High 5 For AI

Quantitative survey results

Detailed Report

REPORT NUMBER OR/26/050_SUBREPORT3

Authors:

Ola Podgorska https://orcid.org/0009-0002-3722-7658

Shwetha Raveendran https://orcid.org/0000-0002-4757-498X

Carl Watson https://orcid.org/0000-0002-9026-7432

Affiliations:

3 - British Geological Survey
4 - British Antarctic Survey
5 - UK Centre for Ecology and Hydrology
6 - National Oceanography Centre
7 - Digital Solutions Hub, University of Manchester

National Oceanography Centre

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKCEH | Environmental Information Data Centre

BGS | NGDC

British Antarctic Survey
UK Polar Data Centre

TABLE OF CONTENTS

METHOD 1
DURATION 1
KEY SURVEY GOAL - to inform product development BY: 1
PROJECT (AND USER RESEARCH) CONTRAINTS 2
SURVEY DISSEMINATION 2
SURVEY RESULTS IN FULL 3

HIGH 5 FOR AI
ABOUT THE SURVEY

METHOD

Due to the limited project duration, for WP2 we decided to carry out a user research 'market' survey as this would provide rapid insights from the potential user base. The survey incorporated feedback from all Work Package Leads in the project, and was published using the Hubble platform, which automatically anonymises user data, to ensure GDPR compliance.

Participants were given the opportunity to provide their email address on the last page, if they wished to be contacted for further moderated user research interviews.

DURATION

The survey was launched live on the **19 Nov 2025.**
It closed on the morning of **19th January 2026 with a total of 41 responses.**
We reopened it on the 27th for two more participants to fill out, **totalling 43 responses.**

Total duration – 2 months (40 business days)

Weekend Days*	18
Holidays	2 (Christmas & New Years Day)
Business Days	40

*Assuming 5 day workweek from Monday to Friday, weekend is Saturday and Sunday in a seven-day week.

KEY SURVEY GOAL - TO INFORM PRODUCT DEVELOPMENT BY:

- Understanding if potential users / audiences are currently aware of, or utilise EDS datasets, and to what extent / how they use them
- Understanding key challenges or pain points related to the access / sourcing and utilisation of datasets
- Discovering which data formats are most and least useful to potential users
- Understanding attitudes and usage of AI methods, and workflows in current day-to-day work

PROJECT (AND USER RESEARCH) CONTRAINTS

The Work Package leads chose to prioritise two primary use cases in the Funding proposal:

Excerpt from EDS DRI Funding Proposal (page 2):

"We will consider cross domain use cases aligned to Government missions. We will particularly focus on 'making Britain a clean energy superpower' and 'an NHS fit for the future'. At this proposal stage, we have provisionally identified two cross-domain societal challenge use cases related to these missions.

Use case 1: Clean energy
The EDS can contribute to UK net zero targets through improving data access to support decarbonisation and energy security initiatives. A proposed focus is based on the siting of offshore windfarms that would bring together data from multiple EDS centres. A recent stakeholder review (https://nora.nerc.ac.uk/id/eprint/538827) identified a clear need for consistent data access (e.g. bathymetric data, sediment composition, wind resource assessment, weather pattern data, sea level, ocean physics) to improve decision making related to offshore renewable assets.

Use case 2: Access to blue-green infrastructure for benefit to public health.
Blue-green infrastructure (e.g. access to open spaces and water) and good air quality play a crucial role in a healthy society. They facilitate potentially preventative solutions to a range of health conditions and therefore support the NHS Fit for Future mission. EDS has a range of datasets (e.g., Landcover map, sea level, air quality, geological maps) that can be integrated to inform decision making that enables the planning and provision of built infrastructure (e.g. houses, parks, access to beaches) that supports a mentally and physically healthy society.

SURVEY DISSEMINATION

Consequently, the survey was designed to place greater emphasis on these use cases, and its distribution was targeted toward individuals who work within, or are likely to engage with, the relevant industries.

To obtain a wide range of responses, the survey was distributed using various methods, including conference sharing, direct emails to individuals, email mailing lists, online platforms (websites and blogs), and LinkedIn.

EDS Website: https://eds.ukri.org/projects/high5-ai
EDC Website: https://ukerc.rl.ac.uk (courtesy Peter Holt, Energy Research Unit)

Email:
• Directly to a custom-collated distribution list of individuals who were selected by Work Package Leads, and the Lead Researcher, based on their alignment to the two chosen use cases that the High 5 for AI project was focusing on. A record of individuals contacted was kept during the outreach, and collated on this spreadsheet.
• Mailout through the JASMIN user list (courtesy Jesse Alexander, NCAS)

Workshops, meetings and conferences:
UK Environmental Exposure (UK-EEX) Hub workshop, MEDIN Monthly meeting, Data Sharing Working Group (x2 meetings), Digital Twins of the Ocean Event,

Posts on LinkedIn shared by Ola Podgorska (NOC), Carl Watson (BGS) & reshared within our networks.

SURVEY RESULTS IN FULL

Total responses as at 27 January 2026 = 43

Q1. Your Role - which of the following best aligns with your current professional expertise or research focus?

Note: This is a multiple-choice question, the percentages won't add up to 100%.

- 40% - *Green spaces* (their impact within communities)
- 37% - *Digital technology*
- 35% - *Blue economy (planning and industry)*
- *26% - Other (self-describe)*

Options	Number of responses (43)	%
Green spaces	17	40%
Digital technology	16	38%
Blue economy	15	33%
"Other" responses - total	11	26%
Marine Spatial Planning	1	2.3%
Research	1	2.3%
Hydroclimate extremes and their impacts	1	2.3%
"We use data sets for land-use planning"	1	2.3%
Soils	1	2.3%
Environmental pollution	1	2.3%
Natural catastrophes	1	2.3%
Broader environmental monitoring	1	2.3%
Water resources	1	2.3%
Academic climate scientist	1	2.3%
Climate and nature interactions	1	2.3%

After grouping the 11 free-form / other responses:

- 44.6% - *Green spaces* – their impact within communities
- 38% - *Digital technology*
- 38% - *Blue economy – planning and industry*
- *12% - Climate Environmental Monitoring*
- *4.6% - Other (self describe)*

Options	Number of responses (43)	%
Green spaces	**19**	**44.6%**
"We use data sets for land-use planning"	Grouped	See above
Soils	Grouped	See above
Digital technology	**16**	**38%**
Blue economy	**16**	**38%**
Marine Spatial Planning	Grouped	See above
"Other" responses - total	11	26%
Climate environmental monitoring	**5**	**12%**
Hydroclimate extremes and their impacts	Grouped	See above
Environmental pollution	Grouped	See above
Academic climate scientist	Grouped	See above
Climate and nature interactions	Grouped	See above
Natural catastrophes	Grouped	See above
Other (too generic to group)		
Water resources	**1**	**2.3%**
Research	**1**	**2.3%**

Q2. Do you use/need environmental data (eg: EDS datasets) in your role?

