

Exploratory analysis of data and code availability statements in Copernicus journals

Code ¹

AUTHOR

Matthias Schlögl  

AFFILIATIONS

[BOKU University](#)

[GeoSphere Austria](#)

Introduction

Reproducibility is a cornerstone of scientific research, yet its implementation across disciplines remains inconsistent. To gain a preliminary understanding of reproducibility practices within the geosciences community, we reached out to two major publishers: the American Geophysical Union (AGU) and the European Geosciences Union (EGU, through Copernicus Publications). Specifically, we sought information on their policies and practices regarding data and code availability.

AGU

The AGU directed attention to their data policy guidelines as part of their publication requirements, referencing the information that is publicly accessible on their website.¹:

*AGU **requires** that the underlying data and/or software or code needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication.*

EGU

The EGU, through Copernicus Publications, publishes a wide range of peer-reviewed open-access journals that cover diverse topics in Earth, planetary, and space sciences².

Copernicus Publications offers metadata and full-text XML files for all published articles³. These files are formatted using the NLM (PMC) standard⁴ and are accessible via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)⁵. This approach ensures that research outputs are easily discoverable and reusable by the scientific community.

Setup

R packages required for the analysis

- [tidyverse](#)
- [ggrepel](#)
- [here](#)
- [english](#)

► Code

Custom functions

► Code

Data

► Code

Overview of publications

We scraped and analyzed the full-text XML-files for 25 Copernicus journals in the field of geosciences for the period 2016-01-01–2024-12-31, resulting in a total number of 32129 articles ([Figure 1](#)). However, due to inconsistencies in the results for 2016, which were likely caused by changes in the XML structure, data from that year were excluded from the analysis of statements.

► Code

doi	journal	volume	pages	year	
<chr>	<fct>	<int>	<int>	<int>	►
10.5194/acp-16-1-2016	acp	16	1	2016	
10.5194/acp-16-21-2016	acp	16	21	2016	
10.5194/acp-16-35-2016	acp	16	35	2016	
10.5194/acp-16-47-2016	acp	16	47	2016	
10.5194/acp-16-71-2016	acp	16	71	2016	
10.5194/acp-16-85-2016	acp	16	85	2016	
10.5194/acp-16-101-2016	acp	16	101	2016	
10.5194/acp-16-123-2016	acp	16	123	2016	
10.5194/acp-16-135-2016	acp	16	135	2016	
10.5194/acp-16-145-2016	acp	16	145	2016	
1-10 of 10,000 rows 1-5 of 6 columns		Previous	1	2	3
				4	5
				6	... 1000 Next

