JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 32 (2026), pp. 67–93 doi:10.5311/JOSIS.2026.32.591

RESEARCH ARTICLE

Improving reproducibility of GIScience publications through novel reproducibility guidelines and revised review procedures

Carlos Granell[1], Frank O. Ostermann[2], Daniel Nüst[3], Peter Kedron[4,5], Eftychia Koukouraki[6], Miguel Matey-Sanz[1], Rémy Decoupes[7,8], Sergio Trilles[1], Anita Graser[9], and Tom Niers[3]

[1]Institute of New Imaging Technologies, Universitat Jaume I, Spain
[2]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, The Netherlands
[3]Department of Geosciences, Technische Universität Dresden, Germany
[4]Department of Geography, University of California Santa Barbara, United States
[5]Center for Spatial Studies and Data Science, University of California Santa Barbara, United States
[6]Institute for Geoinformatics (ifgi), University of Münster, Germany
[7]TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[8]INRAE, Montpellier, France
[9]AIT Austrian Institute of Technology, Vienna, Austria

Received: December 3, 2025; returned: March 13, 2026; revised: June 8, 2026; accepted: June 13, 2026.

Abstract:

Recent research in the field of geographic information science shows that reproducibility and replicability of publications have substantial room for improvement and proposes various actions to improve the situation. However, the impact of these actions remains unclear. This study investigates the combined effect of novel author guidelines and workflow review process, which award badges for successful reproductions, on the potential reproducibility of articles published in the AGILE conference series proceedings over the past decade. While replicating the approach of previous studies, this work expands the scope of prior reproducibility assessments and systematically compares the findings for the AGILE conference with those of the GIScience conference series proceedings, which

has not undergone similar changes to guidelines or procedures. Results indicate that the reproducibility guidelines and the review process measurably improved the potential reproducibility of AGILE publications. The comparison with GIScience papers further suggests that clear and enforced guidance is a key driver for change. Our findings demonstrate the value of institutional policies and community norms in fostering reproducible research in the GIScience field and identify pathways for its ongoing improvement.

Keywords: GIScience, reproducible research, reproducibility, computational science, data science, open science, meta-science, meta-research, science policy, open access

---

1 Motivation and research questions

Computational research relies on the ability to verify, build upon, and critically evaluate published findings. However, across disciplines, researchers have documented persistent challenges when attempting to reproduce computational results [1, 3, 11] from published papers, even when authors provide code and data [5, 22, 30, 42]. In Geographic Information Science (GIScience), a small but growing literature suggests this reproducibility gap may be particularly acute [30, 32]. GIScience research often depends on specific software versions, computational environments, proprietary datasets, and undocumented analytical decisions that are difficult to reconstruct from published descriptions alone. To address these challenges, leading GIScience journals and conferences have adopted data and code sharing policies that establish standards for code availability, data sharing, and computational documentation. The extent to which such guidelines improve the reproducibility of published computational research is an actively researched question [13, 18, 49].

We have laid the groundwork to address this question in a pair of studies that examine publications from the Association of Geographic Information Laboratories in Europe (AGILE) and the GIScience conference series [37, 43]. In these studies, we evaluated the potential reproducibility of the conference publications based on an *assessment rubric*, discussed the state of reproducibility in comparison to other disciplines, and suggested actions to improve reproducibility within the GIScience community. In this rubric, we categorised the availability of input data, methods, and results into four levels: unavailable; documented; available; and available, open, and permanent. These levels document to what degree the necessary conditions for a reproduction were met, but neither study attempted to actually reproduce any work. The second study on the GIScience conference series [43] successfully replicated our first study on the AGILE conference series [37], using the same methodology on a different pool of input data.

Both analyses revealed that "reproducibility and replicability have not been a core concern in the contributions" [43, pp 2:2] in either conference series [37, 43], which led us to make recommendations to improve author guidelines and peer review procedures as important levers capable of improving reproducibility practices around computational workflows [11]. A key outcome of our first study [37] was the development and publication of the AGILE Reproducible Paper Guidelines (hereafter referred to as *the AGILE Guidelines*) in July 2019 [41]. These AGILE Guidelines were revised in late 2020 [41] and have been gradually implemented by a Reproducibility Committee which performs reproducibility reviews, i.e., attempts for an actual reproduction, of all accepted full papers. No similar

guidelines or review processes were adopted by the GIScience conference series, creating an opportunity to evaluate the potential impact of the AGILE Guidelines.

In this work, we use an updated rubric that remains backwards-compatible to our earlier studies to compare the changes in the potential reproducibility of papers published by the AGILE and GIScience conference series, framed within these research questions:

- Question 1: Has the level of potential reproducibility of AGILE full conference papers changed after the introduction of the AGILE Guidelines and reproducibility reviews?
- Question 2: Has the level of potential reproducibility of GIScience full conference papers changed during the same period?
- Question 3: Is there an observable difference between the two conference series in the change in potential reproducibility?

We expect the overall level of potential reproducibility in papers published in the AGILE conference editions after the introduction of the AGILE Guidelines and the corresponding Reproducibility Committee to be higher than the potential reproducibility of papers published before their introduction (Question 1). However, one might anticipate a general trend of increasing reproducibility in the domain, given the increasing adoption of more open and reproducible research practices. We therefore expect that for the GIScience conference series, the level of potential reproducibility also increases (Question 2). Question 3 seeks to determine whether there are significant differences in potential paper reproducibility that might be explained by differences in author guidelines and the review process.

Therefore, this study allows us to discuss whether the implementation of reproducibility guidelines and review processes contributes to making conference publications more reproducible, and to provide empirical evidence for the utility (or lack thereof) of changing publication requirements and processes to increase open and reproducible research practices in the field of GIScience. The remainder of the article is structured as follows. Section 2 provides an overview of related work, followed by a description of the methods (Section 3) and results in Section 4. The article concludes with discussing methods and findings in Section 5 and an outline of future research directions in Section 6.

2 Related work

The AGILE Reproducible Paper Guidelines [41] offer authors extensive advice on open science practices aimed at improving workflow reproducibility of conference papers. Published in 2019, they have been recommended for all authors and reviewers since 2020[1] and a Reproducibility Committee has conducted reproductions of accepted full paper submissions [40]. For the 2021 edition, the updated guidelines introduced a mandatory requirement[2] that all papers submitted to AGILE conference series must include a section titled *Data and Software Availability* (DASA), where authors describe and reference the data and software used in their work, or document reasons for unavailability. AGILE's reproducibility reviewers attempt to reproduce a paper's computational workflow based on

[1] https://web.archive.org/web/20200926160015/https://agile-online.org/conference-2020/call-for-papers-2020

[2] https://web.archive.org/web/20210812211908/https://agile-online.org/index.php/conference-2021/call-for-papers-2021

given documentation, but they do not critically evaluate the whole research in the manner of traditional peer review. Authors have the opportunity to revise their manuscripts and associated resources in response to reviewers' feedback prior to publication.

