NERC Environmental Data Service

Advancing the use of environmental data in AI solutions

Final project report for HIGH5 for AI
Report number OR/26/050

Authors: **(Alphabetical order)** Jesse Alexander[2], Oladimeji Awe[4], Andy Bean[3], Emma Bee[3], Patrick Bell[1], Andy Bevan[3], Gordon Blair[4], Jon Booth[3], Jennifer Bulpett[2], Chris Card[3], Matthew Cazaly[5], Esther Conway[2], William Cross[2], Louise Darroch[5], Thomas Gardner[5], David Philip Green[4], Rachel Heaven[3], Michael Hollaway[4], Joshua Holland[4], Jasmine Hunter[4], Lucy Killoran[2], Diane Knappett[2], Matthew McCormack[5], Sam Pepler[2], Alekasdra Podgorska[5], Shwetha Raveendran[3], Helen Rawsthorne[4], Jennifer Roebuck[4], Roman Roth[3], Helen Snaith[5], Ag Stephens[2], Matthew Watkeys[4], Carl Watson[3], Daniel Westwood[2]

Jesse Alexander [2] https://orcid.org/0009-0006-2877-3197
Emma Bee [3] https://orcid.org/0000-0003-4504-0009
Patrick Bell [1] https://orcid.org/0009-0001-7871-7429
Gordon Blair [4] https://orcid.org/0000-0001-6212-1906
Jennifer Bulpett [2] https://orcid.org/0009-0009-8142-8298
William Cross [2] https://orcid.org/0009-0004-8084-8681
Louise Darroch [5] https://orcid.org/0000-0003-4163-9575
David Philip Green [4] https://orcid.org/0000-0002-3256-8640
Rachel Heaven [3] https://orcid.org/0000-0002-6172-4809
Mike Hollaway[4] https://orcid.org/0000-0003-0386-2696
Jasmine Hunter [4] https://orcid.org/0000-0001-8643-2602
Lucy Killoran [2] https://orcid.org/0000-0002-1907-6078
Diane Knappett [2] https://orcid.org/0000-0002-7596-6729
Matthew McCormack [5] https://orcid.org/0009-0004-2295-024X
Sam Pepler [2] https://orcid.org/0000-0001-5307-911X
Aleksandra (Ola) Podgorska[5] https://orcid.org/0009-0002-3722-7658
Shwetha Raveendran [3] https://orcid.org/0000-0002-4757-498X
Helen Rawsthorne [4] https://orcid.org/0000-0002-6540-8547
Roman Roth [3] https://orcid.org/0009-0004-4439-2365
Ag Stephens [2] https://orcid.org/0000-0002-1038-7988
Matthew Watkeys [4] https://orcid.org/0009-0003-5257-9982
Carl Watson [3] https://orcid.org/0000-0003-4940-5503

Daniel Westwood [2] https://orcid.org/0009-0007-1866-5843

Affiliations:

1 - NERC Environmental Data Service
2 - Science and Technology Facilities Council
3 - British Geological Survey
4 - UK Centre for Ecology and Hydrology
5 - National Oceanography Centre

Acknowledgements:

The project gratefully acknowledges funding by the UKRI Digital Research Infrastructure (DRI) Programme to a value of £500,000.

Executive Summary

Introduction and Purpose of the HIGH5 for AI Project

The HIGH5 for AI project represents a significant step forward in the evolution of the Environmental Data Service (EDS) as a platform not only for long-term stewardship of environmental data, but also as a critical enabler of artificial intelligence–driven research, innovation, and policy support. Funded by the UKRI Digital Research Infrastructure Programme with a total value of £500,000, the project ran from October 2025 to March 2026 and built upon the foundations laid by the earlier AMPLIFY and API4AI initiatives. Its overarching aim was to advance the accessibility, usability, and integration of high-value environmental datasets into AI and data science workflows, thereby broadening their impact well beyond traditional environmental science communities.

At its core, HIGH5 for AI sought to address a persistent and well-documented challenge: while the UK possesses a rich national environmental data estate, much of this data remains difficult to discover, interpret, and use effectively, particularly for researchers, developers, and decision-makers operating outside established environmental science disciplines. The project recognised that this problem is not merely technical, but sociotechnical in nature, encompassing issues of metadata quality, interface design, trust, governance, usability, and skills. Accordingly, the project adopted a user-led, co-creative approach that integrated technical development with structured user engagement, iterative design, and reflective evaluation.

Project Rationale and Strategic Motivations

The motivations behind the project were grounded in growing evidence that data access practices are increasingly misaligned with contemporary analytical processes. Many users expend the majority of their time cleaning, subsetting, and harmonising datasets before meaningful analysis can even begin. These challenges are exacerbated when working across domains or whilst attempting to integrate environmental data into machine learning pipelines that expect standardised, machine-readable formats. As AI techniques become more central to scientific discovery and policymaking, failure to address these bottlenecks risks marginalising environmental data from high-impact interdisciplinary work.

Government Mission–Aligned Use Cases

Against this backdrop, HIGH5 for AI focused on two cross-domain use cases aligned with UK Government missions. The first centred on clean energy, specifically the siting of offshore wind farms, which depends on the integration of diverse datasets such as seabed characteristics, wind resources, ocean physics, and climate observations. The second use case focused on public health and the planning of blue-green infrastructure, drawing on land cover, meteorological, air quality, and coastal datasets to inform decisions that support physical and mental wellbeing. These use cases were not treated as abstract demonstrations, but as concrete problem domains through which technical capabilities, user needs, and system limitations could be explored in depth.

Project Structure and Integrated Work Packages

The project was organised into five interconnected work packages. Project management and communication ensured coordination across partners and maintained a clear narrative of objectives and outcomes. A major strand of work focused on user engagement and co-design, deliberately extending beyond the traditional EDS user base to include stakeholders from industry, government, and interdisciplinary research. Through surveys, workshops, and in-depth interviews, the project team validated real-world use cases, identified barriers to AI adoption, and surfaced nuanced expectations around trust, provenance, and accountability. This engagement led to the development of user archetypes and clarified the need for services that support discovery and sense-making without displacing human judgement.

Development of AI-Ready Data APIs and Metadata Standards

Technical development formed the backbone of the project and was centred on the implementation of robust, standards-based data APIs. These APIs exposed carefully selected high-value datasets from across EDS data centres using existing community standards such as OGC APIs, ERDDAP, and STAC, while extending them through adoption of the emerging Croissant metadata specification. Croissant was chosen because it speaks directly to the needs of the machine learning community, providing structured, machine-readable dataset descriptions that integrate seamlessly with popular AI frameworks. By applying Croissant consistently across heterogeneous data holdings, the project demonstrated a practical route to making environmental datasets “AI-ready” in a meaningful sense.

Cross Centre Technical Implementation and Interoperability

A significant body of work was dedicated to mapping Croissant metadata onto existing EDS infrastructures and ensuring interoperability between standards. Several centres developed or extended tooling to generate validated Croissant descriptors dynamically or as static artefacts, while

an application profile was created to constrain and standardise how Croissant should be applied to EDS datasets. This work helped surface both the strengths and the current limitations of emerging standards such as GeoCroissant, highlighting opportunities for richer geospatial description alongside the need for further maturation of tooling and community support.

Demonstrating Reusable Analytical Workflows and Open Methods

Importantly, the project did not treat APIs and metadata as ends in themselves. A substantial effort was made to document methods, share code openly, and provide worked examples that demonstrate how data from multiple centres can be integrated into reproducible analytical workflows. A new public-facing GitHub repository was established containing narrative Jupyter notebooks that guide users through discovery, access, extraction, and analysis using Croissant-enabled datasets. These exemplars lowered barriers to entry for new users and provided a baseline against which more advanced AI-driven approaches could be compared.

Agentic AI Demonstrator: Objectives and Design Principles

One of the project's most innovative components was the development and evaluation of an agentic AI demonstrator. Rather than simply building a chatbot interface, the project took an explicitly evaluative stance, asking how such systems behave under different design choices and forms of user interaction. The resulting prototype combined retrieval-augmented generation with access to domain-specific tools capable of discovering datasets, extracting subsets, and performing simple analytical comparisons.

Evaluation of Agentic AI Behaviour and System Trustworthiness

The findings from this work underscore the promise of agentic AI for environmental data discovery while also illustrating its risks. The demonstrator showed that AI systems can handle multi-step tasks and work with data systems, but also highlighted weaknesses, security risks, and performance limitations. Crucially, the work reinforced the conclusion that trust in AI systems cannot be assumed or optimised solely through technical tuning. Instead, trust must be earned through transparency, evaluability, governance, and sustained user involvement.

Exploration of the Model Context Protocol and Emerging AI Interfaces

Parallel exploration of the emerging Model Context Protocol further highlighted this tension between opportunity and risk. MCP was recognised as a potentially valuable interoperability layer for connecting AI systems to tools and services, but the project work highlighted the importance of cautious, internal testing before public exposure. Concerns around provenance, attribution, security, and long-term stewardship featured prominently, reinforcing the project's broader emphasis on responsible, mission-aligned innovation rather than technology adoption for its own sake.

Enabling Users Through Onboarding, Documentation, and Community Engagement

A further strand of work focused on enabling users through improved onboarding, communication, and community engagement. Drawing directly on user research, the project developed conceptual designs and practical guidance to help users navigate the heterogeneity of EDS services and make effective use of AI-ready resources. Plans for a post-project hackathon were developed as a

mechanism for stress-testing APIs, engaging new communities, and reinforcing a culture of co-creation within and beyond the EDS.

Integrated Working Across the Environmental Data Service

Across all work packages, a recurring insight was the value of integrated working. The project demonstrated that progress toward AI readiness depends as much on alignment between centres, shared standards, and common tooling as it does on individual technical innovations. Where integration was achieved, efficiencies emerged, duplication was reduced, and a clearer vision of future services became possible. Conversely, areas of inconsistency in workflows, metadata granularity, and documentation highlighted the need for further investment.

Strategic Recommendations and Future Directions

The project highlighted that further work is needed in the consolidation of Croissant usage across EDS, the development of a shared Python-based data loading capability, and the integration of AI-ready metadata into the core discovery catalogue. Equally important is the establishment of robust evaluation and assurance frameworks that allow EDS to make defensible claims about the trustworthiness of AI-enabled services. The project results emphasise that agentic AI will not be suitable for all applications and that careful prioritisation, grounded in user needs and institutional values, is essential.

Conclusion: Positioning EDS for Trusted AI Enabled Environmental Data Services

The HIGH5 for AI project proved a successful foundational step in the journey of providing enhanced AI readiness of trusted environmental data services. It has delivered tangible technical assets, clarified user needs, and generated evidence to inform future decision-making. The project also concluded that the full potential of AI-ready environmental data will require continued funding, deeper integration across centres, and an explicit commitment to sociotechnical assurance. If these conditions are met, the Environmental Data Service is well placed to play a central role in enabling trusted, impactful AI applications that address national and global challenges.

Contents

About the report 8
About the Environmental Data Service 8
About the HIGH5 for AI project 8
Motivations 9
Our use cases 10
Project structure 10
WP2 User engagement and project co-design 11
Objectives 11
Deliverables 12
Selection of Datasets to meet the use cases 12
Objectives 12
Deliverables 13
BODC 13
NGDC 13
CEDA 14
EIDC 14
Findings and conclusions 14
Next steps 15
WP3 Data API implementations 16
Objectives 16
Agreed Technologies 16
ERDDAP (BODC and EIDC) 16
OGC API – Features (NGDC) 16
STAC (CEDA) 17
Croissant 18
Data APIs accessible for AI/ML 19
BODC 19
NGDC 19
CEDA 20
EIDC 20
Expose datasets using Croissant metadata 21
BODC 21
NGDC 22

EIDC 23
Documenting and sharing of methods 24
Findings and conclusions 26
Next steps 28
WP4 Agentic AI capability demonstrator 29
Objectives 29
Deliverables 29
4.1 Environmental Data Agent 29
4.2 Model Context Protocol (MCP) Trial 32
Findings and conclusions 35
Next steps 37
WP5 Enabling users 38
Objectives 38
Deliverables 39
Findings and conclusions 39
Next steps 40
Communicating the project 41
Key stories 42
Increased availability of high value data 42
Value of integrated working 43
Discovering and using our AI-ready resources 44
Current user journey 44
Envisioned future journey 44
Need for a future EDS API (LoadEDS) 45
Trusted AI 48
Overall recommendations and next steps 49
Conclusions 52
Supporting materials 54
Appendix A: List of high value datasets for which APIs were developed 55

About the report

This report provides a high-level summary of the work undertaken during the project, highlights how this work can be developed further (including key recommendations), and signposts several supporting technical reports containing more detailed information.

We believe this report will be valuable to all stakeholders interested in the work of the EDS and its future direction. It outlines how teams across the EDS are collaborating under a shared approach to data provision, enabling a consistent user journey for discovering, accessing, and utilising data in scientific workflows. This report demonstrates how, through the creation of appropriate interfaces and formats informed by detailed user engagement, the EDS is enabling the integration of its data with external sources to address complex interdisciplinary challenges, supporting solutions for the challenges embedded in the Government's key missions. The report summarises the project's achievements, the lessons learned, and the clarified vision for delivering the next steps toward a streamlined and integrated service spanning multiple data types and domain data centres. In addition, it provides a steer as to the future funding requirements to achieve this vision.

About the Environmental Data Service

The Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains: atmosphere and climate, earth observation, polar and cryosphere, marine, terrestrial and freshwater, geoscience, solar and space physics. The EDS is made up of a network of distributed data centres, with domain specific expertise.

Our goal is to ensure that environmental data are made available, accessible and re-usable for the long-term in order to fully realise their value. We are funded by the Natural Environment Research Council (NERC), part of UK Research and Innovation, to advise researchers on how to prepare data for long term curation, management and dissemination.

The EDS is a fundamental part of NERC's digital strategy and works with the Digital Solutions Hub, and other partners, to break down disciplinary barriers and facilitate data sharing beyond academic use. As well as a platform for long-term stewardship of environmental data, the EDS is a critical provider of actionable information to address complex interdisciplinary challenges, supporting solutions for the challenges embedded in the Government's key missions.

About the HIGH5 for AI project

The HIGH5 for AI project was funded by the UKRI Digital Research Infrastructure (DRI) Programme to a value of £500,000. It ran from October 2025 to March 2026. The project builds on the outcomes of the previous AMPLIFY project and API4AI project.

The main goals of the project were to:

- Increase the availability of high-value data in accessible, AI-ready formats through the development of APIs
- Showcase the use of such data provision through the creation of demonstrator AI applications, such as an agentic AI chatbot that would enable easy access and utilisation of complex data through innovative interfaces
- Promote these tools to users within and outside of the EDS
- Learn from and engage with the wider community so the EDS can deliver value beyond the environmental science community by a broad range of disciplines and sectors to address key government mission challenges

Our approach was based on the following principles:

- Co-Creation and User-Led Development: Engage end users early to ensure outputs meet real-world requirements
- Rapid End-User Engagement: Leverage networks and public profiles to identify and engage users in the first two months
- Strategic Data Decisions: Prioritise dataset release to meet timelines and developer capacity

Motivations

The EDS and the UK's national environmental data landscape in general contains a wealth of high value scientific datasets, yet many potential users, both within and beyond environmental science, struggle to access, understand, and fully exploit this information. Despite the significance of these datasets for research, innovation, and government mission challenges, several barriers persist. Users often find it difficult to discover relevant data, interpret complex variables, or work with unfamiliar formats. Even once accessed, preparing these datasets for analysis or integrating them into AI and data science workflows can be prohibitively complex.