► Code

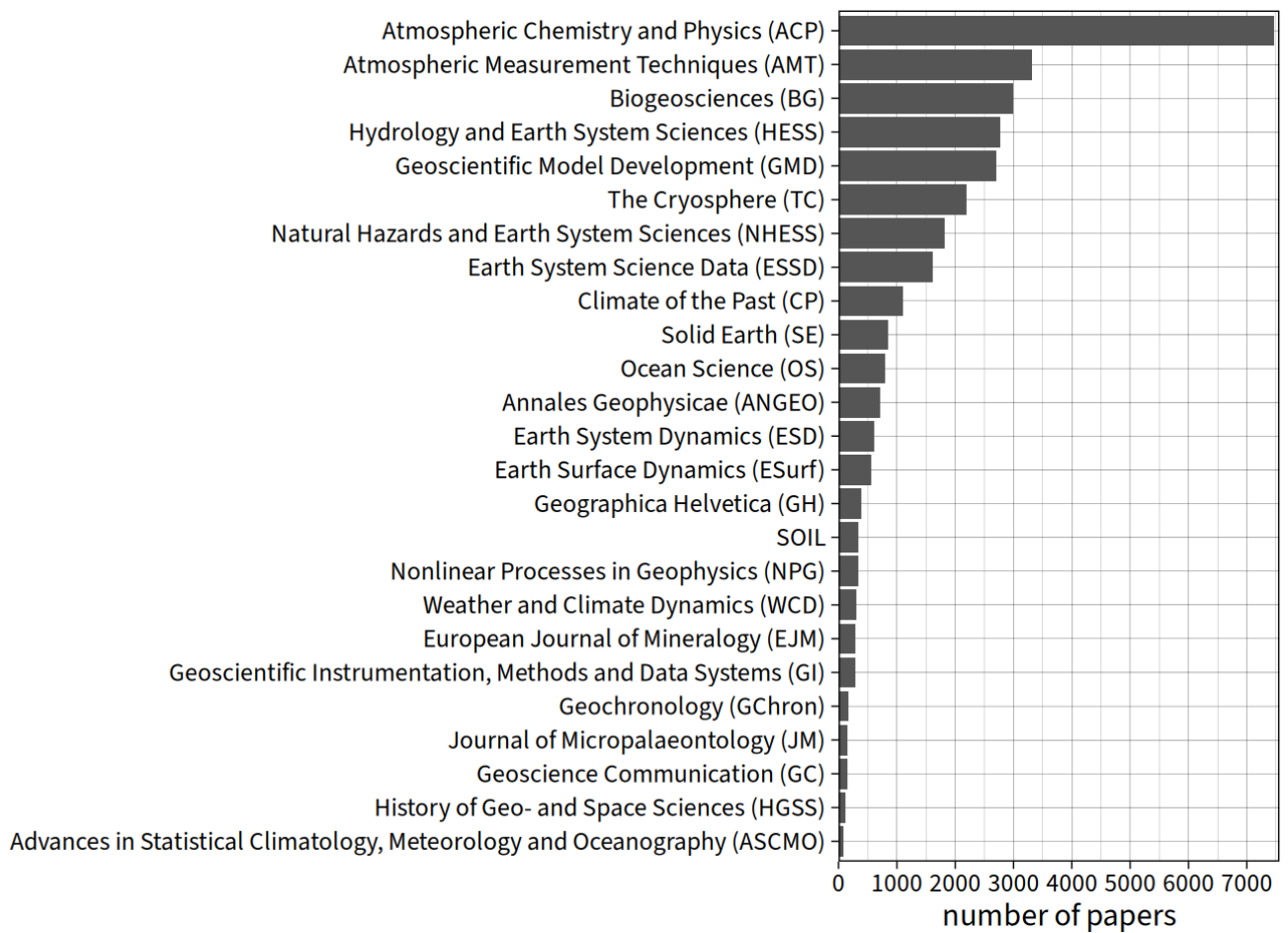


Figure 1: Overview of geoscience journals (published by Copernicus Publishing) and the corresponding number of articles published during the time period between 2016-01-01 and 2024-12-31.

Overview of statements

The section names of paragraphs addressing code and data availability are generally consistent across articles, though a few exceptions with varying section titles were observed. Below is a list of all simplified section name variations that appeared more than once:

► Code

title	n
<chr>	<int>
dataavailability	20971
codedataavailability	5893
codeanddataavailability	181
dataavailabilityanduserguidelines	3
dataformatsandavailability	3
dataandcodeavailability	2
dataavailabilityandfilestructure	2
dataavailabilityandstructure	2
dataavailabilityanduse	2
datausageandavailability	2
1-10 of 10 rows	

► Code

title	n
<chr>	<int>
codedataavailability	5893
codeavailability	4059
codeanddataavailability	181
dataandcodeavailability	2
4 rows	

► Code

The inclusion of data availability statements remains consistently high, with a median of 97% of all papers across journals and years containing such statements. Notably, only 3 journals reported a share below 75% in more than one year ([Figure 2](#)).