In scope and extent, the AGILE reproducibility review is a concrete implementation of the CODECHECK principles [36] for evaluating scientific workflows as part of scholarly communication. The independent execution of full or partial workflows is carried out by *codecheckers*, a specialised reviewer role that gives recognition to other less common profiles in the peer review process, such as early-career researchers or data/software experts. The conversation between authors and codecheckers benefits all parties involved, and increases the availability and transparency of elements crucial to open reproducible research and education. A completed reproduction grants a timestamped CODECHECK certificate (see CODECHECK Register [39]). AGILE calls their certificates "Reproducibility Review"[3] and shows a completed reproduction with a badge on the issues and article pages[4].

More broadly, the AGILE Guidelines and the associated review process exist within an expanding literature on the reproduction and replication of geographic research along several lines of inquiry. *Conceptually*, geographers have begun to investigate the epistemological role of reproduction and replication within the discipline [14, 25, 29] and examine how work in these areas is related to disciplinary research questions [6, 26, 50, 52]. These works emphasised that replicability across space and time must be weak due to the ubiquity of spatial heterogeneity, and stressed how variation in the design of reproduction and replication attempts tests different forms of study validity [24, 28]. *Empirically*, researchers have attempted to identify the causes of the irreproducibility of geographic analyses [27,28] while also attempting to reproduce and replicate selected studies [23,34]. In addition to the research leading to this study [36,37,43], geographers have attempted to replicate analyses of the spatial distribution of COVID-19 [24, 45], reproduce published maps [32, 33], and create "replicability maps" in GeoAI based research [34]. Multi-analyist replication studies [2] are likewise beginning to appear in the literature, signalling the emergence of a new empirical approach to the challenge of replication in geography [4].

These lines of inquiry have revealed significant reproducibility challenges and prompted broader calls for substantial institutional changes within the GIScience community [29], in line with the broader scientific landscape [10, 35]. In response to these challenges, recommendations such as introducing reporting checklists or guidelines and adopting reward badges [12] have been proposed. While investigations suggest such interventions have positive effects across disciplines [4,12,13,19,49], and while some geography journals (e.g., [7]) have already introduced data and code sharing policies and incorporated some level of materials review into their peer review process, to our knowledge, no longitudinal study has yet systematically investigated the effects of policy-related changes in the domain of GIScience. Whereas previous work has focused on the reproducibility of single studies or paper collections, our work evaluates how potential reproducibility changes within a particular publication venue after the introduction of a specific policy and review process. If such policies are ineffective, they would represent a significant waste of author and reviewer efforts.

[3] AGILE-related certificates available at https://codecheck.org.uk/register/venues/conferences/agilegis/

[4] For example, https://agile-giss.copernicus.org/articles/6/index.html and https://agile-giss.copernicus.org/articles/6/2/2025/.

3 Methods

3.1 Preregistration and overall approach

To enhance the open science practices of this work, we published our hypotheses, research objectives, data collection and assessment process, and analysis plans as a preregistration [38]. We further share all data and code (Section 3.5) and the preprint [15]. In case we deviate from the preregistered design, we explicitly mention and justify these deviations.

This study improves on prior studies on the assessment of potential reproducibility [37, 43] in a number of ways. First, it adds a crucial amount of input data to the comparison: Previous studies were focusing on the level of reproducibility of the respective conference's papers, analysing the factors contributing to the observed low levels, and on raising awareness of this problem within the GIScience research community. These studies could not examine an impact of the AGILE Guidelines or workflow evaluation because they preceded their introduction at the AGILE conference series. Second, the prior studies looked at conferences separately while this work uses robust statistical assessment of the two conference series based on three hypotheses. Third, we employ a newly formalised reproducibility assessment protocol (see Section 3.3) based on the assessments of previous studies, thereby increasing chances for successful replication of the protocol. The protocol allows us to combine experienced contributors with a new group of assessors yet stringently ensure consistency and quality of the assessment.

3.2 Data collection process and corpus

The AGILE and GIScience conference series follow two different rhythms. AGILE is held annually, while GIScience is bi-annual. The COVID-19 pandemic led to cancelled conference editions in 2020 and altered publication patterns for GIScience within the observed time frame. The publication of the AGILE Guidelines, the *intervention*, took place in 2020, when the DASA section remained optional for authors. We therefore consider three periods: *pre-intervention* (proceedings published before and during 2019), *transitional* (submitted or published in 2020), and *post-intervention* (published in 2021 and later). Papers from the transitional period were assessed but excluded from the analyses in Sections 4.1 – 4.3.

For a sufficiently large and balanced paper corpus, we matched the post-intervention proceedings from both conferences with a similar number of pre-intervention proceedings: three AGILE and two GIScience post-intervention proceedings were matched with the same number of pre-intervention editions, resulting in coverage of a 9-year period (details below). We considered $n > 15$ papers per year adequate for statistical inference and testing, and all papers from a year were assessed without further sampling. Extending coverage further back in time was deliberately avoided, as reproducibility was not yet on the GIScience agenda a decade ago, which could bias the analysis towards a positive result.

The preregistration initially only considered all full papers published in the AGILE proceedings of the 2017, 2018, 2019, 2021, 2022, and 2023 editions. The AGILE 2024 proceedings became available before the assessment began and were subsequently included in the paper corpus. For GIScience, the 2021 proceedings comprised two volumes: papers submitted and published in 2020 (GIScience 2021 Part I) and those submitted and published in 2021 (GIScience 2021 Part II). GIScience 2021 Part I proceedings were therefore assessed but excluded from the analysis as a transitional year, like the AGILE 2020 papers.

Full texts of the conference papers were accessed directly from the publishers' websites (Springer, LIPICS, Copernicus). Access to copyrighted content from Springer (GIScience 2016, AGILE 2017, 2018, and 2019) required institutional access granted to some authors' universities. Papers were downloaded locally to facilitate assessment and were not publicly disclosed during the study. In total, 224 full articles were collected: 78 from GIScience and 146 from AGILE, as listed in Table 1.

Conference/year	Publisher	Open access	Total papers	Eligible papers
AGILE 2017	Springer	×	20	16
AGILE 2018	Springer	×	19	17
AGILE 2019	Springer	×	19	18
AGILE 2020	Copernicus	✓	22	0
AGILE 2021	Copernicus	✓	16	13
AGILE 2022	Copernicus	✓	22	20
AGILE 2023	Copernicus	✓	14	13
AGILE 2024	Copernicus	✓	14	12
Total			**146**	**109**
GIScience 2016	Springer	×	21	17
GIScience 2018	LIPICS	✓	17	17
GIScience 2021 Part I	LIPICS	✓	16	0
GIScience 2021 Part II	LIPICS	✓	13	13
GIScience 2023	LIPICS	✓	11	11
Total			**78**	**58**

Table 1: Paper corpus. "Eligible papers" column denotes the number of papers finally considered for analysis, excluding conceptual ones and those published in the transitional year (GIScience 2021 Part I and AGILE 2020). Source: `02_methods.qmd` [16].

A summary of the collected data elements is available at [16], which includes the dataset and code used to produce all tables and figures in this study, along with related documentation such as a data sheet to describe each column [5] and an analysis of authorship overlap between both conference series. The dataset contains basic bibliographic metadata of eligible articles and reproducibility assessment scores.