Context provided: The five main NERC EDS data centres are: British Oceanographic Data Centre, National Geoscience Data Centre, Polar Data Centre, Centre for Environmental Data Analysis, Environmental Information Data Centre.

Link: https://eds.ukri.org/environmental-data-service

Options	Number of responses (43)	%
Yes	37	86%
No	6	14%

*Q3. Sources: **how many distinct datasets would you typically use for any specific work task?** Specifically, we are interested in datasets held by the EDS you need to access.*

This question was optional, 38 responded (out of 43)

*A majority **(89%) use 2-5, or more** datasets.*

We have investigated this further during out research interviews and have found that there are distinct use-cases whereby individuals combine many different datasets together, and we explored some of the different formats which require combination.

(See Finding 1, Page 15 in Research Report for more information)

Options	Number of responses (38)	%
1	4	11%
2-5	22	59%
5 or more	11	30%

Q4. Which of these categories of data do you currently source from the EDS?

Consolidating the overlapping answers in "Other":

Options	Number of responses	%
Environmental data	25	68%
Baseline data	4	11%
Ecological data	3	8%
Multidisciplinary	3	8%
Geological data	1	2.7%
Meteorological data	1	2.7%

"Other" answers before consolidation:

Options	Number of responses	%
Baseline data	**4**	**11%**
Environmental data	**25**	**68%**
"Other" responses - total	**8**	**22%**
"I work on Irish MSP, I use all available sources of marine data in Ireland." *(Multidisciplinary)*	1	2.7%
"Meteorological; any data that has GIS relevance" (Meteorological)	1	2.7%
"Mainly land and biodiversity" (Ecological)	1	2.7%
"Both of the above" (Multidisciplinary)	1	2.7%
"A whole range of environmental and other data including biota, water, soil and air chemistry, weather, land use, biodiversity, etc." (Multidisciplinary)	1	2.7%
Minerals data (Geological)	1	2.7%
Ecological data, vegetation data, species distribution data (Ecological)	1	2.7%
Raw met data from NWP-UKV or MIDAS, estimated of ammonia concentrations. Sorry but I don't understand in which of the above categories they fall in. (Ecological)	1	2.7%

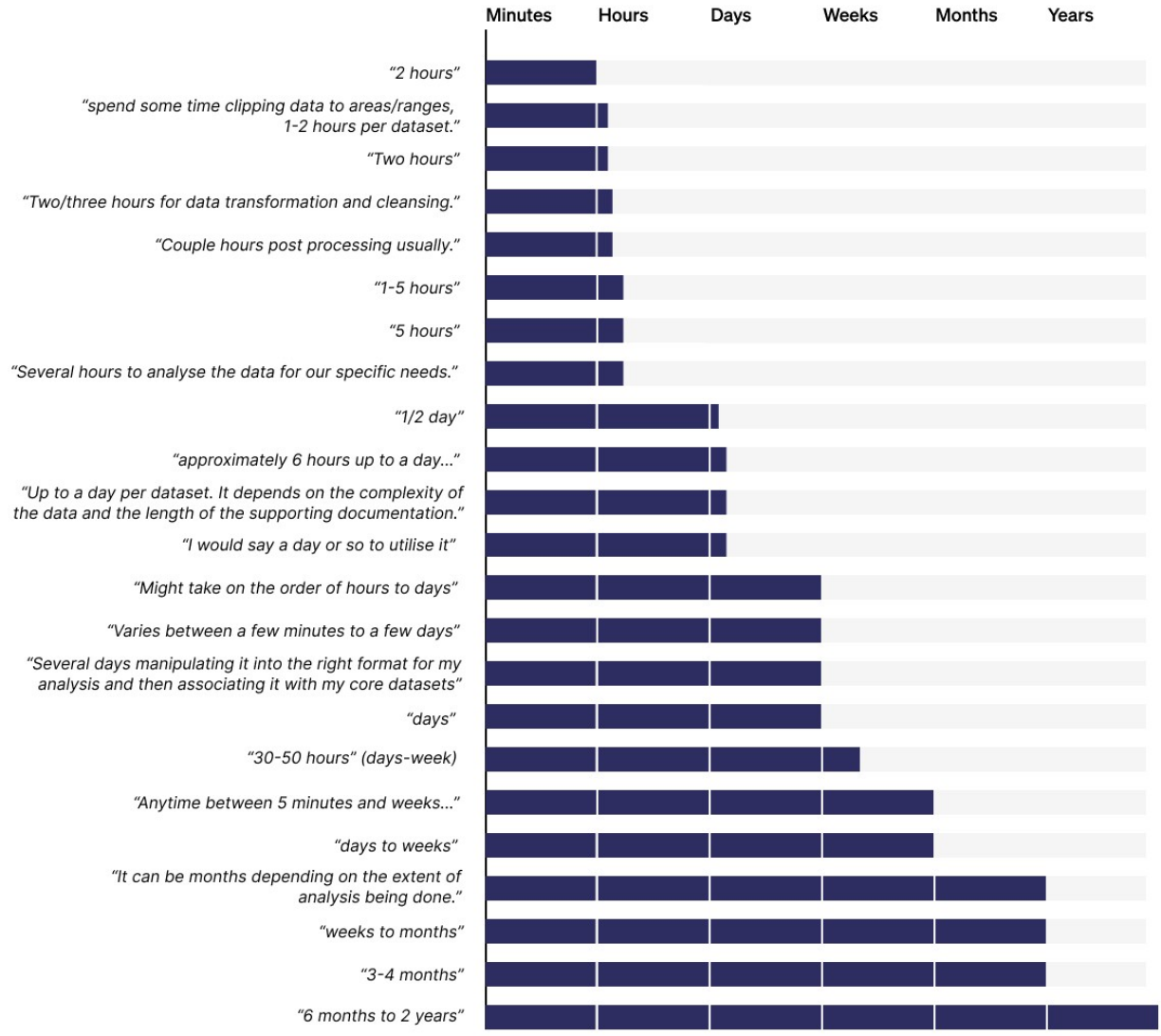
Q5. How much time do you spend on the data once you've sourced this from the EDS? *Eg: Before it's useful to you, or you can utilise it for what you need (e.g: to deliver an economic impact study)?*

Data Processing Time Varies Significantly
Participants reported varying times spent on data processing, ranging from a few minutes to several months, with common tasks involving data transformation, quality checking, and format adjustments.

4 participants did not respond with a definitive answer which could be plotted in time, but reported "significant" time use:
- *"Too much time"*
- *"Significant time to create ML and cloud friendly data."*
- *"Usually lots of time quality checking it, making sure it's in the right format, checking for errors benchmarking etc."*
- *"Often quite some time, firstly to understand the data using the supporting documentation and then integrating data sources."*

Eg: Some participants noted times of 30-50 hours or even days to weeks, indicating a need for streamlining data processing, and reducing the time it takes. Several responses cited the need for extensive formatting and error checking, suggesting a lack of usability in current data formats. This was confirmed in our moderated user research sessions.

