], Elnaz Azmi[1], Balazs Bischof[2], Alexander Dolich[2], Sibylle K. Hassler[2,3], Mirko Mälicke[2], Ashish Manoj Jaseetha[2], Jörg Meyer[1], Achim Streit[1], Erwin Zehe[2]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology;
[2]Institute of Water and River Basin Management, Chair of Hydrology, Karlsruhe Institute of Technology;
[3]Institute of Meteorology and Climate Research, Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology

The quantity and diversity of digital environmental data are growing, but they are often inaccessible due to the lack of metadata, inconsistent formats, and local storage of data. ISABEL aims to solve this problem by advancing the V-FOR-WaTer virtual research environment (VRE), which provides a user-friendly web portal for scientists to access and share data from various sources. The portal includes tools for data processing, scaling, and complex analysis, with contributions from both developers and users. Shareable workflows ensure reproducible analysis, to advance research in hydrology and environmental sciences.

1 Introduction

Observational data serve as a fundamental building block for developing a deeper comprehension of ecological systems, either through data-driven approaches or by comparing the data with model predictions. Nevertheless, a significant portion of this data can be challenging to access and often lacks adequate metadata descriptions. Consequently, the data requires significant effort to be useful for science. Accessing, preparing and (pre)processing of this data can be incredibly time-consuming, particularly when attempting to combine datasets from various sources. In the end, the results are often not reproducible (Hutton et al. 2016; Stagge et al. 2019). The ISABEL project aims to improve the situation by providing findable, accessible, interoperable and re-usable (FAIR; Wilkinson et al. 2016) hydrological data and tools to its users through a single entry-point: the user-friendly V-FOR-WaTer web portal (Figure 1).

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.c18091

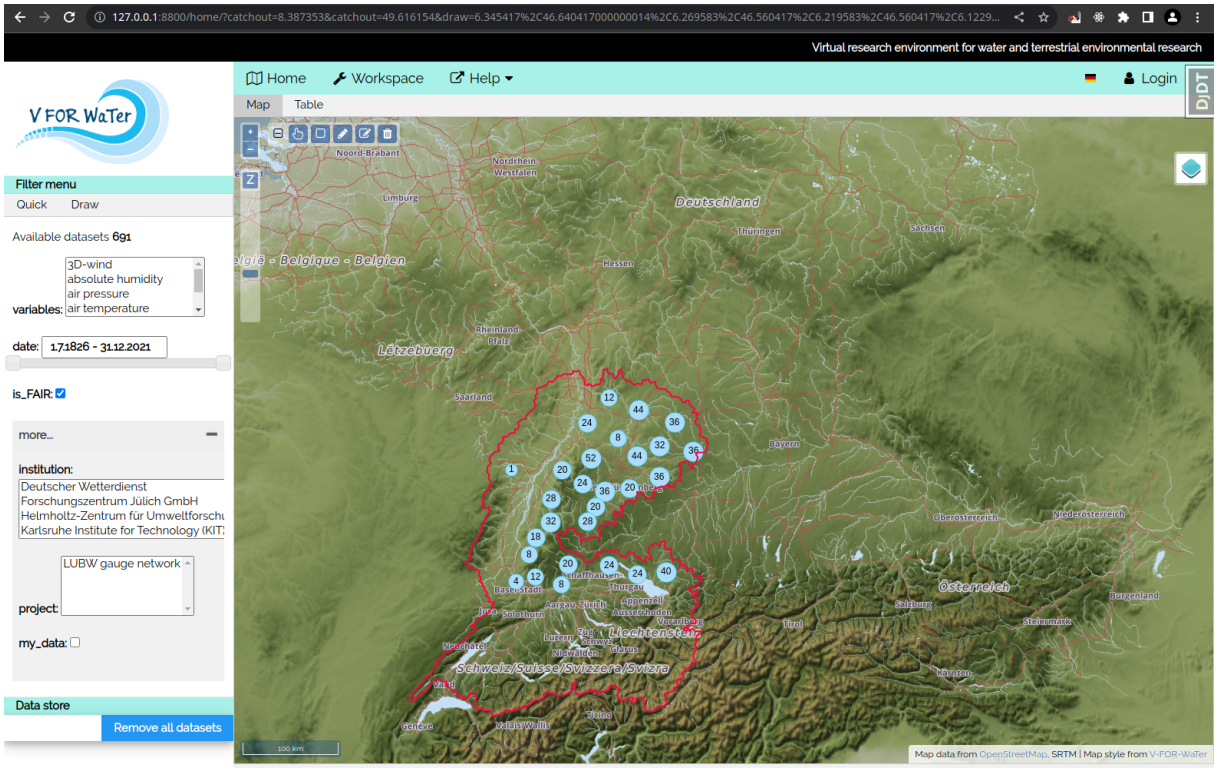

Figure 1: Screenshot of the V-FOR-WaTer web portal. Shown is the filter menu on the left, and data filtered within the upper Rhine catchment on a map.

