Communicating data quality through open reproducible research

Markus Konkol, Research Software Engineer

Want to follow on your machine?
Get a GitHub (https://github.com/) and ORCID (https://orcid.org/) account!

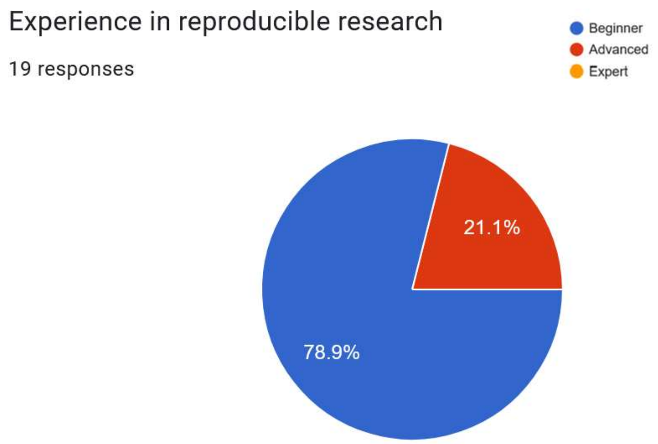
EGU General Assembly 2023

52north

Learning Goals

Upon completion of this short course, you will be able to

- Articulate what Open Reproducible Research is
- Understand which obstacles impede Open Reproducible Research
- Apply Open Reproducible Research principles to your own work
- Choose appropriate tools to publish Open Reproducible Research



Agenda

1. Introduction to Open Reproducible Research
2. Recommendations & best practices
3. Technical obstacles
4. Practical 1: Computational environments
5. MINKE research project
6. Practical 2: Creating and publishing a reproducible workflow

Introduction to Open Reproducible Research

Reproducible Research refers to achieving the **same results** (e.g., tables, figures, numbers) as reported in the paper by using the **same source code and data**.

In **Open Reproducible Research**, these materials are **publicly accessible**.

Replicable Research refers to coming to **similar conclusions** based on the **same analysis**, but **newly collected data**.

Reproducibility & Replicability are both essential for **transparent**, **verifiable**, and **reusable** scientific work.

Further reading: Goodman et al. (2016)

Why is unreproducible research a problem?

- The analysis is not fully transparent and easily understandable.
  - Difficult/impossible to describe analysis in pure text.
  - Access to source code can be a shortcut.
- The analysis is not verifiable.
  - Reviewers need to trust the results.
  - Investigating the analysis in any case a complex task.
- The analysis is not reusable.
  - Waste of time and money (duplication of efforts).
  - Waste of opportunities for collaborations and credit.

Further reading: Gil et al. (2016)

Five ‘selfish’ reasons to do reproducible research

Reason number 1: reproducibility helps to avoid disaster

Reason number 2: reproducibility makes it easier to write papers

Reason number 3: reproducibility helps reviewers see it your way

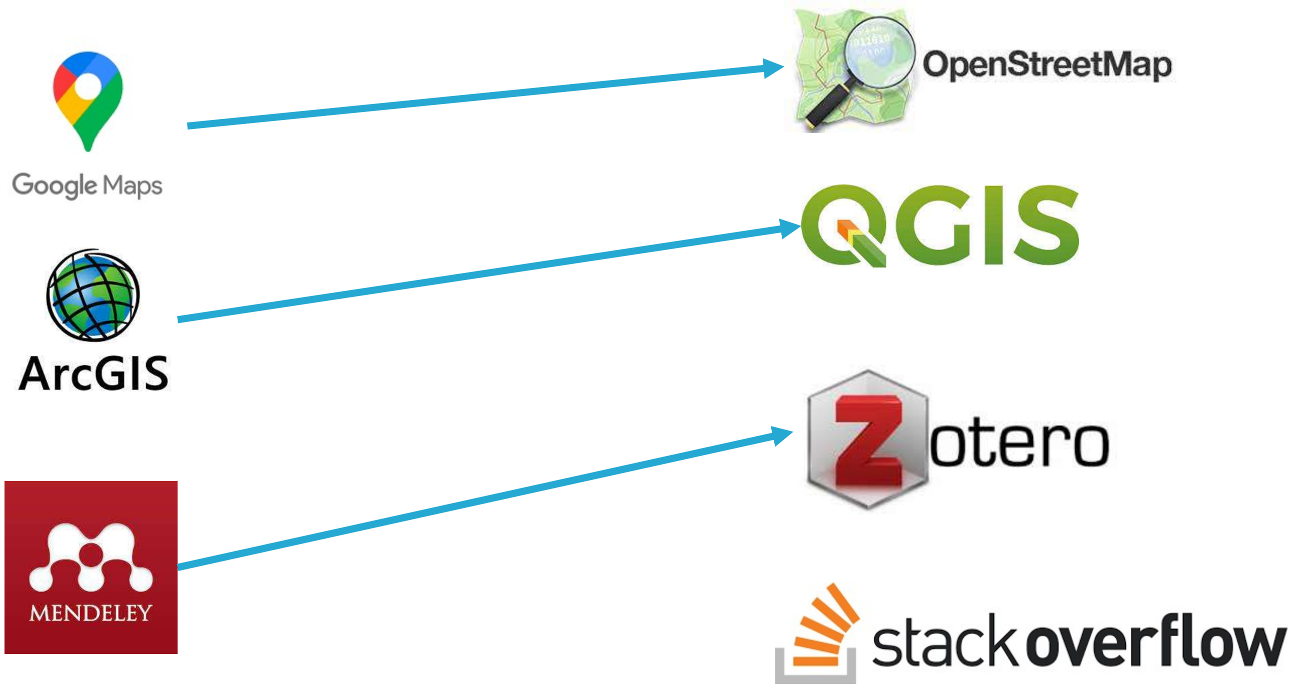
Reason number 4: reproducibility enables continuity of your work

Reason number 5: reproducibility helps to build your reputation

Further reading: Markowetz (2015)

Five recommendations for ORR

Recommendation 1: Use open source software instead of commercial software.