Out of 34 responses, 23 are plotted on the timeline below: *Please note this is just an approximated visual representation, it is not representative of exact time to task.*



7 participants also highlighted the time spent is relative to the task being carried out:

- *"I read through the documentation. It can depend widely depending on familiarity with the data. Some times it will need lots of manipulation, filtering, joining etc. sometimes it's ready to go"*
- *"depends on the task"*
- *"It depends on the project?"*
- *"I may need to re-grid in space, or take annual or climatological means or anomalies, before I can use it."*
- *"...it depends on the formatting. Sometimes all I have to do is to read it or download it to use it. Sometimes I have to prepare an extraction routine, test it, etc. (especially on these GRIB2 files from Met Office, with a not-up-to-date documentation)""*
- *"Couple hours post processing usually. Does depend on what I am using it for. "*
- *"It can be months depending on the extent of analysis being done."*

Out of the 34 responses, only 1 participant uses data "as is".

1 participant had a management role and does not deal with data directly
("I do not do the work myself but rather supervise and encourage")

Recommendation

Create an API, or other digital tool to support improving the speed and efficiency of data processing. For example: cleaning, categorisation and labelling across various datasets, to help introduce consistency across varieties of different formats and categories of data. A step in the journey ***before*** ingestion/download.

Survey Answers in Full:

- *As a principle, I will not alter data that is not mine. I will use data available to me as evidence to inform decision-making. I do spend some time clipping data to areas/ranges, 1-2 hours per dataset.*
- *2 hours*
- *approximately 6 hours up to a day before using it for my work.*
- *do you mean how long does it take to process it? I write some python or bash and this might take on the order of hours to days or how long do I need to hold it for to complete projects? months to years*
- *Significant time to create ML and cloud friendly data.*
- *days to weeks*
- *Several days manipulating it into the right format for my analysis and then associating it with my core datasets*
- *too much time as the data are hard to find and generally not in a format that is useful to me*
- *It can be months depending on the extent of analysis being done.*
- *Usually lots of time quality checking it, making sure its in the right format, checking for errors benchmarking etc.*
- *I read through the documentation. It can depend widely depending on familiarity with the data. Some times it will need lots of manipulation, filtering, joining etc. sometimes it's ready to go*
- *1/2 day*
- *I do not do the work myself but rather supervise and encourage*
- *Often quite some time, firstly to understand the data using the supporting documentation and then integrating data sources.*
- *I would say a day or so to utilise it*
- *6months to 2 years*
- *It depends on the complexity of the data and the length of the supporting documentation. Up to a day per dataset.*
- *1-5 hours*
- *Two/three hours for data transformation and data cleansing.*
- *days*
- *Does depend on what I am using it for. Couple hours post processing usually.*
- *Several hours to analyse the data for our specific needs.*

- *Use as its*
- *5 hours*
- *Two hours*
- *I need those data with coding resources*
- *30-50 hours*
- *3-4 months*
- *depends on the task*
- *varies between a few minutes to a few days*
- *weeks to months*
- *It depends on the project?*
- *I may need to regrid in space, or take annual or climatological means or anomalies, before I can use it.*
- *Anytime between 5 minytes and weeks, it depends on the formatting. Sometimes all I have to do is to read it or download it to use it. Sometmies I have to prepare an extraction routine, test it, etc. (especially on these GRIB2 files from MetOffice, with a not-up-to-date documentation)*

Q6. What's your biggest challenge finding relevant datasets or data across the EDS data centres? Eg: what takes up most of your time?

Participants report significant challenges in finding relevant datasets across EDS data centres, highlighting issues with data formatting, accessibility, and usability. Common themes include difficulties with search functionality, lack of user-friendly interfaces, and the need for better compatibility with analysis tools.

Recommendations:
• Improve search functionality and a way of bringing together different sources and varieties of datasets into one portal (eg: NUAR National Underground Asset Register does this for underground / geological data on land, in the UK – we don't currently have 'one place' for Environmental data in the UK like this.
• Consider introducing auto-download features similar to the RGBIF package to enhance usability. categorization and labelling of datasets to reduce confusion.

Survey answers in full (Note: spelling /grammar have been corrected):

1. *"Ensuring origins and quality."*
2. *"Lack of familiarity with project/dataset names. I know what data I want, but that doesn't always make it easy to find!"*
3. *"Data search itself is the greatest difficulty, especially the hierarchy and naming conventions used for the datasets could have been systematic".*
4. *"One-click data still isn't implemented 10 years after it was promised...."*
5. *"Retrieval. ARCO formats for streaming and subsetting would improve the experience."*
6. *"Different grids and formats"*
7. *"Finding the most up to date dataset that is actually avaialble at the scale I need. Some are too large (weather) and not specific enough."*
8. *"Find the data on BODC, I can't even find data I submitted myself. And then I spend time turning it back into a format I can use."*
9. *"Just to get the data into the format or orientation you need, especially if metadata is not written well."*
10. *"Formats."*
11. *"Most datasets are not designed to be generally applicable, or at least not advertised as such. They are usually super specific; this is more in reference to EIDC data. So, it can be difficult wading through data sets about foraging habits of a species of moth on farmland in York in May 1998 when I'm trying to do something more generic like map UK biodiversity."*
12. *"Making sure data aligns with and is compatible with answering specific research question"*
13. *"Downloading and loading into other software - APIs are essential"*
14. *"Search functions currently don't really work. checking supporting documentation and assessing data quality"*
15. *"Whether or not it has the variables I need how I need them"*
16. *"Knowing that the datasets exist"*

17. *"You can't save your search so when you leave a catalogue, you leave where you have got to. Finding and accessing the most up to date data. Data publication is patchy and people have only been making data available for about the last 20 years so there's a huge lack of legacy data. Some people don't bother to publish their data."*

18. *"Figuring out what's available for any given need/environmental variable. Understanding differences and similarities between datasets"*

19. *"Just finding the most up-to-date datasets - current industry question is revolving around the last three years and future changes. While it is good to have a baseline comparison for 1961 to 2010, the data is not useful in the wider impact assessment"*

20. *"I usually need to know about the data somewhere else before I will think about going to search at the data centres"*

21. *"Would be better for EDS to make itself more user friendly. Maybe an R package to auto-download like how RGBIF package allows access to its mainframe. You'd get a lot more uptake doing that."*

22. *"Gaining access, ensuring licencing agreement allows for it to be used commercially, and downloading to our system."*

23. *"Large files that need to be uploaded"*

24. *"Searching"*

25. *"Data compilation"*

26. *"Harmonising format with climate data"*

27. *"Finding exact tiles for my study area or to resample datasets"*

28. *"Understand them, downloading them , are of interest"*

29. *"Understanding the data and it's structure. README files are often empty or incomplete"*

30. *"Pre-processing into a proprietary form to be used by our local models"*

31. *"Spatial searching, lack of an up to date STAC API"*

32. *"Each has a different interface. Individual measurements are sometimes repeated in different datasets. You often need to already know what you are looking for."*

33. *"On CEDA or EIDC, I sometimes struggle to make sure the dataset I want to access is the one I intend to use. For instance, there are dozens of NWP datasets, and* ***the grouping is hard to understand.***

    I always ***triple check and ask colleagues*** *before to ask for the one I should use as it's not 100% clear. Also the process for asking authorisation to work on a dataset is quite robotic and not so clear from the user end, more human interaction would be welcome. I recently asked the authorization to work on a specific NWP dataset as we had been prevented from using that dataset on similar project, and also because I had received an automatic notification saying I couldn't use another NWP dataset I had never used* ***(turned out the id was for a group of datasets, so I couldn't understand which dataset it was about).*** *It turns out the dataset I wanted was open to all CEDA users. When I asked to CEDA to double check I was indeed allowed to use it, I got an automatic notification for a group of datasets,* ***and no human wrote to explain*** *whether I understood correctly my permissions…"*

Insight: across responses, the dominant pain points are about discoverability, interpretability, and usability. Many respondents describe a workflow where finding, understanding, accessing, and preparing data each represent significant, independent barriers.