3.3 Assessment of potential reproducibility

We recruited eight assessors by email in October 2024, selected based on previous published work on reproducibility and/or former members of the Reproducibility Committee of the AGILE conference series. All assessors became also co-authors of the final study.

All assessments followed a structured Assessment Protocol [44] compiled to integrate new assessors and document implicit knowledge from previous studies [37,43]. The protocol included instructions for determining whether an article contained any computational analysis, with the aim to exclude purely conceptual papers, literature studies, review papers, or opinions, unless they include at least a small case study or descriptive statistics. It also provided a rubric and examples to guide assessors rate the potential reproducibility level of a computational research article.

[5] https://github.com/nuest/reproducible-research-giscience-longitudinal-study/tree/main/data-clean

Criteria	Rubric	Comments
Input Data (*Data* for short)	NA	When data is not required or used in the reported study.
	U	When data is not described (including available upon request) and is not recreatable (even if documented or with metadata).
	D	When data includes metadata and is recreatable (same or similar data can be retrieved from original source).
	A	When data is accessible on non-permanent websites (e.g., no DOI).
	O	When data is openly available and permanently archived (with DOI).
Methods, Analysis, Processing (*Methods* for short)	NA	When methods are not required or used in the reported study (which should be completely atypical).
	U	When methods are a kind of visual dossiers of screenshots poorly documented or described.
	D	When methods are described using text, pseudo-code, workflow description, diagrams, etc. and parameters in the analysis are all well described.
	A	When methods are self-described via online source code, e.g. GitHub.
	O	When methods are fully described as runtime image/container, standardised metadata, open/permissive license.
Results	NA	When results are not produced or described in the reported study (which should be completely atypical).
	U	When results are understandable and context provided, i.e., with reasonable statistical measures or summaries, textual descriptions, tables, maps, etc.
	D	When results are described using text, pseudo-code, workflow description, diagrams, etc. and parameters in the analysis are all well described.
	A	When model outputs, output data, scripted plots/maps are included/cited.
	O	When results are understandable, available, open and permanently archived (with DOI).
Comput. Environ.	True	When hardware configuration, operating system, run time or computational demand, and requirement settings are present.
	False	When at least one of the items above is missing.

Table 2: Criteria and rubric for assessing potential reproducibility of a paper. NA: *Not Applicable*; U: *Undocumented*; D: *Documented*; A: Documented & Available (*Available* for short); O: Documented, Available and Open (*Open* for short). *Documented* should be the minimum level to enable reproducibility.

The rubric in Table 2 comprises four criteria: *Input data*; *Methods, analysis, processing*; *Results*; and *Computational environment*. The first and third criteria remain unchanged from earlier versions [37,43]. The former "Methods" criterion has been consolidated into a single criterion *Methods, analysis, processing* to avoid ambiguity in distinguishing preprocessing from analysis steps. The *Computational environment* is now treated as a separate criterion, reflecting its role in encompassing all code and infrastructure used for both analysis and result presentation. The originally numeric levels have been replaced with descriptive labels and single-letter codes (UDAO scheme, see Table 2) to enhance interpretability and avoid implying linear relationships between levels. We also aggregate levels of potential repro-

ducibility into *Low* (U, D) and *High* (A, O) to highlight the key change from *Documented* to *Available*. For the *Computational environment* criterion, only a binary evaluation (True/False) is applied, given the difficulty of assigning finer-grained levels consistently. For readability, the short names of the criteria and rubric levels can be found in Table 2.

Each paper was independently assessed by two assessors: one who had extensive prior experience in evaluating potential reproducibility using the original protocol [37, 43], and another assessor drawn from a pool of eight researchers with varying levels of experience. From hereonwards, we refer to these two assessors as A1 and A2, respectively. Assessors recorded observations, sources of information, and decisions made in borderline cases, according to the shared Assessment Protocol [44]. Each A1 selected entire conference years to assess (May–July 2024). Then each A2 followed a batch-based strategy assessing papers in four batches ordered by proximity to the intervention year [38], completing each batch before moving to the next. All A2 assessments were completed between November 2024 and February 2025, without knowledge of A1 scores.

Once all assessments were complete, the A1 consolidated the separate evaluation spreadsheets and resolved disagreements by analysing assessors' written comments. Inter-assessor agreement was classified into five categories: *no disagreement*; *borderline conceptual paper*, where A1 and A2 differed on whether a paper was labelled conceptual (and thus to be assessed at all); *annotation inconsistencies*, reflecting obvious scoring errors relative to written notes; *uncertain assessment*, where an assessor explicitly indicated ambiguity; and *significant disagreement*, where A1 and A2 had strong disagreement over multiple criteria. Annotation inconsistencies were resolved directly by A1. Cases of uncertain or significant disagreement were resolved through bilateral or joint discussion between the two assessors, or, if needed, by involving a third assessor. All disagreement resolutions are documented by versioned spreadsheets [16].

We recognise that the assessment process is inherently subject to the subjective interpretation and prior experience of the assessors. To mitigate this, we recruited assessors with varying degrees of experience in reproducibility and replicability across diverse geographical science disciplines, and provided all assessors with the common Assessment Protocol [44] to minimise variability in scoring.

3.4 Analysis of potential reproducibility

To address our research questions statistically, we first considered the characteristics of our dataset. Since the unit of analysis was the published papers, we considered for each paper its group (pre- and post-intervention for AGILE and GIScience), authors and three ordinal (ranked) criteria (*Data*, *Methods*, *Results*). All groups had different sizes, the smallest being 24 (GIScience post-intervention, compare Table 1). Each paper had one or more authors. Several papers had at least one (co-)author who contributed to one or more papers in the other group, thus the groups were not independent[6]. The criteria were ranked labels in the order of U < D < A < O.

To address Questions 1 and 2, we compared the pre- and post-intervention groups within a conference. Since some authors were common to all groups and the unit of interest was the papers, we modelled the authors as random effects using an ordinal mixed effects model: a cumulative link model that represented, for each paper criterion, the odds

[6]For details on the extent of authorship overlap, see notebook `07_authorship.ipynb` in [16]

of achieving a higher rank, cumulative across all rank thresholds and influenced by the random effects of all its authors. For both questions, the hypotheses were similar:

- *H-Null*: The odds of a paper scoring a higher reproducibility rank do not differ significantly between pre-intervention and post-intervention conference editions, after accounting for the non-independence introduced by authors publishing in both periods.
- *H-Alternative*: The odds of a paper scoring a higher reproducibility rank are significantly different in post-intervention editions compared to pre-intervention editions, after accounting for author non-independence.

To address Question 3, we compared the effect sizes of the above analysis, i.e., the odds ratios of having a higher rank for a criterion. However, a cumulative link model assumes that the predictor effect is the same across all thresholds between ranks. To correctly interpret the odds ratios, we also verified whether this proportional odds assumption holds. We adopted a more conservative significance level of 0.01 instead of the 0.05 commonly used in preregistration to reduce the risk of falsely rejecting the null-hypothesis and thus committing a Type I error (false positive).

Full details of the statistical analysis can be found in the corresponding computational notebook (see the following section). In Section 4, we only present the results directly pertinent to our three questions.