As AI, machine learning and advanced computational facilities become increasingly central to scientific and policy driven decision making, these barriers restrict the wider societal and cross disciplinary value that environmental data could deliver. Currently, only a fraction of the potential user community, particularly those outside traditional environmental domains, are able to utilise these datasets effectively.

This project is motivated by the need to transform how environmental data is delivered and used. By making national datasets more discoverable, accessible, and AI ready through high quality APIs and common metadata standards, we can open them up to a much broader set of disciplines, innovators and problem solvers. This aligns with major national ambitions to unlock public data assets for AI, enhance FAIR access, and support mission led innovation in areas such as clean energy, environmental resilience, and public health.

Alongside technical development, the project aims to understand real user needs through engagement, onboarding resources, and hands on events such as hackathons. These activities will

help ensure that new tools and interfaces genuinely support researchers, developers, and other communities in working with complex environmental datasets.
Ultimately, the motivation behind this project is to enable the EDS to empower society to extract greater value from environmental data, expanding its applications across sectors, and showcasing how AI enabled interfaces can unlock new insights and opportunities from the UK's most important environmental national datasets.

Our use cases

The project considered cross domain use cases aligned to Government missions. We particularly focused on **making Britain a clean energy superpower** and **an NHS fit for the future**. At the proposal stage, we had provisionally identified two cross domain societal challenge use cases related to these missions. We also picked out datasets relevant to their resolution that were frequently requested by our users to be provided in more AI accessible and usable formats. At the start of the project, a co design phase with users finalised the use cases and the datasets needed to fulfil them. APIs were created for these datasets, and we demonstrated how these APIs could be used to address these challenges by building a demonstrator agentic AI chatbot solution.

The selected datasets demonstrated how the EDS provided an integrated service that worked across multiple environmental domains to answer Government mission challenges beyond the environmental sphere. The use cases we identified were:

Use case 1: Clean energy.
The EDS contributes to UK net zero targets by improving data access to support decarbonisation and energy security initiatives. A proposed focus was based on the siting of offshore windfarms that brought 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 public health benefit.
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., Land Cover Map, sea level, air quality, geological maps) that can be integrated to inform decision making that enable the planning and provision of built infrastructure (e.g., houses, parks, access to beaches) that support a mentally and physically healthy society.

Project structure

The project was structured around five work packages that collectively aimed to improve access to high value environmental data, develop AI ready services, and enable wider user engagement.

Work Package 1 focused on project management and communication, ensuring smooth delivery by overseeing progress, finances and risks. It included producing monthly progress reports and a final project report, as well as developing communication materials that clearly presented project objectives, achievements and lessons learned to both internal and external audiences.

Work Package 2 delivered internal and external co design activities to refine the project plan, identify government mission use cases, and select the high value, nationally significant datasets to be exposed via APIs. This involved running a launch workshop to finalise the project programme, followed by co-design work with the wider user and stakeholder community.

Work Package 3 implemented the data APIs themselves, providing access to the datasets selected in WP2. This included creating robust, sustainable and standards-driven APIs for a range of gridded and tabular data types, ensuring consistency, interoperability and long-term maintainability. The datasets were exposed using community metadata standards such as Croissant or GeoCroissant, enabling AI ready integration. The project also shared code and documentation openly on the EDS GitHub repository, demonstrating how multiple datasets can be combined in analytical or AI focused workflows.

Work Package 4 developed an Agentic AI capability demonstrator, building a prototype AI Agent that can interact with EDS data, metadata catalogues, and supporting tools. Using the emerging Model Context Protocol (MCP), the WP explored how to link APIs, data catalogues and LLMs, showcasing how basic environmental queries can be handled using AI. This culminated in demonstrators linked to the use cases identified in WP2.

Work Package 5 focused on enabling users beyond the traditional environmental data community, implementing findings from previous AMPLIFY EDS user research to ensure the APIs and associated resources are easy to discover and use. It also included planning a future hackathon (beyond the project timeline) to broaden engagement and gather user feedback, and involved undertaking user research activities to better understand how non-environmental users discover and adopt these data services.
Each of these work packages are described in further detail in the next section of this report.

WP2 User engagement and project co-design

Objectives

The primary objective of the User Engagement and Co-design work stream (spanning WP2 and WP5) was to bridge the gap between the UK's Environmental Data Service (EDS) and non-traditional data users. We set out to:

- **Validate and Refine Use Cases:** Move beyond theoretical applications to ground the project in real-world requirements for "Clean Energy" (offshore wind siting) and "Public Health" (blue-green infrastructure).

- **Expand the Co-design Community:** Deliberately shift the design process from an insular, NERC-centric model to an inclusive framework involving industry, government policy makers, and cross-sector researchers.
- **Identify Barriers to AI Adoption:** Uncover the specific technical and cultural friction points preventing the adoption of environmental data in machine learning workflows.
- **Establish Trust Mechanisms:** Determine what "trustworthy data" means to high-stakes decision-makers to inform the development of AI tools and APIs.

Deliverables

The following outputs were generated through iterative engagement with stakeholders:

Stakeholder Kick-off & Co-design Sessions:

- Conducted initial in-person workshops at the NERC Digital Gathering (DG) in Cranfield to align the team and agree an approach to stakeholder engagement
- A quantitative survey was completed by 49 potential end-users
- A series of 14 moderated user research interviews were held online (led by EDS User Researchers with Stakeholders aligned with the two use cases), iteratively checking in with the project management and technical teams to support across work-package communication [**see Page 11** – more detail inside UX Research Summary report].
- These sessions facilitated the transition from broad project goals to specific, actionable user stories grounded in what was feasible within the limited project timeframe.

Comprehensive User Research Reports (WP 2 & 5):

- Quantitative Survey: Reached ~2,500 potential users across various EDS networks, securing 43 detailed responses across academia, private sector, and government.
- Qualitative Interviews: Conducted 14 qualitative, moderated interviews with key personas including Marine Spatial Planners, Data Harmonisers, and Technical Implementers.
- User Archetypes: Defined a robust taxonomy of users (eg. "Data Detective," "Technical Implementer," "Policy Maker") to guide future product development.

Co-design Expansion Strategy:

- Successfully demonstrated a plan to engage users beyond NERC centres by leveraging external networks (JASMIN user groups, MEDIN, industry LinkedIn networks) and targeting specific pain points relevant to the private sector (e.g., firewall compatibility, licensing clarity)

Selection of Datasets to meet the use cases

Objectives

Assess the datasets owned by EDS centres that are relevant to the use cases identified in the user engagement task, namely offshore wind farm siting and blue green infrastructure for better health.

Select datasets that are not yet readily available for AI/ML use, either due to lack of programmatic access to the data and/or lack of machine-readable descriptive metadata. At least one new dataset per data centre should be selected for publication as a new operational API.

Deliverables

- Targeted datasets that will be published as new API
- Targeted datasets that will be enhanced with provision of Croissant metadata

A detailed list of the datasets and the pre and post project status with respect to improved usability in AI workflows is presented in the appendix. A narrative summary is provided below for each data centre.

BODC

The chosen headline datasets from BODC were the water column CTD (conductivity, temperature and depth) profiles and the processed values of sea level from tidal gauge stations within the UK tidal gauge networks. These datasets provide key information on the water conditions around the coast which could feed into decision making on suitable locations (E.g. access for building, maintenance, etc) for locating new offshore windfarms. Furthermore, information on tidal ranges can help inform planning for locating new parks and/or blue spaces when siting new blue-green infrastructure (E.g. High tidal zones may be suitable for locating new parks). Therefore, these datasets were chosen as being of high value to both use cases adopted for this project.

NGDC

The following high value datasets were chosen because of their relevant to offshore wind farm siting, for example to assess geotechnical suitability of the foundations or laying of subsea cables:

- Offshore digital geological mapping of bedrock and seabed sediments
- Catalogue of geological samples from subsurface boreholes and seabed sediments, linking to associated images and analysis

Another suite of datasets provides access to Strategic Environmental Assessment reports (spatial indexes reports and seabed photos and videos) used for marine spatial planning.
An onshore geochemical dataset providing background concentrations of contaminants in English soils was chosen due to relevance to the siting of new blue green infrastructure.
Datasets of historical and modern recorded earthquakes for onshore and offshore UK are relevant to both use cases: offshore infrastructure siting should prefer sites with lower risk of seismic disturbance, whereas blue green infrastructure siting may prefer higher seismic risk because these would not be suitable for a built environment. A suite of web pages containing offshore geological reports as annotated machine readable text were also targeted as high value information relevant to offshore wind farm siting, and which could also be used in WP5 knowledge base to demonstrate chat bot Agentic AI.

CEDA

Several CEDA datasets were found to be relevant to the blue green infrastructure use case, including soil moisture from the ESA Climate Change Initiative (CCI) dataset and surface wind speed from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 project.

It also has some relevant larger scale datasets with multiple variables including UKCP Projections over the UK 1980-2080 and HadUK Gridded Climate Observations which provides numerous variables including relative humidity, rainfall and ground frost data.

Whilst most CEDA datasets are limited to land, the UK Climate Projections and HadUK-Grid also include wind speed data, useful for the offshore wind farm use case.

EIDC

Two headline datasets were identified by EIDC. These were the Land Cover Map for Great Britain and Northern Ireland (stored as 2 separate datasets) and the CHESS-Met gridded meteorology dataset. These are heavily downloaded datasets from the EIDC with the land cover map being one of the key headline datasets from the centre. These are useful for the blue green infrastructure use case.

Findings and conclusions

What Went Well

- **Validation of the "Data Wrangling" Bottleneck:** Research concluded that users spend 60–80% of their time cleaning and harmonising data rather than analysing it. This validated the project's increased focus toward "AI-ready" data formats and automated cleaning tools.
- **Successful Cross-Sector Engagement:** By focusing on government missions (Net Zero, NHS Fit for the Future), we successfully engaged environmental and non-environmental sectors (insurance, urban planning, energy) on topics that they cared about rather than what was previously viewed by some as irrelevant or overly technical.
- **Identification of the "Single Source of Truth" Need:** The research clearly articulated the user demand for a unified discovery layer with provenance tracking, moving the conversation away from disparate repository searches.

What Could Have Been Improved

- **Recruitment Velocity:** While the survey reached a wide audience, the conversion rate to deep-dive interviews was lower than anticipated in the commercial sector due to confidentiality concerns, time scarcity and a general perception that the information and services available through EDS were not commercially relevant, yet. Future efforts require a clearer commercial pitch and earlier engagement with industry bodies to secure buy-in.
- **Terminology Alignment:** Initial discussions revealed a significant disconnect in the definition of "AI-Ready." Some providers viewed "open access" as sufficient, while users required semantic interoperability and cloud-native formatting. Earlier alignment on these definitions would have streamlined the technical requirements phase.

- **Research discovery and technical delivery alignment:** For future opportunities, shifting from a linear/waterfall handover to a more agile workflow would be recommended. Embedding researchers directly into the development lifecycle (eg: through regular syncs, shared planning ceremonies, and co-facilitated briefings) would ensure that technical implementation is continuously aligned with user evidence.

Issues Encountered

- **The "API Paradox":** A major finding was that while APIs are ideal for discovery and certain machine to machine tasks, they often fail for heavy-lifting analysis due to corporate firewalls, data volume limits, and instability ("API drift"). This may encourage EDS to adopt a hybrid approach (API + bulk cloud-native download) which would satisfy a wider user base but would complicate the technical architecture. This hybrid approach is aligned to the W3C Spatial Data on the Web Best Practices
- **Data Fragmentation:** Users reported that critical datasets are often known only through informal networks ("who you know") rather than searchable catalogues, highlighting a gap in metadata coverage that the project could not solve within its timeframe.

Next steps

- **Target "Unknown" Users:** Launch proactive outreach campaigns targeting sectors currently unaware of EDS data capabilities (e.g. health sector, construction, insurance), leveraging the "Single Source of Truth" narrative with a clear commercial pitch, perhaps exploring which pre-integrated multi-disciplinary datasets industry would be willing to pay for.
- **Integrate Research into Hackathon Design:** Utilise the specific pain points identified (e.g., need for audit trails, cloud-native formats) to design challenges for the proposed WP5 Hackathon, ensuring it solves real industry problems rather than hypothetical ones.
- **Formalise Community Engagement in Governance:** Embed the user archetypes and co-design processes into the permanent EDS governance structure to ensure user feedback loops continue post-project.

Please refer to the following documents for more detail which are published alongside this main report:

• User Research Report
• User Research Summary Report
• Survey Report

WP3 Data API implementations

Objectives

The primary objective was to develop and provide a consistent suite of sustainable application programming interfaces (APIs) and appropriate metadata to enhance access to the datasets selected in WP2. The provision of these APIs have many use cases, but the focus of this project is their use in AI/ML workflows.

We set out to:

- Agree a small number of technologies and formats that would be used to publish data APIs, using technologies already established in the EDS centres to ensure robust and sustainable deliverables.
- Implement methods to expose the high-value datasets using the schema.org based Croissant metadata and GeoCroissant format to promote their use in AI and data science applications
- Simplify and document use of data APIs being loaded into a computing workflow though example code notebooks

During development we identified another useful deliverable to meet the need for simplification of API consumption:

- new python library functioning as a data loading interface utilising croissant metadata and the chosen EDS API technologies.

Agreed Technologies

The technologies agreed for data APIs were ERDDAP, OGC API – Features (via ArcGIS Server, pygeoapi and in house python code), and STAC catalogue. The croissant metadata standard was agreed as the unifying interface to the different API end points.

ERDDAP (BODC and EIDC)

ERDDAP (https://erddap.github.io/) is a free and open source scientific data server developed by the National Oceanic and Atmospheric Administration (NOAA) to give users a simple, consistent way to download subsets of gridded and tabular datasets using common file formats.

This technology was used in this project because EIDC and BODC already run ERDDAP servers to expose numerous gridded and station-based datasets. In addition, ERDAPP already includes an integration that enables datasets to be exported to Croissant format.

OGC API – Features (NGDC)

OGC API – Features is one of the next generation standards from OGC, defining a RESTful web API that delivers data in JSON and GeoJSON formats. The APIs support content negotiation, lightweight data payloads, and are formally described using the standardised OpenAPI specification. First released in 2019, the OGC API - Features standards define modern mechanisms for querying and serving geospatial data, including attributed points, lines and polygons. It is intended as functional successor to the legacy Web Feature Service (WFS) standard, providing similar capabilities through a simpler, more web native architecture. NGDC already publish a large number of spatial datasets in OGC Web Map Services (WMS) and in ArcGIS REST API, but
OGC API Features was chosen as the baseline standard for delivery of new APIs by NGDC in this project because of its

- developer friendly architecture
- vendor agnostic server options
- suitability with respect to the vector (point, line, polygon) spatial feature datasets chosen.
- established use and server technologies in NGDC to ensure sustainability. Such technologies include:
    - Licensed ESRI ArcGIS Server (which simultaneously publishes the ArcGIS REST and WMS formats)
    - Free open source pygeoapi (used for rapid deployment of beta "OGC API – Features" services, and for simple non-spatial datasets)
    - In house data API code (Java and python) developed for more complex data delivery

STAC (CEDA)

SpatioTemporal Asset Catalog is an emerging standard for storing Geospatial assets in JSON-style format, with a convention for metadata attributes to be stored so that STAC records are uniform and standardised across institutions and data centres. The STAC methodology consists of several hierarchical layers for organising geospatial data such that searching and accessing data are made consistent and easy. STAC objects are grouped into four main categories:

- STAC Assets: Lowest level object, a single NetCDF file representing some arbitrary N-dimensional data.
- STAC Items: A grouping of assets that are associated by some attributes. This could be a set of NetCDF files or a Cloud Optimised object like a Zarr Store.
- STAC Collections: A grouping of STAC Items, following a common structure.
- STAC Catalog: Access point for Item browsing all items can be searched via their defined attributes using an API.

