# Comparing and Evaluating LLMs for Efficient and Responsible Data Rescue

Presenter: Shona Ferguson, Environmental Data Scientist, UKCEH Edinburgh

05/06/2026

Co-authors: J Neil Cape, Alan Crossley, Frank Harvey, David Fowler, David Leaver, Christine F Braban

UKCEH

Our planet.
Decoded.

Intro

# Why should we rescue scientific data?

- Unique records that cannot be reproduced
- Foundation for new research
- AI and machine learning demand
- Preservation and accessibility

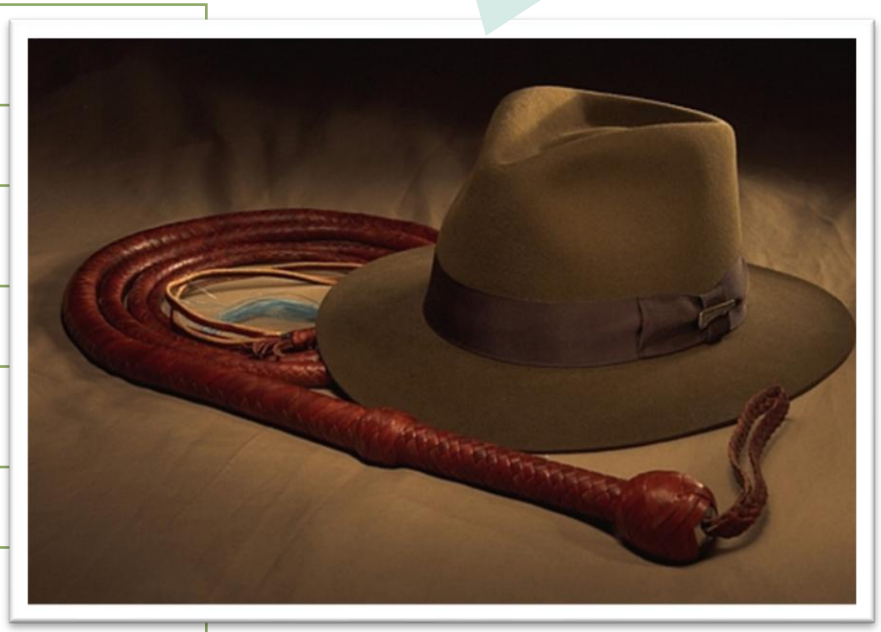
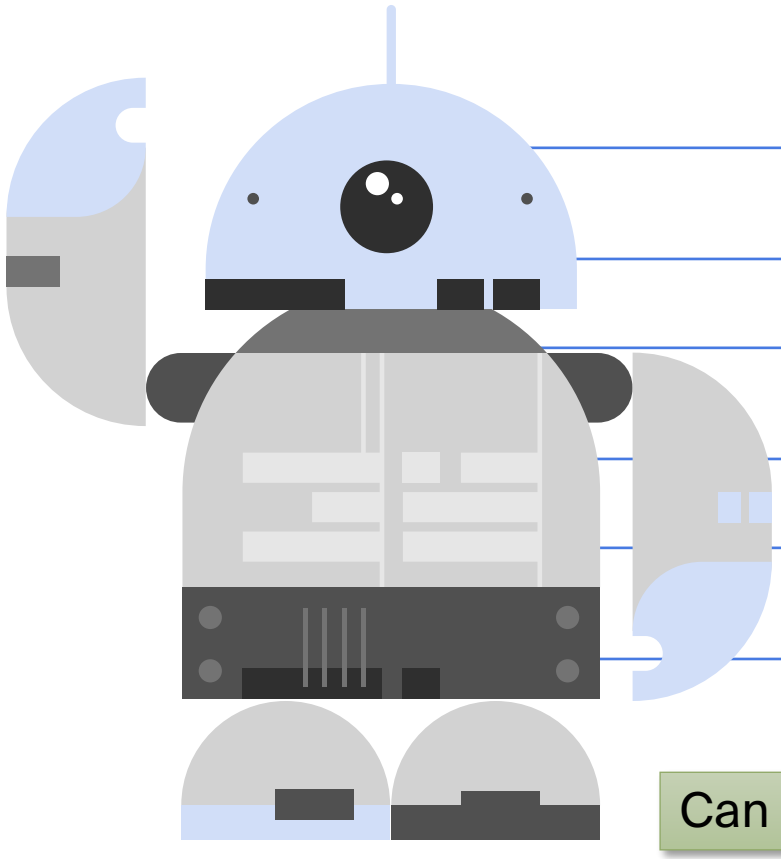



Image credit: Gary Stewart (https://creativecommons.org/licenses/by/3.0/deed.en)

Intro

# Why are we trying AI?



- Manual workflows are slow and expert-heavy
- Accelerate the process
- Reduce subjectivity

Can AI accelerate data rescue without compromising scientific integrity?