Further reading: Hasselbring et al. (2020)

Five recommendations for ORR

Recommendation 2: Learn a scripting language.

- Scripts describe every step of an analysis
- Human-readable description of what the code does
- Others can understand
  - What has been done
  - How it has been done

- Not reproducible
- No step-by-step description
- No control over the algorithms

Further reading: Alston and Rick (2020), Lasser (2020)

Five recommendations for ORR

Recommendation 3: Learn a computational notebook format.

Further reading: Data science in Python & Jupyter, R Markdown Introduction

Five recommendations for ORR

Recommendation 3: Learn a computational notebook format.

```
# Results

This section summarises the results of the survey.
Each subsection reports on the numbers, free text answers, and briefly discusses the results.
An overall discussion is provided in the next chapter.
A link to the dataset and the R Markdown file underlying the results is available in the supplements.

```{r, echo=FALSE, results="hide", message=FALSE, comment=FALSE, warning=FALSE}
likertPlot = function(dataSet, questions, x, y, scale, ordered){
  plotlevels <- scale
  subset = dataSet[,x:y]

  sapply(subset, class)
  sapply(subset, function(x) { length(levels(x)) } )

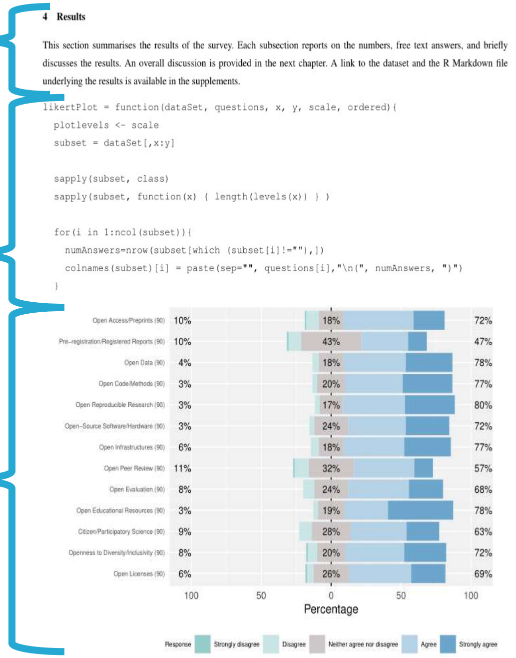
  for(i in 1:ncol(subset)){
    numAnswers=nrow(subset[which (subset[i]!=""),])
    colnames(subset)[i] = paste(sep="", questions[i],"\n(", numAnswers, ")")
  }

  for(i in seq_along(subset)) {
    subset[,i] <- factor(subset[,i], levels=plotlevels)
  }
```

Text

Code

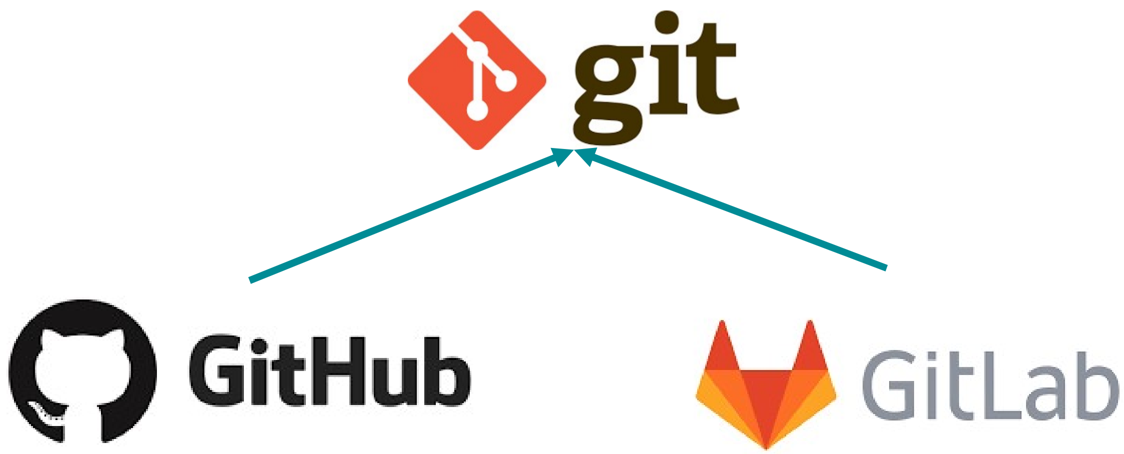
Output



Further reading: [Data science in Python & Jupyter](), [R Markdown Introduction]()
```

Five recommendations for ORR

Recommendation 4: Learn a collaborative software development tool.

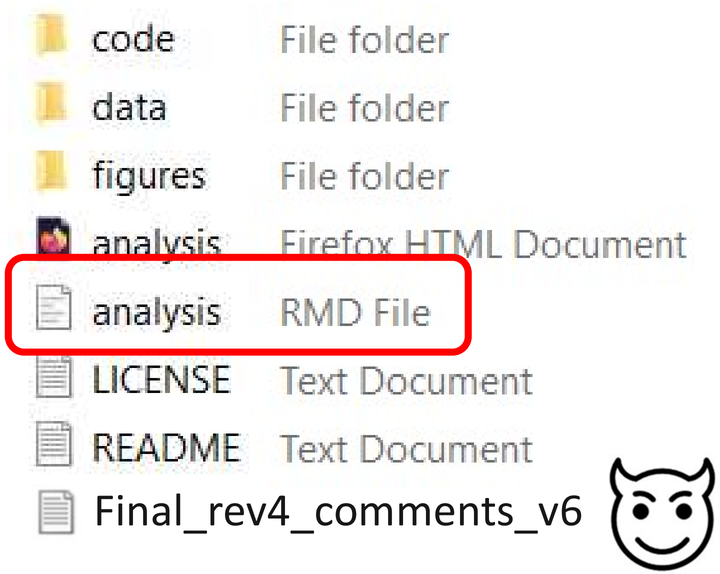


Further reading: Version Control with Git (software carpentry)

Five recommendations for ORR

Recommendation 5: Document your source code.