Read more about STAC conventions and possible extensions/plugins in the STAC Documentation at https://stacspec.org/.

This technology was chosen because CEDA holds a STAC Catalog, for various collections of data, all accessible via a single URL. The convention for storing cloud format records alongside more traditional file types is to create an Item for each dataset object. This item then has one or more assets; the original data files in NetCDF or similar formats plus a reference-file asset which is typically

a Kerchunk or CFA (aggregation) file. This approach would also support Zarr. The actual loading of the cloud object into a dataset happens as lazily as possible - meaning the data is not accessed until directly requested to preserve memory space in the user's environment for as long as possible.

Croissant

A key objective of this project was to ensure that the high value datasets identified were exposed and made accessible to users using consistent and coherent formats that adhered to defined standards. An additional requirement to encourage uptake by the ML community was to make the data as easy as possible to be piped into the popular ML frameworks.
An earlier EDS project, API4AI, investigated and prototyped using the Croissant file format to meet these requirements, and this project decided to continue working with croissant to an operational stage.

Croissant is an emerging standard developed for and commonly used by the ML community.
It is a high-level format for machine learning datasets that combines metadata, descriptions of the underlying data files, structures of the data and default ML semantics and information (E.g. Train/test splits) into one single file. Croissant builds on schema.org and responsible AI principles from mlcommons to make the assets searchable, machine readable and easily interrogatable with ML frameworks such as TensorFlow and PyTorch.

To ensure the high value datasets had maximum discoverability and accessibility in an 'AI ready' format the croissant descriptors must adhere to a common and standardised format. However, at the time of this project, there were no canonical validators for croissant schema, and as such software implementations may vary. Therefore, an application profile (AP) was developed to specify precisely how Croissant v1.1 must be applied to NERC EDS datasets. This AP provided constraints, recommended practices and Linked Open Data (LOD) conventions to ensure that the metadata is interoperable, machine readable and FAIR. This included the standard croissant information on dataset descriptions, location, licence, vocabularies in addition to recommended extra information such as persistent identifiers for the datasets (E.g. Digital object identifiers (DOI)). The final AP describes how to construct the examples of this in practice and lays out the blueprint for how croissant descriptors should be used for other NERC EDS datasets in the future beyond the high value exemplars in this project.

Exploration of the emerging GeoCroissant extension

The GeoCroissant specification has been in draft form for almost 2 years, but was released formally just after the completion of this project, with the specification available from:

https://docs.mlcommons.org/croissant/docs/croissant-geo-spec.html

A set of GeoCroissant recipes is now being released at:

https://zoo-project.github.io/dcai/

Whilst the specification has the potential to significantly improve the descriptive capabilities of a Croissant record, it is yet to be fully supported in software and is therefore mainly useful for adding geospatial *tags* to metadata fields. Within the project, we noted that some aspects of the GeoCroissant record structure were incomplete and some schema-resolution problems that impeded validation. These issues should be ironed out soon by the developers.

The vision for GeoCroissant, and the wider integration of Croissant with EDS data types, is that more data types and formats will be adopted within the standard so that the "mlcroissant" library can be extended to allow loading of datasets relevant to the environmental science community. This is an outcome that will converge neatly with our development of the "loadeds" library (as discussed later).

Data APIs accessible for AI/ML

The API technologies described above were harnessed as follows by each respective data centre to make their high value datasets available:

BODC

An ERDDAP server was deployed as follows to deliver access to the selected datasets:

- A federated ERDDAP service was deployed and made operational providing a more modern and robust approach to BODC's previous ERDDAP deployment. This underpins the longevity of these APIs as sustainable and long-term operational systems.
- All BODC's CTD profile data was made available through this new ERDDAP, making a significant part of the centre's holdings available as well as one of its most frequently requested datasets. ERDDAP facilitates machine access to this data which was previously unavailable.
- Utilised the ability for ERDDAP to serve data in multiple formats to allow the data to be output in parquet format as user research from WP2 indicated that cloud native formats (such as parquet) were desirable for data access. Parquet format also enabled seamless integration of ERDDAP with the mlcroissant library for this dataset.
- Enabled federated access to the tidal gauge data (previously held in ERDDAP) so that it was discoverable and accessible through the same consistent interface as the CTD dataset.

The BODC server for these datasets is accessible from the following link: ERDDAP - Home Page

NGDC

New production APIs were published for the offshore geological mapping and geological samples datasets. ESRI ArcGIS server was used to publish the data as OGC API - Features services alongside

the established legacy formats (ArcGIS REST and OGC WMS). The Earthquakes and Offshore Strategic Environmental Assessment (SEA) datasets were already exposed through existing OGC API - Features endpoints. The Background contaminants in soils dataset was already published as a WMS, but this is a raster layer and not suited for OGC API – Features. Both newly created and pre-existing APIs were integrated within this project by supplementing them with interfaces conforming to the Croissant metadata standard.
The metadata and page urls for the offshore reports web pages were made available through a bespoke (python FastAPI) web API that is also aligned to OGC API Features.

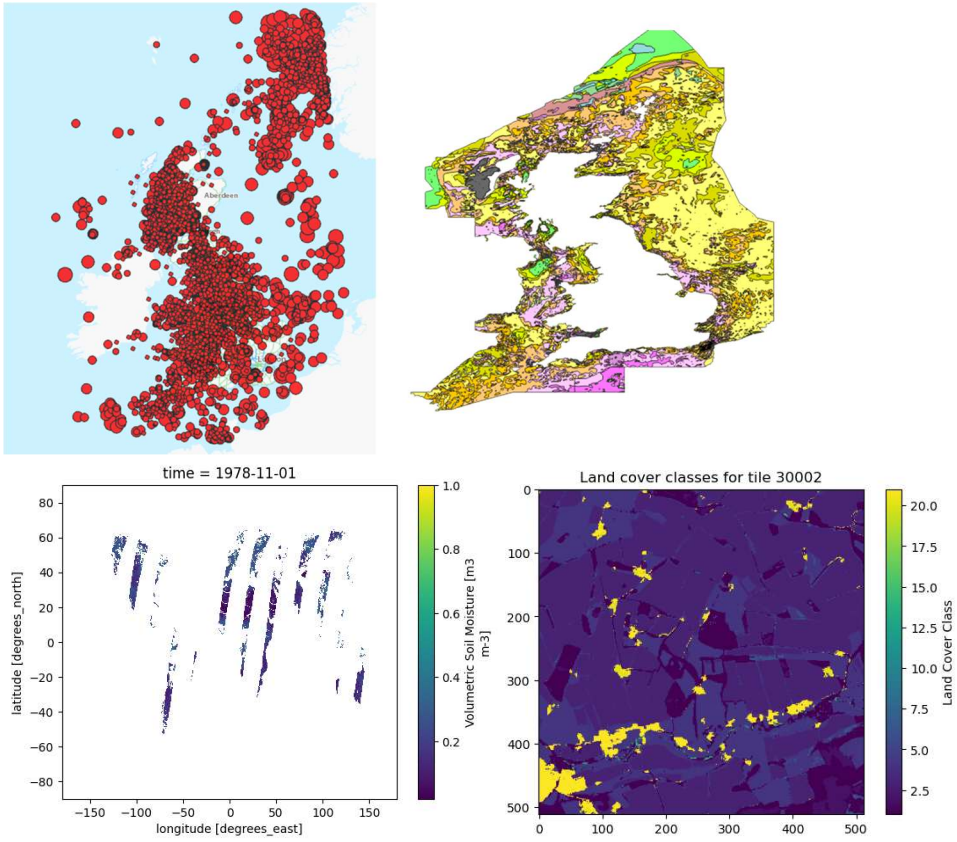
CEDA

The selected datasets are exposed via a range of interfaces, including the main CEDA Catalogue, the CEDA STAC Catalog, and the experimental EDA CCI STAC API.

EIDC

For the CHESS-Met data, supplied as gridded spatio-temporal netcdf files, an ERDDAP server was setup using a similar approach to that utilised by BODC. This allowed users to explore the data in the ERDDAP server and then output a Croissant metadata format for the current data subset and then ingest this directly into a machine learning workflow using the mlcroissant library.

A slightly different approach was taken using the land cover map (LCM) data. Here the 10m raster product for both Great Britain and Northern Ireland were made accessible at a URL for direct download from the EIDC. This was then wrapped up in a custom built dataloader (written in python) that accessed the data, randomly tiled it, prepared it as a torch tensor object (array like object for use in ML models) with appropriate train/test splitting and wrapped it in a dataloader object for use in a model training workflow. This custom loader was developed to work directly from the croissant metadata file so that the user could work directly with the croissant metadata standard to access the data itself.



Example datasets published as APIs across the various Data Centres.

Expose datasets using Croissant metadata

Since Croissant is an emerging standard there is not yet a discovery or search capability provided out-of-the-box. As part of HIGH5, we generated a set of endpoints where Croissant records could be loaded and scanned for relevant content. Initially, these were delivered using the following technologies:

- CEDA & EIDC: through a webserver listing via HTTP(S) endpoint (hosted on JASMIN).
- BODC: via an ERDDAP endpoint.
- NGDC: via a Web Accessible Folder (WAF) endpoint

To enable the respective APIs to be able to operate and integrate with Croissant format files, several croissant generator scripts were created or built upon as part of the project. A summary of each method can be found below with more detail available in the code repositories.

BODC

One key feature of ERDDAP is its built in ability to get the data response from the server in a wide variety of file formats, including the recent addition of Croissant descriptors as one option. Therefore, Croissant descriptors for datasets such as the BODC CDT data have been made available through their ERDDAP server. To ensure that these descriptors validated correctly and integrated seamlessly with the mlcroissant library, the development of a proxy API was required that sat on top of ERDDAP that could dynamically extract the croissant descriptors whilst simultaneously applying corrections to them so that they could successfully deliver the data in cloud native parquet format.

This allowed users to use the mlcroissant library to directly extract the data in a format that could easily be plugged into a machine learning workflow.

NGDC

Unlike ERDDAP, OGC API Feature servers do not have the capability to generate croissant metadata natively, but the OpenAPI specification and the metadata endpoints of the API standard do provide some of the information required. The remaining information is already published in standardised discovery metadata records.
In the previous API4AI project the demonstrator croissant files were semi-manually crafted using these two sources. For this project NGDC wrote two helper software tools to automate the generation.

Croissant4j – A generic Java library for programmatic generation of MLCommons Croissant (CroissantML) dataset descriptors. Designed around the MLCommons Croissant v1.0 specification its focus is exclusively on authoring and serializing Croissant metadata.

Croissant-gen – An internal BGS prototype Java tool for the automated generation of MLCommons Croissant compliant dataset descriptors for the OGC API services. Croissant4j is used to generate compliant JSON-LD and croissant-gen analyses the structure and responses of the OGC API endpoints and combines with externally published discovery metadata to produce the machine-readable croissant descriptors. This can either create croissant descriptors that can be used stand-alone or be used in API mode as a service which allows dynamic generation of croissant descriptors to be requested. For the datasets presented in this project the stand-alone model was used.

Croissant files for the newly published APIs and a range of relevant existing APIs were generated and stored in a Web Accessible Folder (WAF) alongside the existing standard discovery metadata files.

Note that the NGDC discovery metadata server (Geonetwork) is intending to support croissant metadata files in an upcoming release, which could provide an alternative discovery and access endpoint for the croissant metadata.

CEDA

Since STAC provides a similar level of dataset-level metadata to that described using Croissant, the existing CEDA STAC Generator tool was extended to enable the creation of Croissant files alongside their STAC counterparts. This allows both metadata formats to be usable as interfaces to the diverse range of environmental datasets exposed. The work carried out in the “loadeds” library (described later) enables loading of data through the data URIs defined within STAC/Croissant records or a direct extraction of the data from the Croissant record (this latter option is only at a prototype phase at present).

CEDA investigated hosting their created Croissant files on their STAC catalogue as additional metadata files that provide alternate avenues for users to access this data. Despite being beyond the scope of the project, there is a strong consideration given to utilising the STAC catalogue, and its built-in search capabilities, as an interactive mechanism allowing users to find datasets with

Croissant files available. This would also allow auto-generation of STAC and Croissant from the same core metadata, meeting the needs of existing and more AI-centric users and services.

Due to the large nature of CEDA datasets, the use of cloud-optimised formats such as Zarr or Kerchunk is preferred for storing 'chunked' representations of our larger datasets. For the HadUK-Grid observations dataset for example, 40 croissant files were created, one for each Kerchunk file. Each chunk represents different variables and temporal frequencies such as monthly data, annual data, or 30-year climatologies.

The chunking capability allows aggregation over numerous (potentially 10s of 1,000s of) data files, and represents an essential part of the data pipelines required for AI-ready big environmental data. CEDA maintains toolkits for simplifying both the generation of these aggregation files (typically as "Kerchunk" or "CFA" files) and their usage/access (integrated with the Python STAC client):

- Aggregation toolkit (PADOCC): https://cedadev.github.io/padocc/
- Data access library (DataPoint): https://cedadev.github.io/datapoint/

This chunking method allows the subsetting of specific variables from certain datasets, such as the ECMWF ERA5 dataset. For this data, a Croissant file was created referencing the surface wind speed data which had been chunked from the larger dataset. Croissant files were also created for the UKCP dataset and the soil moisture ESA CCI dataset.

The files are served from an HTTP(S) endpoint on the JASMIN infrastructure: https://gws-access.jasmin.ac.uk/public/eds_ai/high5/catalogue/croissant/ceda/

EIDC

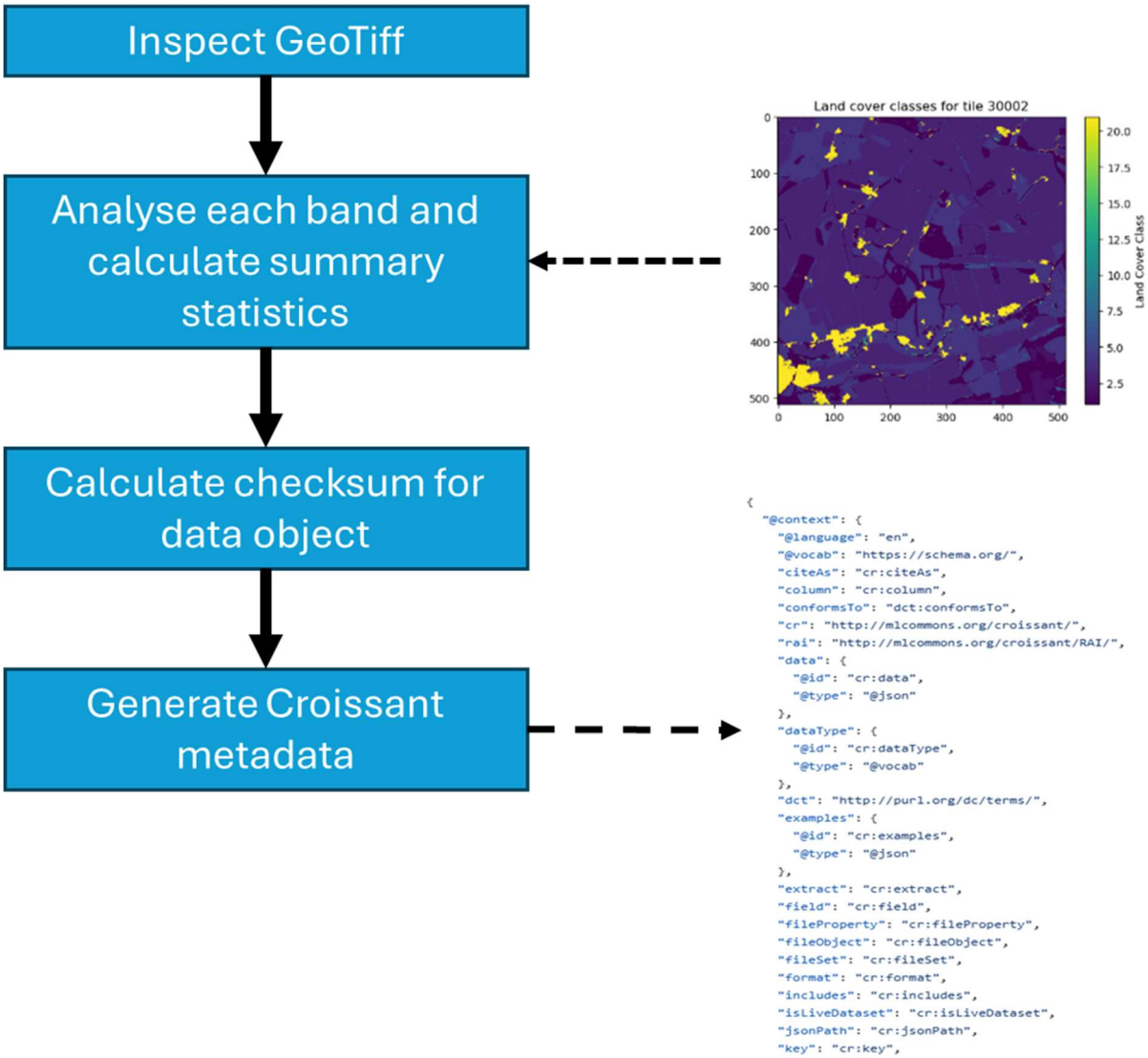
Two different approaches were taken for the EIDC datasets:

CHESS-Met was published using an ERDDAP server using a similar approach to that made at BODC. This allowed Croissant files to be generated dynamically using ERDDAPs built in functionality and accessed using a similar approach to that utilised by BODC.