# Methodology

Methodology

# Overview

Rescued data collected 20 years ago in legacy excel format

Needed cleaned and metadata/supporting documentation added

Data is now published

We used LLMs to repeat this process and see how they compare against my manual process

## Methodology

# Our data rescue project

- Rescued historical cloud and rain chemistry dataset
- Monitoring and modelling
- Strong reuse potential

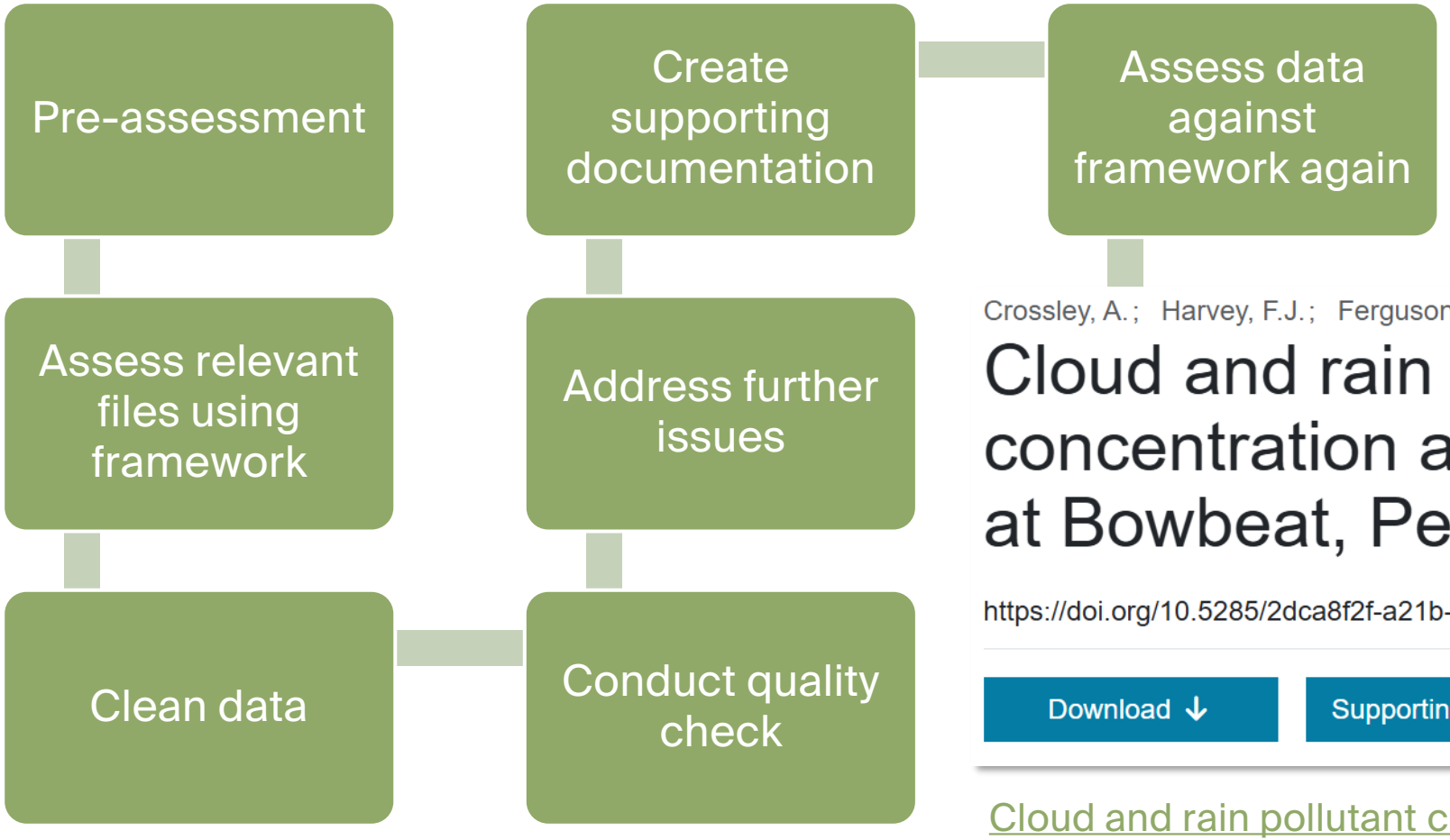
## Methodology

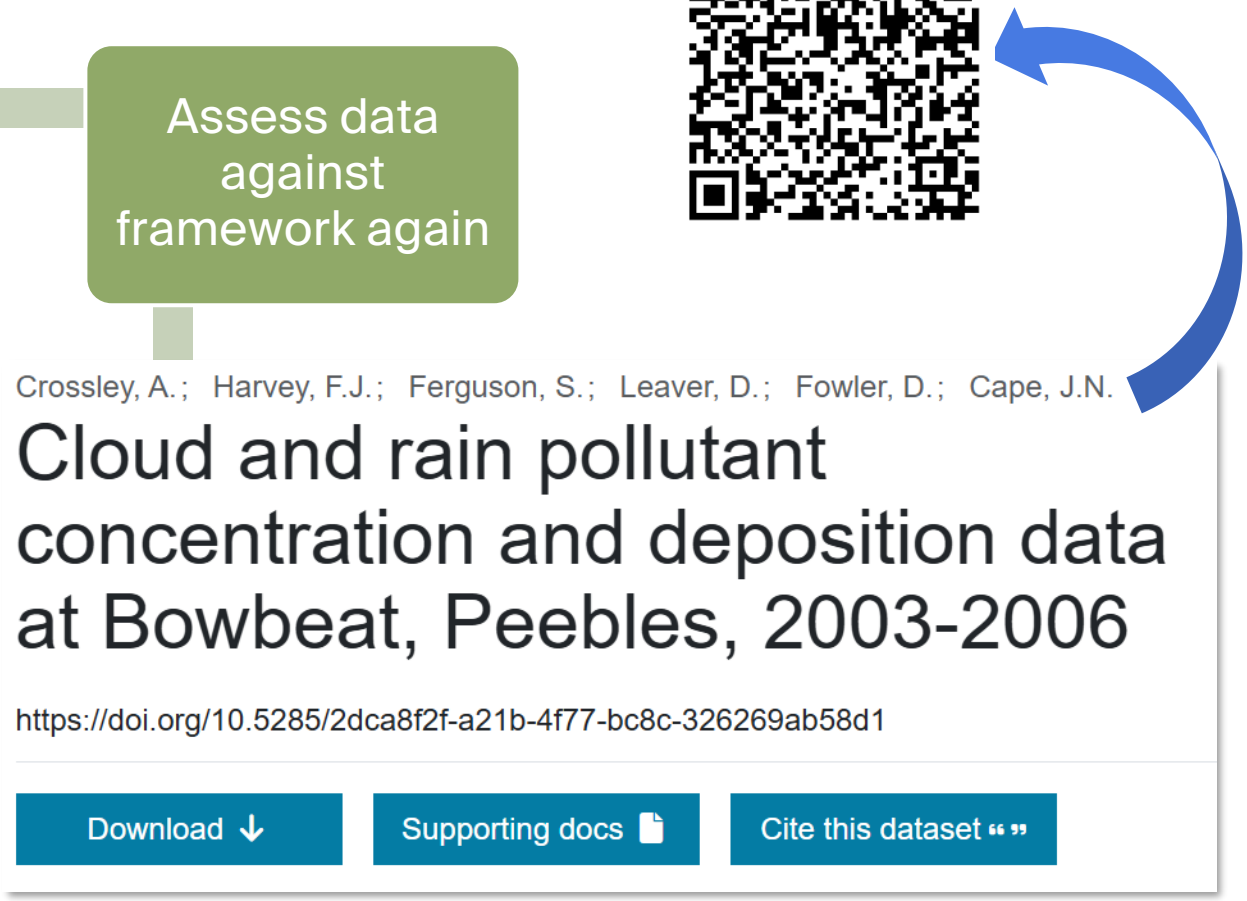
# The data

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC | BBC |
| 2 | COLLECTION DATE | CODE | BOTTLE | COMMENTS | BOTTLE | BOTTLE | SAMPLE | VOLUME | ARCHIVE | ARCHIVE | VOLUME | ph unfiltered | pH filtered | COND | Na | Na | K | K | Ca | Ca | Mg | Mg | NO3-N | NO3-N | NH4-N | NH4-N | Cl | |
| 3 | | | No. | | +SAMPLE | | VOLUME | FOR V.W. | filtered | Unfiltered | mm | | | | mg/L | | mg/L | | mg/L | | mg/L | | mg/L | | mg/L | | mg/L | |
| 4 | | | | | gms | gms | mls | SO4 | | | | | | | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | W/S | | W/S | | W/S | | W/S | | W/S | | W/S | | W/S | |
| 6 | | | | | | | | | | | | | | | <0.10 | | <0.10 | | <0.10 | | <0.10 | | <0.10 | | <0.10 | | <0.10 | |
| 7 | | | | | | | | | | | | | | | <0.02 | | <0.02 | | <0.02 | | <0.02 | | <0.02 | | <0.02 | | <0.02 | |
| 8 | | | | | | | | | | | | | | | sodium | NH4-N | potassium | calcium | magnesiu | Chloride | NO3-N | SO4-S | | | | | | |
| 9 | 07/Jan/2004 | BBC | | | | | | | | | | | | | | 0.00 | | | | | | | | | | | | |
| 10 | 14/Jan/2004 | BBC | 18 | | 2610 | 331 | 2279 | 2279 | 1 | 2 | 72.52 | | | | 14.86 | 1077.64 | 0.67 | 48.61 | 0.70 | 50.68 | 1.44 | 104.36 | 0.97 | 70.39 | 1.63 | 118.16 | 26.14 | 1895.5 |
| 11 | 21/Jan/2004 | BBC | 19 | | 5951 | 321 | 5630 | 5630 | 1 | 2 | 179.19 | | | | 4.97 | 889.99 | 0.42 | 74.75 | 0.36 | 65.17 | 0.52 | 94.06 | 0.88 | 158.54 | 0.30 | 53.59 | 7.32 | 1311.5 |