► Code

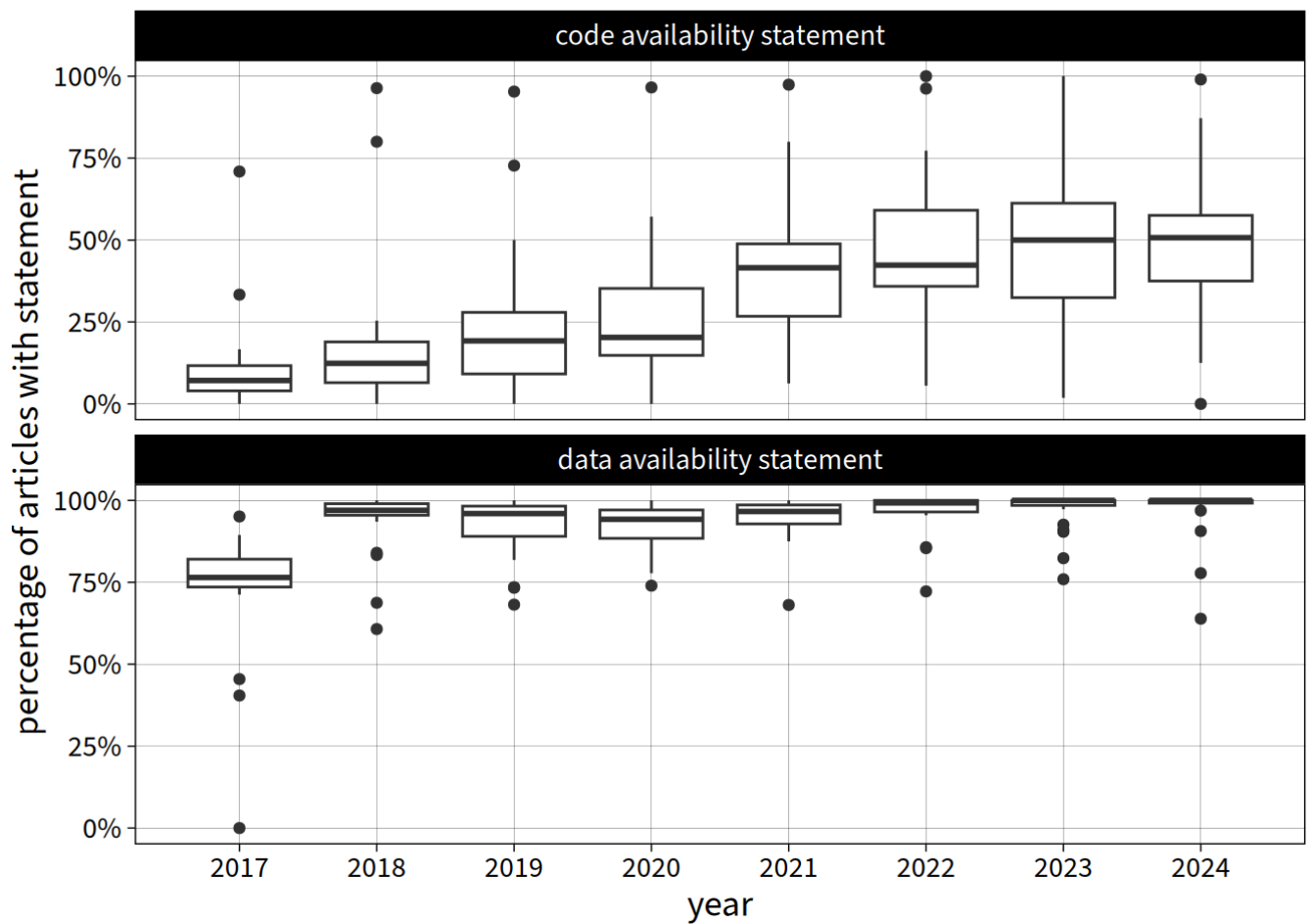


Figure 2: Share of articles published in 25 geoscience journals between 2017-01-01 and 2024-12-31.

The proportion of articles including a code availability statement is substantially lower. Notably, only *Geoscientific Model Development*, being a dedicated outlet for the publication and discussion of numerical models and geoscientific software, consistently includes such statements in nearly all published articles.

