Improving the reproducibility and provenance of urban drainage data and models with RENKU, a platform for sustainable data science

Alfredo Chavarría, Simon Tait, Mathieu Lepot, Jean-Luc Bertrand-Krajewski, João Paulo Leitão, Jörg Rieckermann

Improving the reproducibility and provenance of urban drainage data and models with RENKU, a platform for sustainable data science

A. Chavarría[1], S. Tait[2], M. Lepot[3], J.-L. Bertrand-Krajewski[3], J. P. Leitão[1], J. Rieckermann[1]

[1] *Department of Urban Water Management, Eawag, Dübendorf, Switzerland*

[2] *Department of Civil and Structural Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, United Kingdom*

[3] *Université de Lyon, INSA Lyon, laboratory DEEP EA 7429, F-69621 Villeurbanne cedex, France*

**Corresponding author email: joerg.rieckermann@eawag.ch*

Highlights

- Improve the application of FAIR[1] data principles in use of urban drainage research data and models through the data and code sharing platform RENKU, which tracks the provenance of datasets and derived information.
- Provide a reproducible approach in three use cases from i) sediment research, ii) sensor calibration and data validation using the Urban Drainage Monitoring Toolbox and iii) automated in-pipe defect classification using and open sewer asset data.
- Our insights suggest that RENKU is no panacea, but could be a cornerstone to making our research reproducible in the urban drainage community, to sharing data and models and to track the provenance of derived data.

Where does your open dataset data come from and how has it been pre-processed?

Utilizing existing data to extract fresh insights forms a vital aspect of scientific exploration. By re-examining previous observations with new perspectives and potentially incorporating supplementary datasets, unexplored inquiries can be tackled, leveraging advancements in analysis methods. This approach is particularly beneficial in non-stationary research environments, where access to well-organized historical data proves invaluable. Repurposing models for data refinement, including process-based urban drainage models, offers substantial societal advantages by showcasing adaptability across diverse scenarios, thereby fostering knowledge generation.

Within this framework, the centrality of Open Research Data to scientific advancement is underscored ("Research Parasite Award," 2021). Given the laborious, hazardous, and costly nature of experimentation in wastewater and urban drainage systems, fostering a culture of open data could significantly cut costs and mitigate project risks in both research and industrial R&D. While replicating experiments in urban drainage precisely may not be feasible due to inherent site-specific disparities, the authors advocate for sharing contextualized data and metadata, promoting a new standard for the field.

Open, reliable, and reusable datasets hold immense potential for studying domain-specific processes, constructing predictive models, improving model calibration, and assessing policy impacts across different catchments and countries. Established workflows for enhancing data quality and understanding uncertainty can boost confidence in data analysis methods, fostering greater public trust in models describing urban drainage systems.

[1] FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability (FAIR). The acronym and principles were defined in a March 2016 paper in the journal Scientific Data by a consortium of scientists and organizations.

While funders advocate for FAIR principles in data collection (EC, 2022), the urban drainage field lacks a standardized platform for sharing data, code, and workflows, hindering progress in method standardization and data accessibility. A major challenge lies in linking published data with computational processes, necessitating a comprehensive grasp of inherent uncertainties in observations and simulation outcomes.

This paper proposes an innovative solution leveraging RENKU, a FAIR-compliant platform, to integrate urban drainage models, data, and computational environments. This integration enhances reproducibility, fosters collaboration, and facilitates continuous improvement in analysis tools, promising a more cohesive research landscape in urban drainage

Selecting a platform for reproducible urban drainage research

In our commitment to FAIR principles for managing Co-UDlabs project data, we've sought platforms supporting reproducibility, with Google Colab and Binder emerging as prominent choices. Google Colab simplifies Jupyter notebook execution without setup, while Binder offers transparent notebook execution in containerized setups. Papers with Code links published papers, datasets, methods, and model performance metrics. Researchers can find relevant datasets and compare task performance based on existing literature. However, these platforms lack mechanisms for tracking dataset usage within projects (Roskar et al., 2023).

Renku stands out as an open-source platform for collaborative data science endeavors, offering an encompassing environment for efficient data handling, experiment replication, and iteration monitoring (Roskar et al., 2023). It is currently being developed by the Swiss Data Science Center and key features include workflow management, collaboration, reproducibility, and integration with tools like Jupyter Notebooks, RStudio, and Docker for seamless data science workflows (Chavarria et al., 2023). Available as RenkuLab (web platform) and Renku Client (command-line interface), Renku supports researchers, data scientists, educators, and students by managing data, code, workflows, and computational environments. While it does not directly store data, Renku integrates with storage systems like Zenodo for effective dataset management (Figure 1).

Testing reproducibility on three use cases from urban drainage

To test the different capabilities, we selected three use cases from research on sewer sediments, uncertainty analysis and asset management. While we have very promising results for the first, the others are currently work in progress and will be finished in spring 2024.