| 12 | 28/Jan/2004 | BBC | 20 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13 | 04/Feb/2004 | BBC | 21 | | 4735 | 327 | 4408 | 4408 | 1 | 2 | 140.31 | | | | 13.30 | 1866.22 | 0.76 | 106.99 | 0.63 | 88.93 | 1.31 | 183.35 | 0.80 | 112.77 | 1.86 | 260.53 | 23.26 | 3263.9 |
| 14 | 11/Feb/2004 | BBC | 22 | | 2640 | 324 | 2316 | 2316 | 1 | 2 | 73.71 | | | | 33.15 | 2443.83 | 1.37 | 100.93 | 1.40 | 103.48 | 2.74 | 202.05 | 1.25 | 92.41 | | 0.00 | 51.60 | 3803.3 |
| 15 | 18/Feb/2004 | BBC | 23 | | 1561 | 332 | 1229 | 1229 | 1 | 2 | 39.10 | | | | 4.03 | 157.49 | 0.57 | 22.19 | 1.77 | 69.37 | 0.57 | 22.21 | 2.73 | 106.64 | 1.97 | 77.17 | 5.80 | 226.8 |
| 16 | 25/Feb/2004 | BBC | 24 | | 1011 | 332 | 679 | 679 | 1 | 2 | 21.62 | | | | 31.01 | 670.44 | 1.39 | 30.10 | 1.49 | 32.31 | 2.68 | 57.98 | 2.09 | 45.09 | 3.87 | 83.74 | 57.58 | 1244.5 |
| 17 | 03/Mar/2004 | BBC | 25 | | 165 | 111 | 54 | #REF! | 0 | 0 | 1.72 | | | | 188.54 | 324.04 | 8.83 | 15.18 | 13.25 | 22.77 | 17.07 | 29.34 | 41.04 | 70.54 | 20.60 | 35.40 | 286.97 | 493.1 |
| 18 | 10/Mar/2004 | BBC | 26 | | 1085 | 321 | 764 | 764 | 1 | 2 | 24.32 | | | | 44.47 | 1081.37 | 2.41 | 58.59 | 3.26 | 79.35 | 4.25 | 103.24 | 10.68 | 259.67 | 6.80 | 165.26 | 71.24 | 1732.3 |
| 19 | 17/Mar/2004 | BBC | 27 | | 1227 | 334 | 893 | 893 | 1 | 2 | 28.41 | | | | 28.25 | 802.47 | 1.38 | 39.11 | 2.45 | 69.56 | 2.63 | 74.68 | 8.97 | 254.77 | 5.73 | 162.66 | 41.94 | 1191.4 |
| 20 | 24/Mar/2004 | BBC | 28 | | 2159 | 324 | 1835 | 1835 | 1 | 2 | 58.41 | | | | 20.77 | 1213.18 | 0.70 | 40.88 | 0.79 | 46.15 | 1.69 | 98.89 | 0.56 | 32.43 | | 0.00 | 31.96 | 1866.5 |
| 21 | 31/Mar/2004 | BBC | 29 | | 1714 | 333 | 1381 | 1381 | 1 | 2 | 43.96 | | | | 15.61 | 686.16 | 0.97 | 42.85 | 1.29 | 56.83 | 1.58 | 69.32 | 4.24 | 186.53 | 6.27 | 275.53 | 27.66 | 1215.8 |
| 22 | 07/Apr/2004 | BBC | 30 | | 2951 | 319 | 2632 | 2632 | 1 | 2 | 83.77 | | | | 10.38 | 869.85 | 1.19 | 99.64 | 1.31 | 109.78 | 1.13 | 94.58 | 6.18 | 517.74 | 7.57 | 634.54 | 18.76 | 1571.3 |