Content of statements

While the presence of data and code availability statements is a positive indicator, it does not necessarily reflect practical reproducibility. To gain deeper insights into the actual content of these statements, we conducted a targeted search for specific keywords and phrases within them.

► Code

It is important to note that this exploratory analysis likely includes double-counts, as approximately 6080 statements combine both data and code availability information.

Data availability

► Code

Data is most commonly shared through the generalist repository [zenodo](#), followed by [PANGAEA](#), a specialized repository for Earth and environmental sciences ([Figure 3](#)). Approximately 48% of all statements include a URL⁶, about 18% reference a citation likely found in the reference list⁷, and about 10% contain the string `doi` (case insensitive). Approximately 4 percent include the literal string `from the author`, suggesting that some data may only be available upon request from the authors.

► Code

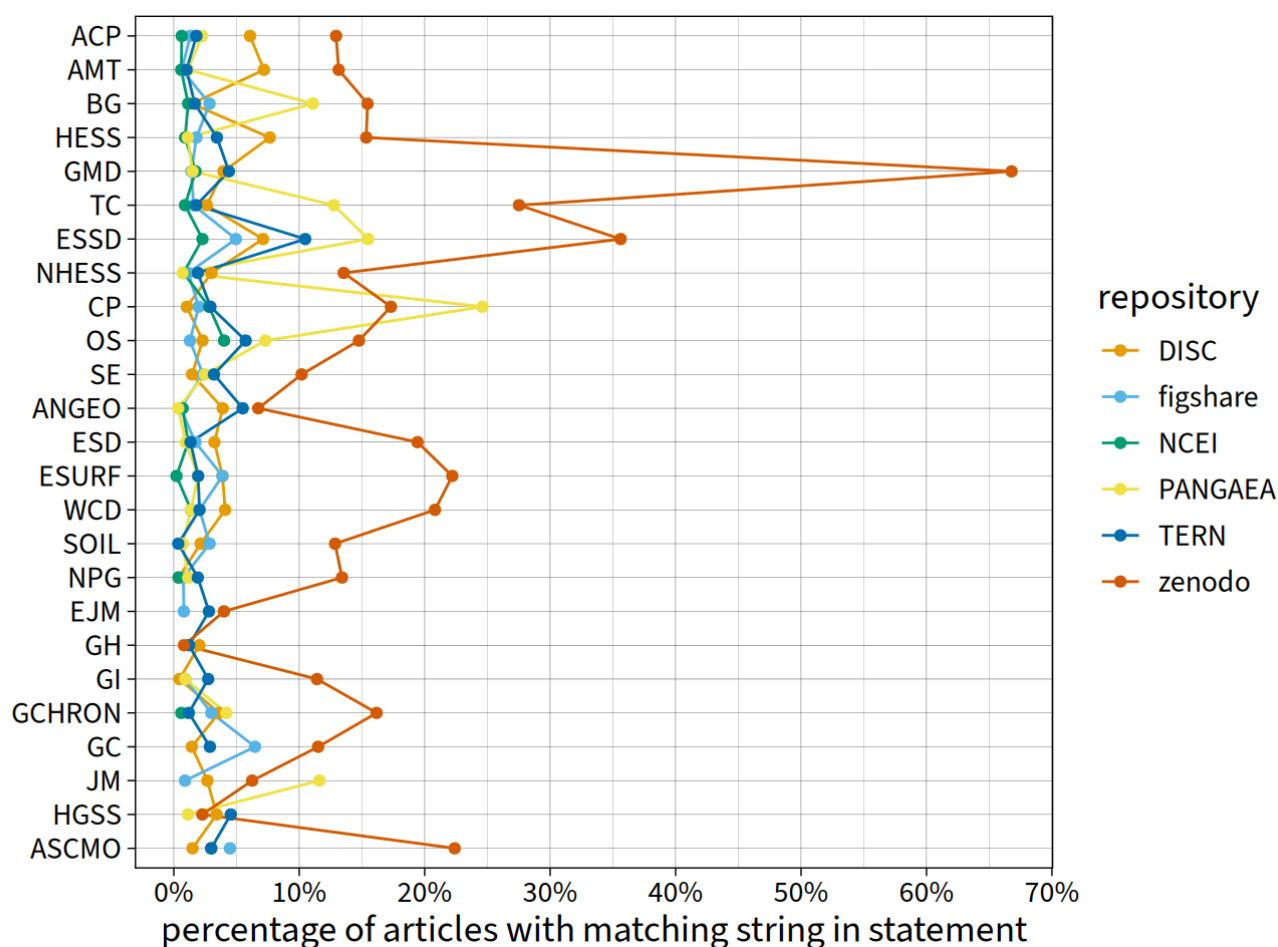


Figure 3: Occurrence of selected strings in data availability statements. Journals are ordered according to the absolute number in descending order.

Code availability

Code is primarily shared via [zenodo](#) and [GitHub](#) (Figure 4). However, due to the relatively lower number of code availability statements, the potential impact of double-counting zenodo entries may be more pronounced in this context.

