CAMELS-DE PLUS

Mirko Mälicke, Alexander Dolich and Ralf Loritz
Institute for Water and Environment, Hydrology, Karlsruhe Institute for Technology (KIT), Karlsruhe, Germany.

Date of submission: 20.06.2024

Abstract:
CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) datasets are well established in the hydrological research community, especially in the field of machine learning and data-driven hydrology as benchmark datasets. Although Germany is one of the data-richest countries in the world, such a dataset (CAMELS-DE) and consistent pre-processing workflows were previously lacking.
CAMELS_PLUS aims to enhance and update the CAMELS-DE benchmark dataset through the development of a community standard for structuring and describing containerized scientific tools. This project addresses the challenges of manual data processing workflows by introducing a standardized, metadata-rich approach that improves reusability and provenance. The initiative will implement and test the standard using CAMELS-DE, promoting interoperability and accessibility within the Earth System Science (ESS) community. Deliverables include an updated CAMELS-DE dataset, extended community engagement through camels-de.org, and the release of the tool specification standard compatible with the NFDI4Earth Knowledge Hub, thereby facilitating broader adoption across ESS subdomains.

1. Introduction

1.1. Which of the track(s) are you applying to?
We apply to tracks 2,3 and 4 with a strong focus on track 3.

1.2. What is the data-challenge you face and what is the current state?
The first version of CAMELS-DE was processed by numerous scientists over the course of two years and we plan to extend CAMELS-DE and update it in the future. Using a manual workflow to make updates and enhancements of CAMELS-DE and create similar datasets is laborious and error-prone. Furthermore, combining pre-processing from different actors is not straightforward, and CAMELS-DE might lose its provenance context slowly over time. Metadata about used tools and workflows is missing. Additionally, the original data provided by authorities is neither standardized nor of common format.
The immediate challenge is that scientists from other ESS subdomains, especially meteorology and forestry, seek to extend CAMELS-DE right now, but we currently lack standardized workflows and procedures that would enable someone outside of our domain to easily contribute.

1.3. What is your vision for your community and their RDM workflows if your challenge would be solved?
We envision an easy-to-implement minimal standard for scientific data pre-processing workflows and their metadata, that can be directly applied to the enhancement of the CAMELS-DE dataset. We already tested different versions of the envisioned metadata and workflow schema

and a number of tools in the virtual research environment V-FOR-WaTer[1].
This schema offers significant benefits: it enhances the description of pre-processing tools, boosting their interoperability, accessibility, and reusability. When paired with a metadata standard, it could be directly linked to the NFDI4Earth Knowledge Hub. ESS developers would gain a community standard for easily sharing workflows and tools. Those aiming to extend CAMELS-DE would benefit from a ready-to-use processing container integrated into the CAMELS-DE pipeline, allowing them to focus on contributing data rather than navigating CAMELS-DE's storage and access complexities.

2. Pilot description

2.1. What is the proposed solution to your thematic track?

We propose a new community standard on how to structure and describe containerized scientific tools to enhance reusability. The current lack of standardized workflows and metadata leads to fragmented and error-prone processes, hindering efficient collaboration and data sharing across different domains. Moreover, without a unified standard, tools lose their provenance and become less accessible and reusable over time. We will develop and test the implementation by enhancing and updating the community-driven CAMELS-DE benchmark dataset. This standard includes metadata about the tool, citing information, a common layout, description of input and output parameters and data along with a common entry-point.
Implementations of this standard make tools more useful from a FAIR perspective but also help the author of the tool directly. We demonstrate the standard and a possible implementation by extending and updating CAMELS-DE, from data acquisition to publication on the GFZ repository. The connection of CAMELS-DE with its tools on the NFDI4Earth Knowledge Hub is envisioned to promote CAMELS-DE beyond the hydrology community.

2.2. What is the technological backbone you rely upon?

At its core, the proposed standard applies to containerized software following OCI standards[2]. This implies that our approach is limited to use-cases in which Docker (or similar technology) is a viable option. The metadata standard is implemented as a YAML file. We provide a growing number of templates (hosted on Github) to make implementation as easy as possible. We already implemented a proof-of-concept for Python, R, Octave and NodeJS. Tool parameterizations are stored as JSON files, citation information as CFF[3]. We already contributed open-source packages that read and write the parameters into native data structures for Python, R, Octave, Matlab, Type- and Javascript.

2.3. What are the standards and interoperability approaches used in the pilot's context?

The metadata for the tool itself will be further developed to be compliant to The NFDI4Earth Knowledge Hub, as soon as the Knowledge hub makes tools accessible.