| 23 | 14/Apr/2004 | BBC | 31 | | 1994 | 323 | 1671 | 1671 | 1 | 2 | 53.18 | | | | 3.68 | 195.90 | 0.47 | 25.25 | 0.65 | 34.66 | 0.50 | 26.82 | 1.54 | 82.12 | 2.11 | 112.12 | 5.85 | 310.9 |
| 24 | 21/Apr/2004 | BBC | 32 | | 3301 | 333 | 2968 | 2968 | 1 | 2 | 94.47 | | | | 2.57 | 243.22 | 0.38 | 36.30 | 0.44 | 41.52 | 0.35 | 33.37 | 0.93 | 88.11 | 0.60 | 57.14 | 3.78 | 357.4 |
| 25 | 28/Apr/2004 | BBC | 33 | | 1086 | 111 | 975 | 975 | 1 | 2 | 31.02 | | | | 17.85 | 553.55 | 0.87 | 27.06 | 1.46 | 45.22 | 1.78 | 55.35 | 3.04 | 94.28 | 3.14 | 97.35 | 30.07 | 932.6 |
| 26 | 05/May/2004 | BBC | 34 | | 5919 | 323 | 5596 | 5596 | 1 | 2 | 178.11 | | | | 5.18 | 922.88 | 0.52 | 93.10 | 0.61 | 108.37 | 0.59 | 105.48 | 1.57 | 280.37 | 1.88 | 335.32 | 8.16 | 1453.1 |
| 27 | 12/May/2004 | BBC | 35 | | 3160 | 326 | 2834 | 2834 | 1 | 2 | 90.21 | | | | 1.36 | 122.84 | 0.52 | 47.15 | 0.94 | 85.04 | 0.30 | 26.88 | 2.86 | 258.12 | 2.68 | 242.11 | 1.87 | 168.7 |
| 28 | 19/May/2004 | BBC | 36 | | 2153 | 331 | 1822 | 1822 | 1 | 2 | 57.98 | | | | 4.58 | 265.56 | 0.49 | 28.65 | 0.81 | 46.68 | 0.52 | 29.86 | 1.65 | 95.78 | 2.48 | 143.53 | 6.96 | 403.2 |
| 29 | 26/May/2004 | BBC | 37 | | 517 | 119 | 398 | 398 | 1 | 1 | 12.67 | | | | 39.23 | 497.14 | 1.28 | 16.27 | 2.37 | 30.06 | 3.27 | 41.48 | 3.76 | 47.63 | | 0.00 | 55.61 | 704.8 |
| 30 | 02/Jun/2004 | BBC | 38 | | 744 | 112 | 632 | 632 | 1 | 2 | 20.12 | | | | 11.88 | 238.94 | 1.08 | 21.79 | 2.88 | 58.01 | 1.40 | 28.08 | 7.55 | 151.80 | 5.94 | 119.47 | 16.37 | 329.3 |
| 31 | 09/Jun/2004 | BBC | 39 | | 1079 | 127 | 952 | 952 | 1 | 2 | 30.31 | | | | 7.76 | 235.30 | 0.66 | 19.86 | 0.76 | 22.97 | 0.79 | 23.83 | 1.25 | 37.94 | 2.33 | 70.69 | 12.65 | 383.4 |
| 32 | 16/Jun/2004 | BBC | 40 | | 3561 | 240 | 3321 | 3321 | 1 | 2 | 105.68 | | | | 5.66 | 598.47 | 0.50 | 52.55 | 0.80 | 84.35 | 0.61 | 64.42 | 1.03 | 108.70 | 1.30 | 137.58 | 8.38 | 885.4 |