3.5 Data and software availability

The data and code in this work are published under permissive licenses in the GitHub repository (github.com/nuest/reproducible-research-giscience-longitudinal-study). The repository is archived on Zenodo [16] at `10.5281/zenodo.21097308` and in Software Heritage with SWHID `swh:1:rev:ec0bd7a2c0b60e290b6183d3e073d992d901b827` [17]. The repository contains data at various stages of processing and computational notebooks in R (Quarto) and Python (Jupyter) for the analyses and visualizations. The corresponding notebook file is referenced in each table or figure.

4 Results

Of the initial 224 papers in the corpus, after removing non-computational papers and those from the transition year, 168 papers remained: 109 AGILE and 58 GIScience (see the “Eligible papers" column in Table 1 for the breakdown by year).

4.1 AGILE papers

Figure 1 shows the distribution of the potential reproducibility levels for each criterion for the AGILE conference series, divided into two groups: 51 papers from three conferences (2017–2019) published before the intervention are in the upper row, and 58 papers from four conferences (2021–2024) published after the intervention are in the lower row. AGILE 2020 papers were excluded as they belong to the transitional year when the AGILE Guidelines were introduced but still optional. The figure also shows aggregate levels *Low* and *High*.

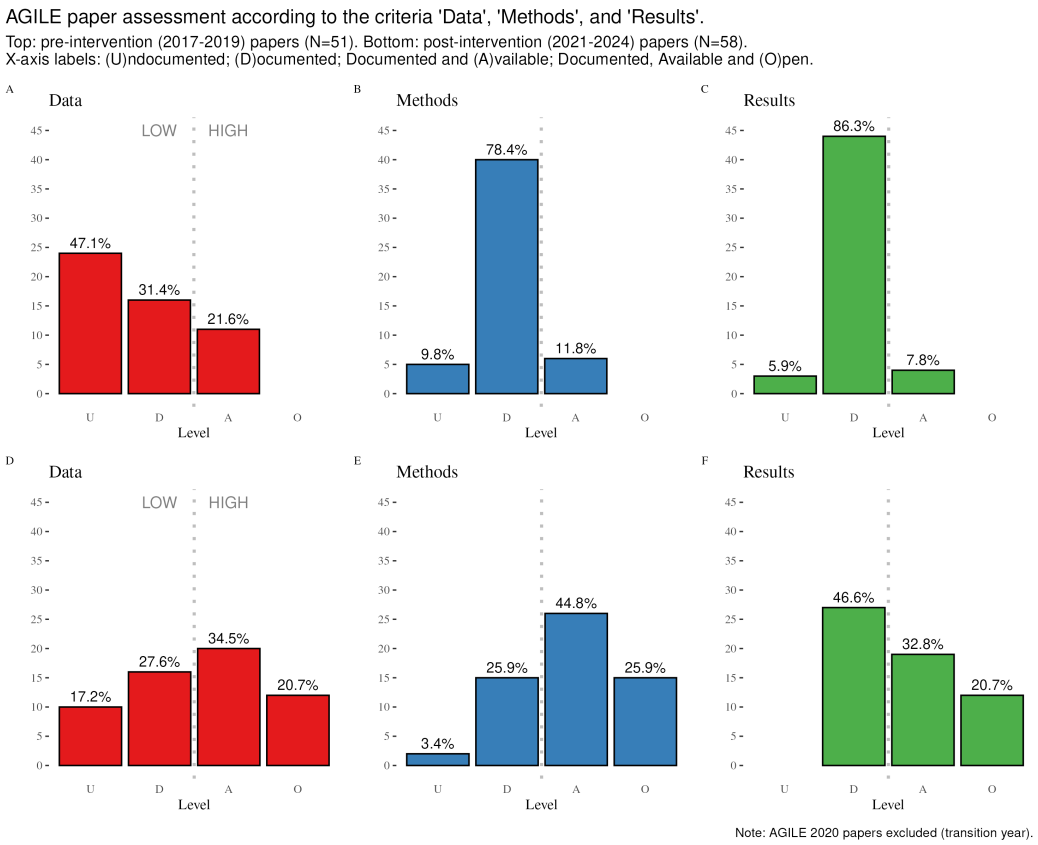


Figure 1: Evaluation of the potential reproducibility level of AGILE papers (pre- vs post-intervention) for each criterion: *Data* (A and D), *Methods* (B and E), and *Results* (C and F). Source: `03_results_reprolevels.qmd` [16].

The pre-intervention results (Figure 1, top row) match the results of a previous study evaluating the best papers from 2010, and 2012–2017 [37]. None of the pre-intervention papers reached the highest level of *Open* on any criterion. An issue encountered in the previous study [37] persists in the current larger pool of pre-intervention AGILE papers: the high proportion of papers (24, or 47.1%) with level *Undocumented* for *Data*, meaning that input datasets were unavailable and cannot even be re-created today from the information provided. By contrast, the *Methods* and *Results* criteria show a remarkably high proportion of *Documented* papers: 40 (78.4%) and 44 (86.3%) respectively. These percentages suggest that authors generally provide sufficient documentation to follow their analysis. Compared to [37], a slightly larger share of papers now reach the *Available* level for these two criteria, though this difference is based on a small number of studies and should be interpreted with caution.

The post-intervention results (Figure 1, bottom row) present a completely different picture. Most notably, a considerable number of papers reached the highest potential repro-

ducibility level of *Open* for all criteria, a level never achieved in the pre-intervention period. For the *Data* criterion, the frequency distribution tends to be right-skewed towards higher levels: *Available* is the most frequent level (20 papers, 34.5%), and together with *Open* (12 papers, 20.7%), the *High* aggregate exceeds 55% of post-intervention papers. This represents a substantial improvement over the pre-intervention period, where nearly 80% of papers were categorised as *Low*. However, 16 papers (27.6%) only reached the *Documented* level and 10 papers (17.2%) remained *Undocumented*, indicating that there is still room for improvement when it comes to reproducing input data sets.

For *Methods*, *Available* was again the most frequent level (26 papers, 44.8%), and together with *Open* (15 papers), approximately 70% of papers are categorised as *High*. The *Documented* level, which dominated in the pre-intervention period (8 out of 10), dropped to only a quarter of post-intervention articles, a four-fold reduction suggesting a clear improvement in reproducibility.

For *Results*, the most common potential reproducibility level remains *Documented*, followed by the two higher levels (*Available*, *Open*). While the two highest levels combined account for only 7.8% of pre-intervention papers, the percentage considerably grew to 53.5% of post-intervention papers. This seven-fold increase indicates a clear improvement in reproducibility for the post-intervention AGILE papers.

The statistical analysis supports this assessment. Concerning the comparison between the pre- and post-intervention groups (Question 1), the pairwise group comparison per criterion offers an intuitive approach to understanding the results, which are quite clear: For each criterion, the groups have different ranks at our chosen significance level ($p < 0.0001$), confirming the descriptive statistical analysis that the groups differ significantly.

