UK Centre for Ecology & Hydrology
National Capability for UK Challenges
#CC02 - 9th July 2025
1.00pm – 3.00pm
55 participants

Community Conversation

Co-Designing Digital Infrastructure for Environmental Science

Report on #CC02: Enhancing Discoverability & Access to Methods

Eunice Agyei (Lead author: eunagy@ceh.ac.uk), Helen Rawsthorne, Jeremy Carter, Matthew Nichols, Fiona Seaton, Michael Tso, Susan Jarvis, Michael Hollaway, Carolynne Lord, David P. Green, Kelly Widdicks, and Gordon S. Blair. UK Centre for Ecology & Hydrology. 2025.

What did we do?

Our second in a series of online **Community Conversations** focused on Enhancing Discoverability & Access to Methods, as part of the National Capability for UK (NC-UK) Challenges programme delivering integrated monitoring, modelling and data for the UK environment. We shared the Environmental Data Science Toolbox, which is designed to make trusted, reusable methods easier to share, find and apply. With this, we aim to strengthen the UK's national capability to tackle complex environmental challenges. The session was opened by Dr. Kelly Widdicks, the NC-UK Digital and Data Integration Co-Lead. Dr. Susan Jarvis and Dr. Michael Hollaway, joint task leads for the Environmental Data Science Toolbox, chaired the session. There were four presentations, followed by an interactive **Q&A session**, and small group **breakout discussions**. We conducted a series of **polls** throughout on discovering and accessing data science methods. Afterwards, we invited participants to complete an opinion **survey**.

Presentations:




Dr. Fiona Seaton Quantitative Ecologist (UKCEH)	**Dr. Matthew Nichols** Data Steward – Technical Specialist (UKCEH)	**Dr. Helen Rawsthorne** Data Scientist – Semantic Specialist (UKCEH)	**Dr. Jeremy Carter** Environmental Data Scientist (UKCEH)
Example data science methods (1): R Package for Joint Species Distribution Modelling in Stan.	**Example data science methods (2):** Accessible File Check Tool for promoting data reusability.	The importance of **Metadata for Methods** to underpin discoverability and accessibility.	The prototype: **Environmental Data Science Toolbox** for discovering and accessing the methods.

Additional panellists: Dr. Susan Jarvis, Dr. Mike Hollaway, Prof. Gordon Blair, and Dr. Kelly Widdicks.

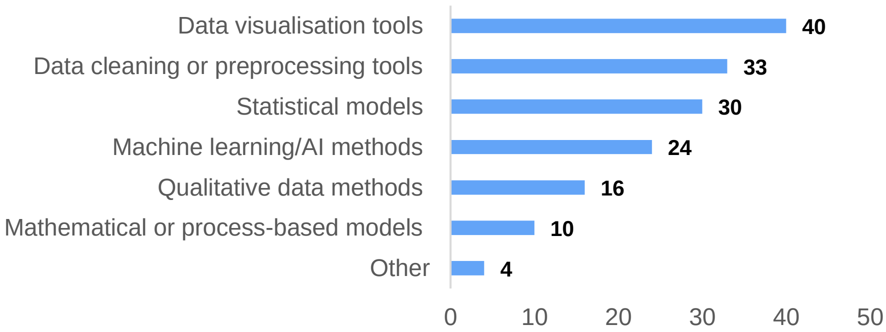
UK Centre for Ecology & Hydrology NATIONAL CAPABILITY FOR UK CHALLENGES

What did we find?

The event aimed to deepen understanding of how users discover, explore, and adopt emerging data science methods. It also served as a platform to gather feedback on current developments, helping to refine designs and deliver solutions to align with the evolving needs of the digital research infrastructure (DRI) community. We identified and discussed several **challenges, opportunities,** and **needs** related to improving discoverability and access to data science methods.