| 33 | 23/Jun/2004 | BBC | 41 | | 4598 | 337 | 4261 | 4261 | 1 | 2 | 135.60 | | | | 2.31 | 313.83 | 0.24 | 32.99 | 0.29 | 39.44 | 0.22 | 29.91 | 0.85 | 115.06 | 0.81 | 110.13 | 3.67 | 497.0 |
| 34 | 30/Jun/2004 | BBC | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 35 | 07/Jul/2004 | BBC | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 36 | 15/Jul/2004 | BBC | 42 | | 5955 | 333 | 5622 | 5622 | 1 | 2 | 178.91 | | | | 6.90 | 1234.42 | 0.48 | 85.89 | 0.42 | 75.21 | 0.86 | 153.72 | 0.88 | 157.69 | 1.02 | 182.31 | 11.02 | 1971.8 |
| 37 | 21/Jul/2004 | BBC | 43 | | 530 | 111 | 419 | 419 | 1 | 1 | 13.33 | | | | 12.27 | 163.53 | 0.95 | 12.67 | 1.16 | 15.44 | 1.68 | 22.45 | 2.68 | 35.77 | 2.68 | 35.70 | 18.69 | 249.1 |
| 38 | 28/Jul/2004 | BBC | 44 | | 1194 | 128 | 1066 | 1066 | 1 | 2 | 33.94 | | | | 9.91 | 336.39 | 0.64 | 21.66 | 0.72 | 24.58 | 1.26 | 42.62 | 1.13 | 38.49 | 1.57 | 53.27 | 16.20 | 549.8 |
| 39 | 04/Aug/2004 | BBC | 45 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 40 | 11/Aug/2004 | BBC | 46 | | 3638 | 336 | 3302 | 3302 | 1 | 2 | 105.10 | | | | 0.95 | 100.31 | 0.24 | 25.06 | 0.77 | 81.21 | 0.07 | 7.75 | 2.49 | 261.39 | 2.10 | 220.91 | 1.57 | 165.3 |
| 41 | 18/Aug/2004 | BBC | 47 | | 4373 | 333 | 4040 | #REF! | 1 | 2 | 128.59 | | | | 3.53 | 454.47 | 0.38 | 48.39 | 0.33 | 42.80 | 0.33 | 42.14 | 2.30 | 296.21 | 1.76 | 226.36 | 5.56 | 714.9 |
| 42 | 25/Aug/2004 | BBC | 48 | | 2010 | 321 | 1689 | 1689 | 1 | 2 | 53.76 | | | | 5.26 | 282.52 | 0.41 | 21.93 | 0.54 | 29.22 | 0.57 | 30.90 | 2.07 | 111.23 | 1.11 | 59.43 | 7.52 | 404.0 |
| 43 | 01/Sep/2004 | BBC | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 44 | 08/Sep/2004 | BBC | 49 | | 3035 | 331 | 2704 | 2704 | 1 | 2 | 86.07 | | | | 19.03 | 1637.72 | 0.81 | 70.01 | 1.05 | 90.72 | 2.28 | 195.93 | 1.74 | 149.94 | 2.23 | 191.91 | 33.19 | 2856.2 |