(See Finding 2 in Research Report, Page 18)

"Challenge" categories identified:

1. **Location & Discovery**
   Locating datasets, even when users know broadly what they want. Challenges in obtaining the data once identified (findability / discovery, inconsistency in naming conventions, search experience eg: difficulty finding exact tiles or study-area-specific subsets). Usability and support issues were also mentioned (overwhelm of volume of availability, too many different interfaces/dashboards, perception that platforms could be more user-friendly).
   Insight: Lack of a trusted, curated 'single source' significantly increases time and effort to locate, introduces risk - eg: missed datasets which would be better suited for projects.

2. **Fit for purpose**
   Challenges in understanding what a dataset contains and whether it is fit for purpose (lack of visibility of what they contain, insufficient metadata, documentation, limited reusability, specificity of data not broad enough for use, challenges understanding differences and similarities between related dataset). Currency of data (difficulty identifying the most up-to-date, lack of digitised legacy data, unclear versioning, data availability not recent enough for policy/industry). ***Insight: Data value is constrained by limited reusability and poor alignment with user decision-making needs.***

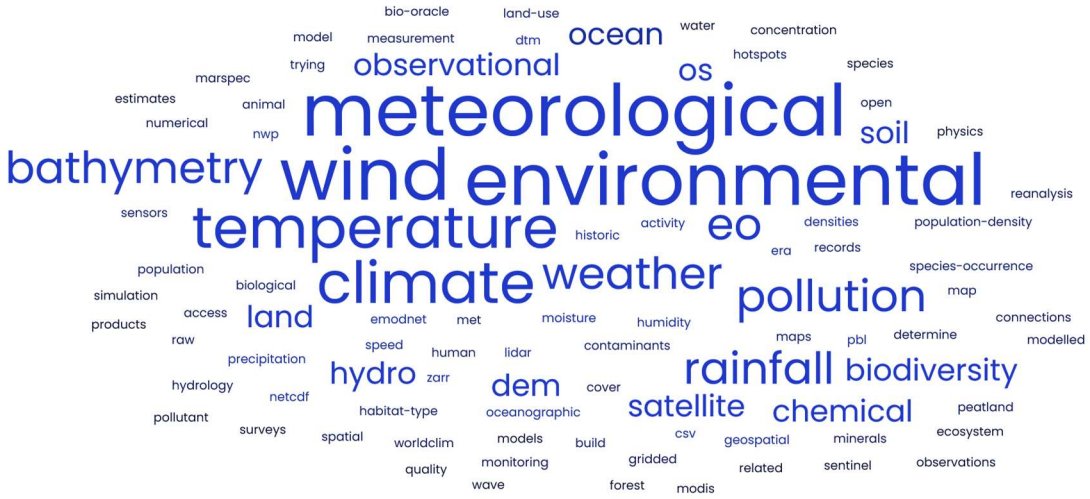
3. **Access & governance**
   Even when data is identified and permitted, it is difficult to retrieve and use efficiently due to licensing, permissions and other issues. ***Insight: Governance processes introduce friction that discourages uptake and wastes user and provider time.***

4. **Data formats and interoperability / compatibility across systems**
   e.g: Multiple grids, projections, file formats, orientation and structural mismatches across datasets, lack of standards/formats such as STAC or ARCO and language-native tooling e.g: R or Python packages for automated access…
   Insight: "Format multitude" prevents scalable, automated, and repeatable use of data.

5. **Readiness for use**
   …which leads to a significant effort in work required before data can be analysed or ingested into models or to harmonise datasets to use in conjunction with each other (data not analysis-ready "out of the box"). ***Insight: Disproportionate time on data wrangling rather than insight generation.***

Q7. What kind of data do you currently utilise most often, in your role

Participants who took part in the survey had a strong focus on hydrometeorological, climate, and environmental data. Categories include weather data, satellite imagery, and various environmental measurements.



Survey answers in full, arranged into three categories:
(Note: spelling /grammar have been corrected):

Type of data / model

- *Wind speed data*
- *Many, but broadly: Bathymetry, Oceanographic, Activity Densities*
- *Meteorological data (PBL, Temperature, Wind, Humidity, etc.) and LiDAR data*
- *Satellite EO data*
- *Geospatial*
- *NWP (Numerical Weather Prediction)*
- *Weather data - rainfall and wind and temperature*
- *Historic environmental data*
- *Hydrometeorological data (precipitation, temperature, rainfall, soil moisture).*
- *DEM/DTM/Bathy and Meteorological data*
- *Biological Records*
- *Biodiversity species occurrence data*
- *Weather data*
- *All aspects of land and biodiversity data - we also deposit data*
- *Soil and environmental data*
- *Met data*
- *Environmental data, measurements from sensors, map data*
- *Environmental measurement data overlaid on a map with population density, or land use, or habitat type. Looking at different contaminants (trying to build up a map of chemical/pollution hotspots) with human or animal population data to see whether they are connected.*
- *Weather data/reanalysis data*
- *It varies, but mostly climate, meteorological, and hydro data.*
- *Chemical pollution and hydrology data, as well as land cover and modelled gridded data*
- *Anything open access. All the usual areas, numerical, spatial.*
- *Climate model data.*
- *Minerals data, OS, historic OS maps*
- *Environmental monitoring*
- *Peatland related, water quality, forest, surveys*
- *Wind, Rainfall and Temperature*

- *Climate simulation data and the observational data necessary to evaluate historical simulations.*
- *Ecosystem and species data*
- *will use Pollution data*
- *Satellite data (EO)*
- *wind/wave/ocean/bathymetry*
- *climate data*
- *Ocean physics data, from models and observations and observational products.*
- *Meteorological data, pollutant concentration estimates*