- Create a clean workspace with a hierarchical folder structure and name files properly.
- Include a README text file to explain the code.
  - What does the software?
  - How can I install it?
  - Are there any computational requirements (e.g., operating system)?
  - How can I use it?
  - How long does the analysis take?
- Add a LICENSE, e.g., MIT License, APACHE License, or GNU.

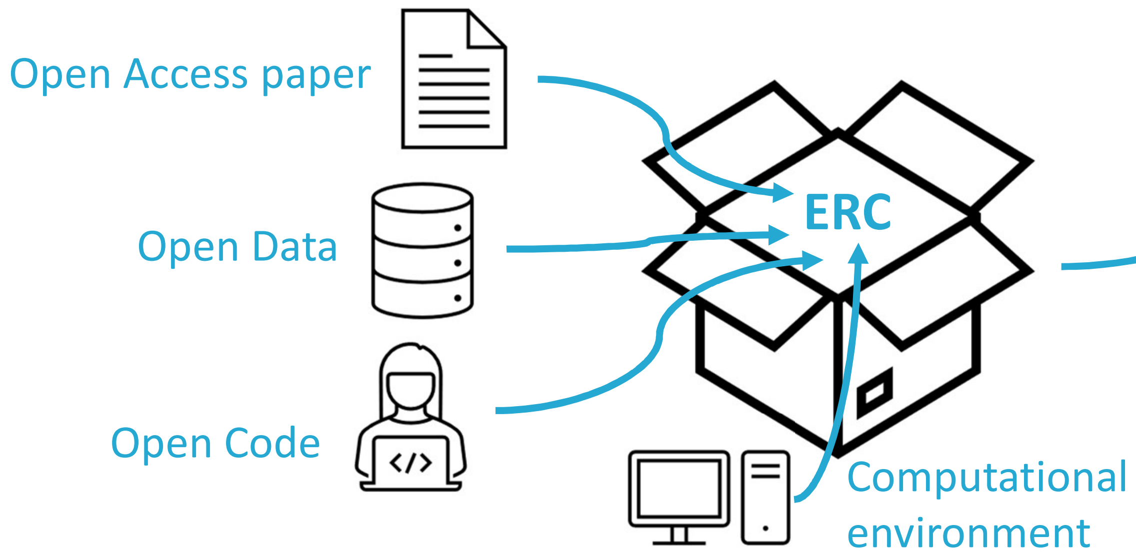


Further reading: Hettne et al. (2020)

Best practices

Share the **scientific paper**, **research data**, **source code**, and details of the **computational environment** that generate published findings in **open trusted repositories**.

- Such a package is also known as *Executable Research Compendium* (ERC).



Further reading: Nüst et al. (2017), Stodden et al. (2016), Sandve et al. (2013)

Best practices

Insert a persistent identifier (e.g., DOI) in the published article that links to the data and source code underlying the results

- Example: *“Research data and source code supporting this publication is available on [name of the repository] and accessible via the following DOI: [doi to repository]”*

If legitimate reasons to restrict access to the materials apply to your work, mention it.

- Example: *“Research data and source code supporting this publication is not available due to [indicate reasons, e.g., licenses, data on human subjects, private or sensitive data etc.]*

Further reading: Nüst et al. (2021), Stodden et al. (2016), Sandve et al. (2013)

Best practices

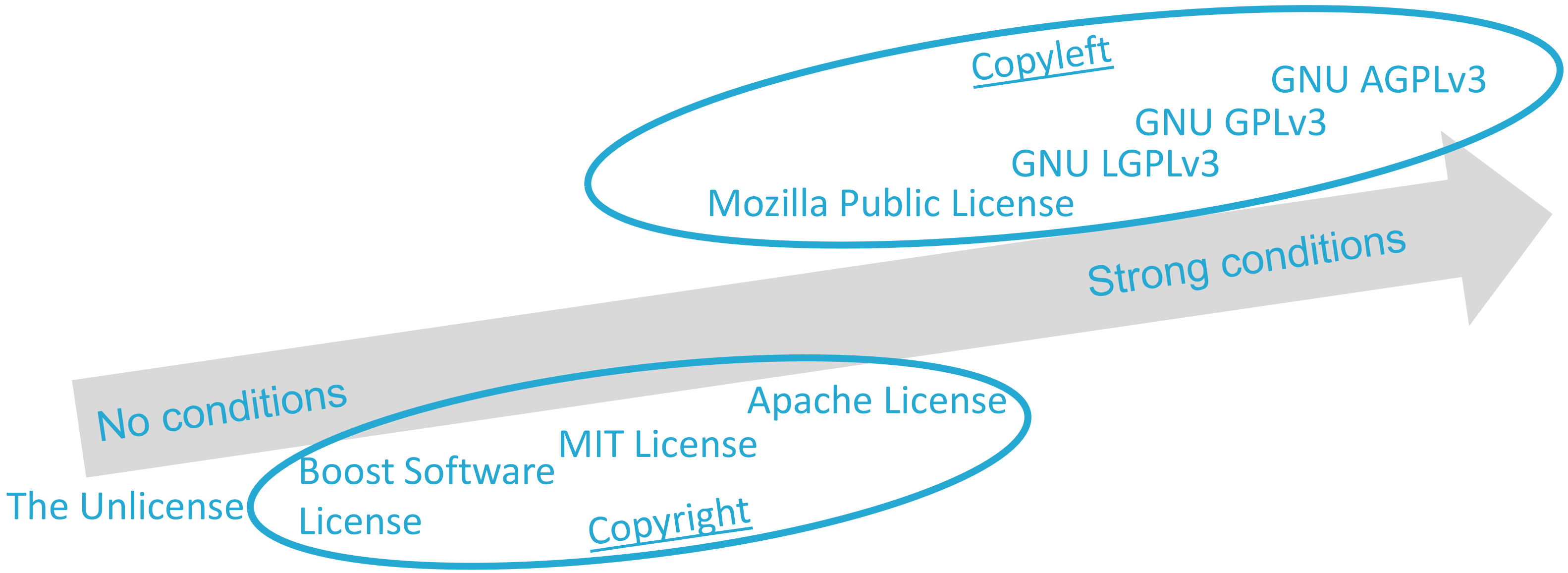
To enable credit for shared materials, citation should be standard practice.