2 The V-FOR-WaTer portal

Starting as part of the E-Science initiative of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry for Science, Research, and Arts Baden-Württemberg), V-FOR-WaTer has been developed to foster access and management of diverse hydrometeorological data and provide tools for preprocessing, standard hydrological procedures, and more sophisticated analyses (Strobl et al. 2022). Within the DFG-funded project ISABEL, we further develop the virtual research environment to (i) considerably expand its scientific scope, the toolbox and its user-friendliness, (ii) broaden the spectrum of hosted data to include data from state offices, complex data structures and important remote sensing products and (iii) provide access to data and tools in a modern, secure, and responsive web portal with GIS functions and a drag and drop functionality to connect tools and data for building workflows.

The portal already incorporates data from a variety of sources, including state offices and university projects. New data is gradually being added to the portal by the ISABEL team, and the data schema is continuously extended to accommodate new data types. Furthermore, an interface for open data repositories is being developed to make the most important datasets accessible in the Virtual Research Environment (VRE). This way, the web portal becomes a comprehensive resource for accessing hydrological data. The portal offers various features that facilitate the sharing of data in a metadata scheme. Users can

share their data in different common file formats, facilitating import and export of new data. The metadata can also be shared following ISO 19115 in a standardised way. In the current version, data is provided in CSV and XML formats. However, we are also preparing to support additional file formats for exportation, such as Shapefile, NetCDF, and JSON. Access management is implemented to protect critical data and maintain the ownership of unpublished data, ensuring that only authorized users can access it. In the final version, the portal will have interfaces with existing data repositories, allowing scientists to publish their data directly from the portal. To meet standards for data publication, we maintain a close collaboration with the GFZ Data Services repository to work on interfaces, both for accessing their published data and for enabling publication of data with a Digital Object Identifier (DOI) from V-FOR-WaTer in their repository. In the productive version, these features will facilitate the secure and convenient sharing of data with other members of the scientific community.

Data processing within the portal is facilitated through integrated tools for pre-processing and scaling of the data. Currently, the tools are implemented by the portal developers; in future versions, users will have the possibility to include their own tools as well, making the toolbox development a collaborative community effort. The toolbox already contains processes for geostatistics (Mälicke 2021), variogram analyses and kriging tools, hydrostatistics and visualization. More tools such as uncertainty package, GIS tools, data scaling, evaporation and Eddy covariance tools are being added within the scope of the ISABEL project. Workflows can be composed and customized easily via drag and drop. Users can also store their workflows, and the upcoming feature to share workflows will ensure reproducible data analysis. These features render the portal's data processing efficient and user-friendly.

In 2022 and 2023, we started two projects as case studies to actively use the portal to access data and contribute to the toolbox for testing. The aim is to ensure that V-FOR-WaTer covers a wide range of hydrological research and practical applications. These projects require integration of a variety of functions, data types and several user-developed packages. Their scientific focus is (i) hydrological model evaluation and associated uncertainty estimation (Manoj Jaseetha et al. 2023) and (ii) evaluating the potential of machine learning to support hydrological modelling. Given the challenging nature of these case studies, the processing and analysis workflows should be adaptable to a wide range of use cases.

3 Technical aspects

The design of the V-FOR-WaTer web portal follows well-known Geographic Information Systems (GIS) such as ArcGIS or QGIS, as the handling of such systems is intuitive among environmental scientists. The portal includes map-based operations, sophisticated data filters, workflows, and data visualization. The system provides an advanced metadata catalogue based on PostgreSQL (Mälicke and Dolich 2023), a fine-grained user and authentication management, workflows and tools for data visualization and data analysis. Under the hood, we put emphasis on a modular design through containerization (Figure 2), allowing for easy exchange of components and extensions of the portal. The

whole system is composed of open-source projects and is itself open source as well. The central building block is the secure and scalable Python web framework Django, which is well documented and actively supported by a large community. Interaction with the map is handled with the JavaScript library OpenLayers.