Stored in GeoTIFF format, the Land Cover Map (LCM) required a slightly different approach to construct the Croissant descriptors. A large amount of the required metadata was stored in the file itself and therefore a custom croissant file was required that successfully validated using mlcroissant. Therefore, a custom script was developed that performed this required task.

1) Inspect the GeoTiff
2) Analyse each band (land classification in pixel and associated confidence) and calculate summary statistics
3) Calculate the Checksum (a piece of data that checks for changes in a digital object) for the file

4) Generate the appropriate metadata in croissant format that will accurately describe the data and validate using the mlcroissant library. This included creating the appropriate data sets, record sets and fields along with locations where the data can be accessed.



EIDC Land Cover Map Croissant generation workflow

The output is a croissant descriptor file for use with the LCM dataloader to enable the data to be read directly into an AI ready format for use in an AI/ML workflow along with exposing all the corresponding metadata in a machine-readable format. For this project the script was demonstrated using the 2024 editions of the GB and NI land cover maps however it can operate with other editions of the data that represent different time stamps and generate corresponding croissant records.

Documenting and sharing of methods

To promote ease of access to the resources produced during High5 we created a new public facing GitHub repository in the NERC EDS account (NERC_EDS_AI_Workflows). This repository contains a series of Jupyter notebooks explaining how to use the APIs, how to set a call to retrieve the data

using the croissant descriptors and demonstrate how to utilise the datasets in a machine learning workflow. An important feature of these notebooks was to utilise clear narrative to show how to construct a call to the individual API types and explain the reasoning behind settings in each call. The following examples were generated:

1) ERDDAP: A notebook developed by BODC to show an example of reading in Sea Level data using croissant and ERDDAP and then applying an ML based linear regression.
2) OGC: A notebook developed by NGDC to demonstrate using an OGC API to read in UK historical Earthquake data and explore the contents of the resultant GeoJSON
3) STAC: A notebook demonstrating the use of the STAC and Croissant routes for accessing data from the CEDA services.
4) LCM_direct: A notebook developed by EIDC to demonstrate using croissant and the dataloader to directly read the land cover map (LCM) into a torch dataloader ready for use in a machine learning workflow.

These were setup to show examples for specific datasets for each API type though demonstrate workflows that could be adapted to other datasets for which a croissant file and associated API can be made available.

Finally, to demonstrate utilisation of data from across the centres to answer real world use cases, two notebooks were developed to show how croissant and the APIs could be utilised to integrate datasets from multiple sources. These notebooks were based around the 2 headline user cases of this project:

1. Siting of new Blue-Green Infrastructure for better health

2. Siting of new offshore windfarms

Both notebooks followed a similar format and provided a simple example of the following workflow:

- Choose 2 locations and a time frame
- Identify datasets and their associated croissant descriptors that match the search terms
- Utilise the croissant files to extract the required datasets and store locally as GeoJSON
- Compare the two locations against pre-defined thresholds for suitability for our use case
- Provide tabulated summary of comparison of suitability of locations for addressing the chosen use case and deciding the best location for establishment of a new site (either offshore windfarm or new blue-green infrastructure).

The purpose of this proof-of-concept workflow was to demonstrate the process of bringing in the data and therefore used simple pre-defined thresholds in the suitability comparison summaries. However, they clearly demonstrate the ease and reusability of using the APIs to bring in data from across the centres, in formats ready to be used to address real world examples. This helps demonstrate how the APIs can potentially address the bottleneck of wrangling multiple sources of data and allow more focus on analysing the data itself.

These notebooks also serve as exemplar workflows that can be used to evaluate the exploratory agentic AI system being developed in WP4. They can serve as a benchmark to see how an agentic AI system (with access to the APIs) performs at addressing the use cases compared to just asking the base large language model (LLM) without access to the data. This was implemented in the Agentic AI package within WP4, which makes use of the "data-tools" package available from:

https://github.com/NERC-EDS/NERC_EDS_AI_Workflows/tree/main/data-tools/

These tools, defined as Python functions, were created as follows:

1. Discover environmental data (exposed as Croissant)
2. Extract data (based on a given requirement)
3. Compare data at locations given a set of thresholds (for a given environmental use case such as siting and Offshore Wind Farm).

Within the agentic AI component, these were *wrapped* using both Langchain (a Python library for natural language processing and LLM interactions) and MCP (as explored in D4.2). This enabled the agents built within the agentic package to make use of the same data tools that were being demonstrated in the standard Python notebooks provided. This parallel capability allows a side-by-side comparison of how humans and AI-driven systems might utilise the tools.

Findings and conclusions

What went well?

Several successes were achieved during the API development work of this project, which included:

- Generation of standardised cross-EDS croissant descriptors for high value datasets to facilitate consistent access to a wide variety of data types. This included ensuring that all datasets had their metadata correctly and accurately described adhering to accepted vocabularies and standards.
- Specifications for an application profile (AP) to indicate how Croissant v1.1 must be applied to NERC EDS datasets.
- Enhanced discoverability of EDS high values datasets through machine readable formats such as the croissant standard.
- Creation of notebooks demonstrating simple workflows to utilise the croissant files to import data and produce formats that are of use in machine learning and or data science pipelines.
- Production and sharing of exemplar notebooks showing how to discover datasets relevant to specific use cases, find and access the corresponding croissant records, extract and integrate the data and answer example questions of the combined datasets.
- Sharing of knowledge, methods and tools between the EDS partners was a major outcome from the project. Plans are already afoot to explore re-using software and capabilities to increase our exposure through established services.

- GeoCroissant is an emerging extension to Croissant that will help define spatial and temporal characteristics to improve future interoperability. Whilst still in its infancy, we expect GeoCroissant to add richness to the Croissant metadata standard, and with it there are opportunities to connect greater functionality to the "mlcroissant" library for loading large environmental datasets. We should continue to track its development.
- A key outcome of the project was a clearer shared understanding of where users expect help and guidance within the data discovery process, and why delivering this consistently is challenging within a federated service model.

What could have been improved?

Overall, the development work in this project was successful though some areas of improvement remain.

- Due to the multiple underlying technologies (STAC, OGC and ERDDAP), as identified through the onboarding work in WP5, there was differing levels of granularity in each centres API services and usage documentation for that specific service. It was necessary to document these individual technologies so that users could understand and operate with them as they see fit however this made is difficult to generate a one size fits all set of documentation to importing the data. However, using the croissant standard and the mlcroissant library did facilitate a consistent workflow to using croissant to read in the datasets themselves albeit with different underlying technologies. Therefore, there could be improvement in harmonising the different workflows across the centres.
- An evaluation exercise was conducted across the EIDC LCM and BODC ERDDAP notebooks which identified areas to improve the clarity for users. The main headline areas for improvement include clarity in the definition of what 'AI ready' data means, better wording when describing the technical abilities of the APIs and more context setting for non-technical users who wish to follow the workflows.

Issues Encountered

ERDDAP:
One major issue encountered during the project that effected the efficiency and consistency of the workflows was the interaction of ERDDAP with croissant in the sense that if the user wanted to generate a subset of any dataset they had to do this in the ERDDAP server and then export the corresponding croissant descriptor for that subset. This was then used to extract the data itself using the API and mlcroissant workflow. However, having to negotiate two different services increased the technical expertise required to extract the data and reduced the efficiency of the workflow and user experience whilst also having implications for the curation of dynamically generated croissant records. This would require a solution that allows the user to specify an appropriate subset from the whole dataset croissant record and dynamically extract this subset whilst also storing a corresponding croissant descriptor for the subset to record the appropriate metadata. This would facilitate reproducibility and reuse in data science and ML workflows in the future.

Croissant:
A key challenge encountered during the use of MLCommons Croissant specification and the associated tooling was inconsistency between the published specification, the reference examples, and the Python implementation. In practice, the mlcroissant Python library appears to have evolved more rapidly than the formally published specification, resulting in cases where Croissant descriptors that were valid according to the specification were rejected by the library. This made the process of authoring a compliant Croissant descriptor significantly more difficult, as validation depended not only on adherence to the published standard but also on the specific expectations of the library implementation.

These issues were compounded by limited diagnostic feedback from the mlcroissant Python library. Several runtime and validation failures produced cryptic, incomplete, or apparently unrelated error messages, making it difficult to identify the true source of schema or metadata problems. As a result, debugging often required inspection of the library source code and comparison against examples in the MLCommons Croissant GitHub repository rather than reliance on the published documentation alone.

Next steps

Next steps for the API development work include the following:

- Utilise the exemplar services generated during this project to recommend a roadmap for extending the APIs to expose other key headline datasets from across the EDS to enable discovery of more assets in a data science/AI ready format.
- Ensure services built to access datasets across the EDS are harmonised as much as possible and incorporated into a single software package to enable users to seamlessly discover a dataset (or datasets), read in the corresponding croissant records and then use these records to extract the data. This can help define a roadmap to produce an initial EDS software package for easier use and analysis of its data holdings (see need for future EDS API section below). This will extend the reach of EDS assets beyond the environmental science domain including but not limited to the AI and ML community.
- Share the common interfaces and tools within our EDS teams and identify opportunities to re-use these capabilities in our business-as-usual activities and future EDS features/tools.
- At the very end of the project version 1.0 of the GeoCroissant standard (GeoCroissant v1.0) was released. To incorporate geospatial metadata standards into the EDS croissant descriptors the developed AP should be updated to specify how GeoCroissant v1.0 must be applied to NERC EDS datasets. This will ensure that the metadata is consistent, interoperable, machine actionable, FAIR whilst also accounting for any vocabularies and standards in relation to the geospatial properties of the data. This further facilitates automated discovery and reuse of the data.
- Investigate how to extend croissant descriptor to use controlled vocabulary identifiers for dataset field descriptions, utilising planned improvements in the publication of NERC vocabularies.

- Investigate the upcoming support for croissant within the Geonetwork metadata catalogue software, currently used by NGDC and for the EDS data catalogue. This could consolidate the discovery metadata and technical croissant metadata in a single discovery platform.
- Enhance the API delivering text so it delivers RAG-ready semantically chunked and indexed text

WP4 Agentic AI capability demonstrator

Objectives

WP4 of the HIGH5 for AI project aimed to explore whether agentic large language models (LLMs) could meaningfully improve access to environmental data for users who sit beyond the traditional environmental science community. This work was shaped by earlier HIGH5 and AMPLIFY-EDS user research, which identified strong interest in AI-enabled interfaces that could remove barriers to data discovery by offering plain-English explanations, early-stage interpretation, and lightweight analytic support.

Crucially, those same users also expressed a strong lack of trust in information produced autonomously by AI systems. Accuracy, reliability, and transparency were consistently cited as concerns, and users across a range of backgrounds preferred a human-in-the-loop approach: where AI assists with discovery and sense-making but does not act as an unaccountable decision-maker. In response to this, HIGH5 WP4 was framed not as a technology demonstration alone, but as an early contribution toward understanding how to build AI-enabled environmental data services which warrant user trust.

The main objectives of WP4 were therefore to evaluate **how different design choices affect AI system behaviour under realistic patterns of use**, while also considering **how those systems interface with external tools, services, and standards**. This included both:

- An **ablation study** examining agentic system configurations, and
- Exploratory engagement with **Model Context Protocol (MCP)** as a potential standard for tool integration

These strands of work are complementary: the ablation study examined how systems behave, what design choices matter most, and where risks and limitations lie; while the MCP exploration considered how such systems might interact with external services and infrastructure.

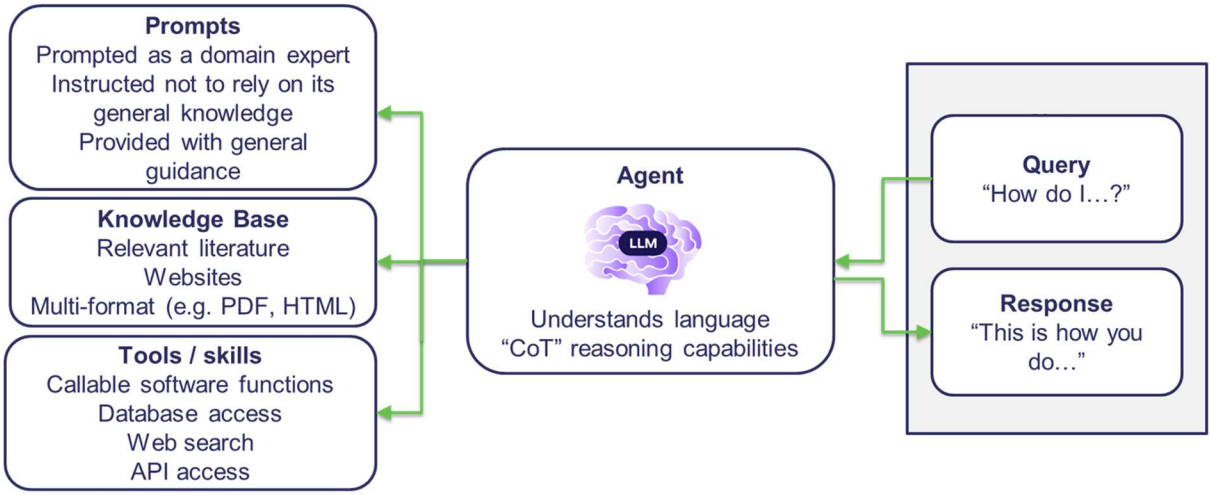
Deliverables

4.1 Environmental Data Agent

The principal deliverable of WP4 was a prototype agentic AI demonstrator designed to support environmental data discovery through natural language interaction. The system was implemented as

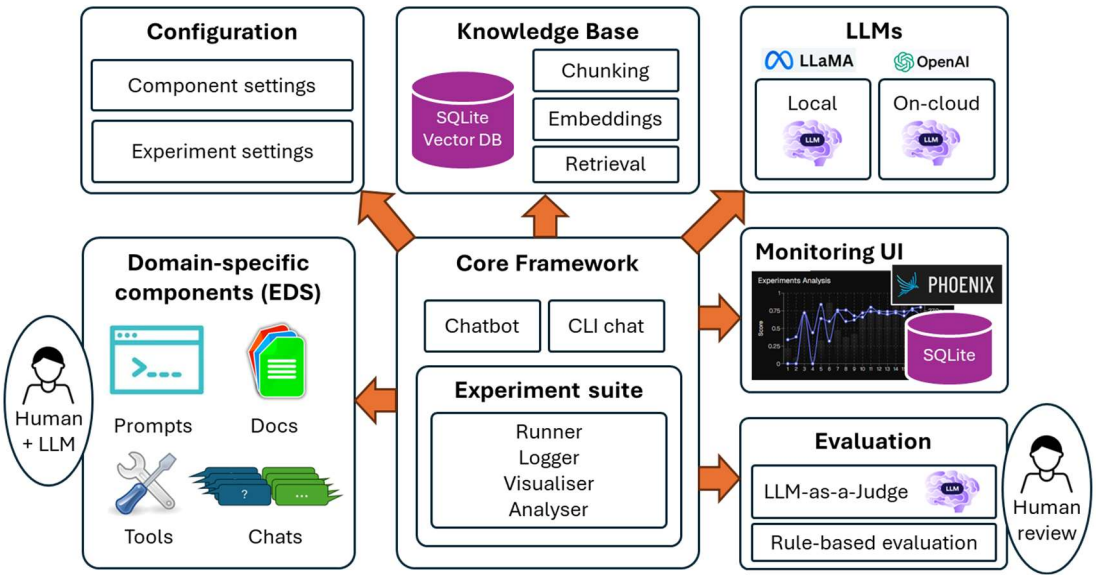
a modular codebase driven by a set of configuration settings enabling experiments to be run that compare the capabilities of a range of 'agents' (or, software programs designed to perform specific tasks autonomously). The figure below shows how each agent is constructed: using a knowledge base (a curated set of documents), retrieval techniques, tools, and a prompted LLM.