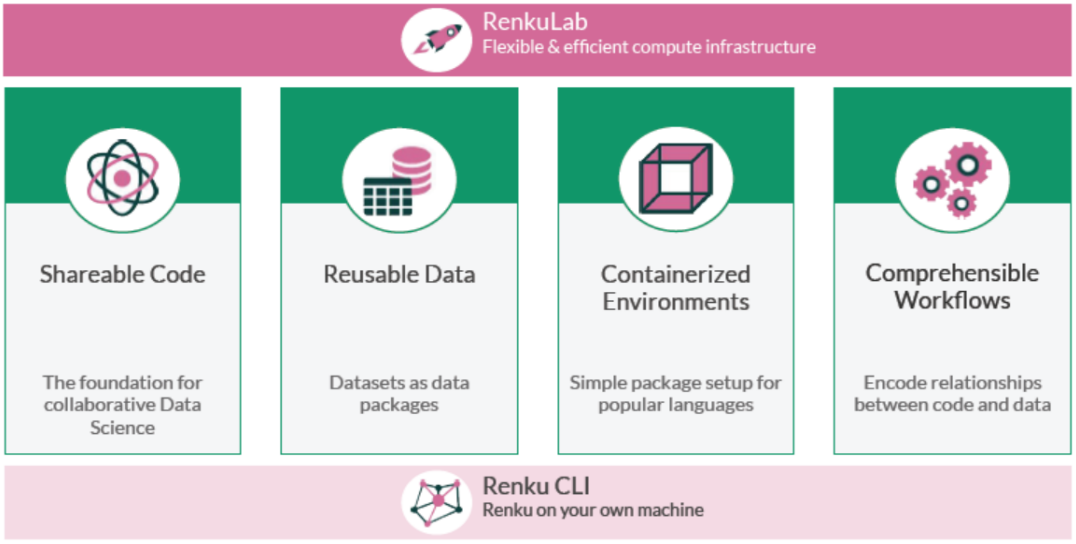

Figure 1. A summary of the components of RENKU (from Roskar et al. (2023))

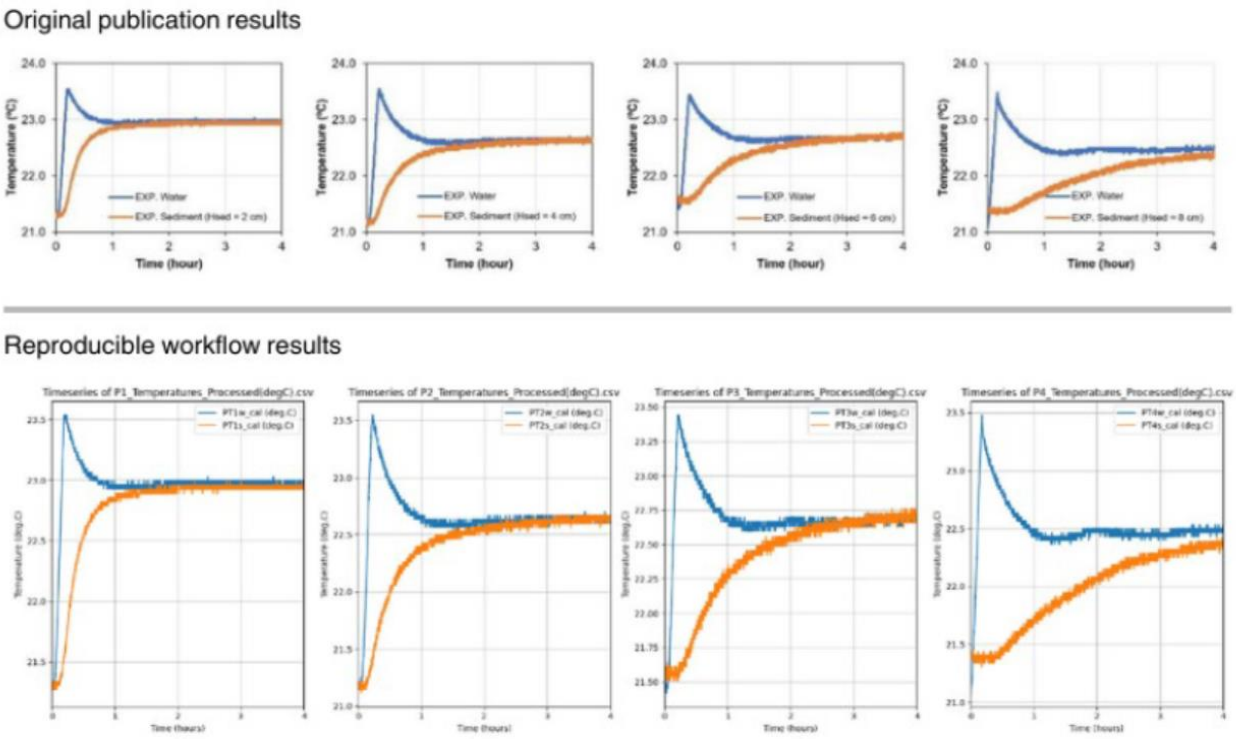

Figure 2. Comparison of results from the original publication and the reproducible workflow in Renku (Source: Chavarría, 2023)

1. Identifying sediment deposits from temperature signals

We tested Renku with an open dataset from Co-UDlabs project[2], published on Zenodo (Figure 2, top row). This dataset explores sediment deposit identification from difference in temperature signals without sediments (blue) and with sediments (red). It comprises CSV files containing cycle-based experiment data, sensor calibration information, and standard methodologies, which are published under a CC-BY 4.0. Detailed metadata is available in a 40-page data collection report.