Data formats

- *Netcdf or Zarr or csv formats.*
- *Raw data*

Data sources

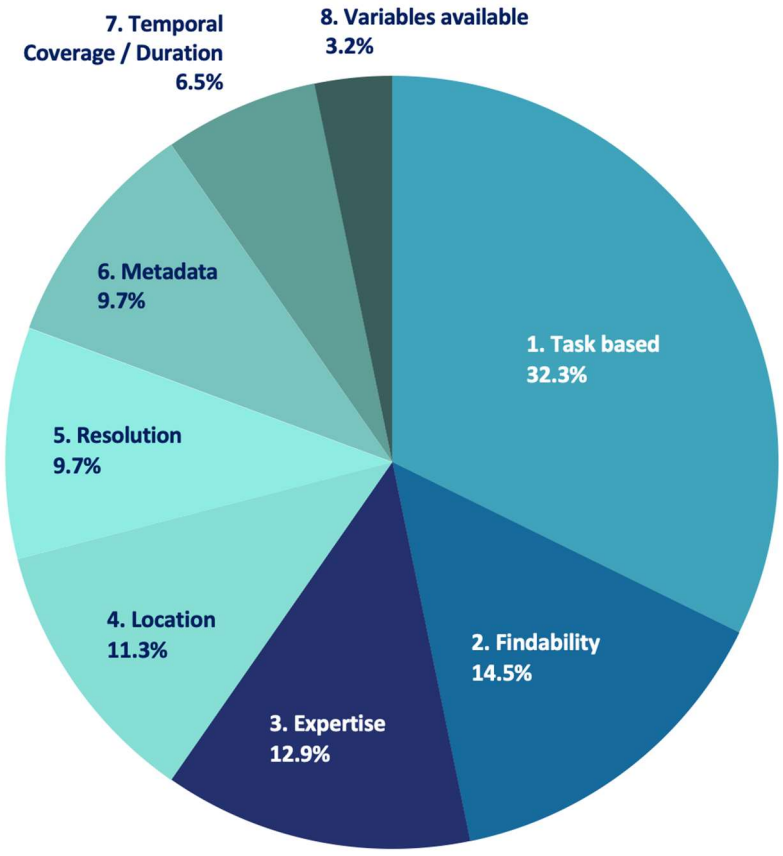
- *EMODnet*
- *ERA5*
- *WorldClim, Bio-ORACLE, MARSPEC recently as they have easy pipelines to code for download.*
- *MODIS DEM 30m Sentinel*
- *Earth Observation (Satellite)*

Q8. How do you choose which datasets are relevant to you?

Participant's key consideration criteria for selecting relevant datasets (order priority by number of mentions, 62 answers from 40 participants):

1.	Task relevance (fitting the datasets to their specific research questions and project requirements)	**20 responses**
2.	Ease of findability and availability of data (using search to find data)	**9 responses**
3.	Expertise and existing knowledge (individual or community subject matter expertise)	**8 responses**
4.	Proximity to location and geographical coverage (specific area relative to focus of research)	**7 responses**
5.	Spatial and temporal resolution / extent	**6 responses**
6.	Metadata/Documentation quality & completion	**6 responses**
7.	Temporal coverage and completion of dataset (eg: from year X-Y, no gaps in data)	**4 responses**
8.	Variables available	**2 responses**

The responses based on the categories are visualized below:
Please note this is just an approximated visual representation, it is not relative to the number of individuals who filled in the survey, as some of the responses listed multiple aspects.



Variability between the methodology to determine 'relevance'
We had answers on both sides of the spectrum ("We develop a protocol to search for data, then review lists of what we found to decide whether to include/exclude" and "I take advantage of the CF conventions, the DRS conventions used by CMIP " – compared to "pot luck")

One individual mentioned *"Customer demand"* which is also important to note as different User "Archetypes" will have different reasons for working with data. **(See Research Report, Page 34)**

A user flow/journey diagram has been created based on this feedback (Page 13 in Research Report)

Survey answers in full, mapped to categories above
(40 participants answered, 3 N/A responses)

1. Relevance (20)

- *"Project's requirements…"*
- *"Relevance to the query…"*
- *"Fit to problem/task"*
- *"Depending on work"*
- *"(is it useful for my use-case)"*
- *"Extent, resolution, duration"*
- *" …and the data itself"*
- *"…whether the environmental parameters meet the science questions requirements"*
- *"Relevancy"*
- *"Depends on the task"*
- *"Relevance"*
- *"The one that suit my need"*
- *"Customer demand"*
- *"Datasets that relate to my research questions"*
- *"Depends upon task at hand"*
- *"Always related to the question I need to investigate"*
- *"Use for validation and future projections. Latest CMIP models for future climate work"*
- *"Depends on the project…"*
- *"Satellite, etc"*
- *"From CEDA and EIDC catalogues pages (source)"*

2. Proximity of location and geographical coverage (7)

- *"Proximity to the location…"*
- *"Modelling fishing activities in the Mediterranean"*
- *"Proximity to the location…"*
- *"Proximity and geographical coverage…"*
- *"Usually filter by spatial area…"*
- *"Based on the region"*
- *"Looking for the right location…"*

3. Variables (2)

- *"Depending on the variables that they contain…"*
- *"Availability of variables…"*

4. Spatial and Temporal resolution / extent (6)

- *"…and the spatial resolution and temporal resolution I need"*
- *"Resolution of the data, temporally and spatially.*
- *"Resolution, time periods…"*
- *"Spatial and temporal resolution"*
- *"Spatio-temporal resolution…"*
- *"Spatial extent --> temporal extent --> resolution"*

5. Temporal coverage and completion of dataset (4)

- *"Look for matching years and…"*
- *"Recency, completion, bounded areas"*
- *"Usually filter by date…"*
- *"Timescales I'm investigating…"*

- ***6. Metadata/Documentation related (6)***
- *"Hoping they have good metadata..."*
- *"Precipitation, air temperature, wind speed and direction."*
- *"Based on the description written and the metadata details."*
- *"From the description"*
- *"Title, supporting documentation…"*
- *"and using the documentation to understand what the datasets hold…."*

7. Ease of findability / dataset availability (9)

- *"Searching for keywords"*
- *"Usually I search for the dataset in catalogue search bar and check the metadata of the dataset displayed in the result and download them if useful"*
- *"Availability and ease of use/access"*
- *"We develop a protocol to search for data, then review lists of what we found to decide whether to include/exclude."*
- *"Pot luck"*
- *"Availability of data"*
- *"I don't have station data for most of environmental data in the horn of Africa. So I fill the data demands by your products"*
- *"I take advantage of the CF conventions, the DRS conventions used by CMIP…"*
- *"--> licensing --> ease of use"*

8. Expertise / existing knowledge (8)

- *"Based on research"*
- *"My expertise + co-design"*
- *"Experience..., community confidence, validation and uncertainty data in reports, journal papers."*
- *"Usually expert knowledge and in association with taxa study groups determines what is needed."*
- *"Scientific expertise"*
- *"…and my own domain knowledge."*
- *"Collaborations with experts and data providers, relevant literature"*
- *"Advice from others, own research, academic papers, CEDA archive"*

Q9. What data formats do you most often require or use?

Note: This is a multiple-choice question; the percentages won't add up to 100%.