To determine the direction and size of the effect, we look at the Odds Ratios (OR) in Table 3: For the reference criterion of *Data*, the OR for the post-intervention group is 19.271. This means that post-intervention papers have 19 times higher odds of having a higher rank compared to pre-intervention ones. This is a large, significant improvement for *Data* after intervention. The effect of the intervention does not differ significantly in the other two criteria: For *Methods*, the OR is 20.485, indicating almost the same effect, while for *Results* the OR is 12.333, indicating a somewhat weaker but still significant effect.

Criteria	p-value	OR	logOR
Data	<0.0001	19.271	2.959
Methods	<0.0001	20.485	3.020
Results	<0.0001	12.333	2.512

Table 3: Differences between pre- and post-intervention periods for AGILE conference. OR: odds ratios. Source: `04_results_hypotheses.ipynb` [16].

However, testing also indicates the proportional odds assumption is violated. This means that we cannot guarantee equal effects across all rank transitions and have to interpret the effect sizes carefully. Nevertheless, a rejection of H-Null still indicates that group membership is a statistically significant predictor of reproducibility rank within the cumulative link mixed model framework. In summary, we can conclude that:

- H-Null can be rejected because the odds of a paper scoring a higher reproducibility rank are significantly different in post-intervention AGILE proceedings compared to pre-intervention proceedings, after accounting for author non-independence.

- The OR are large and positive, indicating higher ranks in the post-intervention group, suggesting that post-intervention AGILE papers are significantly more likely to reach higher reproducibility levels.

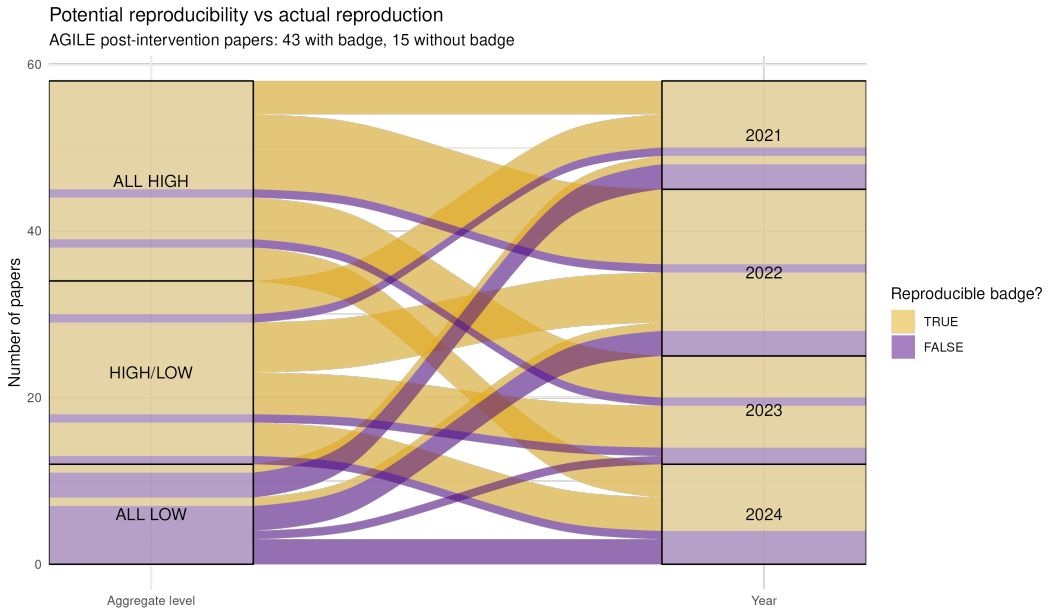
We recognise, however, that assessing *potential* reproducibility without *actually* reproducing the study is only an indicator (or *preproducibility* [48]). Given that statistical testing above confirmed potential improvement, this raises the question of whether high potential reproducibility aligns with successful reproduction in practice. We therefore conducted a badge analysis, using the AGILE reproducibility badge as a proxy to compare our assessment against actual reproductions reported by the AGILE Reproducibility Committee for the 2021–2024 period. For this analysis, post-intervention AGILE papers are grouped into three categories based on their scores across all criteria: *All High* (*Available* or *Open* on all criteria), *All Low* (*Undocumented* or *Documented* on all criteria), and *High/Low* (any combination). Figure 2 shows that potential reproducibility is a fairly good indicator of actual reproducibility, because AGILE papers rated *All High* were consistently awarded a reproducibility badge following the reproducibility review process, while those rated *All Low* largely were not. Of the 58 post-intervention papers, 43 earned a badge and 15 did not. This supports that our reproducibility rubric is a good indicator of whether an article provides a sufficient basis for successful reproduction.

4.2 GIScience papers

Figure 3 shows the distribution of potential reproducibility levels for each criterion for the GIScience conference series, divided into two groups: 34 papers from two pre-guidelines conferences (2016 and 2018) are in the upper row, and 24 papers from two post-guidelines conferences (2021 and 2023) are in the lower row. The figure also shows aggregate levels *Low* and *High*. Recall that the GIScience conference series never implemented a reproducibility review process; the pre/post division is made only for comparison with the AGILE conference series.

For the pre-guidelines results (Figure 3, top row), the *Data* criterion shows an especially high proportion of *Undocumented* papers (19, or 56%), with no papers reaching the higher level *Open* on any criterion. The *Documented* and *Available* levels each share around 20%, mirroring the results of a previous study focused on the GIScience conference series alone and with different assessors [43]. We identify the same underlying problem in the more recent conference editions: the large proportion of papers with undocumented input datasets represents a significant barrier to reproduction. Input data are not only unavailable but cannot be recreated from the information provided. By contrast, the *Methods* and *Results* criteria show a very different distribution: 32 papers (94%) reached the *Documented* level for both criteria, indicating that authors generally meet the minimum standard of publication and provide sufficient documentation for reviewers to follow their analysis and results. These results again align with earlier reproducibility assessments of GIScience articles [43].

The post-guidelines results (Figure 3, bottom row) show a different picture. A few post-guidelines papers achieved the highest potential reproducibility level of *Open* on any criterion (not necessarily in the same paper), though half of the papers (12) still reached the *Undocumented* level for the *Data* criterion, remaining a clear barrier to reproducibility. For the *Methods* criterion, *Documented* remains the most frequent level (12, 50%), as in the pre-guidelines period, but the number of *Available* papers increased significantly from 1 (3%) to



Aggregate Level	With reproducible badge		Without reproducible badge	
	Number of papers	Proportion	Number of papers	Proportion
ALL HIGH	22	51.2%	2	13.3%
HIGH/LOW	19	44.2%	3	20.0%
ALL LOW	2	4.7%	10	66.7%

Figure 2: Potential reproducibility levels as indicators of reproduction in AGILE post-intervention papers. Figure shows connection between aggregated levels and reproducible badges earned. Levels *Undocumented* and *Documented* on all criteria are considered *All Low*, levels *Available* and *Open* on all criteria are considered *All High*. Table below shows actual numbers and proportions for context. Source: `06_discussion.qmd` [16].

9 (37.5%). For the *Results* criterion, *Documented* also remains the most common potential reproducibility level, but the distribution has shifted notably compared to the pre-guidelines period: four papers are *Available* (16.7%), two are *Open* (8.3%), and no papers are *Undocumented*. Taken together, these changes clearly show a trend towards higher reproducibility levels in the post-guidelines GIScience conference editions.