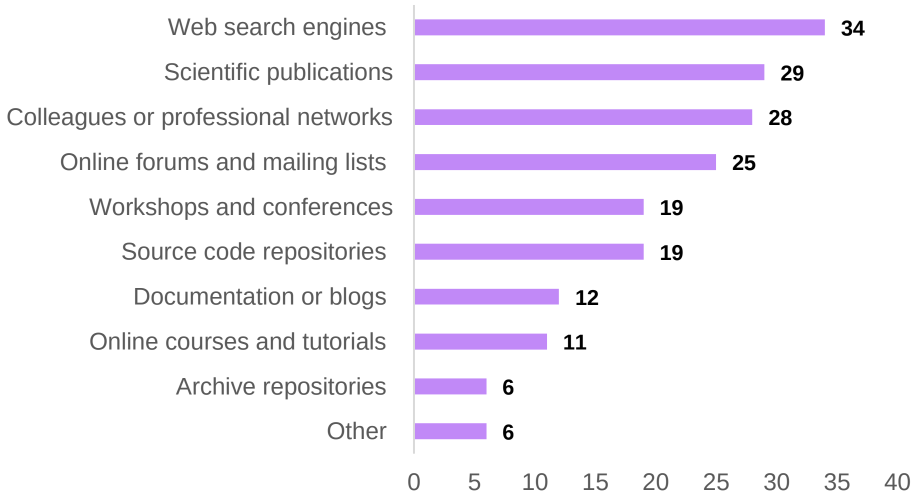
Many types of methods are available and used in environmental science.

Many different data science methods and techniques are applied in environmental science, with new approaches often emerging. As examples, we shared two contrasting methods for the Toolbox that may appear at different points in the data science pipeline (e.g. data cleaning, analysis and visualisation). We then asked participants **“Which of the following do you currently use in your work?”.** The most common response was ‘Data visualisation tools’ (n=40; total = 45), and on average, participants use three to four method types across their work process.



These methods are scattered across platforms.

We asked participants “**What are the primary ways you currently discover and explore data science methods?”** and found that they discover methods from many sources.



On average, participants shared that they use four to five different ways to discover methods. Web search engines (n=34; total = 41) were the most popular, but the top four sources show that participants use a diversity of sources from peer reviewed literature to unmoderated forums.

Given the many platforms to find methods, one participant said: “Decentralisation of methods is a problem. There is no single place to find them.”

Detailed method descriptions are important metadata.

Following an introduction to metadata and our work to date on method metadata, we asked participants, **“What metadata do you wish others would always include with methods?”**. ‘Detailed description’ (n=13; total = 37), ‘Dates’, and ‘How to cite’ were the most popular responses. 'Other' metadata suggested included of time requirement for the method runtime, licencing and restrictions on use, versioning, and location.

But there are challenges in creating metadata and making it fundamental to science.

In breakout room discussions, we asked participants “**What challenges do you face in creating or using metadata for methods?”** Key challenges raised include:

- Limited awareness of why metadata for methods matters.
- Time and resources required to generate metadata is limited and not accounted for.
- Making metadata readable by machines for automation and by humans for understanding.
- Lack of metadata standardisation across different domains in environmental science.

On the challenge of creating metadata, one participant said: “Lack of standard way to create metadata [for methods].”

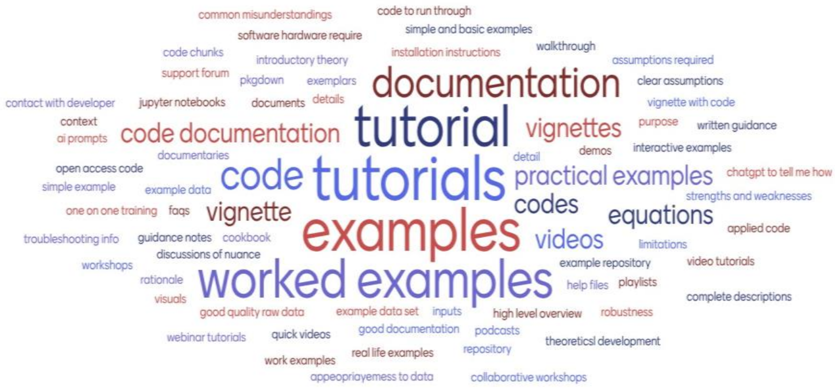
UK Centre for Ecology & Hydrology
NATIONAL CAPABILITY FOR UK CHALLENGES

Methods are adopted when they are designed for users and their communities.

In an interactive poll, we asked participants, **“What criteria do you use when choosing a method?”.** Participants indicated that they would prioritise ease of use, documentation, suitability among other criteria when choosing a method.

Worked examples can enhance the uptake of new methods.