Note: Further contextual feedback as to why certain formats are preferred over others was discovered during our moderated user research sessions. **(User Research Report – Page 16 and 17)**

Options	Number of responses	%
CSV / ASCII / Text	30	70%
Shapefile / GeoJSON / KML	27	63%
NetCDF (Network Common Data Form)	24	56%
GeoTIFF	14	33%
GRIB (Gridded Binary)	7	16%
HDF/HDF5 (Hierarchical Data Format)	6	14%
Cloud formats (eg: Zarr, COG, Parquet)	6	14%
BUFR (Binary Universal Form for the Representation of meteorological data)	2	4.7%
Other responses (free text field)		
Proprietary formats	1	2.3%
FASTQ and GFF	1	2.3%
EA API	1	2.3%

Q10. Do you prefer to download data or access via an API?

Insight:
Direct data downloads are still the most popular option among users surveyed, closely tied with the use of both (downloads and APIs).

During user research, the reason behind this was often cited as to how the data needs to be utilised (on internal systems, or due to the data size, for processing, for example).

(User Research Report Finding 2: Page 18 and Finding 6: Page 27-28)

Options	Number of responses	%
Download data directly	19	44%
Use and API to access the data	15	35%
Other responses (free text field)		
Both *("Both - download and use APIs", "Both - depends on project", "both", "Browser-based for the first time, API for repeated downloads")*	4	9.2%
"Some read-only copy like Datalabs asset repo, or Google colab or Kaggle"	1	2.3%
"Use data in the cloud"	1	2.3%
"As long as what I need to do is clear, and as long as it is reproducible, I don't mind."	1	2.3%
"Depends on size of data and use case."	1	2.3%
"I use the CEDA connected to Jasmin a lot otherwise directly (download)"	1	2.3%

Q11. What takes up most of your time?

The majority of participants reported that 'Cleaning the data' takes up most of their time, followed by 'Analysis' and 'Sourcing the data'.

"Converting to proprietary format" was also notably added by one respondent.
(See Question 6 – "Challenge Nr 4", Page 12 in this report)

Recommendation: 'Cleaning the data' is the most time-consuming task for participants, indicating a possible need for improved data cleaning processes or tools. Consider exploring automation for data cleaning to enhance efficiency.

Options	Number of responses (43)	%
Cleaning	17	40%
Analysis	12	28%
Sourcing	10	23%
Other responses (free text field)		
"depends on the dataset"	1	2.3%
"All of the above"	1	2.3%
"cleaning and analysing"	1	2.3%
"converting to proprietary format"	1	2.3%

Based on the "other" responses being consolidated into votes:

Options	Number of votes
Cleaning	19
Analysis	14
Sourcing	11
Other responses (free text field)	
"depends on the dataset"	1
"converting to proprietary format"	1

Q12. AI Readiness. What does "AI ready data" mean to you?

Participants provided diverse definitions of 'AI ready data'. Many responses reflected scepticism regarding the practical implications of the term, highlighting confusion and varying interpretations. **Themes mentioned (in order of mentions):** (1) Structure/Formatting & AI/ML compatibility, (2) Metadata/Semantics & Interoperability, and (3) Data Quality/Cleanliness (the requirement for standardised metadata and accessible formats) – reinforcing the responses from Question 6 (*See Challenges – Question 6, Page 12 in this report).* Another key theme (4) which emerged was around concerns and scepticism around the topic of AI (or the term itself being 'too general' to qualify).

We delved into this further in our 1-2-1 user research interviews for context.
(See Report Finding 5: Page 25-26)

How people interpret the term "AI-read data":

(1) Structure, format & AI/ML compatibility	(2) Skepticism, ambiguity and buzzwords	(3) Metadata, semantics & interoperability	(4) Data quality and cleanliness
"Datasets with sufficient context and labelling to allow application of machine learning algorithms."	"It puts me off wanting to use the data… actively avoid GenAI"	"Semantic interoperability provided by high quality metadata (incl. artefacts like dictionaries, controlled vocabularies, taxonomies, ontologies, knowledge graphs)"	"Clean, complete data in an efficient and widely compatible format"
"Standardised data shape and metadata… streamed for exploration… collocated with compute"	"that's context dependent"	"Data which is well documented and available in suitable formats for the applications that ensure interoperability."	"Providing clean, non-messy data. Easy to use, with a few or no errors."
"Clean, no missing data values, gridded, fast for reading it from disc (Zarr)"	"Not sure actually"	"common terminology for easier access"	"Quality controlled, checked, cleaned and in a readable format."
"An AI process has summarised the data for me in the way I want it?"	"Nothing."	"Clean, FAIR, machine readable"	
"I can build an AI agent on it without tearing my hair out…"	"Marketing hype and the implication that whoever has labelled the data as such has no requirements for accuracy."	"The data is stored in files… metadata about coordinates and variable definition… standardization"	
"compatibility with models through standardised structure and forma"t	"as a statement it could mean many things"	"Metadata catalogued in something like STAC and in a cloud native format like COG"	
"Data that can be easily plugged into a ML model or an AI-based tool"	"Not much… when I hear "AI ready data" I more think about LLMs…"	"Documented, with standard metadata incl on the data quality"	
"data that a machine can understand"	"It is sources of information"		
"Data that are ready to use in a model"	"Nothing"		
"AI-ready data is related to analysis-ready data… high-quality, cleaned, structured, labeled… context"	"I don't know how that is applicable to me… interested to see a concrete example"		
"Structured and accessible for an AI pipeline."	"no idea"		
"Data that can be interrogated by AI"	"Meaningless buzzwords"		
"ready for AI"			
"Data that is available from the AI"			
"ready for running AI related models"			
"Labelled data to traine models"			
"help in narrowing what I am looking for… initial analysis"			
"Data available on cloud-based storage with a flexible API."			
Total: 18	***Total: 12***	***Total: 7***	***Total: 3***

Responses in full:

- *semantic interoperability provided by high quality metadata (incl. artefacts like dictionaries, controlled vocabularies, taxonomies, ontologies, knowledge graphs)*
- *Data which is well documented and available in suitable formats for the applications that ensure interoperability.*
- *Clean, complete data in an efficient and widely-compatible format*
- *Providing clean, non-messy data. Easy to use, with a few or no errors.*
- *Datasets with sufficient context and labelling to allow application of machine learning algorithms.*
- *It puts me off wanting to use the data, because it's almost certainly not in a suitable format for anything I'm likely to use it for. I do use some classical ML methods, but actively avoid GenAI*
- *Standardised data shape and metadata. Available in a format that is easily subset on various coordinates. Able to be streamed for exploration and then retrieved to disk collocated with compute for analysis.*
- *Clean, no missing data values, gridded, fast for reading it from disc (zarr)*
- *An AI process has summarised the data for me in the way I want it?*
- *that's context dependent*
- *Not sure actually*
- *Nothing*
- *Marketing hype and the implication that whoever has labelled the data as such has no requirements for accuracy.*
- *I can build an AI agent on it without tearing my hair out*
- *compatibility with models through standardised structure and format*
- *as a statement it could mean many things*
- *common terminology for easier access*
- *Clean, FAIR, machine readable*
- *Data that can be easily plugged into a ML model or an AI-based tool*
- *data that a machine can understand*
- *Data that are ready to use in a model*
- *Not much, as the data that I use is often too sparse to use machine learning or AI. While we develop deep learning application and use neural networks in parts of our work, when I hear "AI ready data" I more think about LLMs and that's not what we use currently. But generally speaking, the better the data is prepared and cleaned, the easier it is to use for any application beyond AI.*
- *AI-ready data is related to analysis-ready data, as they both need to be prepared in such a way machines can consume readily. Analysis-ready data is more like a harmonised data cube that can be fed into a statistical or machine learning model directly. It is a result of integrating and QC many source datasets. They need to be high-quality, cleaned, structured, labeled. They needs to be 'large sample' as well as lots of models don't work well if the number of data points is too small. A prime example is the CAMELS-GB dataset for hydrology, which just published v2 in EIDC. For AI-ready data, I think the additional requirement is information on the context of the dataset. This is well understood in the modeller, but it may be difficult for an AI model to guess.*
- *Structured and accessible for an AI pipeline.*
- *Quality controlled, checked, cleaned and in a readable format.*
- *Data that can be interrogated by AI*
- *ready for AI*
- *Data that is available from the AI*
- *ready for running AI related models*
- *It is sources of information*
- *The data is stored in files which with metadata about coordinantes and variable definition (including methods of averaging, if any) to the highest possible level of standardisation.*
- *Nothing*
- *Labelled data to train models*
- *help in narrowing what I am looking for and with some initial analysis to see if what I am looking for is fit with what i want to investigate further*
- *I don't know how that is applicable to me. I'd be interested to see a concrete example*
- *no idea*

- *Metadata catalogued in something like STAC and in a cloud native format like COG*
- *Documented, with standard metadata incl on the data quality*
- *Data available on cloud-based storage with a flexible API.*
- *Meaningless buzzwords*

Q13. Do you currently use your data in conjunction with any AI platforms or tools?

The usage of data in conjunction with AI platforms is evenly split, with 22 respondents (50%) reporting use and 22 respondents (50%) indicating no use.

Options	Number of responses (43)	%
Yes	21	49%
No	22	51%

Q14. Have you considered whether AI might simplify/improve/speed up the steps in your current workflows?

- AI shows promise in simplifying tasks like coding and data cleaning: 11 out of 22 (50%) responses explicitly mentioned its usefulness.
- Some participants caution against relying on AI due to concerns over output reliability – the need for transparency and replicability of workflows was
- Participants expressed a clear interest in further integrating AI into workflows, especially for low-risk and well-defined tasks.

Answers in full:

Considered:

- *"Many ways, testing necessary"*
- *"I use AI to help write data-processing code, but don't yet embed AI within the pipeline itself"*
- *"Data access command creation via API or cli."*
- *"Yes, it can, but it can also reduce transparency."*
- *"Yes (5 responses)"*
- *"I think for clearly defined, low risk, verifiable tasks it could be super helpful"*
- *"Yes. should be used to simplify and speed up data cleaning and processing".*
- *"Already use AI tools to simplify and improve current code and workflows"*
- *"Yes, but currently there is little opportunity."*
- *"Yes, I'm using CNN models"*
- *I use AI to help me write code, that speeds up the coding part but not the processing time*

"For sure. I already know it can and want to develop such integration. Its best usefulness would be to drag net and clean multiple OA sources."

Not considered:

- *Typically AI can help with coding a specific section of a code, however, this comes then with the issue of provenance, when this code might be published on GitHub. There is a speed up in helping to source data, but it is also not reliable, which is why, the benefit to workload ratio is not high.*
- *Not particularly. I use copilot to help sometimes but rarely*
- *No*

***Q15. Please specify which tools, platforms or models you currently use** (eg: pre-existing LLM like ChatGPT, or building your own ML models?)*

Respondents primarily use a variety of tools including ChatGPT, Copilot, and custom ML models for processing and preparing data. **Common preprocessing methods include** dimensionality reduction, meta/data standardisation, and transforming data into simpler, AI-ready formats.

- 22 respondents provided insights across a mix of proprietary and pre-existing AI models, tools and data preparation methods. Most (
- Diverse preprocessing techniques are used based on the specific AI or ML model, **indicating a tailored approach.**

Pre-existing platforms or tools (15)	**Bespoke / Build own (4)**
I'm using CNN models for image segmentation and use spectral indices along with histogram threshold to classify data that needs to be provided as training data for ML *models (**Note**: use of pre-existing AI frameworks with domain-specific customisation)*	I write bespoke ML models for use within modelling workflows. Mostly I work with gridded data which is already fairly "AI ready" by my definition
ChatGPT for ad-hoc help with coding e.g. for data processing, but not as part of the pipeline itself.	Own ML. Any tool that can deliver numpy arrays is fine for most toolkits.
I use individual models from within the python library scikitlearn on ocean data, mostly unsupervised classification tools. I need to filter my data to create a training set that is as evenly sampled in space and time as possible.I then need to apply dimensionality reduction via PCA analysis before it is ready to use with an unsupervised classification tool. *(**Note**: all mentioned are existing tools and techniques – Python, established algorithms, known techniques such as PCA)*	Yes, modified GPTs that are in line with data privacy regulations. Usually, the data should be "AI ready" depending on the problem I want to solve. There is a difference between classification and prediction, so the data has to adhere to the simple data format principles that are used for a GAN, or Graph Neural Network, or a Bayesian classifier. There is no one solution fits all. *(**Note:** Using some pre-existing AI tools but with configuration/fine-tuning, advanced use)*
ChatGPT, GitHub Copilot	
The prediction using ML models is very important which is considering available data to train the activities happening.	
Llms	
ChatGPT	
Depends on the system. Bespoke LLM is best, but eventually be good if accessible portage was built more formally for cGPT and CoP.	
Existing LLMS, i.e. ChatGPT, Gemini and Claude. I clean the data and format it in a way that makes my querying and conversation easier	
copilot and chatgpt. I just give data to these tools	
ChatGPT, CoPilot. Generally speaking having a simplified structure is paramount for AI processing.	
I generate summaries that can be easily digested by LLMs (e.g. generating a comma separated species list for a location, rather than giving it a dataframe of lat and longs and species). Or give the AI tools to be able to generate these summaries.	
LLm for coding to prep/sort/clean/understand data. *(**Note:** no mention of building or training model themselves)*	

LLMs, software in-built analysis tools, I turn them into science ready datasets	
pre-existing LLMs	
Anemoi datasets. Or cloud native format conversion. Metadata standardisation using custom code which changes depending on the data source type. *(**Note:** ETL logic, not AI/ML algorithms)*	

Q16. What's the biggest challenge you've come across in your AI workflow creation?