- Example: *Statistics were done using R 3.5.0 (R Core Team, 2018), the rstanarm (v2.13.1; Gabry & Goodrich, 2016) and the psycho (v0.3.4; Makowski, 2018) packages. The full reproducible code is available in Supplementary Materials.*

Further reading: Makowski (2018)

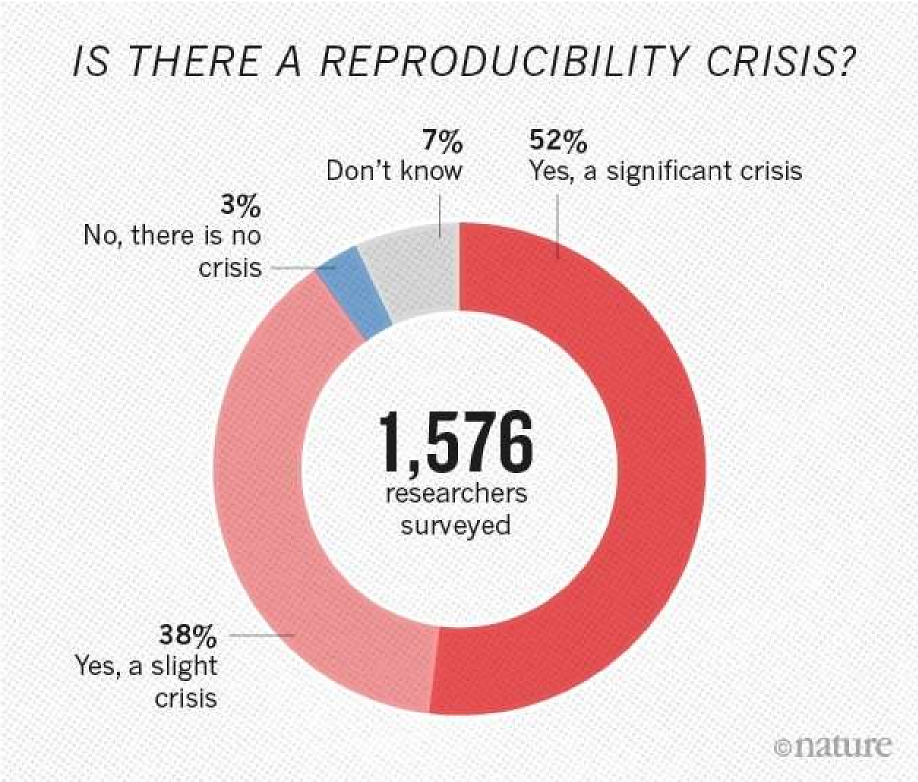
Best practices

Use open licensing when publishing source code.



Further reading: Stodden et al. (2016), Choose an open source license, Open Source Initiative

The Reproducibility Crisis



~~**Reproducible Research** refers to achieving the **same results** (e.g., tables, figures, numbers) as reported in the paper by using the **same source code and data**.~~
~~In **Open Reproducible Research**, these materials are **publicly accessible**.~~

Further reading: Baker (2016)

The Reproducibility Crisis


Research Articles

Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study

Markus Konkol, Christian Kray & Max Pfeiffer

Pages 408-429 | Received 09 Apr 2018, Accepted 30 Jul 2018, Published online: 13 Aug 2018

Download citation https://doi.org/10.1080/13658816.2018.1508687 Check for updates

Full Article | Figures & data | References | Supplemental | Citations | Metrics | Licensing | Reprints & Permissions

EPUB

ABSTRACT

Formulae display: MathJax

Reproducibility is a cornerstone of science and thus for geographic research as well. However, studies in other disciplines such as biology have shown that published work is rarely reproducible. To assess the state of reproducibility, specifically computational reproducibility (i.e. rerunning the analysis of a paper using the original code), in geographic research, we asked geoscientists about this topic using three methods: a survey (n = 146), interviews (n = 9), and a focus group (n = 5). We asked participants about their understanding of open reproducible research (*ORR*), how much it is practiced, and what obstacles hinder *ORR*. We found that participants had different understandings of *ORR* and that there are several obstacles for authors and readers (e.g. effort, lack of openness). Then, in order to complement the subjective feedback from the participants, we tried to reproduce the results of papers that use spatial statistics to address problems in the geosciences. We selected 41 open access papers from *Copernicus* and *Journal of*

Related rese

People also read

Practical Reprod Geosciences

Daniel Nüst et al.