► Code

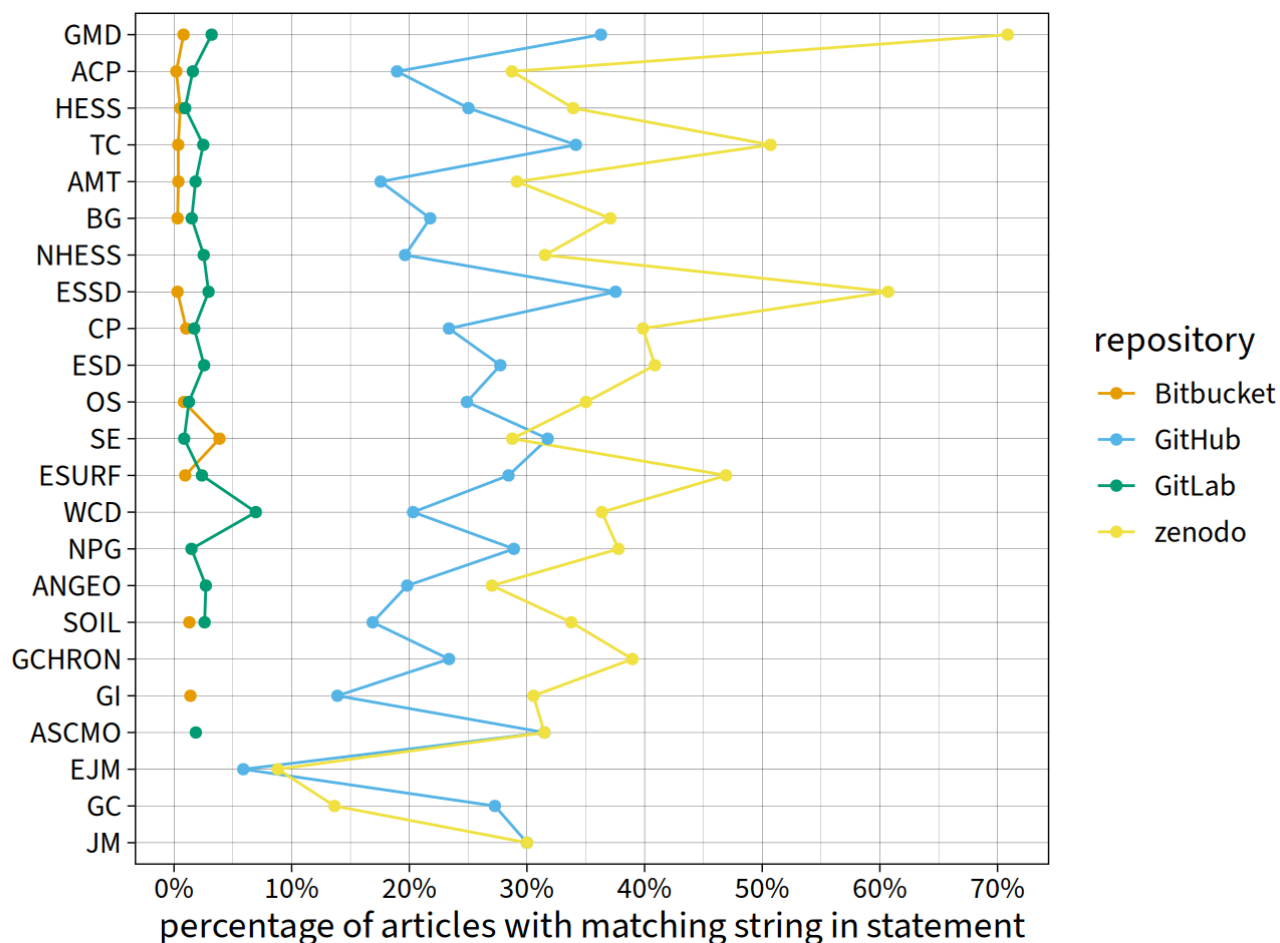


Figure 4: Occurrence of selected strings in code availability statements. Journals are ordered according to the absolute number in descending order.

A joint visualization of the share of data availability statements against the share of code availability statements for the selected journals provides a visual representation