We asked participants: **“What information do you need to apply new methods?”.** Examples, tutorials, code, and documentation were the most popular response among participants.



Of the Toolbox, one participant said: “[It's] a good way to package [...] citable things that aren’t papers”

Interactive content and support features will be valuable to a method repository like the Toolbox.

The NC-UK Environmental Data Science Toolbox showcased and discussed in the session goes beyond a method repository, integrating rich metadata and offering an interactive and community-driven space for discovering, learning and contributing data science methods. It aims to provide a foundation for an open, interoperable and community-sustained DRI platform. We asked participants to share what additional features they might find valuable for this. Suggestions given included:

- Interactive content allowing users to learn by doing, offering hands-on experience with the tools and workflows.
- Tutorials and examples to support method understanding, as well as clear documentation and help resources.

- A diverse range of high-quality methods across multiple domains of environmental science.
- Access to the method source code and relevant data.
- Artificial Intelligence (AI) tools to support coding, interactive querying, and enhanced search functionality.

What would make a difference?

Environmental data science methods should be easy to discover and use everywhere. As such, overcoming barriers such as **location, usability, trustworthiness, and sustainability of methods** is essential. Reflecting across the polls, Q&A and breakout discussions, in this section we outline four key community needs and areas of work to further enhance the Toolbox, its contents, and its metadata. By building on the work we showcased, these will help address current barriers and make a positive difference in method discovery and access.

Enable easy discovery and access to data science methods across environmental science domains.

Easier access to methods across environmental science domains would help researchers to find and use methods efficiently across repositories, hubs, and portals. The Toolbox could act as a curated platform for organising key methods in one place with clear metadata, descriptions, and example workflows to make them usable. To complement this, search engines or other unified search interfaces could be introduced to enable users to locate methods across multiple sources using advanced queries and filtering. Additionally, a federated hub system could link multiple repositories through standardised metadata and shared protocols to connect data, methods, and other research objects – providing seamless discovery across repositories. AI could also be leveraged for easy search and discovery of methods.

Design intuitive methods for effortless use and instant adoption.

Method developers, or other stakeholders should design methods and their associated resources to be intuitive for people and optimised for systems to facilitate sharing and reuse. To achieve this, methods should be well documented – detailed with explanations of the method's purpose, workflow, assumptions, examples, and stories of its use. This would allow researchers to understand, reproduce, and apply methods correctly. Ideally, methods should be modular, interoperable, and integrate seamlessly into different stages of the data science pipeline and evolving workflows. To reduce the time, burden and resource costs of making methods more accessible and reusable, additional support should be investigated and offered for method developers.

Apply an effective metadata strategy to enhance the discoverability of methods.

A good metadata strategy should be applied to balance lightweight and rich metadata and ensure efficient discovery, understanding and use of methods. While lightweight metadata provides minimal descriptors like title, author, date, and basic keywords to facilitate quick discovery methods, rich metadata provides structured and contextual information, including provenance, required inputs and outputs, software dependencies, and DOIs (Digital Object Identifiers) for methods. Such a strategy should also incorporate training and awareness to support consistent metadata practices across diverse disciplines. AI and automation tools could streamline the creation of metadata by extracting, tagging, and classifying methods to reduce the manual effort and time needed for method sharing.

Strengthen the Environmental Data Science Toolbox by making it trustworthy and sustainable.

The success of the Toolbox depends on the diversity of high-quality methods in it. This will require clear governance to guide what gets included, and support for community contributions and ongoing curation. This may also include training or assistance from technical experts (e.g., Research Software Engineers) or user forum features within the Toolbox to enable cross-community support. For such support and curation, it is important to secure reliable funding to sustain the Toolbox and ensure that it evolves to meet the needs of the community.

What's next?

We were delighted to showcase NC-UK work to date on the Toolbox, its contents, and method metadata in this Community Conversation, and were encouraged by the participants' support for this. To ensure this becomes an established community resource, our next steps include:

- Enhancing the Toolbox by introducing new features based on the insights above, including advanced and AI-supported searching.
- Incorporating intuitive methods into the Toolbox developed by the community, supported by further engagement with stakeholders.
- Encouraging use of a standard, structured metadata template for methods, with basic minimum requirements and other optional fields that suit a wide range of method-type research outputs.
- Co-designing the governance of the Toolbox including licensing, curation and citation to support community trust and ownership.
- Integrating the Toolbox with other DRI platforms, enabling stakeholders to connect methods with other relevant research objects (e.g., data) in environmental science.
- Disseminating this report as a valuable community resource in outlining challenges, opportunities and needs for method sharing.