The knowledge base documents are indexed into a vector database which supports retrieval-augmented generation (RAG): advanced search capabilities which allow the system to identify and retrieve content based on the underlying meaning of user queries rather than simple keyword matching. Domain-specific tools are also provided to allow the system not only to retrieve and summarise contextual information from the knowledge base, but also to perform actions such as data discovery via Croissant metadata files, data extraction, and basic threshold-based assessments.



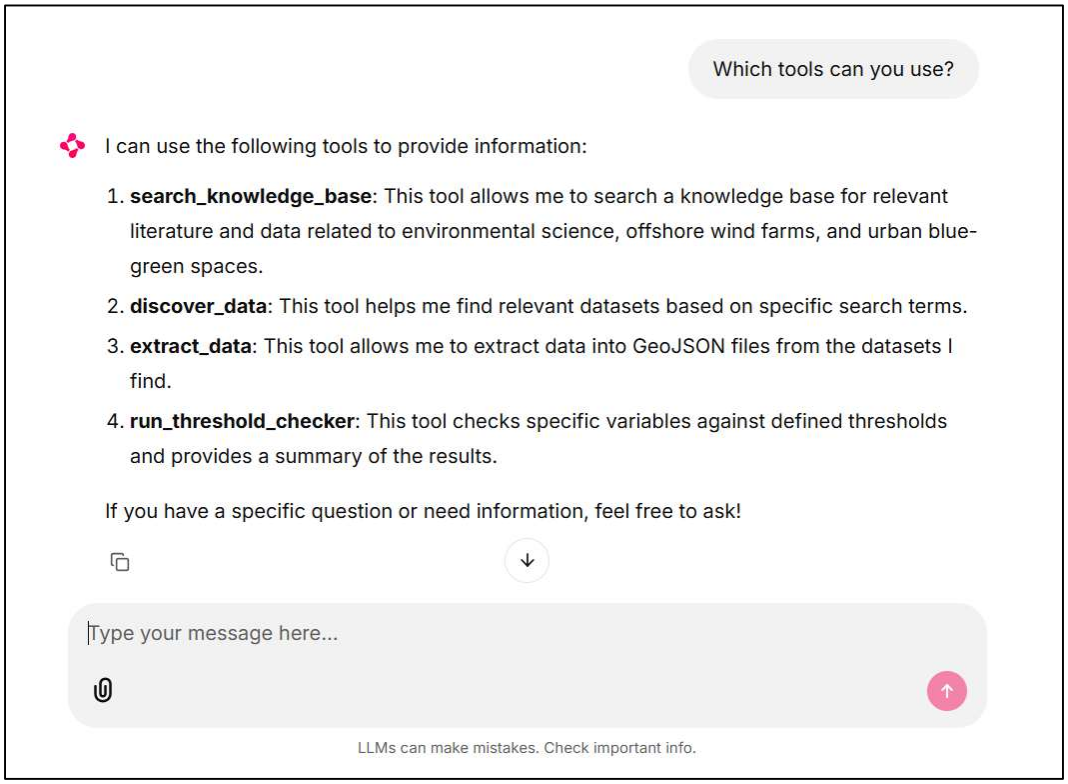
Components of a basic single-agent system, and an example chat (query-response pair).

The agentic evaluation framework (see: https://github.com/NERC-EDS/HIGH5forAI_AgenticAI/tree/main/agentic-package) consists of a set of open-source tools, along with LLM integrations either running locally or from remote APIs, as shown in the figure below.



The agentic evaluation framework is composed of a selection of supported open-source tools, code components and integrations with LLMs.

As part of the development process, a password-protected version of the prototype was deployed on Amazon Web Services (AWS) as a proof-of-concept web-based chatbot, as shown in the figure below.



A chatbot user interface exposing the RAG and tool capabilities of a particular configuration of the agentic package. This was deployed under password protection on AWS and used the Chainlit Python library.

The agentic AI demonstrator was grounded in HIGH5's two use cases to ensure that development and evaluation took place in scenarios that are representative of real-world contexts of use. The system's underlying knowledge base consisted of 70 vetted documents relevant to each use case, combining open access academic literature with domain-specific technical resources; and the evaluation dataset focused on testing the system's responses to use-case-relevant questions. This evaluation dataset formed the basis of WP4's structured evaluation framework.

The evaluation dataset comprised 50 question-response pairs spanning five question types—factual, numeric, ambiguous, adversarial, and data tool questions—across both HIGH5 use cases. Question types were treated as design hypotheses for different kinds of user interaction, with the intention of simulating the diversity of real-world queries that a user-facing EDS AI tool might receive. This design choice was based on the rationale that any future EDS AI tool must be able to support a range of users with varying disciplinary backgrounds and levels of technical proficiency. The evaluation produced a results dataset of 450 generated responses, together with metadata such as tool usage and response generation time. While the full dataset could not be exhaustively analysed within project constraints, it represents a substantial additional output.

These additional deliverables are documented in the supplementary report 'Evaluating the HIGH5 agentic AI demonstrator'. This report is therefore also a core deliverable, not only as documentation but as an analysis of the system 'in action', rather than presenting the demonstrator in isolation. The report provides a concrete foundation for future work and records design decisions, explains evaluation choices, analyses observed behaviours, and situates the findings within broader discussions of trusted AI, governance, and sociotechnical system design.

4.2 Model Context Protocol (MCP) Trial

Deliverable 4.2 focused on how the EDS might employ the emerging Model Context Protocol (MCP) standard for interfacing systems and services with LLMs. This was tackled in two ways within WP4:

1. The Agentic Package codebase was adapted to support and serve its 'tools' as either Langchain tools (a Python implementation) or via an 'MCP server' This work exists as a branch on the main WP4 GitHub repository ready for adoption if required.
2. A NERC-EDS meeting was convened around the theme of MCP. Details of this meeting and its outcomes are provided here.

MCP meeting: content, context, aims and themes

A meeting was convened by the NERC Environmental Data Service (EDS) within the context of the 'High5 for AI' project, which provided dedicated resources to explore emerging AI-enabling technologies relevant to environmental science and data services. The specific aim of the meeting was to explore the potential relevance, value and risks of the Model Context Protocol (MCP) for the EDS, and to capture collective insight that could inform future recommendations. The meeting brought together 27 representatives from across EDS partners and related organisations (including STFC, BGS, NOC and others) with a mix of technical, data stewardship, governance and strategic perspectives. The session combined presentations, demonstrations and an extended open discussion.

MCP was introduced as an open, language-agnostic protocol originating from Anthropic [Anthropic, 2025] and designed to standardise how AI systems (particularly large language models (LLMs) and agents) interact with tools, services and data sources. MCP aims to reduce the complexity of bespoke integrations between models and services by providing a common interface and shared conventions for tool discovery, description and usage. A recurring theme in the introduction was that MCP should be seen primarily as an interoperability and transport layer rather than as a solution to semantic, vocabulary or data-quality challenges.

Concrete experience with MCP was presented from the High5 for AI project, where existing Langchain (a Python library for interfacing with LLMs) tools had been refactored and exposed via MCP. This included examples of converting analytical and data-extraction functions into asynchronous MCP tools, standing up simple MCP servers, and demonstrating how an AI agent could discover and use those tools dynamically. Related environmental science examples were discussed, including experiments at NOC around quality control workflows, BGS examples of exposing map services and the existence of MCP servers within the NASA Earthdata ecosystem [NASA, 2025].

Much of the meeting focused on discussion rather than technology demonstration. Key themes repeatedly surfaced:

- The distinction between MCP as a low-level protocol and user-facing AI services or interfaces.
- The difference between internal, exploratory use of MCP versus publicly exposed services.
- The relationship between MCP and existing EDS priorities around APIs, standards, interoperability and ‘AI-ready’ data.
- Strategic uncertainty around how user access to data through intermediate AI systems could affect trust, provenance, attribution and the long-term role of EDS.

The discussions were exploratory, explicitly acknowledging uncertainty, rapid technological change and evolving funder and policy expectations around AI.

Potential opportunities of adopting MCP

Several potential opportunities associated with MCP were identified. At a practical level, participants noted that MCP is relatively lightweight to deploy and experiment with. This makes it attractive as a means of rapid prototyping, internal experimentation and skills development. Several contributors emphasised that MCP could act as a common wrapper for existing tools, allowing teams to catalogue, test and combine functionality that currently exists in fragmented forms across organisations.

There was broad interest in MCP as an internal capability first. Internally, MCP could support experimentation with agentic workflows for data discovery, extraction, quality control and lightweight analysis, particularly in secure or sandboxed environments. This was seen as a way to prototype and trial AI-enabled approaches, improve shared understanding, and identify real risks before considering public exposure.

Strategically, MCP was seen as one possible route to making EDS assets more accessible to AI-driven workflows in a world where external users may increasingly expect to interact with data via conversational or agent-based systems. Some participants argued that ignoring MCP-like approaches risks EDS falling behind broader technological trends, particularly given increasing pressure from funders around AI readiness.

Potential risks of adopting MCP

A wide range of risks and concerns were also articulated. One major concern relates to trust, misuse and misrepresentation. If MCP services expose data or functionality directly to AI systems, there is limited control over how outputs are generated, interpreted or presented to end users. This raises the risk of incorrect analyses, loss of context, or spurious authority being attributed to results generated through EDS-backed tools.

Intellectual property, attribution and branding were persistent concerns. Participants questioned how EDS provenance, citation requirements and reputation could be preserved when data are accessed and processed through third-party AI tools. There was anxiety that large technology providers could absorb EDS outputs into commercial products with little visibility of origin or benefit returning to the UK research ecosystem.

Governance and security risks were also prominent. Concerns included where prompt context and data are sent, how authentication and authorisation would work for MCP services, and whether MCP becomes an additional maintenance burden layered on top of existing APIs. Some questioned whether MCP would remain relevant as models become better at consuming standard API documentation directly.

Finally, several participants cautioned against technology-led development without clear use cases. Past experience within the EDS highlighted the risk of investing heavily in infrastructure that does not ultimately align with user needs or strategic priorities.

Potential recommendations and plans for the EDS

While no formal decisions were taken, several provisional recommendations and directions emerged from the discussion, as follows:

1. **Experimentation and prototyping:** there was strong support for treating MCP initially as an internal, experimental capability rather than a public-facing service. A lightweight internal MCP testbed or sandbox could allow EDS partners to expose a small number of well-understood tools (for example, catalogue search or simple data sub-setting) and explore how AI agents interact with them under controlled conditions.
2. **Focus on low-risk use cases:** experimentation with MCP should be anchored in clearly defined, low-risk use cases. Examples suggested included internal data discovery, metadata interrogation, or comparison of existing tools rather than complex analytical pipelines. Risk-based experimentation was seen as more valuable than aiming immediately for high-impact services.
3. **Alignment with EDS governance/strategy:** the group highlighted the need to integrate MCP exploration with broader governance discussions already underway in EDS and partner

organisations. Rather than creating a parallel governance structure, MCP-related work should inform shared understanding of AI risks, attribution requirements, ethical considerations and institutional constraints.

4. **Trusted/branded EDS capabilities:** several participants suggested that EDS should focus on offering trusted, branded capabilities rather than attempting to compete directly with generic AI platforms. This could include curated MCP services or AI-enabled systems that emphasise *provenance*, *domain expertise* and *transparency*, even if they utilise third-party models under the hood.
5. **Internal collaboration and alignment:** it was agreed that further synthesis and communication are needed to make best use of our existing/future resources and expose our capabilities effectively. For agentic systems to work correctly (i.e. to make the correct use of information and tools), we must provide clear unambiguous documentation and build clean APIs/services within the EDS. A closer level of alignment and communication between the NERC Centres is essential to achieve these goals.

Findings and conclusions

What went well

One of the strongest outcomes of WP4 was demonstrating the feasibility and potential value of agentic AI for environmental data services. 4.1's prototype showed that such systems can support analytic tasks by retrieving and synthesising information from a domain-specific knowledge base, identifying relevant datasets via Croissant metadata, and extracting and exporting data according to user requirements. Importantly, the system was also able to attempt multi-step workflows that resemble realistic user interactions, in addition to answering more isolated conceptual questions.

One of 4.1's clearest positive findings is that question type strongly drives system behaviour. Across all nine system configurations, factual questions consistently produced responses that converged most closely with the reference baseline, while similarity decreased for numeric, ambiguous, and adversarial questions. This pattern was stable across configurations, indicating that it is a strong high-level determinant of system behaviour. This shows that considerations of the ways users might ask questions must be built into system design alongside technical design decisions.

This result is important because it demonstrates that future EDS AI tools must be **evaluated against different slices of input, rather than relying on aggregated performance metrics**. When similarity scores were averaged across all question types, differences between configurations appeared marginal. However, when results were analysed by question type, strong and interpretable patterns emerged. This demonstrates the value of slice-based evaluation for systems intended to support heterogeneous user groups.

4.1's study also showed that system configuration and prompt strategy do matter, albeit in more nuanced ways. For example, adaptive system prompts tended to converge with the reference baseline for numeric and adversarial questions more often than safety-first or specialist prompts. These effects were not universal, but they revealed meaningful interactions between prompt design and question type. There may have been larger differences in result between technical

configurations if different LLMs were compared, so future work should establish the impact of different LLMs on system behaviour as a priority. The main insight produced here is that any **future deployment scenario will likely require negotiating trade-offs between different aspects of system behaviour** rather than identifying a single optimal configuration.

From a functional perspective, the agentic workflow for data discovery and extraction appeared largely successful. Inspection of responses strongly suggested that target datasets were located and that GeoJSON files were generated in most cases. Deviations from strict response templates often took the form of short explanatory sentences clarifying relevance, which added value but may require stricter controls in scenarios where this is not acceptable. These results strongly suggest that **agentic LLMs can support complex, multi-step interactions with environmental data infrastructure**.