Methodology

# Manual methodology



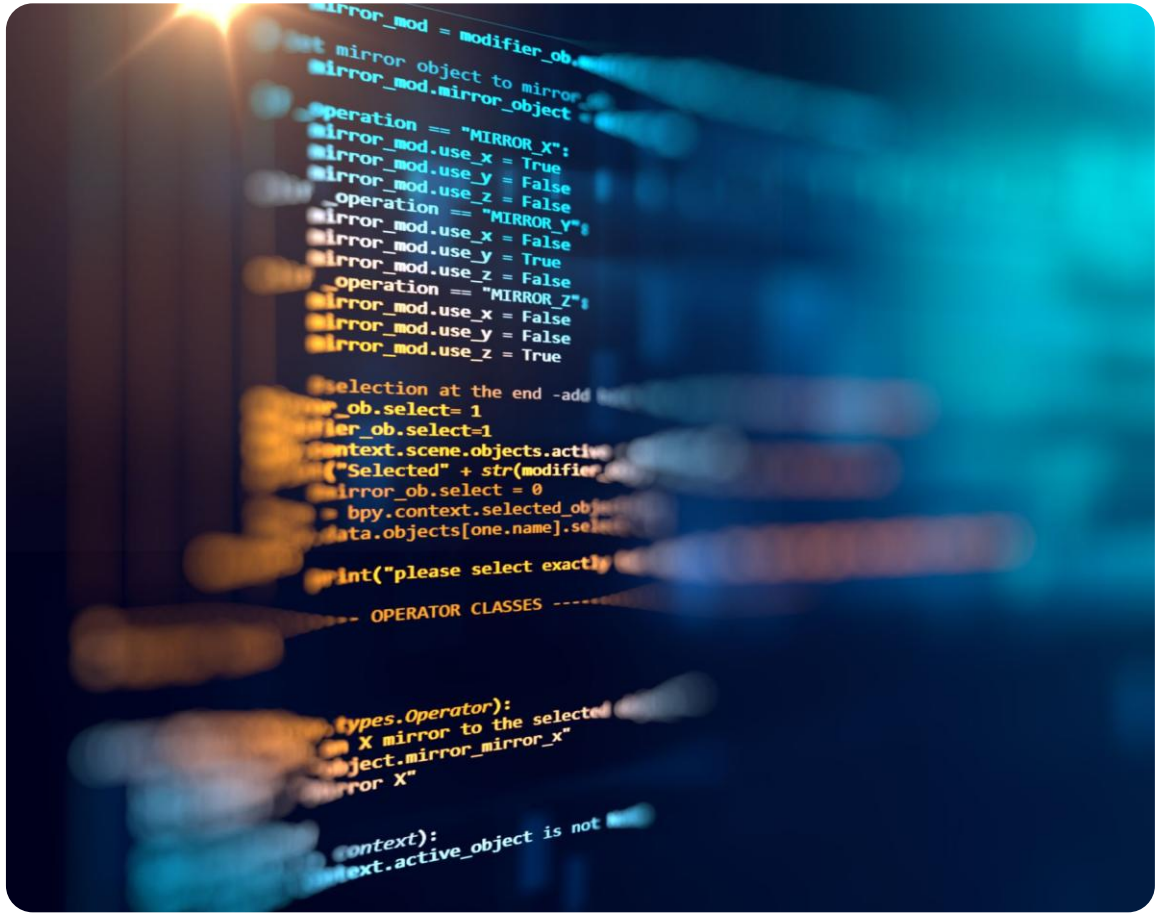




Cloud and rain pollutant concentration and deposition data at Bowbeat, Peebles, 2003-2006 - EIDC

Methodology

# AI methodology

- Created some LLM prompts based on each step of the manual methodology
- Fed prompts to three LLMs
- Created a grading system to score the LLMs based on 4 metrics for each task



Methodology

# The AI prompts

Code generation

Comprehensive assessment of the data before and after rescue

Searching the internet for associated publications and for the most appropriate repository

Writing supporting documentation

Methodology

# Evaluation metrics

Accuracy – how correct is the LLM output based on the manual findings?

Completeness – how complete is the LLM output?

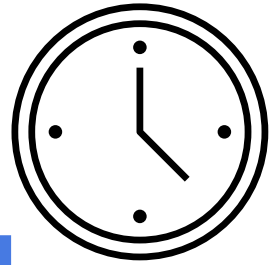
Efficiency – how much time was saved compared with the manual effort?

Usability – is the LLM output usable without major rework?

# Key Findings

Key Findings

# Estimated efficiency gains



| | Estimated Time Taken |
|---|---|
| Manual | 134.5 hours (~ 18 days) |
| Hybrid LLM/human QC | 92 hours (~ 12 days) |

**The estimated total amount of time saved is 42.5 hours (32%)**

- Approximately 6 days of full-time work

**Gains were most evident in the following tasks:**

- Constructing a framework-based assessment of the dataset
- Writing supporting documentation drafts
- Code generation

## Key Findings

# Best LLM is...

- It depends!

Key Findings

# Evaluation metrics scoring

| Metric | Copilot | ChatGPT | Claude |
|---|---|---|---|
| Accuracy | 3.3 | 4.1 | 4.5 ★ |
| Completeness | 3.6 | 4 | 4.5 ★ |
| Efficiency | 2.8 | 3.2 | 3.3 ★ |
| Usability | 3.5 | 4.5 ★ | 4.1 |

Key Findings

# Performance varies significantly between models

Claude: More **accurate and scientifically robust**

ChatGPT: Most **consistent, structured, usable outputs**

Copilot: More variable; sometimes insightful but less reliable

## Key Findings

# Structured frameworks reduce LLM variance

### Before rescue

| assessment_factor | manual | copilot | chatgpt | claude |
|---|---|---|---|---|
| Extent | 4 | 4 | 4 | 4 |
| Data Objects | 4 | 5 | 4 | 5 |
| User Communities | 5 | 4 | 4 | 4 |
| Stakeholders | 5 | 4 | 4 | 5 |
| Reuse Value | 4 | 4 | 5 | 5 |
| Reusable Objects | 5 | 4 | 5 | 5 |
| Historical Value | 5 | 4 | 4 | 4 |
| Historical Objects | 5 | 3 | 5 | 4 |
| Completeness | 4 | 3 | 4 | 3 |
| Sensitivity | 5 | 5 | 5 | 5 |
| Access and use constraints | 3 | 5 | 4 | 5 |
| Rarity or uniqueness | 5 | 5 | 5 | 4 |
| Reproducibility | 3 | 3 | 4 | 4 |
| Relevant collections | 3 | 3 | 4 | 4 |
| Associated publications | 5 | 2 | 2 | 3 |
| Fit for purpose | 3 | 3 | 4 | 3 |
| Obstacles to recovery | 3 | 3 | 4 | 3 |
| Priorities | 5 | 5 | 5 | 5 |
| Total | 76 | 69 | 76 | 75 |