Please sign up to NC-UK mailing list via NC-UK Website for the latest updates on DRI development, as well as for wider updates on the programme and opportunities to engage (e.g., through future Community Conversations).

Special thanks to the **organisers, speakers, facilitators and notetakers** (mentioned above as well as David Green, Marika Glasby, Mary Preston, Michael Tso, Carolynne Lord, Sophie Moran, Tom Bell, Jasmine Hunter and Rod Scott), but most of all, thank you to the **participants** for making this a rich and engaging conversation, with valuable contributions to shape this work.

Appendix

Responses to Questions & Answers

1. What will happen to outdated methods? E.g., how will they be identified and how will they be 'retired'?	Various potential approaches are being discussed for this, for example, badges like repostatus.org can highlight how current a method is. In general, when methods become outdated, the method can either be kept in the Toolbox with some warning badges and/or links to updated versions, or the method can be removed/archived from the next published version of the Toolbox. Since the Toolbox itself is housed in a GitHub repository, any removed methods will always be traceable. We're also exploring the option of including standard, machine-readable metadata tags to indicate whether the code is outdated.
2. Who will be responsible for keeping the Toolbox up-to-date, whether that's ensuring links remain working (e.g. to datasets) or whether that's updating the structure and software behind the Toolbox?	Over the next 4 years in the NC-UK programme, UKCEH will help keep the Toolbox up to date. We will also co-design guidance and governance processes with the community for self-management by the community, including suggesting changes and making amendments through Git pull requests (contributing guidelines have been created for this, and the process will continue to be developed for ease of contribution). The hope is that the Toolbox will be low maintenance, and processes will be put in place to make contributing to the Toolbox simple. The Toolbox is based on the well-known Jupyter Book framework, which has a lot of documentation and is expected to remain relevant into the future.
3. Why would we use this over just putting a Jupyterlab or R notebook as the front page for a GitHub/GitLab repository?	The Toolbox serves as a structured, centralised hub to enhance the discoverability and accessibility of methods. Including links to interactive notebooks in GitHub repositories themselves is still encouraged. The Toolbox aims to add to this by collecting these resources and offering a more approachable and consistent interface for non-developer users (e.g. policymakers, analysts) who may find GitHub intimidating or fragmented. In GitHub, relevant repositories are also often buried among many others. The Toolbox aims to provide an easier route to finding and using data science methods specifically for environmental science.

4.	What are the key differences between this and the Environmental Science Book? If similar, how will any duplication between the two resources be handled?	The NC-UK Environmental Data Science Toolbox, developed by UKCEH, shares common goals and design principles with the Environmental Data Science Book developed by the Turing Institute. The key differences will be in the exact scope of methods included and finer elements of the design details – such as incorporating AI search engines, etc. The aim is that these resources support each other, and to avoid duplication, relevant signposting between the resources can be made. The goal is *interoperability, not uniformity*: both resources are valuable, and linking them effectively will strengthen the UK environmental data science ecosystem.
5.	If I've developed a tool in the past that has a stand-alone vignette, how easy is the integration of this into the Toolbox? Will it require re-working/formatting?	We're exploring ways to make this as seamless as possible, as we want to avoid duplication of effort. Currently, the process involves: using Quarto to convert the file that generates the vignette into a Jupyter notebook; applying some light reformatting to improve consistency with the Toolbox design; and use of MyST markdown. We'll develop some guidelines to add to the contributing page to specifically address incorporating R vignettes and markdown files.
6.	Can the Jupyter book consume an existing Git repo or do I have to recreate all the documentation and maintain in in two places?	The Jupyter Book can't automatically consume documentation from a GitHub repository, although it does use similar file types. In the case where there's a lot of existing documentation, our advice is to include a distilled version of that in the NC-UK Environmental Data Science Toolbox, which gives the user a good feel for the method and then signposts to the repository for further information – helping improve discoverability of the method. The aim is to make this process as easy as possible, and that content from files used in the existing documentation can be utilised with minor formatting adaptations required for consistency. In terms of maintaining the documentation in two places, the version stored in the Toolbox is expected to be stable and mostly static, i.e. it will not need to be fully up to date with the repository but should indicate which version it relates to. Updates are always encouraged, but the effort involved in this is recognised.
7.	Have you explored the way others internationally are doing this such as ESA, NASA and Long Term Ecological Research (LTER) programmes for example?	This is a good point, currently we've not explored in detail the ways ESA, NASA and LTER are approaching sharing data science methods. Interoperability and alignment with what's being done internationally are active considerations in the Toolbox design as the platform matures.
8.	I wonder who the ideal user of the Toolbox is. Is it	The Toolbox should be useful for stakeholders across a spectrum of technical expertise. It supports technical users (such as researchers, data scientists, etc.) who want direct access to the