At a higher level, the work demonstrates the promise of agentic LLM RAG systems for **meeting user demand for new types of interaction** that support data discovery, verification, and early-stage sensemaking. It provides concrete evidence that such systems can operate beyond solely conversational roles and engage directly with data services in ways that are potentially useful.

In terms of exploring MCP, we concluded that **MCP is neither a silver bullet nor something that EDS can ignore**. It represents one component in a rapidly evolving AI ecosystem that the EDS should engage with cautiously, pragmatically and in a way that reinforces its core values of trust, evidence and scientific integrity. The improvement of the EDS core capabilities is a higher priority than rapid adoption of new technologies, but MCP may be a means of exposing those capabilities to this new paradigm of agentic systems.

What could have been improved

4.1's study identified key limitations in observability and experiment tracking. Tool routing was not recorded reliably, which constrained the analysis of tool use. In many cases, qualitative inspection suggested that tools were called successfully or that calls were attempted but failed, yet the absence of reliable logs meant these patterns could not be confirmed systematically. This points to the need for stronger instrumentation as part of future work.

4.1's study also relied on a synthetic evaluation dataset, which was intentionally created as a low-noise baseline rather than a gold standard. While this was an appropriate choice given the exploratory nature of the work, it limits how far conclusions can be generalised. Because of the characteristics of the evaluation dataset, a similarity metric was intentionally selected over an accuracy metric. Before accuracy-based metrics can be meaningfully applied, **future work will require the definition of criteria for 'correct' answers**, ideally through expert consensus

In addition, the analysis was necessarily high-level. No formal statistical modelling was undertaken, and only a subset of the available results dataset could be examined in depth. More extensive analysis—such as examining how responses diverge semantically or factually from the reference baseline—would doubtless surface additional insights, particularly for ambiguous and adversarial questions.

Finally, while question type proved to be a powerful explanatory factor, it also likely bundles together **multiple unmeasured dimensions such as difficulty, ambiguity, and user intent**. Future work could refine this abstraction by developing more granular characterisations of user behaviour, informed by the user archetypes already developed in HIGH5 WP2 and WP5.

Issues encountered

The most significant technical issues encountered during 4.1 arose at the interfaces between system components. In particular, tools that required strictly formatted inputs proved vulnerable to minor mismatches in LLM-generated strings. The threshold checker tool failed in the majority of cases because the LLM supplied plausible but incorrect use case identifiers, rather than the exact strings expected by the tool.

The specific failures encountered are solvable, but they illustrate a broader point: **agentic workflows are brittle without careful interface design and validation**. Small discrepancies between natural language inputs and machine-interpretable parameters can cascade into tool failures, even when the system's underlying plan is correct.

The study also highlights risks associated with adversarial prompting. All configurations showed divergence from the reference baseline for adversarial questions, and while the precise nature of this divergence was not analysed, **poor understanding of system behaviour under adversarial conditions represents a significant security risk for the EDS**. Addressing this may require accepting trade-offs in performance elsewhere, reinforcing the importance of explicit prioritisation as part of system design.

Next steps

Progress in this area will require a sociotechnical approach, rather than a sole focus on technical optimisation. Agentic AI offers real opportunities for the EDS to meet user demand for improved data discovery, but users should not be asked to trust such systems until they are shown to be useful and trustworthy. A key next step is to **identify which core EDS capabilities are suitable for enhancement through agentic AI**, and which are not. This requires clear prioritisation of goals, metrics, and acceptable trade-offs, as well as the sustained involvement of EDS users.

Building on this, the EDS should **develop an assurance methodology that can support claims about AI system trustworthiness**. Approaches such as argument-based assurance are highlighted in the supplementary report as promising ways to connect technical evidence with accountability, governance, and user confidence. Assurance processes should also involve the **investigation of NERC or UKRI capacity for running open-weight and/or open-source LLMs on UK research infrastructure**. Doing so could reduce reliance on commercial providers, address ethical and privacy concerns, and support a more sustainable, transparent foundation for future EDS AI services.

The exploratory evaluation approach presented in the supplementary report should be **expanded into a robust, reusable evaluation framework**. This includes improved observability and experiment tracking, richer evaluation datasets, and eventual support for more complex system architectures

such as multi-agent systems. Continued horizon scanning and engagement with the rapidly evolving research literature will be essential to achieving this. There is also scope to lower barriers to participation by developing low-code or guided tools that help diverse users engage with this framework without programming expertise.

Regarding MCP, we propose that the EDS should engage with cautiously, pragmatically and in a way that reinforces its core values of trust, evidence and scientific integrity. The EDS should **continue to monitor the evolution and utility of MCP**, and potentially adopt it to expose core capabilities if it gains relevance in our domain.

In summary, WP4 should be viewed as a foundational step that demonstrates feasibility, exposes critical challenges, and provides an evidence base for future development. The next phase should focus on consolidation and assurance, **ensuring that any AI services delivered by EDS are not just innovative but genuinely trustworthy and responsible**, and that there is **a credible, responsible path for advancing work in this area**.

WP5 Enabling users

Objectives

The "Enabling Users" work stream (WP5.1 and WP5.2) focused on removing technical and cognitive barriers to data access. Our objectives were to:

- **Provide practical guidance:** Translate the findings from WP2 and WP5.3 (moderated user research interviews/survey) into guidance for the internal technical work packages as well as tangible onboarding materials for external users.
- **Enhance Discoverability via Metadata:** Understand how the **Croissant** metadata standard could address real-world user pain points and design how this information is communicated to EDS users.
- **Foster Innovation through Community Events:** Plan and design a hackathon to advance EDS aims, engage external users and potentially stress-test the new APIs and AI tools with a diverse group of developers and data scientists.
- **Support and Explain EDS Heterogeneity:** Design resources that acknowledge and support users to navigate the current variability in EDS data, onboarding help, discovery metadata and Croissant metadata.

Prior to this work, there was a reasonable assumption that the natural place for users to access practical guidance when discovering datasets would be via some sort of information and link presented alongside datasets and services in the EDS discovery metadata portal. However, the project found that this is not straightforward to implement is a consist or efficient manner. EDS metadata are created at different stages, with different levels of granularity, and published through different systems across the member data centres. As a result, simple “one size fits all” onboarding

guides (for example, one guide per API or data type) are not currently viable without further harmonisation of workflows.

Deliverables

WP5.1: Onboarding Materials & User Guides:

- Developed a user-friendly shop window 'Environmental Data and AI' linking to the most relevant resources and "getting started" guides based on the specific friction points identified in user interviews. Linking through to the user focused Python notebooks developed in WP3.
- To assess the feasibility and user experience options for future improvements to our guidance in a way that supports wider EDS (not project specific) AI ready deliverables.

WP5.2: Hackathon Plan:

- Produced a comprehensive Hackathon Proposal Document outlining three distinct event models:
  1. **Community-Attached Event:** Leveraging existing environmental science gatherings, e.g. NERC Digital Gathering.
  2. **Industry/Startup Focus:** A competitive, innovation-driven event targeting the private sector, e.g. Geovation challenges.
  3. **Government Mission-Aligned Event:** Themed around "Clean Energy" or "Public Health" to engage policy stakeholders.

WP5.3: User research report:

Comprehensive analysis and delivery of the **UX Research Report** which triangulates the findings from the moderated user research (**UX Research Summary**) and the WP2 quantitative survey (**Survey Results Report**). The Reports feature relevant user-research findings, along with:

1. Recognition of gaps in the initial hypothesis, recommendations on how to address these
2. Outline of research impact within the project
3. Definition of three strategic areas of opportunity to transform the EDS from a siloed asset into a shared resource and recommendations to position the EDS as a trusted broker for defensible decision-making.
4. User Archetypes
5. Ecosystem Service Map
6. User stories

Limitation: evidence is currently limited to two primary use-cases, highlighting the need for further user research across diverse scientific areas to ensure a more inclusive Environmental Data Service view.

Findings and conclusions

What Went Well

- **Research-led design:** The onboarding materials were significantly more effective because they were derived directly from user research qualitative data and behaviours (eg. the "Data Detective" problem), rather than assumed technical needs.
- **Flexible Hackathon Framework:** The development of a modular hackathon taxonomy allowed the project team to visualise how different event structures could yield different outcomes (e.g. prototypes vs. policy insights), providing a strategic tool for future engagement and very likely to be added to the EDS co-creation toolkit.
- **Croissant as a Bridge:** The adoption of Croissant metadata has provided a promising option to address the user demand for "machine-readable" data, once launched, this needs to be promoted and monitored to understand if it is successful in an operational context.

What Could Have Been Improved

- **Granularity of Onboarding:** While general guides were created, the varying levels of granularity of EDS data and metadata from across different Data Centres made it difficult to produce a single, unified "one-size-fits-all" onboarding experience. Users still face confusion when moving between resources that have come from different centres and data formats.
- **Visibility of Standards:** While Croissant records were created, their visibility within the broader metadata catalogue remains a challenge. The metadata catalogue needs to be populated in a consistent way for all data centres and data formats with information necessary to highlight "Croissant-enhanced" records during discovery.

Issues Encountered

- **Inconsistent Centre Workflows:** A significant hurdle was the lack of standardisation in internal workflows across the five NERC Data Centres. Each centre has different levels of granularity in their metadata and different publication pipelines, complicating the delivery of a unified user experience.
- **Resource Constraints for Hackathon Execution:** While the *plan* was successfully delivered, the actual execution of the hackathon is scheduled for post-project. Ensuring continuity and resource allocation for this event remains a risk.

Next steps

- **Iterative Onboarding Updates:** Establish a feedback loop and online monitoring where usability reports, user support tickets and hackathon outcomes directly inform updates to the onboarding documentation, ensuring it evolves with the API capabilities. The webpages, notebooks and technical documentation produced in this project have been designed to meet user needs not project reporting obligations, they should evolve along with EDS data and services.
- **Standardise Croissant Visibility:** Work with the technical teams to ensure Croissant metadata is prominently displayed and filterable within the central EDS catalogue, making "AI-Ready" datasets instantly identifiable.
- **Harmonise Workflows:** Initiate a cross-centre working group to align onboarding granularity and metadata publication workflows, reducing the cognitive load on users navigating between repositories.

- **Execute the Hackathon:** Secure funding and partners to deliver the planned hackathon(s) within 12 months of project end.

Please refer to the following documents for more detail which are published alongside this main report:

User Research Report
User Research Summary Report
Survey Report
Hackathon report

Communicating the project

The goal was to communicate the project outputs to users both within and outside of the EDS. To easily share information about the project, a dedicated page was created on the EDS website. This displays a high-level overview of the project as well as specific notices such as the user research survey. It provided a central place to point people towards in order to find out more about the project outputs, barriers, and why it matters in the wider EDS context.

A master slide deck was created that can be tailored for different audiences and meetings, ranging from a quick introduction to in-depth progress updates on specific work packages. This was used throughout the project to share updates with groups across and beyond the EDS, and continues to be used for events & conferences after the end of the project. A High5 for AI overview poster was also produced, which is used to present a visually engaging summary of the project and introduce people to the concepts and goals of the project. The poster and presentation can be presented at future conferences and meetings to promote the project.

Finally, a webinar was planned that took place on the 8th of June 2026 after the project had concluded. It was a great opportunity to spotlight the project as a whole, showcasing the tools we built, and highlight the supporting materials produced. It included a general overview of the EDS and the High5 project, before delving into the specific sections of the project: User research, technical developments, Agentic AI exploration, and Outreach & Outcomes.

The webinar was promoted widely both within the EDS and externally to relevant organisations and companies. It was a huge success, Over 110 participants attended, and around 20 of those were from external organisations and government bodies including the Met Office, Historic Environment Scotland, the Joint Nature Conservation Committee & Floatation Energy – an offshore wind farm developer.

The attendees were really engaged with the talk, asked poignant and thoughtful questions, and were eager to hear more about future developments in this area. For those who couldn't attend, the webinar was recorded and uploaded to the EDS YouTube page as then shared across the same channels that were used to promote the webinar such as the EDS site page & the NERC newsletter.

Unfortunately, there wasn't enough time to respond to all the questions, so these were answered offline and posted alongside the recording.

Key stories

The project has resulted in a number of achievements, findings and conclusions leading to recommendations for next steps. Some highlighted key stories from the work undertaken are presented.

Increased availability of high value data

A central achievement of the HIGH5 for AI project has been the substantial increase in the availability, accessibility, and operational readiness of high value environmental data services across the EDS. This outcome reflects the project's overarching aim to transform how national environmental datasets are delivered and used, ensuring they are not only preserved for long term stewardship but also made readily usable within modern analytical and AI driven workflows as well as by other uses such as inclusion in data portals.

The project delivered this improvement through the coordinated development of robust, sustainable, standards based APIs that expose priority datasets identified through user engagement and government mission aligned use cases. The project ensured that high value datasets, spanning marine, atmospheric, terrestrial, and geoscience domains, are now accessible through consistent, machine readable interfaces. This represents a significant shift from legacy, heterogeneous access mechanisms toward a unified, interoperable service layer capable of supporting both traditional scientific analysis and emerging AI applications.

A key factor in this increased availability was the project's focus on selecting datasets of national importance that directly support cross domain challenges. By prioritising these datasets and exposing them through operational APIs, the project demonstrated how the EDS can provide actionable data services that meet the needs of policymakers, industry, and interdisciplinary researchers.

The project's emphasis on open methods and reproducible workflows has strengthened the operational value of these data services. The creation of narrative Jupyter notebooks, shared code repositories, and cross centre examples has demonstrated how the newly available APIs can be used in practice, lowering barriers for new users.

Overall, the increased availability of operational data services represents a major step forward for the EDS. The EDS will build on this platform within its business as usual practices and continue to identify additional high value datasets that can be published through the pipelines established in this project.

Value of integrated working

Integrated working between the project's work packages was fundamental to the success of HIGH5 for AI, enabling the delivery of outputs that were more coherent, more user centred, and more technically robust than would have been possible through isolated development. The project's structure created a framework in which each work package informed and strengthened the others. Where integration was achieved, efficiencies emerged, duplication was reduced, and a clearer vision of future services became possible. This collaborative approach ensured that the project's outputs were aligned with real user needs, interoperable across data centres, and grounded in a shared vision for AI ready environmental data services.

Close collaboration between the user focused work packages (WP2 and WP5) and the technical development work (WP3) ensured that the APIs, metadata standards, and data loading tools were shaped by genuine user requirements. This integration meant that technical decisions, such as adopting Croissant metadata, implementing Parquet based data delivery, and designing the "loadeds" Python package, were grounded in evidence rather than assumptions. The result was a set of technical outputs that addressed real user needs and were immediately usable by the communities the project aimed to support. A key learning point for the project was the value of embedding user centric scientists within the technical work packages to facilitate this interaction Integrated working also strengthened the consistency and interoperability of the project's technical outputs. WP3's work on APIs and metadata standards encouraged cross centre coordination which reduced duplication of effort and ensured that metadata, documentation, and API behaviours were consistent across the EDS. Such alignment was essential for demonstrating how heterogeneous datasets could be combined into reproducible analytical workflows which was a core objective of the project.

The development of the agentic AI demonstrator (WP4) further illustrates the value of integrated working. The demonstrator relied on the APIs, metadata, and workflows developed in WP3, while its design and evaluation were informed by the user needs identified in WP2 and WP5. This integration enabled the demonstrator to perform realistic tasks (such as dataset discovery, subsetting, and simple analysis) using real EDS services. It also allowed the project to evaluate the behaviour, risks, and trust implications of agentic AI systems in a grounded, domain specific context. The insights gained from this work directly informed the project's recommendations on AI assurance, governance, and responsible innovation.