Annals of the Amer

Published online: 13

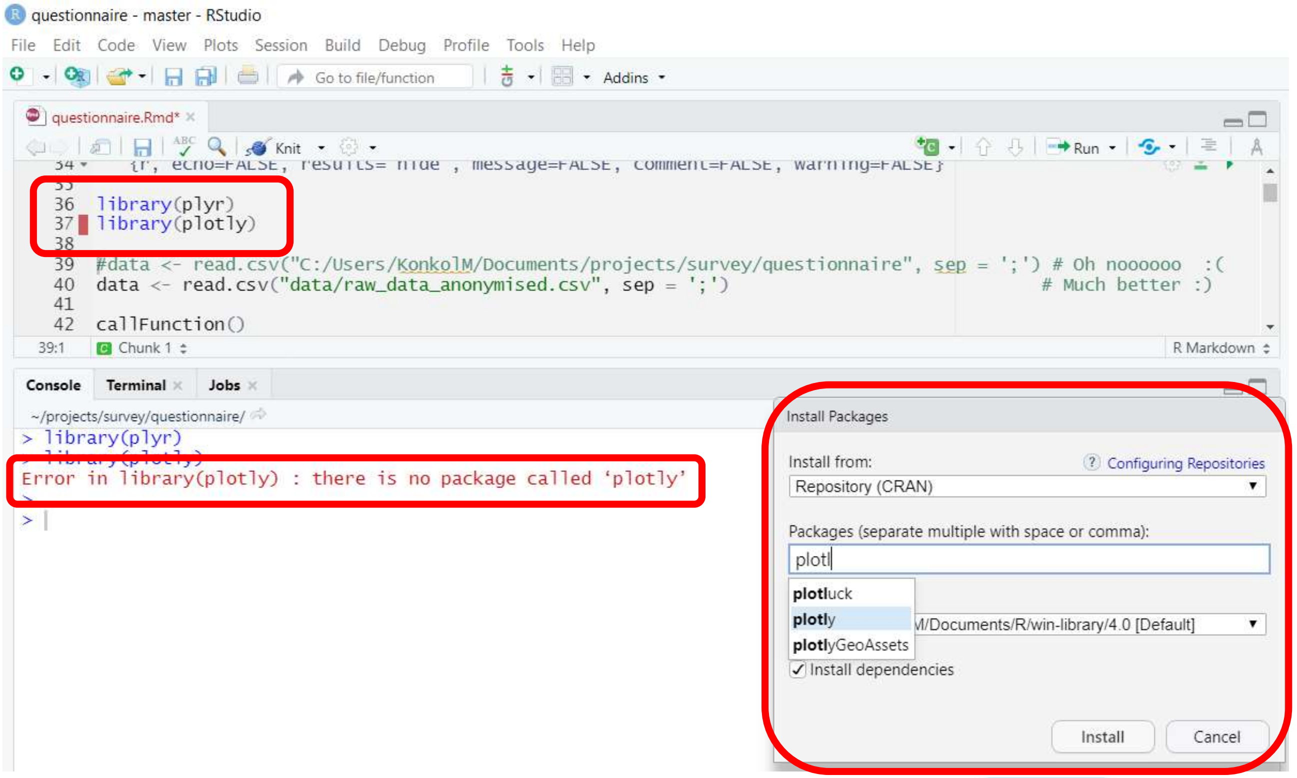
Reproducibility and challenges f

- Checked 41 papers that had code and data attached for executability and reproducibility.
- 2 out of 41 papers were executable + reproducible.
  - Several technical issues.
  - Several content-related differences.
  - More on that later…

Further reading: Konkol et al. (2018)

Technical obstacles impeding ORR

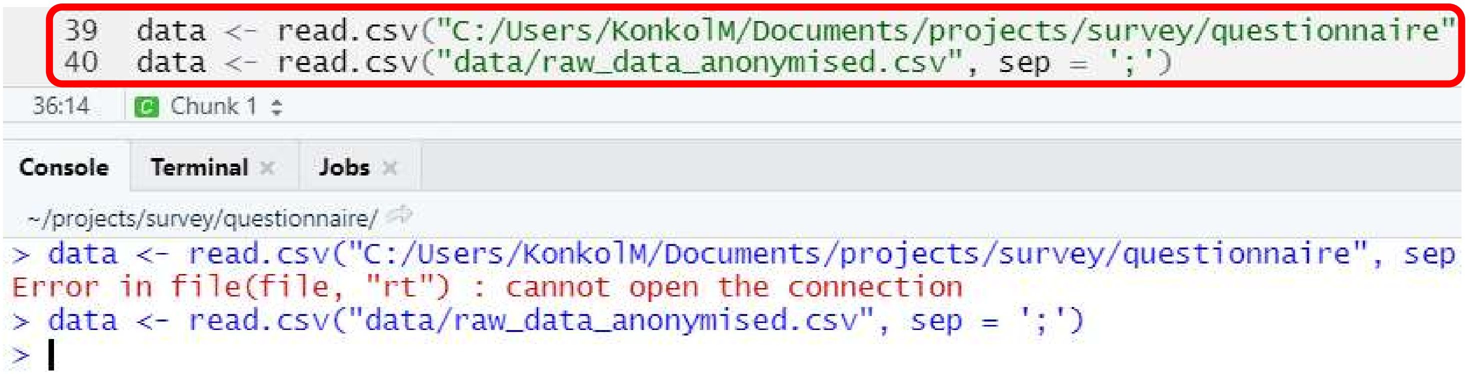
- Three categories of technical issues.
- *Minor issues* rather easy to solve.
- Example error: Library not found but available in repository.



Further reading: [Konkol et al. (2018)]

Technical obstacles impeding ORR

- *Substantial* issues require more effort.
- Example error: Wrong file directory.
- Solution: Use relative instead of absolute file paths.