of how these two aspects correlate across journals ([Figure 5](#)).

► Code

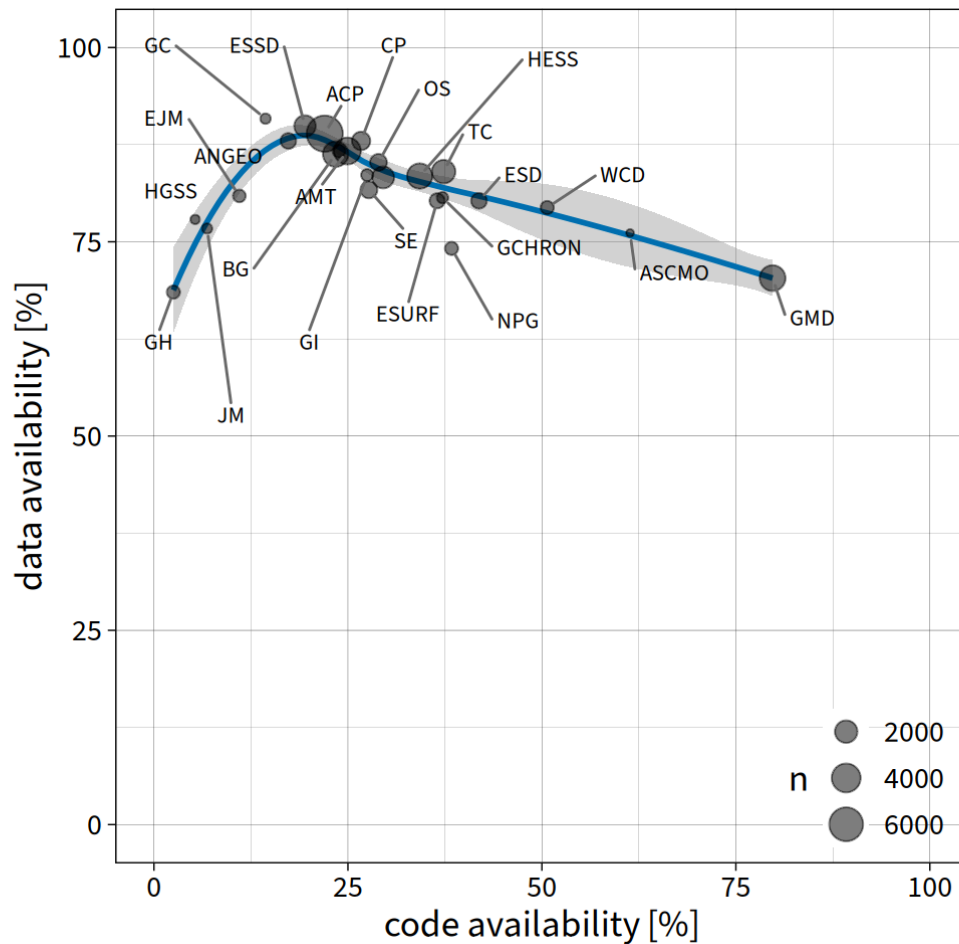


Figure 5: Share of data and code availability statements across journals. The point size represents the total number of articles published per journal. The smoothed line is a LOESS curve weighted by the number of observations for each journal.