### After rescue

| assessment_factor | manual | copilot | chatgpt | claude |
|---|---|---|---|---|
| Extent | 5 | 4 | 5 | 5 |
| Data Objects | 5 | 5 | 5 | 5 |
| User Communities | 5 | 5 | 5 | 4 |
| Stakeholders | 5 | 5 | 5 | 5 |
| Reuse Value | 5 | 5 | 5 | 5 |
| Reusable Objects | 5 | 5 | 5 | 5 |
| Historical Value | 5 | 5 | 4 | 4 |
| Historical Objects | 5 | 4 | 5 | 4 |
| Completeness | 4 | 4 | 5 | 5 |
| Sensitivity | 5 | 5 | 5 | 5 |
| Access and use constraints | 5 | 5 | 5 | 5 |
| Rarity or uniqueness | 5 | 5 | 5 | 4 |
| Reproducibility | 4 | 4 | 5 | 4 |
| Relevant collections | 3 | 4 | 5 | 4 |
| Associated publications | 4 | 4 | 5 | 4 |
| Fit for purpose | 5 | 5 | 5 | 5 |
| Obstacles to recovery | 5 | 5 | 5 | 5 |
| Priorities | 5 | 5 | 5 | 5 |
| Total | 85 | 84 | 89 | 83 |

## Key Findings

# Structured frameworks reduce LLM variance

### Chemistry Dataset: Hoffman Point Grade Assessment

Collection: CHEMISTRYMar08JNC_SFcopy.xlsx | Sites: Black Burn Cloud/Rain & Auchencorth Moss | Period: 2003–2007

*Grades 1–5 follow Hoffman Point Grade Descriptors. Grade 5 = most favourable; Grade 1 = least favourable. Starred (*) criteria are flagged as priority in the descriptor framework.*

| # | Criterion | Assessment | Grade Rationale | Grade |
|---|---|---|---|---|
| 1 | Extent | The collection consists of a single .xlsx workbook (~1MB) with 9 sheets spanning 2003–2007. Sheets contain between 66–172 rows and 21–81 columns of weekly atmospheric chemistry data. Two site clusters are covered: Black Burn Cloud (BBC) and Black Burn Rain (BBR), plus Auchencorth Moss. A summary sheet aggregates 2003–2006. The data are already in digital form and partially processed: volume-weighted means, sums, and sea-salt-corrected concentrations have been calculated. However, the workbook is incompletely documented and contains known formula errors (#REF!). | Collection is small and partially processed. | 4 |
| 2 | Data Objects | All data exist as structured tabular values within an Excel (.xlsx) workbook — a machine-readable, open format. Variables include collection date, bottle/sample metadata, precipitation volume, pH (filtered/unfiltered), conductivity, and ion concentrations (Na, K, Ca, Mg, NO3-N, NH4-N, Cl, SO4-S) in mg/L and µmolar, alongside sea-salt corrections and volume-weighted means. Data are standalone in the spreadsheet, not embedded in reports or field notebooks. Some columns contain derived values (e.g. Log10 uMolar, non-marine fractions) computed within Excel formulas. | Data are in a machine-readable, open format (.xlsx). | 5 |
| 3 | User Communities | The dataset is directly relevant to atmospheric chemists, environmental scientists, and biogeochemists working on precipitation chemistry, acid rain, and critical loads research. Potential user communities include researchers studying sulphur and nitrogen deposition in Scotland and the UK more broadly, policymakers concerned with air quality and ecosystem impacts, and long-term monitoring networks such as the UK Acid Deposition Monitoring | There are identifiable user communities that can reuse these data, though no contact appears to have been made. | 4 |

Key Findings

# Where do LLMs add value?

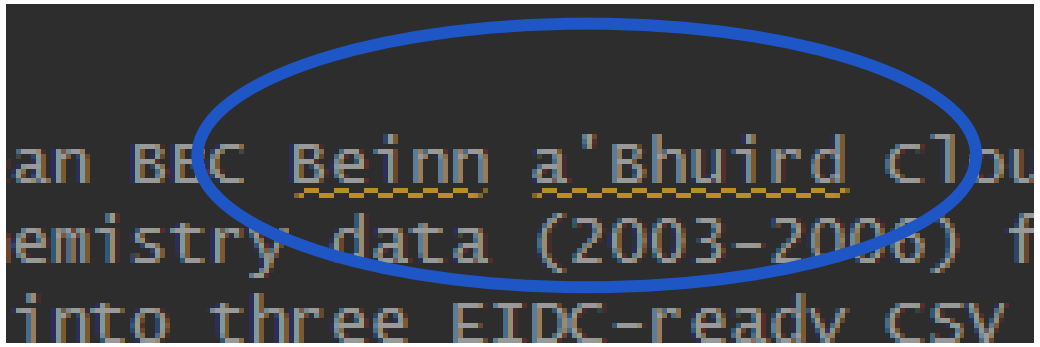
- Summarisation and structuring
- Metadata drafting
- Identifying file issues
- Generating code scaffolds
- Interpreting tabular patterns