Following the approach from the previous section, the pairwise group comparison for GIScience (Question 2, see Table 4) shows that at our chosen significance level, ranks for the *Data* criterion are not significantly different ($p = 0.0683$) between pre- and post-intervention groups, but those for *Methods* and *Results* criteria are ($p < 0.0001$).

The interpretation of the effect size for the reference criterion *Data* is negligible or no improvement (OR = 1.995). Again, *Methods* and *Results* are different from *Data*: The OR for *Methods* is 8.365 and for *Results* is 5.329, indicating a clearly positive change for the post-intervention group.

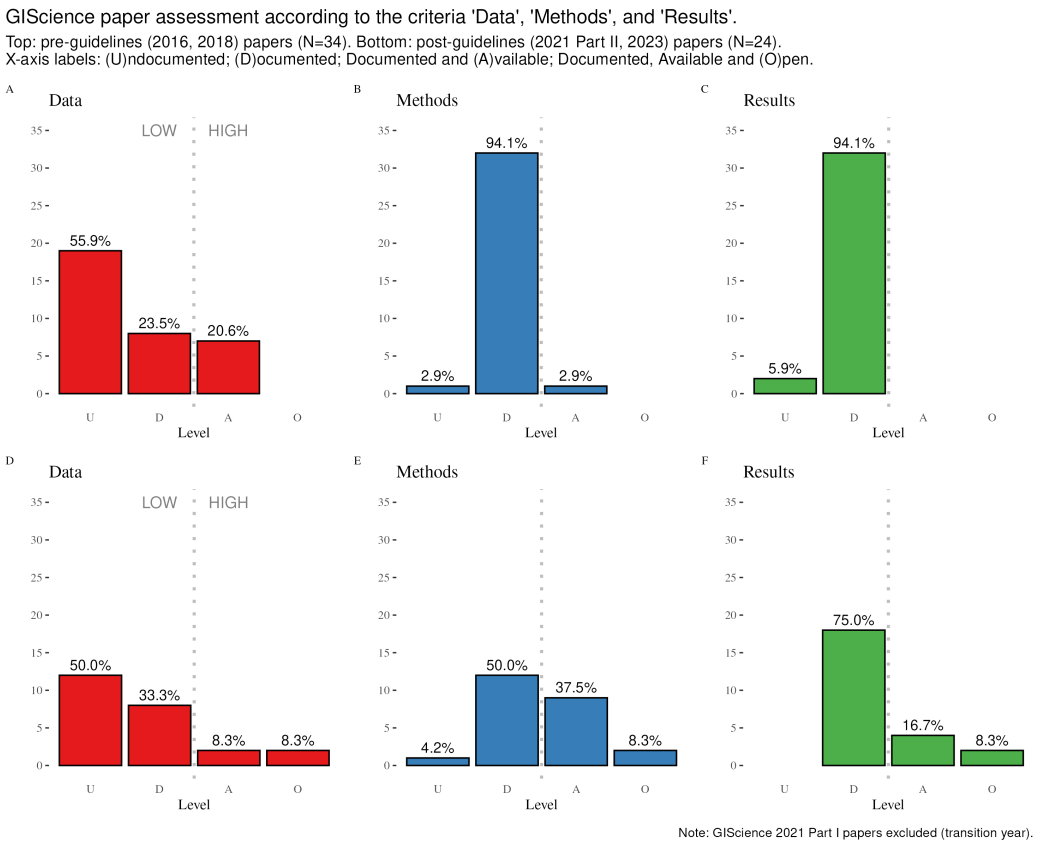


Figure 3: Evaluation of the potential reproducibility level of GIScience papers (pre- vs post-guidelines) for each criterion: *Data* (A and D), *Methods* (B and E), and *Results* (C and F). Source: `03_results_reprolevels.qmd` [16].

Criteria	p-value	OR	logOR
Data	0.0683	1.995	0.691
Methods	<0.0001	8.365	2.124
Results	<0.0001	5.329	1.673

Table 4: Differences between pre- and post-intervention periods for GIScience conference. OR: odds ratios. Source: `04_results_hypotheses.ipynb` [16].

Like for AGILE, the proportional odds assumption is violated, meaning that we have to interpret actual numbers with care. However, the OR for the three criteria are sufficiently distinct to allow us to summarise the analysis as follows:

- H-Null can be rejected only for two criteria. The odds of a paper scoring a higher reproducibility rank are significantly different in post-intervention GIScience proceed-

ings compared to pre-intervention proceedings for the *Methods* and *Results* criteria, but not for *Data*, after accounting for author non-independence.
- Like for AGILE, the OR are clearly positive for *Methods* and *Results*, so the change is towards higher ranks for the post-intervention group.

4.3 Comparison between conferences

The assessment results of the pre-intervention AGILE papers (Figure 1, top) show similar patterns to the results of the pre-intervention GIScience conference series (Figure 3, top). In other words, the top rows of both figures are quite similar and show continuity with the reproducibility assessments of our previous studies in preceding years. Lower rows in both figures, however, present right-skewed distributions, favouring *High* levels of potential reproducibility. Despite this overall improvement, post-intervention AGILE papers show a more pronounced shift towards the *High* level across all criteria than post-guidelines GIScience papers.

To better illustrate the change after the introduction of the AGILE Guidelines between both conferences, Table 5 shows the change in absolute percentage points in potential reproducibility levels for each criterion and conference. For simplicity, we compare the aggregate reproducibility level *High* (A, O). To do so, we determine for each criterion whether the aggregate level *High* has increased or decreased in absolute percentage points before and after the intervention. Positive percentage points in Table 5 indicate an improvement in reproducibility towards *High* level, while negative values indicate a decrease of *High* level in post-intervention (or post-guidelines) papers, suggesting no progress towards reproducibility in a given criterion. Overall, the improvements are very noticeable for AGILE compared to GIScience across all criteria, but are particularly notable with respect to *Data*, given that GIScience post-intervention papers do not improve on that criterion.

Criteria	AGILE	GIScience
Data	33.6	-3.9
Methods	58.9	42.9
Results	45.7	25.0

Table 5: Absolute change in percentage points between GIScience vs AGILE after the intervention for the aggregate level *High*. Source: `03_results_reprolevels.qmd` [16].

The violation of the proportional odds assumption makes a direct, quantitative comparison of the changes between the two conferences more difficult. We have considered to use partial proportional odds models or using the binary *High/Low* classification. However, neither solves the problem of the limited sample size and distribution of ranks, which results in very few (or none at all) observations for high ranks (*Available* or *Open*) in the pre-intervention groups for both conferences.

Instead, we decided to interpret the outcomes conservatively (Question 3): We observe overall notably lower OR for improved potential reproducibility at GIScience than for AGILE, with no significant change in the *Data* criterion for GIScience, while AGILE shows significant and strong improvements for all criteria. Combined with the fact that many post-intervention GIScience papers have at least one (co-)author who published also at AGILE,

we are confident to state that, based on the statistical testing, the improvement of potential reproducibility at AGILE is larger and broader than the improvement at GIScience.