Finally, integrated working enhanced the project's communication, documentation, and user enablement outputs. By developing shared onboarding materials, narrative notebooks, and open source examples, the work packages collectively ensured that users encountered a coherent and accessible experience across the project's outputs. This consistency reinforced the project's broader aim of lowering barriers to entry and supporting a unified user journey across the EDS.
Overall, integrated working enabled the HIGH5 for AI project to deliver outputs that were technically aligned, user driven, and strategically coherent. It created the conditions for cross centre collaboration, reduced duplication, and established a shared foundation for future AI ready environmental data services.

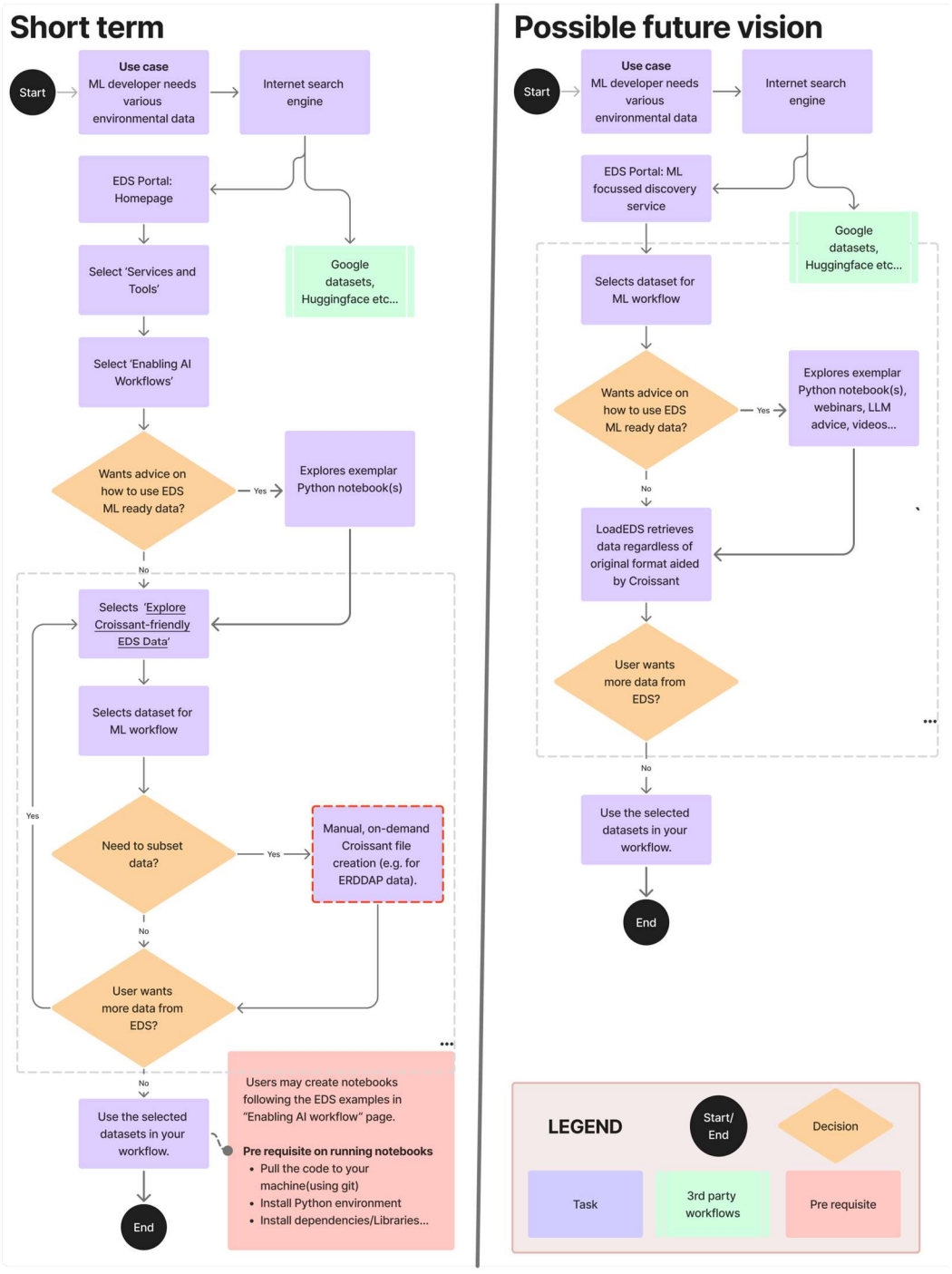
Discovering and using our AI-ready resources

Current user journey

As a result of HIGH5, users can now discover AI ready datasets via centre specific services (e.g. ERDDAP, OGC APIs, STAC catalogues) and use Croissant metadata to integrate those datasets into data science and machine learning workflows. Worked notebooks and examples provide practical guidance, which are accessed via a new set of EDS webpages. However, these pages are manually created and maintained and users are still required to navigate differences in metadata structures, publication workflows, and access mechanisms between data centres.

Envisioned future journey

Beyond the scope of this project, the team envisages a more streamlined experience in which users are guided consistently from discovery to use, regardless of data centre or data type. This would involve the creation of data discover services that are automatically updated as new ML ready data is created, along with clearer signposting from metadata records to onboarding guidance, reusable data loaders, and more harmonised publication workflows across EDS.



Need for a future EDS API (LoadEDS)

As High5 For AI evolved, alongside its sister project FRAME-FM, a recurring theme emerged around data access using Python tools. Multiple scenarios required loading of EDS, and external, datasets into a common Python environment, typically utilising the Xarray [Hoyer & Hamman, 2017] library which has become the leading package for operating on spatiotemporal datasets. This functionality was needed in to:

- Underpin the high-performance Extract-Transform-Load (ETL) caching system (FRAME-FM).

- Enable loading of soil moisture and reanalysis data in (High5 / FRAME-FM).
- Load geospatial datasets discovered through Croissant files (High5).
- Load and convert high-profile EDS data to GeoJSON for the use case comparisons (High5).
- Provide data tools provided for the agentic AI developed in WP4 (High5).
- Support recipes and examples of data loading, analysis and ML workflows (High5 / FRAME-FM).

This led to a clear opportunity to reduce duplication and develop a single Python package to support the above scenarios and cater for future projects with similar requirements. From that, the "loadeds" (i.e. load-EDS) package was developed as a prototype used by various parts of the project to support data discovery, loading, sub-setting and processing. The case for "loadeds" is compelling, including the following factors:

- Across environmental science (and other disciplines), Xarray is heavily used in the Python world, and can interface effectively with ML packages (TensorFlow, PyTorch, etc.).
- Python data services (such as web UIs and data APIs) need a data-loading capability, leading to lots of duplicated functionality.
- The EDS needs to consolidate its capabilities to ensure relevance and achieve its aims of wider data accessibility.
- As an open-source package, built on Xarray, "loadeds" is an effective single interface to support:
    - Multiple data backends
    - Multiple client tools, services and users
    - EDS datasets
    - External (non-EDS) datasets
    - Larger-than-memory datasets that require aggregation across thousands of data files (e.g. large climate and oceanographic gridded model simulations)
    - Local and remote data (including HTTP(s) and S3 interfaces)
    - A large variety of data formats, covering observational and model-derived spatiotemporal datasets

The "loadeds" library was prototyped with the following repository during the HIGH5 project:

https://github.com/NERC-EDS/HIGH5forAI_DataAPIs/tree/main/loadeds [PRIVATE]

The publicly available version will be initially launched in the repository:

https://github.com/NERC-EDS/NERC_EDS_AI_Workflows/tree/main/loadeds

Users can make use of "loadeds" through the Python library, e.g.:

```
import loadeds as eds
ds = eds.load_data("https://example.com/data/data1.json")
```

Data can be loaded from a Croissant file:

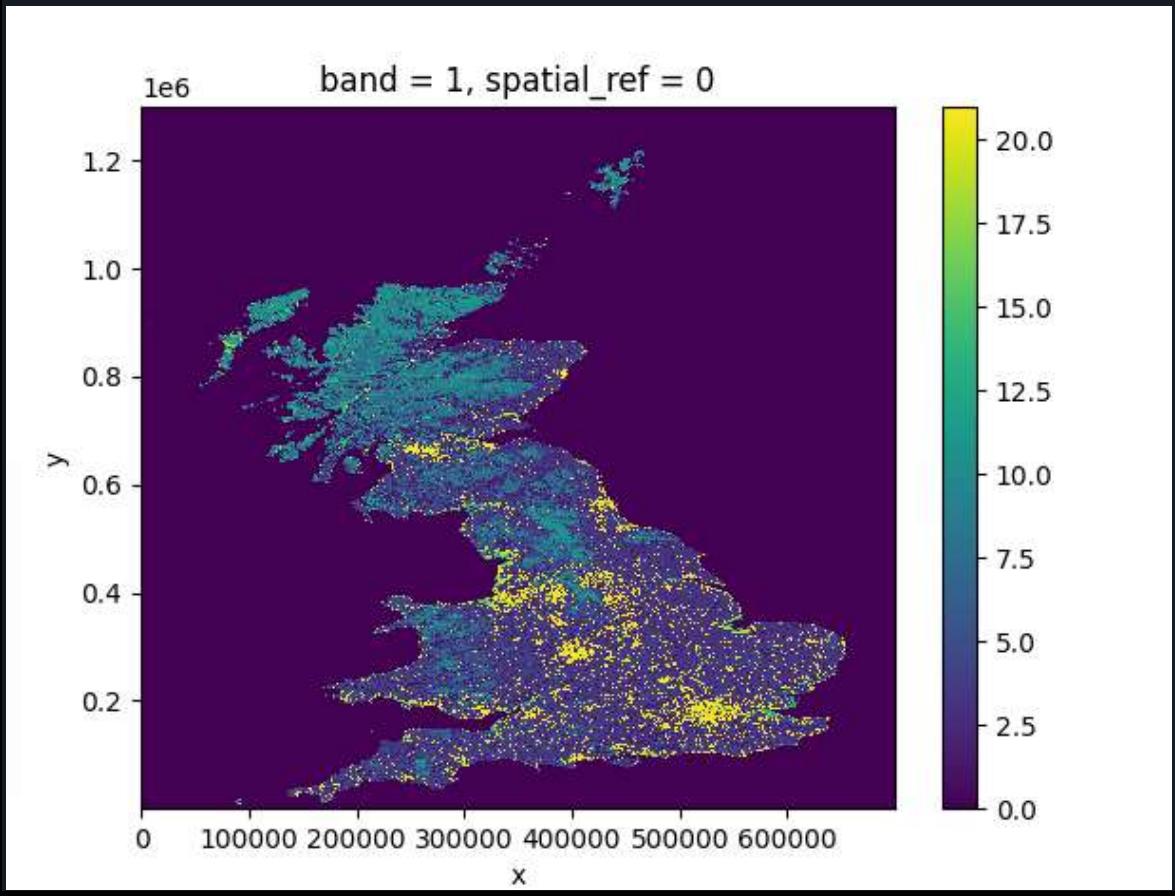
```
import loadeds as eds

ceh_dset_url = "https://gws-access.jasmin.ac.uk/public/eds_ai/high5/" + \
               “catalogue/croissant/ukceh/dataset_GB_remote.json"

da = eds.load_data_from_croissant(ceh_dset_url,
                                  record_set_id="land_cover_classes")

# Subset every 100th pixel in x and y, select first band, then plot

da.plot()
```



Or “loadeds” can be used via a simple Command-Line Interface (CLI), using the `-i` option to enter an interactive Python shell once the file is loaded:

```
$ loadeds load -i loadeds/tests/fixtures/precip_2010.nc

...

Loading data from loadeds/tests/fixtures/precip_2010.nc...

...

Entering interactive mode. The loaded dataset is available as 'ds'.

...

(InteractiveConsole)

>>> max_precip = ds.pre.isel(time=0).max().item()
>>> print(f"Maximum precipitation = {max_precip:.3f}{ds.pre.units}")

Maximum precipitation = 467.223mm
```

Trusted AI

HIGH5 WP4 shows that trust in AI for environmental science depends on how systems behave in real-world contexts, how they are governed, and how visible their limitations are.

To achieve trusted AI, **human-in-the-loop design must be the default**. EDS users consistently prefer AI tools that support discovery, explanation, and early-stage sensemaking, while clearly deferring decisions and accountability to humans. These types of interaction patterns must be central to future work. Agentic capabilities are a good match for human-in-the-loop interactions but introduce additional complexity. Brittle interfaces, choice of LLM tone and guidelines, and adversarial prompting present clear risks. This means that interface validation, controlled flexibility, and explicit prioritisation of trade-offs will be a necessary part of future work in this area.

To support heterogeneous user groups to benefit from agentic workflows, evaluations must focus on **understanding the drivers of system behaviour** on **intentional slices of test datasets** using **appropriate metrics**. WP4 demonstrated that question type is a dominant driver of behaviour and risk: an insight which is obscured by aggregated metrics. Trusted systems therefore require evaluation frameworks that explicitly test different types of interactions which are shown to be relevant to real user groups, and careful matching of metric with behaviour measured.

Finally, **trust requires sociotechnical assurance, not technical optimisation alone**. Argument-based assurance, sustained user involvement, and exploration of open models on trusted research infrastructure can connect technical work with governance, accountability, and long-term user confidence. Observability and transparency are essential technical foundations of this. Strong instrumentation, logging, and experiment tracking will be crucial to understanding technical capabilities, diagnosing failures, and supporting assurance claims.

Together, these recommendations outline a credible path toward AI systems that are **not only innovative, but demonstrably trustworthy for environmental science**.

Overall recommendations and next steps

The following consolidated recommendations bring together the project's findings across all work packages and outline the strategic actions required to sustain momentum and realise the full value of the EDS.

1. Consolidate and standardise AI-Ready metadata across the EDS

A consistent approach to metadata is essential for discoverability, interoperability, and machine-readiness.

Recommendations

- Fully adopt and standardise the use of **Croissant v1.1** across all EDS centres, supported by the application profile developed in the project.
- Update the application profile to incorporate **GeoCroissant v1.0**, enabling richer geospatial description.
- Ensure Croissant metadata entries are **discoverable and filterable** in the central EDS catalogue making "AI-Ready" datasets instantly identifiable.
- Establish conventions for **naming, structure, persistent identifiers, and vocabularies** to ensure metadata is FAIR and machine-actionable.
- Investigate Geonetwork metadata catalogue capabilities for croissant metadata

2. Develop and release a unified EDS python data-Loading package

The project repeatedly identified the need for a single, coherent mechanism for loading EDS datasets into analytical and AI workflows.

Recommendations

- Publish a baseline release of the **"loadeds"** Python package, ensuring:
    - Support for key datasets across all centres.
    - Full documentation and examples.
    - Testing and validation by each EDS partner.
- Integrate Croissant, STAC, OGC, ERDDAP, and cloud-native formats into a **single user-facing interface**.
- Create a roadmap for long-term maintenance, community contributions, and integration into EDS services.

3. Harmonise workflows and improve cross-centre integration

The project highlighted inconsistencies in metadata granularity, onboarding materials, and API behaviours.

Recommendations

- Establish a **cross-centre working group** to align:

    - Metadata publication workflows.
    - API documentation standards.
    - Onboarding granularity and terminology.
- Reduce duplication by sharing tools, code, and expertise across centres.
- Continue developing a unified user journey for discovering, accessing, and analysing AI-ready datasets.

4. Strengthen user engagement, onboarding, and community building

User research revealed persistent barriers to discovery, interpretation, and trust.