Acknowledgements

Thanks to Dr. Johannes Wagner from Copernicus Publications for his valuable assistance in accessing the code and data availability statements of Copernicus journals.

R session info

► Code

– Session info

setting	value
version	R version 4.5.1 (2025-06-13)
os	Ubuntu 24.04.3 LTS
system	x86_64, linux-gnu
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Etc/UTC
date	2025-10-27
pandoc	3.8.1 @ /usr/bin/ (via rmarkdown)
quarto	1.7.32 @ /usr/local/bin/quarto

– Packages

! package	*	version	date (UTC)	lib	source
P bit		4.6.0	2025-03-06	[?]	RSPM (R 4.5.0)
P bit64		4.6.0-1	2025-01-16	[?]	RSPM (R 4.5.0)
P cli		3.6.5	2025-04-23	[?]	RSPM (R 4.5.0)
P crayon		1.5.3	2024-06-20	[?]	RSPM (R 4.5.0)
P digest		0.6.37	2024-08-19	[?]	RSPM (R 4.5.0)
P dplyr	*	1.1.4	2023-11-17	[?]	RSPM (R 4.5.0)
P english	*	1.2-6	2021-08-21	[?]	RSPM (R 4.5.0)
P evaluate		1.0.4	2025-06-18	[?]	RSPM (R 4.5.0)
P farver		2.1.2	2024-05-13	[?]	RSPM (R 4.5.0)
P fastmap		1.2.0	2024-05-15	[?]	RSPM (R 4.5.0)
P forcats	*	1.0.0	2023-01-29	[?]	RSPM (R 4.5.0)
P generics		0.1.4	2025-05-09	[?]	RSPM (R 4.5.0)
P ggkabeito		0.1.0	2021-10-18	[?]	RSPM (R 4.5.0)
P ggplot2	*	3.5.2	2025-04-09	[?]	RSPM (R 4.5.0)
P ggrepel	*	0.9.6	2024-09-07	[?]	RSPM (R 4.5.0)
P glue		1.8.0	2024-09-30	[?]	RSPM (R 4.5.0)
P gtable		0.3.6	2024-10-25	[?]	RSPM (R 4.5.0)
P here	*	1.0.1	2020-12-13	[?]	RSPM (R 4.5.0)
P hms		1.1.3	2023-03-21	[?]	RSPM (R 4.5.0)
P htmltools		0.5.8.1	2024-04-04	[?]	RSPM (R 4.5.0)
P jsonlite		2.0.0	2025-03-27	[?]	RSPM (R 4.5.0)
P knitr		1.50	2025-03-16	[?]	RSPM (R 4.5.0)
P labeling		0.4.3	2023-08-29	[?]	RSPM (R 4.5.0)
P lattice		0.22-7	2025-04-02	[?]	CRAN (R 4.5.1)
P lifecycle		1.0.4	2023-11-07	[?]	RSPM (R 4.5.0)
P lubridate	*	1.9.4	2024-12-08	[?]	RSPM (R 4.5.0)
P magrittr		2.0.3	2022-03-30	[?]	RSPM (R 4.5.0)
P Matrix		1.7-3	2025-03-11	[?]	CRAN (R 4.5.1)
P mgcv		1.9-3	2025-04-04	[?]	CRAN (R 4.5.1)
P nlme		3.1-168	2025-03-31	[?]	CRAN (R 4.5.1)
P pillar		1.11.0	2025-07-04	[?]	RSPM (R 4.5.0)