2.4. How will your pilot produce impactful showcases of usability within the German Earth System Science community?

For the hydrological community, CAMELS datasets have already proven impactful in the past. We use the contribution and extension of the largest CAMELS dataset so far, to drive and promote the usefulness of provenance and reusability of tools in the ESS community. CAMELS-DE itself is well suited to bridge the gap between different ESS disciplines, as actors from other domains are already reaching out to us. First on the camels-org.de website, then through the NFDI Knowledge Hub, the dataset along with the standards for reusable scientific

[1] V-FOR-WaTer data portal website: https://portal.vforwater.de; Project website: https://vforwater.de
[2] Open Container Initiative: https://opencontainers.org/release-notices/v1-0-2-image-spec/
[3] Citation file format: https://citation-file-format.github.io/; used ie. by Zenodo

containers can be transferred to any other (ESS) community, for which the implementation of containerized (pre-processing) tools is helpful.

3. Relevance for the NFDI4Earth (Explains integration and uptake potential.)

3.1. What are expected users and stakeholders and how do they benefit from your solution?

The extension of the CAMELS-DE dataset enlarges the group of benefitting ESS communities beyond hydrology, ie. to geo-hydrologists, ecologists, biologists, meteorologists and water resource managers.

The standardized workflows offer predictable system requirements to system integrators and infrastructure providers. Their implementation lowers the barrier to use our pipelines for other data curators or even university teachers to make the data, and their pre-processing available to undergraduate students.

3.2. What measures are planned to support the uptake of your solution in your community, and to engage with prospective users?

We plan to satisfy the hydrological community's demand for a CAMELS-DE dataset by publishing a data description paper. Beyond that, we plan to publish a description paper about the standard and actively promote the usage of both through scientific contributions. CAMELS-DE datasets will be published on the GFZ data repository, which makes them findable through the NFDI4Earth Knowledge Hub and therefore gives outreach to the general ESS community beyond hydrology. We need an NFDI4Earth pilot to extend www.camels-de.org and guide our audience to the NFDI4Earth Knowledge Hub.This extension will provide a platform where tool workflows are published alongside the dataset, allowing for exploration and engagement beyond simple archiving.

3.3. What is the potential for other ESS subdomains?

The standard for containerized tools can be transferred to other ESS subdomains that are not limited by running (pre-)processing and analysis tools on non-distributed systems. CAMELS datasets can also be of great use for ESS subdomains other than hydrology.

3.4. How does your pilot enact FAIR principles?

The CAMELS-DE updated dataset is published following FAIR principles. Beyond this, the standard for containerized tools directly contributes to the accessibility and interoperability of data processing tools.

3.5. What challenges in managing earth system research data does your pilot address?

CAMELS datasets in general act as best practice for large sample datasets in water related research. The developed standard for data-preprocessing containers directly contributes to dataset provenance and improves the data quality through consistent and comprehensive metadata.

3.6. Contributions to engage with the NFDI4Earth

We want to link camels-de.org and tools using the standard for reusable scientific containers to the Knowledge Hub. By integrating and promoting the standard, we demostrate a way for scientists to make their software reusable and findable.

4. Deliverables

4.1 Technical operability of the pilot

The following deliverables are planned:

I) We publish an updated version of CAMELS-DE on an open data repository (GFZ)

II) We extend camels-de.org to enhance engagement with the ESS community by providing an interactive platform for exploring the used workflows along with the different CAMELS versions
III) We release the standard specification for reusable scientific containers after feedback from the community
IV) We make our tool specification standard compatible with the NFD4Earth knowledge hub and therefore make the tools in principle accessible through NFDI4Earth.

4.2 Material or actions for dissemination of knowledge/data
All deliverables will be comprehensively documented. The CAMELS-DE dataset and the standard will be presented at scientific conferences and possibly an article in a peer reviewed journal.

4.3 Roadmap document for the community after the pilot ends
The roadmap "Future directions of CAMELS" will be a living document published through the CAMELS-DE website (https://camels-de.org).

5. Work plan & requested funding

5.1 Milestone plan

MS	**Task**	**Months**[4]											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
M1	Release tool specifications	1	1							1			
M2	Tool implementations					1	1	1	1				
M3	Additional precipitation sources			1	1								
M4	Extend camels-de.org										1	1	1

5.2 Expenses

Position	Person months	Expenses[5]
Postdoktorand*in	12	86.100€
wissenschaftliche Hilfskräfte		6.000€
Total personnel expenses	12	**92.100€**
EGU 2025 + Tag der Hydrologie 2025 + NFDI Plenary	2 x	1000€
Total travel expenses		**2000€**
Total		**94.100€**

[4] The numbers detail the planned person months for each project month and milestone
[5] The personnel cost is based on the "Personalmittelsätze DFG 2024" table.