Key challenges in AI workflow creation include data labelling, consistency in outputs, and clarity in dataset requirements. Participants also expressed concerns over trustworthiness and reliability of models. ***Option 2 (Data collection, preparation and management) is most mentioned out of the 19 responses collected.***

AI Lifecycle Stage	Challenges	Responses
1. Problem definition, method selection	Unclear goals, methods, workflows	3
2. Data collection, preparation, management	Quality, scale, formats, labelling, misuse	6
3. Model development and prompting	Skills gaps, prompting, coding	3
4. Evaluation, validation, reproducibility	Inconsistent, unreliable outputs	3
5. Deployment, operations, constraints	Cost and usage limits	1
6. Governance, trust, human factors	Time, trust, accountability	3

Full list of responses:

1. Problem Definition (3)

- *Clarity of dataset requirements from customers.*
- *i don't think i have an "AI workflow", what does that even mean?*
- *Figuring out which method is most suitable.*

2. Data collection, preparation, management (6)

- *"Generally cross-platform jumping"*
- *Compute and shaping of massive datasets.*
- *Disparate data types. Changing formats. Changing API's (over time), lack of documentation, lack of example data.*
- *Having enough input data that the outcome is reliable :)*
- *People who don't use CF data, or use it wrongly, so tools don't work as expected and we have to code around it.*
- *Data Labelling*

3. Model development and prompting (3)

- *Data analysis*
- *prompting*
- *Coding*

4. Evaluation, validation, reproducibility (3)

- *Unpredictable and irreproducible results when using an external LLm service like gemini, gpt5 etc*

- *Consistent output*
- *The model coming out with rubbish answers*

5. Deployment, tooling, operations, constraints (1)

- *Limited credits*

6. Governance, trust, human factors (3)

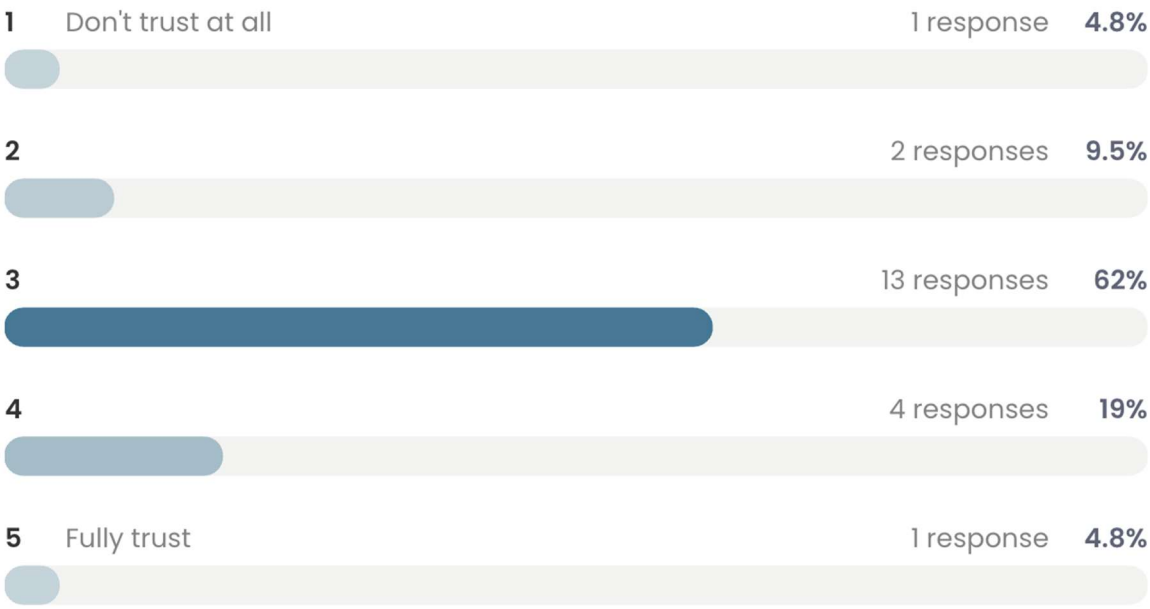
- *trustworthiness*
- *Time away from teaching to do it. Duh.*
- *lack of computing knowledge, increased work with RSEs that aren't subject matter experts, which results in standard based data pipelines but delays due to lack of understanding of subject*

Q17. How much do you trust the output of your AI workflows?

[Format: Linkert Scale rating from 1-5]

The majority of participants (63.6%) rated their trust in AI workflows as a neutral 3 on a 5-point scale, with a weighted average score of 3.09. Responses showed some clustering around the middle of the scale, indicating mixed levels of trust.

- 4 out of 22 (18.2%) respondents rated their trust as 4, indicating a reasonable level of confidence.
- 1 out of 22 (4.5%) rated their trust as 5, showing very high trust is rare.
- A notable segment (2 out of 22, 9.1%) expressed low trust with a rating of 2, and 1 respondent (4.5%) rated their trust as 1.



Q18. What would give you peace of mind that the results you receive are trustworthy? What do you need for assurance?

From 22 responses out of the 43 participants, 14 (63.6%) emphasized the need for explicit processing steps or an audit trail for trustworthy results. Only a small portion mentioned the provision of data sources or overall transparency.

- The clear preference for transparency through explicit processing steps indicates a requirement for methodological rigor and reproducibility.
- There is opportunity to enhance trust by consistently providing detailed documentation, visualisation of data pathways and processing.

Options	Number of responses	%
Explicit processing steps / "audit trail" is visible	13	62%
Data sources / pathways provided	3	14%
Both *("Both"; "all of the above to some extent, context dependent")*	2	9.6%
Other responses (free text field)		
as much transparency and available testing results as possible	1	4.8%
Explicit steps and data sources provided. Proof with snippets of the data and evidence from code/outputs provided	1	4.8%
Recognised post-model parameter testing. But use these anyhow.	1	4.8%

Additional comments left at the end of the survey:

"I'm surprised some AI tools could access datasets I am not even allowed to access on relevant research projects...:

"I firmly believe that 'AI' systems are nothing more than a 'get rich quick' scheme for the companies that develop them. LLMs are non-deterministic, frequently inaccurate, reinforce existing biases in datasets, and have a far higher environmental impact than comparable systems. No responsible organisation should be considering their use."

''It depends what kind of AI you're using. classical AI? (e.g. statistical ML methods); deep learning? GenAI?"

Q19. Would you be interested in co-design or collaboration on the 'High-5 for AI' project?

Out of 43 respondents, 13 provided their email contacts, and we have reached out to them for user-research interviews. The findings from these will be available in the User Research Report once completed.

High 5 for AI is an Environmental Data Service Project.
For more information: eds.ukri.org/projects/high5-ai

If you have any questions regarding the survey, please get in touch with:
Ola Podgorska, Lead Digital Design Researcher – olapod@noc.ac.uk

National Oceanography Centre
Joseph Proudman Building, 6 Brownlow Street, Liverpool, L3 5DA
United Kingdom +44 (0)151 795 4800

National Oceanography Centre is a company limited by guarantee,
set up under the law of England and Wales, company number 11444362.
National Oceanography Centre is registered as a charity in England and Wales,
charity number 1185265, and in Scotland, charity number SC049896.