P pkgconfig	2.0.3	2019-09-22	[?]	RSPM (R 4.5.0)
P purrr	* 1.1.0	2025-07-10	[?]	RSPM (R 4.5.0)
P R6	2.6.1	2025-02-15	[?]	RSPM (R 4.5.0)
P RColorBrewer	1.1-3	2022-04-03	[?]	RSPM (R 4.5.0)
P Rcpp	1.1.0	2025-07-02	[?]	RSPM (R 4.5.0)
P readr	* 2.1.5	2024-01-10	[?]	RSPM (R 4.5.0)
renv	1.1.5	2025-07-24	[1]	RSPM (R 4.5.1)
P rlang	1.1.6	2025-04-11	[?]	RSPM (R 4.5.0)
P rmarkdown	2.29	2024-11-04	[?]	RSPM (R 4.5.0)
P rprojroot	2.1.0	2025-07-12	[?]	RSPM (R 4.5.0)
P scales	1.4.0	2025-04-24	[?]	RSPM (R 4.5.0)
P sessioninfo	1.2.3	2025-02-05	[?]	RSPM (R 4.5.0)
P stringi	1.8.7	2025-03-27	[?]	RSPM (R 4.5.0)
P stringr	* 1.5.1	2023-11-14	[?]	RSPM (R 4.5.0)
P tibble	* 3.3.0	2025-06-08	[?]	RSPM (R 4.5.0)
P tidyr	* 1.3.1	2024-01-24	[?]	RSPM (R 4.5.0)
P tidyselect	1.2.1	2024-03-11	[?]	RSPM (R 4.5.0)
P tidyverse	* 2.0.0	2023-02-22	[?]	RSPM (R 4.5.0)
P timechange	0.3.0	2024-01-18	[?]	RSPM (R 4.5.0)
P tzdb	0.5.0	2025-03-15	[?]	RSPM (R 4.5.0)
P vctrs	0.6.5	2023-12-01	[?]	RSPM (R 4.5.0)
P vroom	1.6.5	2023-12-05	[?]	RSPM (R 4.5.1)
P withr	3.0.2	2024-10-28	[?]	RSPM (R 4.5.0)
P xfun	0.53	2025-08-19	[?]	RSPM (R 4.5.0)
P yaml	2.3.10	2024-07-26	[?]	RSPM (R 4.5.0)

[1] /builds/Rexthor/reproducibility-in-geosciences/renv/library/linux-ubuntu-noble/R-4.5/x86_64-pc-linux-gnu

[2] /root/.cache/R/renv/sandbox/linux-ubuntu-noble/R-4.5/x86_64-pc-linux-gnu/25ebdc09

* — Packages attached to the search path.






P — Loaded and on-disk path mismatch.

System info

► Code

```
root@runner--azerasqr-project-62759060-concurrent-0
OS: Ubuntu 24.04.3 LTS x86_64
Kernel: Linux 5.15.154+
Packages: 641 (dpkg)
Shell: R
CPU: AMD EPYC 7B13 (2) @ 2.45 GHz
Memory: 1.05 GiB / 7.77 GiB (14%)
Disk (/): 11.03 GiB / 25.36 GiB (44%) - overlay
Disk (/builds): 11.03 GiB / 25.36 GiB (44%) - ext4
Locale: en_US.UTF-8
```

Footnotes

1. <https://www.agu.org/publish-with-agu/publish/author-resources/data-and-software-for-authors> 
2. <https://www.egu.eu/publications/> 
3. https://publications.copernicus.org/services/xml_harvesting_and_oai-pmh.html 
4. <https://pmc.ncbi.nlm.nih.gov/pub/filespec/> 
5. <https://www.openarchives.org/pmh/> 
6. i.e., match the regular expression
`https?:\\/[\\^\\s/$.??#].[^\\s]*|www\\. [^\\s/$.??#].[^\\s]*.` 
7. i.e., match the regular expression
`\\b[A-Z][a-z]+(?: et al\\.)? \\(\\d{4}[a-z]?\\)|\\([A-Z][a-z]+(?: et al\\.)?, \\d{4}[a-z]?\\)` 