Recommendations

- Implement an **iterative onboarding feedback loop**, using user support tickets, analytics, and hackathon outputs to refine documentation.
- Expand outreach to "unknown users" in sectors such as health, construction, and insurance. Launch proactive outreach campaigns targeting sectors currently unaware of EDS data capabilities.
- Further investigate the demand for a **"single source of truth"** when designing and prioritising the provision of EDS services.
- Execute the planned **hackathon**, using real user pain points (e.g., audit trails, cloud-native formats) to shape challenges.
- Embed user archetypes and co-design processes into **EDS governance**, ensuring sustained user involvement.

5. Advance trusted AI through assurance, evaluation, and governance

WP4 demonstrated that trust in AI systems depends on transparency, behaviour under real-world conditions, and robust governance. The report emphasises that *"trust must be earned through transparency, evaluability, governance, and sustained user involvement"*.

Recommendations

- Develop a **formal assurance methodology** (e.g., argument based assurance) to support claims about AI system trustworthiness. The EDS requires a mechanism to support claims that its AI capabilities are useful and trustworthy. Future work should develop and adopt appropriate methodologies with clear, rigorous and communicable procedures; for example, the approach of argument-based assurance recommended by the Alan Turing Institute (Department for Science, Innovation and Technology, 2023; Burr et al., 2024).
- Build a **robust evaluation framework** for AI-enabled EDS AI tools and services. Such a framework will be essential for future work in this area, and continued surveying and adoption of suitable methodologies, alongside grounding in user research, will be necessary. Given the rapidly evolving nature of the field, this should include iterative literature reviews, horizon scanning, and further technical experimentation. The framework will also require

enhanced tooling for experiment tracking and observability, as well as expansion to support multi agent systems. The framework should include:

- Slice based evaluation across question types.
- Improved observability, logging, and experiment tracking.
- Support for multiagent systems.
- Provide guidance on connecting diverse users to the evaluation framework. The evaluation framework into a low-code platform which reduces barriers to access for agentic systems. This platform would provide clear guidance on creating representative, high-quality evaluation datasets, including gold standard data developed through expert consensus. This effort should be co-created with early adopters to ensure an intuitive and practical experience for users.

- Explore running **open LLMs** on UKRI/NERC infrastructure to reduce reliance on commercial providers and improve privacy and reproducibility. Future work should establish the impact of different LLMs on system behaviour as a priority; however, dependence on large commercial LLMs is an ongoing issue for the scientific community due to both ethical and privacy concerns. Impressive open-source LLMs can be run on powerful GPU clusters, but this is currently underexplored as a future avenue for EDS research, development, and deployment. The EDS should investigate setting up an LLM-as-a-service approach that ensures no data leakage to outside entities and reduced-cost access to powerful open-source LLMs.
- Provide EDS documents for Agentic AI knowledge bases in RAG-ready chunked and indexed formats
- Prioritise **human-in-the-loop** design for all AI-enabled interfaces.

6. Continue cautious, internal exploration of the Model Context Protocol (MCP)

The project found MCP promising but risky, noting concerns around provenance, attribution, and security.

Recommendations

- Treat MCP as an **internal experimental capability**, not a public-facing service.
- Focus on **low-risk use cases** such as metadata interrogation or dataset discovery.
- Align MCP exploration with EDS governance, ensuring consistency with institutional values and risk appetite.
- Use MCP selectively to expose **trusted, branded EDS capabilities**, not as a replacement for core APIs.

7. Maintain and expand the EDS technical roadmap

To ensure long-term sustainability, the project recommends a shared roadmap.

Recommendations

- Publish a **public EDS Technical Roadmap** via GitHub to demonstrate transparency and cross-centre collaboration.
- Identify priority datasets for future Croissant/API development.
- Maintain a curated set of **portable, CI-tested notebooks** demonstrating EDS data and capabilities that run across multiple platforms and are resilient to systems updates.
- Identify core EDS capabilities that can be enhanced by the use of agentic AI. Agentic AI will not be suitable for all applications. Future work should prioritise which capabilities to build for, which metrics to measure these capabilities by, and which trade-offs are acceptable.

8. Secure continued investment and strategic support

The report concludes that sustained progress requires ongoing funding and institutional commitment.

Recommendations

- Secure funding to:
    - Extend API coverage to additional high value datasets.
    - Mature the “loadeds” package into a production-ready service.
    - Deliver hackathons, onboarding improvements, and user research.
- Position EDS as a **trusted national provider** of AI-ready environmental data aligned with government missions.

Conclusions

The HIGH5 for AI project marks a pivotal step in the evolution of the Environmental Data Service, demonstrating that meaningful progress toward AI ready environmental data requires not only technical innovation but sustained sociotechnical alignment across the entire EDS ecosystem. Over the course of the project, the consortium delivered a suite of tangible assets—robust APIs, Croissant based metadata, reusable analytical workflows, user centred onboarding materials, and an agentic AI demonstrator—that collectively establish a foundation for the next generation of trusted, interoperable environmental data services. As the report notes, “the HIGH5 for AI project proved a successful foundational step in the journey of providing enhanced AI readiness of trusted environmental data services”.

A central achievement of the project was the demonstration that high value datasets from across the EDS can be exposed through consistent, standards driven interfaces that meet the expectations of modern AI and data science workflows. By adopting Croissant and exploring the emerging GeoCroissant extension, the project showed how machine readable metadata can bridge the gap between environmental data infrastructures and the needs of interdisciplinary users. This work was complemented by the development of the prototype loadeds library, which offers a unified Python based mechanism for loading, subsetting, and analysing diverse datasets—an important step toward reducing duplication and enabling a coherent user experience across centres.

Equally significant was the project's emphasis on user centred design. Through extensive engagement, the team validated real world use cases, identified persistent barriers to adoption, and clarified the expectations of users working in domains such as clean energy and public health. These insights shaped both the technical direction and the supporting materials to ensure that the project produced AI ready data services that were discoverable, interpretable, and usable by communities far beyond traditional environmental science. As the report highlights, "users often find it difficult to discover relevant data, interpret complex variables, or work with unfamiliar formats" —a challenge HIGH5 directly addressed.

The agentic AI demonstrator provided a forward-looking exploration of how AI enabled interfaces may support future EDS services. Its evaluation underscored both the promise and the risks of agentic systems: while capable of multi-step reasoning, data discovery, and basic analysis, such systems require careful governance, robust evaluation frameworks, and transparent design choices to earn user trust. The project's findings make clear that trusted AI cannot be achieved through technical optimisation alone; it must be grounded in accountability, provenance, and human in the loop design.

Looking ahead, the project sets out a clear strategic direction. Continued investment is needed to consolidate Croissant usage, mature the loadeds package, harmonise workflows across centres, and integrate AI ready metadata into the core EDS catalogue. Equally important is the development of an assurance methodology and a robust evaluation framework to guide future AI enabled services. If these next steps are realised, the EDS will be well positioned to play a central role in enabling trusted, impactful AI applications that address national and global challenges—fulfilling its mission to unlock the full value of its environmental data assets.

Supporting materials

Anthropic (2025). Model Context Protocol Specification. https://modelcontextprotocol.io/specification/2025-06-18

Hanley, M., & McCormack, M. (2025). BODC Data Submission: designing a new app. Zenodo. https://doi.org/10.5281/zenodo.15114777

Hoyer, S. & Hamman, J., (2017). xarray: N-D labeled Arrays and Datasets in Python. Journal of Open Research Software. 5(1), p.10. DOI: https://doi.org/10.5334/jors.148

NASA (2025). Earthdata MCP Server. https://github.com/datalayer/earthdata-mcp-server

Appendix A: List of high value datasets for which APIs were developed

This appendix provides a summary of the high value datasets from across the NERC EDS data centres for which APIs were developed through the project to make the data more easily accessible and discoverable for AI workflows.

Data Centre	Dataset Title	Data type	Why selected?	Pre-HIGH5 Status	Post-HIGH5 Status
BODC	(Conductivity, Depth, Temperature) CTD Profiles Link to Dataset	Netcdf or ascii	Most requested BODC dataset Currently only access via clicking a GUI. Useful for Wind farms for knowing baseline measurements	No API or AI access	All data available through ERDAPP server including Croissant descriptors.
BODC	Tide gauge station Link to Dataset	Time series	Information on sea level from tidal gauges – useful to inform tidal conditions for both wind farms and blue green infrastructure use cases.	API – needs AI adding	All data available through ERDAPP server including Croissant descriptors.
CEDA	UK Climate Projections (UKCP18) 5km gridded projections Link to Dataset	NetCDF, (STAC), (Kerchunk)	High temporal and spatial resolution of UK climate projections from 1980-2080 Use cases: Offshore (sea level) / Blue-green	STAC, Kerchunk, created croissant file & discoverable by API,	Croissant file created: Link
CEDA	HadUK-Grid weather observations Link to Dataset	NetCDF	High temporal and spatial resolution of UK weather variables from 1836-2024. Same resolutions as UKCP18.	Needs STAC, Kerchunk,recipes and maybe authz	Croissant files created: Link

			Use cases: Blue-green		
CEDA	ESA Sea Surface Temperature Climate Change Initiative (SST_cci): Level 4 Analysis product, version 3.0 Link to Dataset	NetCDF, (STAC), (Kerchunk)	High-resolution SST dataset, combined from multiple sources. 1980-present, global, 0.05°resolution (~5km) Use cases: Offshore	NetCDF, STAC, needs aggregation format, and recipes.	Croissant files created: Link
CEDA	ECMWF ERA5: surface level analysis parameter data Link to Dataset	NetCDF	ERA5 Weather variables on the surface. Global data, 1979-present. Used extensively for training weather and geospatial Foundation Models. Use case: FRAME-FM encoder	Needs STAC, Kerchunk, recipes.	Wide speed croissant files: File 1 File 2
EIDC	Land Cover Map GB Link to Dataset	Tiff	High resolution (10m) pixel data representing the land classification of GB for 2024. Selected as it provides existing land cover information to inform planning of blue/green infrastructure. One of EIDC's most downloaded datasets.	No API or AI access	Croissant file created: Link API user tutorial: Link
EIDC	Land Cover Map NI Link to Dataset	Tiff	High resolution (10m) pixel data representing the land classification of NI for 2024. Selected as it provides existing	No API or AI access	Croissant file created: Link

			land cover information to inform planning of blue/green infrastructure. One of EIDC's most downloaded datasets.		
EIDC	CHESS Link to Dataset	NetCDF	Gridded daily meteorological variables over Great Britain for the years 1961-2019 at 1 km resolution. Selected as could help provide baseline meteorological data for informing blue/green planning. Could also link to some windfarm cases in coastal regions potentially.	No API or AI access	Data and croissant files available through ERDAPP server – created as part of project: Link
NGDC	Offshore bedrock 250K, 4 datalayers Link to Dataset	4 Vector	Provides information relating to foundations suitability for wind farm infrastructure	Existing shapefile, geopackage and OGC WMS	4 Modernised OGC API - Features layers created within a consolidated "offshore small scale digital mapping" service OGC API - Features:https://map.bgs.ac.uk/arcgis/rest/services/SDDS/BGS_offshore_small_scale_digit

					al mapping/OGCFeatureServer • Offshore Bedrock 250k: Structural Geology (0) • Seabed Hard Substrate (1) • Offshore Bedrock 250k: Lithostratigraphical units (2) • Bedrock Summary Lithologies (6) Croissant descriptor created:: Link

					ArcGIS REST API created for portfolio consistency: Link
NGDC	Seabed sediments 250K, 4 data layers Link to Dataset	4 Vector	Provides information relating to foundations suitability for wind farm infrastructure	Existing Shapefile, Geopackage, OGC WMS	4 Modernised OGC API - Features layers created within a consolidated "offshore small scale digital mapping" service • Seabed Sediments 250k (3) • Quaternary Deposits Summary Lithologies (4) • Quaternary Deposit

					s Thickness (5) • Offshore Quaternary 250k UK (8) Croissant descriptor created: Link OGC WMS created (Link) ArcGIS REST API created Link
NGDC	Predictive Seabed Sediments (UK),4 data layers Link to Dataset	1 Vector and 3 Raster.	Provides information relating to foundations suitability for wind farm infrastructure	Existing shapefile/geopackage download.	1 new OGC API - Features created for vector layer, within consolidated "offshore small scale digital mapping" service: Predictive Seabed Sediments

					OGC WMS created for all 4 layers: Link ArcGIS REST API created for all 4 layers : Link Croissant descriptor created: (vector data only): Link
NGDC	Offshore regional guides Link to Dataset	Annotated HTML Text	Provides contextual information Describing offshore geology and data sources used to develop offshore mapping	HTML pages online	Report metadata added to OGC API Features aligned API (Report content Ingested to knowledge base for WP4.)
NGDC	Onshore Borehole materials Link to Dataset	**Vector (point)**	Catalogue of physical samples from the subsurface obtained from boreholes, which can be analysed to inform foundation suitability. Samples are from	Existing web search form, Existing OGC API Features for	APIs created as part of this project:

			onshore boreholes; data extrapolated offshore using knowledge of stratigraphy to fill data gaps. Part of collection is thin section samples which also have associated images.	part of dataset (thin section samples and images, published for API4AI project)	OGC API - Features created: Link Croissant descriptor created:: Link OGC WMS created: Link ArcGIS REST API created: Link
NGDC	Offshore samples, 9 data layers Link to Dataset	Vector (point)	Catalogue of physical samples sampled from the seabed or subsurface obtained from offshore boreholes (hydrocarbons wells), which can be analysed to inform foundation suitability for windfarms. Online photographs included	Existing Web search application Existing beta OGC API – Features, organised as 5 layers Link	APIs created as part of this project: Production OGC API Features created, organised as 9 layershttps://arcgis.bgs.ac.uk/arcgis/rest/services/SDDS/Offshore_Sample_Data/OGCFeatur

					eServer/collections/3 • Offshore samples: metadata (borehole samples) (0) • Offshore samples: metadata (grab samples) (1) • Offshore samples: metadata (other samples) (2) • Offshore samples:

					geology (borehole samples) (3) • Offshore samples: geology (grab samples) (4) • Offshore samples: geology (other samples) (5) • Offshore samples: geotechnical data (6) • Offshore samples:

					particle size analysis (7) • Offshore samples: geochemistry (8) OGC WMS created: Link ArcGIS REST API created: Link Croissant descriptor created:: Link
NGDC	Offshore Hydrocarbon Wells Existing dataset link and API access: Link to Dataset	OGC API -> GeoJSON	Provides information on offshore hydrocarbon wells for information when assessing location for wind farm suitability.	Existing beta OGC API Features	Croissant descriptor created: Link
NGDC	Historic Earthquakes Existing dataset link and API access: Link to Dataset	OGC API -> GeoJSON	Earthquakes in UK and UKCS recorded prior to 1970	Existing beta OGC API Features	Croissant descriptor created: Link

NGDC	Modern Earthquakes Existing dataset link and API access: Link to Dataset	OGC API -> GeoJSON	Earthquakes in UK and UKCS recorded since 1970	Existing beta OGC API Features	Croissant descriptor created: Link
NGDC	Offshore Strategic Environmental Assessment (SEA), 5 data layers	OGC API -> GeoJSON	Datasets obtained from strategic environmental assessments, including sea floor photographs, reports and datasets. Useful for assessing seabed conditions and environmental constraints in siting of offshore windfarms.	Existing Web application: Link Existing beta OGC API Features: • Event Line API • SEA Event Point API • SEA Offshore Sea Areas API • SEA Offshore bounding boxes API • SEA Polygon data API	Croissant descriptor created (Each API distribution within): Link

NGDC	Background concentrations of contaminants in English soils	Raster	Could be used to inform the blue-green space use-case by indicating areas of contaminated land.	Existing ArcGIS REST API Existing OGC WMS: Link	Notebook for data access created: Link