4.4 Assessment disagreements

The analysis of disagreements during assessment considers the total number of papers (224), including those from the transitional year, as they provide useful data for examining in detail assessor agreement. Assessors agreed completely on the potential reproducibility of papers in only 28% of cases (*no disagreement* in Table 6), while some form of (minor) disagreement occurred in over 70% of cases (rest of disagreement types, see Table 6). This high rate of disagreement was not anticipated at the time of preregistration and was therefore not included as a planned analysis. However, the finding is consistent with previous studies [28, 30], which highlight the heterogeneity of perceptions of reproducibility and replicability, often shaped by assessors' expertise, disciplinary background, and subjective interpretation. The observed disagreement rates also align with disagreement causes and patterns reported in earlier reproducibility studies of these conference series [37, 43].

Disagreement types	**#**	**%**
Uncertain assessment	100	44.6%
No disagreement	63	28.1%
Significant disagreement	42	18.8%
Borderline conceptual paper	14	6.2%
Annotation inconsistencies	5	2.2%
Total	**224**	

Table 6: Distribution of disagreement types as introduced in Section 3.3. Source: `05_results_assessprocess.qmd` [16].

After excluding cases of *no disagreement*, approximately 53% of discrepancies categorised as *borderline conceptual paper*, *uncertain assessment*, and *annotation inconsistencies*, were due to divergent interpretations of the assessment criteria, suggesting that further refinement of the Assessment Protocol could reduce such inconsistencies. The remaining *significant disagreements* (almost 20%) reflected deeper issues, often related to insufficient detail in the assessed papers combined with imprecise guidance in the Assessment Protocol, underscoring the difficulty of attributing discrepancies to a single factor.

Analysis across the two conferences revealed no remarkable differences in the overall distribution of disagreement types. AGILE papers exhibited a slightly higher proportion of *significant disagreements* (20.5% vs. 15.5%), while GIScience papers showed a slightly higher proportion of *borderline conceptual paper* (7.7% vs 5.5%) and *no disagreement* (29.5% vs 27.4%). A plausible explanation could be that GIScience papers tend to be more theory-oriented, whereas AGILE contributions are generally application-driven. Theoretical papers, which typically involve fewer computational components and datasets, may be easier to assess under our reproducibility criteria, reducing the number of ambiguous cases. Applied research papers, in contrast, typically involve more complex methodological and code- and data-related components, increasing the potential for disagreement. While the differences are modest, they underscore the importance of tailoring assessment protocols (and tools) to the characteristics of the research community being evaluated.

Focusing on the top 3 disagreement types in Table 6, the frequency of *no disagreement* was lower in more recent years, reaching its lowest percentage in the final year of each series, suggesting that consensus was easier to reach for earlier conference editions (AGILE 2017, GIScience 2016, GIScience 2018). Two factors may explain this. First, authors have increasingly adopted best practices for documenting research resources, whereas earlier papers more commonly omitted such documentation, leading to more assessments with lower reproducibility levels (e.g. *Undocumented*). Second, the increasing number of computational papers in recent editions may make it harder for assessors to reach full consensus on all criteria.

For *uncertain assessment* and *significant disagreement*, *Data* consistently accounts for the largest proportion of discrepancies across both conferences (Table 7). In other words, the *Data* criterion clearly remains the most contentious for both conferences. For *uncertain assessment*, this may stem from subjective interpretations of the Assessment Protocol, stressing the need for more nuanced guidelines to help reviewers and readers evaluate input datasets and authors describe them less ambiguously. For *significant disagreement*, *Data* again ranks highest, though the *Methods* criterion closely followed, especially for AGILE papers (70%). These findings suggest that while *Data* were the single most prevalent cause of discrepancies when analysing uncertain assessments, all criteria, but especially *Data* and *Methods*, are essential to understanding the causes behind significant disagreements.

Uncertain assessment									
Conference	**Data**			**Methods**			**Results**		
Series	Disagr?	#	%	Disagr?	#	%	Disagr?	#	%
AGILE	no	23	35.4%	no	46	70.8%	no	36	55.4%
AGILE	yes	42	64.6%[1]	yes	19	29.2%[2]	yes	29	44.6%
GIScience	no	7	20.0%	no	24	68.6%	no	28	80.0%
GIScience	yes	28	80.0%[1]	yes	11	31.4%	yes	7	20.0%[3]
Significant disagreement									
AGILE	no	6	20.0%	no	9	30.0%	no	13	43.3%
AGILE	yes	24	80.0%[4]	yes	21	70.0%	yes	17	56.7%[5]
GIScience	no	3	25.0%	no	5	41.7%	no	5	41.7%
GIScience	yes	9	75.0%[1]	yes	7	58.3%[6]	yes	7	58.3%[6]

[1]Higher percentage of *uncertain assessment* in both conferences.
[2]Lower percentage of *uncertain assessment* at AGILE.
[3]Lower percentage of *uncertain assessment* at GIScience.
[4]Higher percentage of *significant disagreement* in both conferences.
[5]Lower percentage of *significant disagreement* at AGILE.
[6]Lower percentage of *significant disagreement* at GIScience.

Table 7: Distribution of *uncertain assessment* (N=100) and *significant disagreement* (N=42) per conference and criterion. Source: `05_results_assessprocess.qmd` [16].

5 Discussion

5.1 Impact on potential reproducibility

The potential reproducibility of publications in both conference series during the 2010s was generally low, but especially low for the *Data* criterion. In the past five years, a positive trend toward higher levels of potential reproducibility was observed in both conferences, but AGILE saw a greater absolute change than GIScience and also had the higher overall (average) potential reproducibility. The fact that membership in the AGILE post-intervention group is a strong and clear predictor for improved reproducibility does not imply causality. However, the analysis of our data clearly shows an increase in potential reproducibility that coincides with the introduction of the AGILE Guidelines, and is stronger for the AGILE conference. This makes it very unlikely that the AGILE Guidelines and the reproducibility review had a negative effect on the development of paper reproducibility. To the contrary, it seems very likely that they had a positive impact on improving reproducible research practices, especially for the *Data* and *Methods* criteria.

Our data, especially Figure 3, which shows the distribution of reproducibility levels among pre- and post-guidelines GIScience articles, and Table 5, which shows the absolute percentage point change in potential reproducibility, demonstrate a trend toward higher reproducibility levels (aggregate level *High*) in the post-guidelines GIScience papers as well. This trend is likely to continue, and the inclusion of more recent GIScience papers (from 2025 onwards) could show a further increase. This overall trend may be partly due to greater awareness of reproducibility within the research community, as confirmed by a recent survey in the fields of GIScience and Geography [28].

However, this improvement at GIScience might also be partly attributable to an informal spill-over effect from authors who also published in AGILE, where AGILE Guidelines and review process were already in place. This is plausible given the known overlap in authors, reviewers, and program committee members between the two communities [43]. Of the two GIScience papers that reached the *Open* level on all criteria [20, 43], one had an author who had previously published at AGILE conference series after the intervention [21,47], and the second was co-authored by members of the present study. We consider that a more systematic investigation of authorship patterns with respect to both conference series beyond our initial analysis of (see `07_authorship.ipynb` in [16]) is beyond the scope of this article, but we believe it would certainly be a valuable further step.