- Scientific correctness
- Context-specific interpretation
- Edge cases in messy data
- Literature searches

## Key Findings

# Watch out for hallucinations!

**Caltech Active Strand Cloudwater Collector (CASCC)**



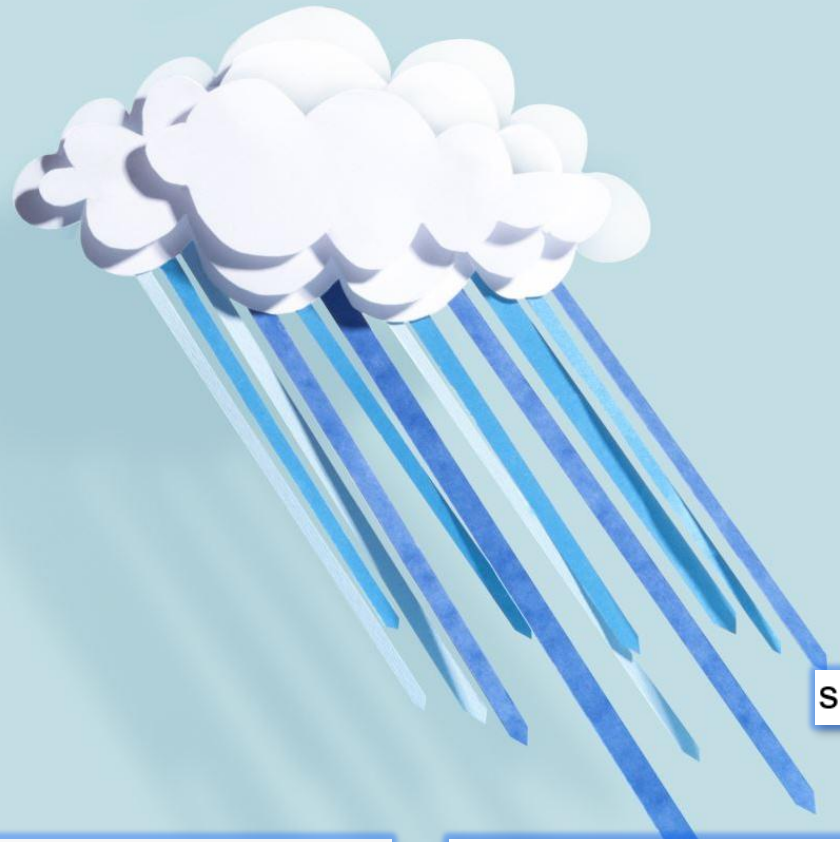

```
cloud_sheets <- c(
    "Cloud 2003",
    "Cloud 2004",
    "Cloud 2005",
    "Cloud 2006"
)

adjusted_cloud_sheets <- c(
    "Adjusted Cloud 2003",
    "Adjusted Cloud 2004",
    "Adjusted Cloud 2005",
    "Adjusted Cloud 2006"
)
```

: Black Burn Cloud (BBC) and Black Burn Rain (BBR),

The dataset's **temporal resolution is "event-based" (individual precipitation events)**

Key Findings



# Scientific understanding is uneven

- LLMs can mimic expertise—but be careful in assuming that they have true domain reasoning

support the development of seeder-feeder cloud deposition models for the UK.

Non-marine (excess) sulphate concentration expressed as sulphur, calculated by subtracting the marine fraction estimated from Na concentration

Cloud water was collected using a passive string (harp) collector, a design widely used in UK upland cloud deposition research. The collector intercepts cloud droplets on vertical strings arranged in a frame; collected water drains into a sample bottle. The equivalent depth (mm) of cloud water collected is calculated by dividing the sample volume (mL) by the effective collection area of the strings ($m^2$), converted to mm.

Key Findings

# Human expertise remains essential

- Validating scientific interpretations
- Resolving ambiguous variables and units
- Interpreting missing data vs structural absence
- Confirming metadata and provenance
- Final QC

Key Findings

# What have I learned about LLM prompting?

- Give as much context as you can
- Provide LLMs with frameworks and structured systems
- Specifically request that the LLM does not “guess” and uses evidence to back up its answers
- Be very clear with exactly what type of output you would like

Key Findings

# Challenges and Limitations

- Freely available Claude has limit of how many messages you can exchange within a 5-hour window
- Mix of consistency with output formats

- Only one small dataset tested and LLM prompts were only run once – can't draw conclusions on consistency of LLM outputs
- Only the free versions of the LLMs were tested

# Conclusions and further research

Conclusions and further research

# Conclusions

LLMs are powerful accelerators—but not replacements

~30% time saving achievable

LLMs perform strongly on structured tasks but have weak scientific certainty

Hallucinations are subtle and significant

Model choice matters

Conclusions and further research

# Best practice: hybrid “human-in-the-loop workflow

This gives:

- Efficiency gains
- Maintained scientific integrity
- Improved reproducibility

## Conclusions and further research

# Impact

- Data rescue is critical for preserving legacy datasets
- LLMs make it:
  - Faster
  - more scalable
  - more accessible

## Conclusions and further research

# Further research

- Assess consistency across different data formats and domains
- Using LLM APIs, paid LLMs, or task specific AI tools
- Carbon footprint/sustainability

# Thank you.

## For more information please contact:

enquiries@ceh.ac.uk
ceh.ac.uk
@uk_ceh

Data:

Paper:

UKCEH

Our planet.
Decoded.