Further reading: Konkol et al. (2018)

Technical obstacles impeding ORR

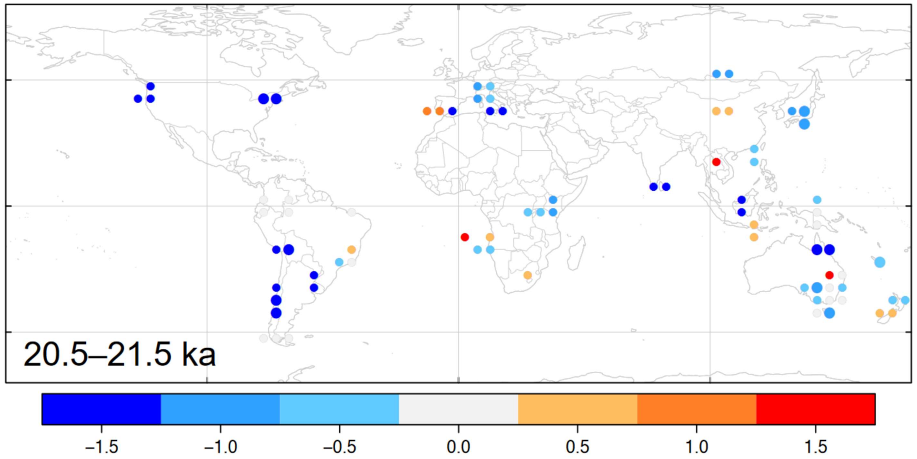
- *Severe* issues require time, knowledge about the programming language, and understanding of the source code.
- Example error: *cannot open file dataABC.csv. No such file or directory.*
  - Was the file available in the folder? No ☹
  - Was the file created by the source code? No ☹
  - Contact author → get **missing** source code snippet that produced dataABC.csv.
- Solution: Use tools like *Binder* or *The Whole Tale*.

More on that later!

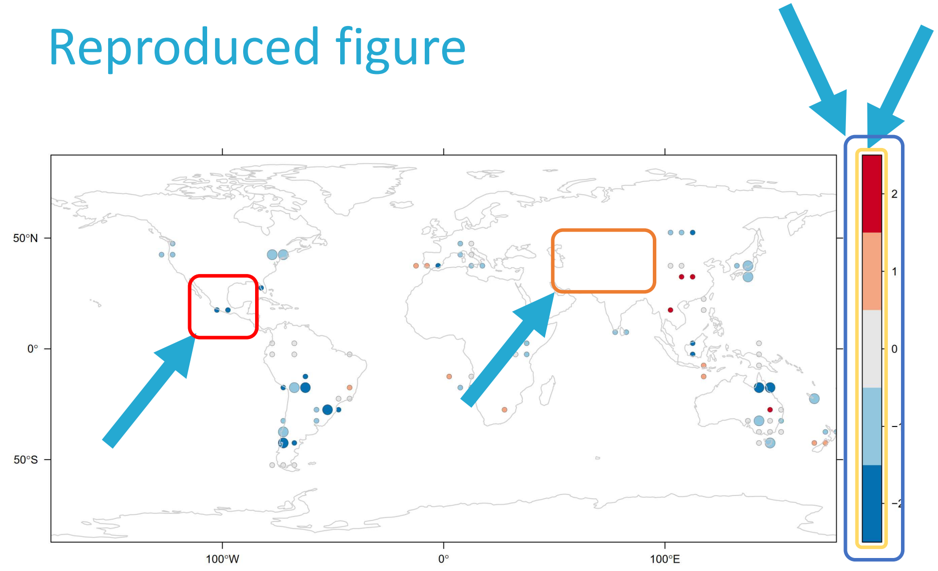
Further reading: Konkol et al. (2018)

Technical obstacles impeding ORR

Original figure



Reproduced figure



Further reading: [Konkol et al. (2018)]()

52north

Technical obstacles impeding ORR

- *System-dependent* issues are issues related to the computational environment...

Further reading: Konkol et al. (2018)

Practical 1

Understanding the importance of computational environments https://github.com/MarkusKonk/egu_shortcourse, see branches map1, map2, and calc

MINKE

Metrology for Integrated marine maNagement and Knowledge-transfer nEtwork

INFRAIA-02-2020: Integrating Activities for Starting Communities

Markus Konkol, Simon Jirka, Christian Autermann - 52°North Spatial Information Research GmbH
Joaquin Del Rio Fernandez, Enoc Martínez - Universitat Politècnica de Catalunya

EUROPEAN COMMISSION
Project funded by the European Commission within the Horizon 2020 Programme (2014-2020)
Grant Agreement No. 101008724

MINKE

CONSORTIUM

PARTNERS

 # 10 countries

 # 22 organisations

The project

PROGRAMME: **H2020-EU.1.4.1.2. - Integrating and opening existing national and regional research infrastructures of European interest**

CALL: INFRAIA-02-2020-1 . **Topic:** *Integrating Activities for Starting Communities*

Integrating Activities shall combine, in a closely co-ordinated manner 3 types of activities:

- **Networking Activities (NA)**, to foster a culture of co-operation between research infrastructures, scientific communities, industries and other stakeholders as appropriate, and to help develop a more efficient and attractive European Research Area;
- **Trans-national Access (TNA) or Virtual Access (VA) Activities**, to support scientific communities in their access to the identified key research infrastructures;
- **Joint Research Activities (JRA),** to improve, in quality and/or quantity, the integrated services provided at European level by the infrastructures.

The main goals

MINKE will integrate key European **Marine Metrology Research Infrastructures**, to coordinate their use and development and propose an innovative framework of ***quality of oceanographic data***

What to measure ?