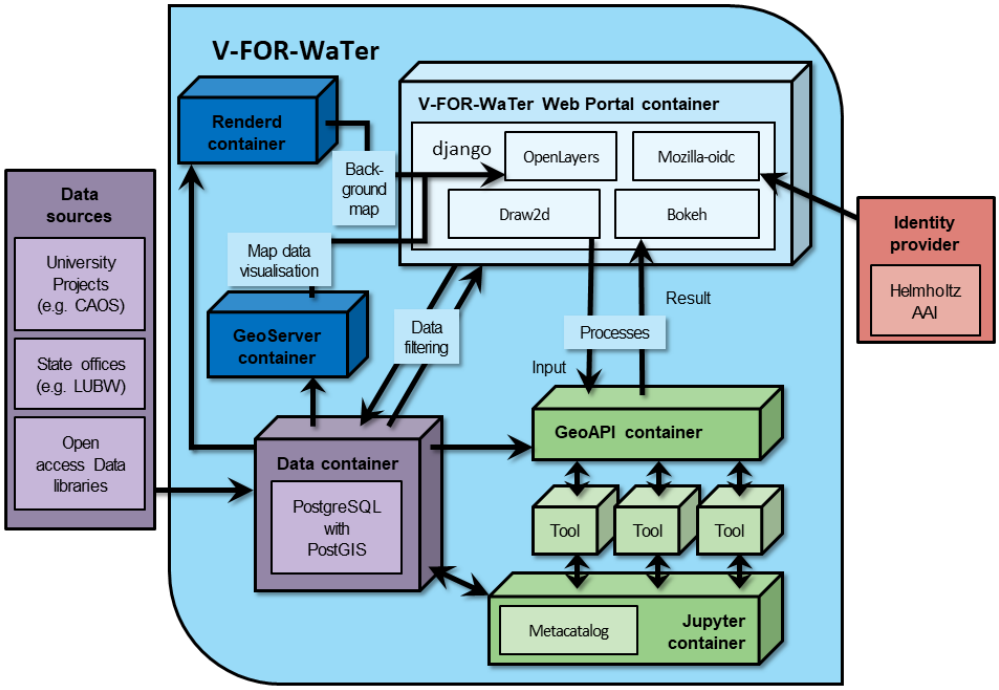

Figure 2: Current architecture of the V-FOR-WaTer portal. All components (3D boxes in the image) run in separate docker containers at Karlsruhe Institute of Technology (KIT).

While V-FOR-WaTer is intended to provide easy and open access to data and tools, restricted data access is necessary in some cases. This can be due to sensitive information contained in the data or a maximum 2-year embargo period imposed by the data owner for completing data analyses and publication. Consequently, no direct access to the data is provided. Instead, access requests are verified and redirected in Django. The authentication is facilitated by the federated Helmholtz Authentication and Authorization Infrastructure (AAI). Access to the metadata of all datasets is open for everyone and happens on the Web Portal through Django and GeoServer. The latter is used especially to visualize the position and extent of datasets on the map through a Web Feature Service (WFS).

The collection of tools provided in the V-FOR-WaTer web portal are accessed as API – Processes of the Open Geospatial Consortium (OGC), a common standard for web-based geo-applications. Using the OGC API – Processes standard ensures that the portal can be easily expanded with new tools and also enables direct access to the tools, e.g., from Jupyter Notebooks. In the Python backend, the V-FOR-WaTer toolbox already contains a set of example tools and packages, from simple hydrological signatures to comprehensive

variogram analyses. For the creation of shareable workflows we have developed a model builder, based on Draw2d.js, offering a drag-and-drop functionality to connect data and tools. A test instance for demonstrations is currently up and running at `https://portal.vforwater.de` (*Last accessed on May 12th, 2023*).

4 Conclusions

The V-FOR-WaTer web portal provides a centralized platform for scientists to access relevant data and tools, thereby greatly assisting them in searching, preparing, analysing, and publishing of data. By streamlining these processes, the portal facilitates the advancement of scientific knowledge and fosters reproducibility in research. In the future, the code of the web portal could be reused in other fields, where spatial information of their data is required and the visualization on a map is mandatory.

Acknowledgements

The ISABEL project is being funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – STR 1231/4-1, ZE 533/21-1.

References

Hutton, Christopher, Thorsten Wagener, Jim Freer, Dawei Han, Chris Duffy, and Berit Arheimer. 2016. "Most computational hydrology is not reproducible, so is it really science?" *Water Resources Research* 52 (10): 7548–7555. DOI: `https://doi.org/10.1002/2016WR019285`.

Mälicke, Mirko. 2021. *VForWaTer/hydrobox: Version 0.2.* Visited on September 6, 2023. DOI: `https://doi.org/10.5281/zenodo.4774860`.

Mälicke, Mirko, and Alexander Dolich. 2023. *VForWaTer/metacatalog: v0.8.0.* DOI: `https://doi.org/10.5281/zenodo.7643117`.

Manoj Jaseetha, Ashish, Franziska Villinger, Mirko Mälicke, Ralf Loritz, and Erwin Zehe. 2023. "Representative Hillslope Approach for Modeling Flash Flood Generation in Ungauged Catchments". DOI: `https://doi.org/10.5194/egusphere-egu23-6096`.

Stagge, James H., David E. Rosenberg, Adel M. Abdallah, Hadia Akbar, Nour A. Attallah, and Ryan James. 2019. "Assessing data availability and research reproducibility in hydrology and water resources". *Scientific Data* 6 (1). DOI: `https://doi.org/10.1038/sdata.2019.30`.

Strobl, Marcus, Elnaz Azmi, Sibylle K. Hassler, Mirko Mälicke, Jörg Meyer, Achim Streit, and Erwin Zehe. 2022. "V-FOR-WaTer – a virtual research environment for environmental research". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 394–398. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13755.

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.