In Renku, we ensured project reproducibility through key steps. Initially, we established a Python environment, imported data from external sources, and set up necessary requirements for data wrangling and cleaning. Using Jupyter Notebook and Python, we created transparent data pipelines. We maintained data lineage by tracking provenance through workflows. Finally, we shared the completed project, ensuring accessibility and transparency throughout its lifecycle.

Figure 2 illustrates the comparison between results from the original publication and those generated via the reproducible workflow. Reproducing these findings enhances their reliability, allowing other users to validate and potentially reduce errors. During our process, we identified a structural error in a CSV file, highlighting how error detection and correction enhance outcome integrity.

2. Assessing the uncertainty of monitoring data

The Urban Drainage Metrology Toolbox (UDMT) is a Co-UDlabs developed software available as a web app[3] and downloadable executable. It aims to support best practices in monitoring urban drainage systems.

[2] https://zenodo.org/records/7783173
[3] https://tinyurl.com/UDMT2023

UDMT access is free without registration. Although it is Matlab-based, we can potentially establish a Matlab environment in Renku, given that we can provide suitable licenses. Implementing UDMT in Renku offers benefits: processed data includes method and parameter information, retained even if the code is modified. In the future, users can easily 'roll-back' the to preferred UDMT versions. We are progressing on implementing some methods into Renku.

3. Automated in-pipe defect detection in CCTV sewer surveys

This use case is the most complex one, because it implements a Co-UDLabs-developed deep-learning framework for automated in-pipe defect detection in CCTV sewer surveys (XREF_MS16_Report). Utilizing the Ultralytics YOLO v8 model, it streamlines defect identification by eliminating manual feature extraction, especially effective for challenging sewer pipe defects. The code, along with sample images and software support, is publicly accessible for easy usage. We're currently adapting a Python-based version of this method in RENKU, expecting that this will facilitate the utilization of different pre-trained YOLO models. Offering this in a functional environment aims to encourage reporting of findings by those using the code and images.

Ensuring Open Research data is license-free (CC0) is crucial to avoid hindering future developments. RENKU aids reusability of open data and processed data interpretability, though it is not a universal solution.

Conclusions and future work

A culture promoting open research data in urban drainage holds immense potential, offering benefits such as new results, improved models, and enhanced data quality, thereby reducing costs and project risks in research and industrial R&D. Despite increasing availability of open research datasets in this field, linking published data with various computational processes for seamless reusability remains a significant challenge. While existing data repositories ensure data findability and accessibility, achieving interoperability is limited, constraining data and analysis approach reusability. Our investigation explores the potential of addressing these challenges using the Renku platform for sustainable data science. Through an example involving sewer sediments, we showcase how Renku streamlines dataset preparation and improves record-keeping for easier dataset usage. Renku's integration with datasets on Zenodo, with the UDMT and Automated Detection of In-Pipe Defects from images signifies a substantial leap forward in urban drainage research. This approach fosters dataset reproducibility and enables continuous data utilization, fostering a more collaborative research environment. Ensuring Open Research data is license-free (CC0-license waiver[4]) is crucial to avoid hindering future developments.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101008626. And from the ETH Domain ORD program. We thank all partners in the Co-UDLabs project for their contributions and has limitations for complex data pipelines.

References

Chavarria, A., Rieckermann, J., Schellart, A., Brüggemann, T., Bertrand-Krajewski, J.-L., 2023. CoUDLabs - D2.1 Report on Data Harmonization - Periodic Technical Report (M18) (1st intermediate report).

EC, 2022. Data Guidelines [WWW Document]. Open Res. Eur. - Data Guidel. URL https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#fairdata (accessed 3.31.22).

Research Parasite Award, 2021. . Wikipedia.

Roskar, R., Ramakrishnan, C., Volpi, M., 2023. Renku: a platform for sustainable data science. Presented at the NeurIPS | 2023 Thirty-seventh Conference on Neural Information Processing Systems.

[4] Commonly adopted licenses for scientific content are typically under Creative Commons. Typically, opting for a dedication to the public domain through CC Zero (CC0) is advisable for datasets and databases. CC0 allows scientists, educators, artists, and other creators who hold copyright- or database-protected content to relinquish their rights in their works. This action aims to place the works entirely in the public domain, enabling others to freely build upon, improve, and reuse the works without encountering constraints under copyright or database law. Conversely, a CC BY license (which mandates attribution only) serves as a suitable choice for works like articles, books, working papers, and reports.