Identifying the **Essential Ocean Variables** (EOVs) as the key parameters to monitor

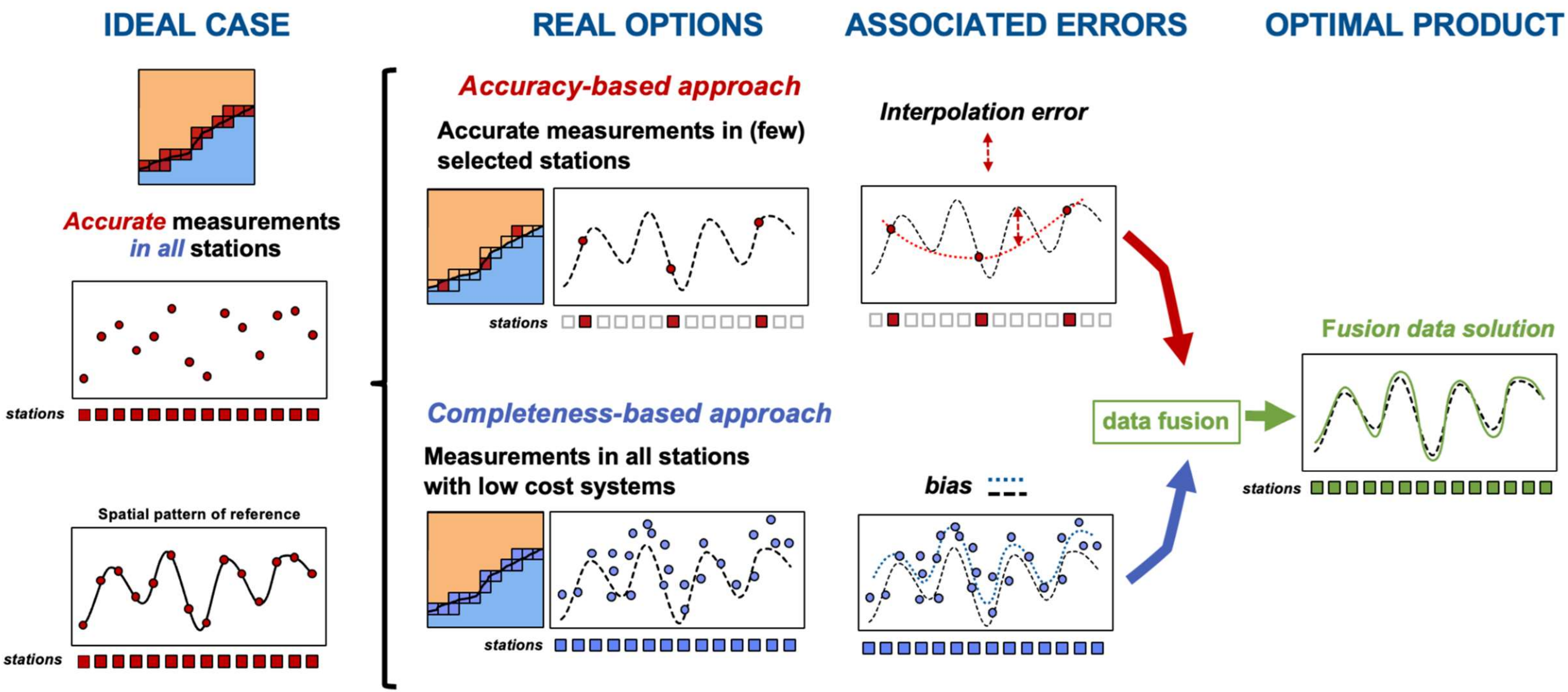
How to measure them ?

Adopting a multidimensional framework of data quality:

- **Accuracy:** Minimising the **measurement errors**
- **Completeness:** Minimising the **interpolation errors**
- **Timeliness:** Providing the observations as fast as requiered

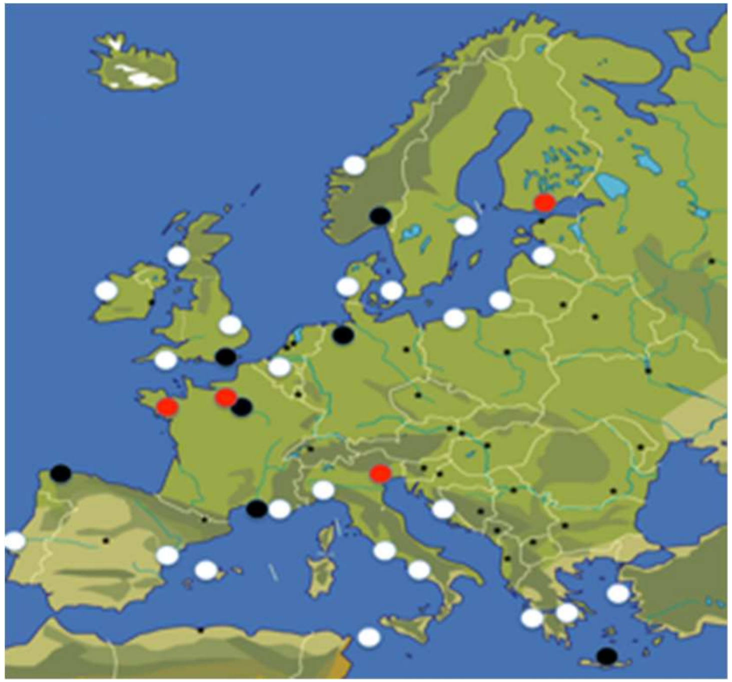
Purpose: *To retrieve (at least) the large scale features, both temporal and spatial, of the EOVs*