other researchers with preexisting technical skills? Could we make the methods and tools in the Toolbox accessible to stakeholders with no skills?	code and the relevant background repositories. It also aims to support those without strong technical coding skills, providing approachable walkthroughs and method explanations. Additionally, some methods like the file-checking tool demonstrated in the Community Conversation are intentionally user-friendly, aimed at users without coding skills. There's still a conversation to be had about how much introductory content to build in, but if stakeholders see real value in that, we can absolutely integrate more 'best practice' and entry-level resources alongside the technical tools. We also aim to explore potential opportunities in the future for Toolbox methods to be run in live interactive environments with minimal setup requirements, if there is a desire for this from the community.
9. There is often still a pressure to "do the thing" and then move on, without spending time on making methods FAIR, which can be hard to justify. How do we as a community ensure we support time for FAIR?	There is a need to recognise the importance of FAIR (Findability, Accessibility, Interoperability, Reusability) digital objects (e.g., data, methods, etc.) at a community level. Rather than placing a focus on the individual researcher to 'find' time, this needs to be created by projects and funders who acknowledge this. Funders must ensure funding is available for these purposes, include it as a requirement in their policies and establish appropriate incentives (e.g., that researchers are accredited for their FAIR methods). Meanwhile, projects must be designed with dedicated time to make both data and methods FAIR. At the technological level, there is a need to ensure that it is as easy and as quick as possible to make methods FAIRer, to ensure that this does not become another time-intensive burden. We are currently developing guidelines for method authors on producing structured metadata from a template (using CFF) and making it available for their methods. AI tools that automatically generate structured metadata from method files or existing documentation about the method could be helpful in reducing the burden on researchers.
10. Project time is very limited already, how would you make sure that people do fill out/create the metadata for their methods? I think it needs to be as easy as possible (as you mentioned AI assist?)	See response to Q9.
11. This is essentially a centralised publishing platform, is there a QC check, or standards	There will be some form of QC (Quality Control) check and curation of methods for the Toolbox. The process for this still needs to be developed and will be shared once it has been developed and agreed with the community. Some features being considered for QC checking include flags to highlight where

	that need to be met, or will anything be accepted? If so, what's the review process, who will that be done by?	methods meet certain standards. As for curation, currently the focus is on including methods that are generalisable across multiple areas of environmental research. Practically speaking, this applies to a lot of approaches, with many methods having components relevant to a range of fields. The contributor is encouraged to emphasise the important components of the method and how they could be relevant for multiple applications.
12.	Will there be a requirement for all packages used in code in the data Toolbox, and all methods presented there, to be free and open source? What license do you envisage?	The repositories housing methods included in the Toolbox will require licenses so that the approaches can be adopted by the community. The specific license used will likely be optional to the method developer. Currently, some guidelines around using CFF metadata templates are being developed to make sure the methods included are citeable, and as part of this, licenses are mentioned (see what is a CFF here!). Licenses for code and methods are becoming more common, and we will follow community best practice in recommending licenses. Packages used by the methods will generally be open source; for cases where this is not true, a discussion could be held around whether the method is suitable for inclusion. The governance process for the methods in the Toolbox will be developed and agreed upon with the community.
13.	Is there a policy for the data Toolbox on AI generated code or code reproduced from elsewhere (such as source forge)?	A lot of research code development is now happening with assistance from LLMs (Large Language Models). Any code, whether generated from LLMs or not, should be reviewed and tested by the contributing developer. Researchers should also be aware of potential licensing issues of copying code reproduced elsewhere. As part of the Toolbox, we aim to develop some guidelines around best practices for method development, alongside the governance processes for the Toolbox; issues of AI-generated methods will be included in these.
14.	How are you planning on managing contributions? I imagine lots of us are using similar methods with slight intricacies, this duplication might be overwhelming for a user trying to pick a method.	Many data science methods are complex, difficult to understand and describe. There can be a lot of overlap between different methods, and it can be overwhelming trying to navigate the vast number of options available for solving a particular issue. We're thinking about various ways of mitigating situations in whereby users feel overwhelmed. One approach is to provide some overarching context about what methods are included and how they relate to each other. Another approach could be to have an LLM chatbot that helps the user directly query the entire Toolbox and navigate to the parts most relevant for their specific scientific question. There's also the possibility of creating direct comparison notebooks between methods for specific example problems, showing how you might evaluate the performance and what sort of considerations factor into decisions around adopting new methods. We're interested to hear any ideas about innovative ways to tackle this and what would be most useful to you as a stakeholder/user of the Toolbox.