Data quality approach

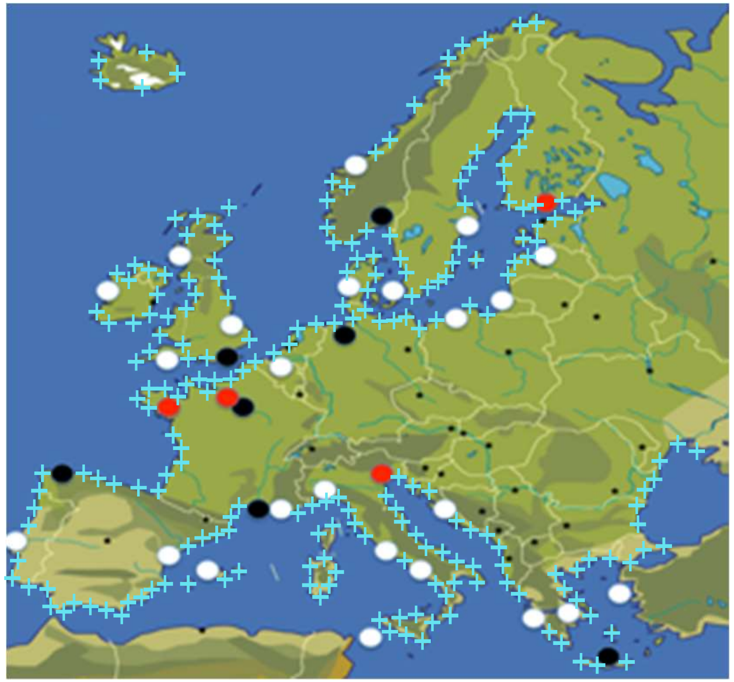


Vision

Accuracy



Accuracy + Completeness

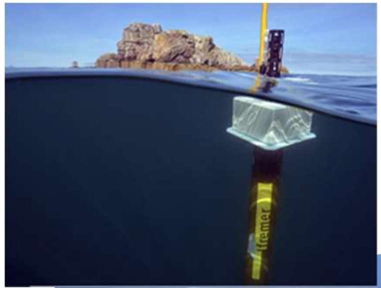


- Primary reference nodes
- Secondary reference nodes
- Scientific users Operators
- Participatory nodes

MINKE Research Infrastructures

MINKE

Accuracy
Advanced instrumentation & Calibration centres





Completeness
Citizen observatories & Fablabs





Test reports

ISTITUTO NAZIONALE
DI OCEANOGRAFIA E DI GEOFISICA SPERIMENTALE – OGS
SEZIONE OCE - CTMO

MINKE

Test Report
Temperature & Conductivity

SBE 37 SMP MicroCAT
Serial Number: 3287

Table. Results of the "as-received" test for temperature and conductivity following the cleaning operation described on page 2 of this report.

T Ref. (°C)	T Inst. (°C)	C Ref. (S/m)	C In (S/
25.0466	25.0451	5.63769	5.63
20.1492	20.1480	5.10228	5.09

Old temperature calibration coefficients[1]:

a0 = -4.078553e-05
a1 = 2.878170e-04
a2 = -3.197355e-06
a3 = 1.795368e-07

ITS-90 Temperature = 1/ {[a0 + a1 [*ln* (n)] +

T Ref. (°C)	Inst Output (n)	T
2.0351	577392.7	2
5.2141	502147.0	5

New temperature calibration coefficients:

a0 = 7.2754375e-06
a1 = 2.7631671e-04
a2 = -2.2808406e-06
a3 = 1.5520126e-07

ITS-90 Temperature = 1/ {[a0 + a1 [*ln* (n)] + a2 [*ln* ^{2} (n)] + a3 [*ln* ^{3} (n)]} – 273.15 (°C)

T Ref. (°C)	Inst Output (n)	T Inst. (°C)	T Inst. - T Ref.* (°C)
2.0351	577392.7	2.0352	0.0001
5.2141	502147.0	5.2140	-0.0001

Quality Flags

Flag	Description
Pass=1	Data have passed critical real-time quality control tests and are deemed adequate for use as preliminary data.
Not evaluated=2	Data have not been QC-tested, or the information on quality is not available.
Suspect or Of High Interest=3	Data are considered to be either suspect or of high interest to data providers and users. They are flagged suspect to draw further attention to them by operators.
Fail=4	Data are considered to have failed one or more critical real-time QC checks. If they are disseminated at all, it should be readily apparent that they are not of acceptable quality.
Missing data=9	Data are missing; used as a placeholder.

Figure 6 - QARTOD / UNESCO IOC 54:V3 flagging scheme (source: U.S. Integrated Ocean Observing System, 2020a)

Metadata on quality

Sensor

- ***<u>Accuracy</u>***: +/- 0.002 °C
- Precision: +/- 0.002 °C
- DetectionLimit: -5 to 45°C
- BatteryCharge: 30%
- MeasurementRate: 1/s
- Coordinates: 52.1234, 7.456
- Placement: <text>
- ***<u>QualityLevel</u>***: checked
- TestReports: [TestReport]
- SensorUncertainty: QualityFlag(?)

Test report

- AsReceived: <text>
- Condition: damaged
- Photographs: [Photos]
- Activities: repaired
- Workflow: <text>
- TestType: NewCalibration
- Procedure: <text>
- Date: Date
- AmbientConditions: °C, %, etc.
- MeasuredValues: [values]
- ReferenceValues: [values]
- Deviations: Measured – Reference
- MeanDeviation: 0.0002
- Satisfactory: pass

Observation

- TimeStamp: Date
- Measurement: 20°
- Validity: inconsistent
- ***<u>DataProcessing</u>***: Adjusted
- Provenance: Code
- ObservationUncertainty: SensorUncertainty + Validity + Processing = QualityFlag(?)

Practical 2

Creating and publishing a reproducible workflow
https://github.com/MarkusKonk/egu_shortcourse