15. Is the scope potentially all methods that exist, and if so what degree of exhaustivity are you envisioning? Or a more or less local sample, focused on CEH-related resources and based on evolving need?	This initiative is intended to be a shared resource for the wider environmental science community. While initially developed by UKCEH, its purpose is to highlight methods created across a diverse range of people and organisations. It's fundamentally about creating a valuable community that is continually shaped by feedback and designed to evolve. The scope of methods to be incorporated is currently quite broad, with an emphasis on including methods that are generalisable/useful for multiple domain areas within environmental science, and relevant for a broad collection of stakeholders. There is space in NC-UK for developing new methods driven by the current needs of the community, so if you have suggestions, please do let us know.
16. Regarding metadata - we all have a vision of the 'ideal', which is often not the reality. Do you have an approach in mind to completing the metadata of methods that get contributed?	We are currently developing guidelines for method authors on producing structured metadata from a template (using [CFF](#)) and making it available for their methods. A basic level of metadata will be mandatory – title, contributors, etc., and other fields will be optional, recommended or mandatory depending on the type of method in question. We do acknowledge that this is not a one-size-fits-all solution, and hence we need to be open-minded and flexible with the template. Another approach would be to find or develop a tool that would generate metadata automatically from the method files/existing documentation. This could be part of a GitHub Action, for example. It could also ask the user to validate the automatically generated metadata and to fill in only the fields that could not be filled in automatically.
17. I was advised by a data steward not to rely on GitHub as longevity is not assured. What is the better alternative?	We currently recommend that code be made available through GitHub but also deposited as a snapshot of the code in a long-term repository. GitHub and similar code repositories are not suitable for long-term preservation of code, as the code can change at any point in the future, and there is no guarantee of future access. For long-term preservation and citation, we currently recommend Zenodo or similar repositories to obtain a DOI, which ensures that code remains citable and accessible. Zenodo uses concept DOIs, which serve as stable, persistent links that always point to the latest version while connecting all individual versions. This allows users to access the complete version history of a method. We will be exploring this question further with the community as part of our co-design process for the Toolbox's governance.
18. Are you going to have defaults in methods? Hard to keep reproducibility if defaults keep changing. Need to make methods FAIR.	Methods are encouraged to have environment files that detail what versions of certain packages were used at the time of development. This helps with retaining reproducibility as packages get updated. Even with this information, however, it can be difficult to reproduce the exact environment, as it can depend on things like your specific machine and operating system. Guidelines around this will be released, and potential avenues such as creating and maintaining containerised environments for running the code will be explored. Clear, user-friendly

		documentation is also essential for maintaining reproducibility as methods evolve. Methods and documentation can be updated, similar to how other software packages release version updates. This ensures that users have access to the latest guidance.
19.	If we're doing lifecycle badges etc, isn't this really just reproducing GitHub? I think there's value to collecting things into a central place that makes it easy for people to find things, but isn't the repository the key unit here where things should reside?	The Environmental Data Science Toolbox aims to add value beyond GitHub by transforming raw code into a curated, discoverable, and reusable resource. Its lifecycle indicators/badges are designed to be clear and standardised indicators of methods' developmental status (e.g., experimental, actively maintained or production-ready). These badges enhance the discoverability of a method and confidence to assess, adopt and reuse methods.