Is it then "mission accomplished" for the GIScience domain as a whole? While the improvements are certainly noteworthy and very encouraging, our investigation also showed that level of *Available* for *Data*, *Methods*, and *Results* criteria is often insufficient a few years after publication, because links to project web pages or code/data repositories, for example, may become unavailable for various reasons in the medium to long term. One overarching objective of open science and computational reproducibility, however, is to ensure that studies remain available for reproduction (and replication), fostering longitudinal research. Clearly, more work and effort are needed to encourage more published studies to fully reach an *Open* level.

5.2 Insights from the assessment process

Our work supports the conclusions on computational reproducibility of AGILE and GIScience conference papers drawn by [37] and [43]. Our systematic review of papers published at both conferences confirms that the revised Assessment Protocol, which comprises criteria, reproducibility levels, and recommendations for the assessment procedure, is robust and effective for assessing the potential reproducibility of scientific papers. Furthermore, the assessment of new papers and re-assessment of earlier papers by a largely new group of assessors produce highly consistent results.

Despite our efforts to provide assessors with an updated, consistent, and concise Assessment Protocol, supported by onboarding and training sessions, the analysis revealed a considerable number of inter-assessor disagreements. Our initial rationale of the protocol was to keep instructions brief enough to enable assessments while defining a robust implementation as a safety measure (see Section 3.3). However, even with clear criteria, the diversity and heterogeneity of geospatial research means that some papers might fall into grey areas open to interpretation. For example, the potential reproducibility levels are ordinal and assume that a higher level includes the lower ones. Yet we encountered situations where used data and methods were hosted in a public repository – thus *Available* – but were so poorly documented that reuse would be doubtful. This aligns with the observation that openness enables reproducibility but does not guarantee it [9]. Should such cases be rated *Undocumented* or *Available*? We resolved these from the perspective of the potential user, rigorously rating them as *Undocumented*.

Fortunately, most disagreements were relatively straightforward to resolve, as assessor's notes made their reasoning transparent. In some cases, a note explicitly stated that an assessor would also support a different level of potential reproducibility, which matched the other assessor's rating (recall that A1 and A2 assessors were unaware of each other's assessments at this stage). In other cases, notes hinted at a simple mistake: for instance, when the reasoning clearly argued for *Available* but only *Documented* was assigned.

Our recommendations for reducing discrepancies in future reproducibility assessments address two aspects: the assessment process itself and the provided instructions, i.e., the Assessment Protocol. Regarding the process, a key principle is that assessments must remain independent, that is, assessors should not be aware of each other's outcomes until all assessments are completed. This would not preclude asking clarifying questions about individual papers. Regarding the protocol, we recommend providing clearer guidelines for writing notes. Specifically, we propose adding the following guidance: "The assessor may include as many comments and notes as they deem necessary. However, a brief explanation or justification is required when the assessment deviates from what the rubric would suggest; e.g., a study claims that all data and methods are in a repository, yet the assessment is only *Documented*. All *Undocumented* ratings must also be justified". This is a critical point, as studies may appear reproducible at first glance but lack the descriptive elements necessary for full verification. For example, recent work has shown that researchers often share code for their proposed models but not for the baselines, making independent validation impossible [46]. A call for authors to adopt more rigorous reproducibility standards would not only raise the quality of reproducible papers but also reduce discrepancies in their assessment.

6 Conclusion

A shift in Open Science practices can be introduced and observed even within a small community conference, given the support of community leadership and collective openness to adopting them. While the AGILE Guidelines offer discipline-specific guidance, the underlying criteria, rubric, and reproducibility process (or a CODECHECK) are largely cross-disciplinary and therefore transferable to other communities. Our results provide strong motivation for other disciplines organised around specific journals or conferences to improve reproducibility practices, whether by introducing author and reviewer guidelines, or by implementing code execution checks as part of peer review. The significant improvements in the potential reproducibility of the AGILE papers published after the introduction of the AGILE Guidelines demonstrate a clear and lasting impact on the GIScience research community; an impact that can be leveraged by related conference venues such as the GIScience conference series and journals such as JOSIS.

Future research could expand on this study in several ways:

- Initiating conversations within the GIScience community to promote the adoption of the AGILE Guidelines across its conference series.
- Exploring the perceptions of authors and reviewers in both communities regarding their familiarity with Open Science practices, understanding of the AGILE Guidelines, and the motivators and values driving their adoption.
- Tracking progress longitudinally through the regular addition of publications from new conference editions (e.g., every 4 years), to monitor the adoption of reproducible practices and identify emerging community needs.
- Complementing the current assessment of potential reproducibility with actual reproduction studies.

Given the emergence of Large Language Models (LLMs) in research across disciplines, GIScience is no exception. A recent study [51] highlights the opportunities LLMs offer across a wide range of geospatial application areas, from urban planning and environmental analysis to education. It can therefore be expected that wider uptake of LLMs will introduce new challenges for sharing and reproducing research. We thus question how LLMs might shape future versions of this study. A recent experiment found that ChatGPT (GPT-3.5) was ineffective at predicting the reproducibility of scientific articles by analysing only their methods sections [8]. For LLMs –and Vision Language Models (VLMs) [31]– to be useful in assessing the potential reproducibility of scientific articles, these authors call for the development of available benchmarks to systematically evaluate their ability to assess potential reproducibility in scientific papers. This study and follow-up studies may be steps in this direction.

Positionality statement

The authors report no conflict of interest. However, DN, CG, and FO were the initiators of the AGILE Reproducibility initiative[7], which resulted in the author guidelines and reproducibility review process that this paper refers to as "the intervention". Most of the remaining authors have participated in the reproducibility review process in one or more

[7] https://reproducible-agile.github.io/

years. We acknowledge that this may result in a stronger than usual bias at least for DN, CG, and FO to see the reproducibility guidelines "succeed" (in the sense of improved potential reproducibility). We have mitigated this potential bias by a very detailed assessment protocol, paired assessments, documented decision-making, and robust quantitative analysis. We believe that through these measures, the impact of our (unconscious) bias is not stronger than that of any researcher who wants their work to be successful. Our desire is that our results apply not only within the AGILE community but also within the broader geographic information science research community of which we are part. Furthermore, this study is as open and reproducible as possible, and we explicitly invite others to challenge or verify our findings and extend, update, or modify the research design and paper corpus.

Author contributions

The contributions of all authors are based on the Contributor Roles Taxonomy (CRediT [8]). All authors have read and approved the final version. No generative AI tool was used in the conceptualization of the study or writing of the paper, except for assistance of coding as indicated in the repository [16].

Carlos Granell: Conceptualization, data curation, investigation, methodology, software, visualization, writing – original draft, writing – review & editing
Frank Ostermann: Conceptualization, data curation, investigation, formal analysis, methodology, software, writing – original draft, writing – review & editing
Daniel Nüst: Conceptualization, methodology, software, data curation, investigation, writing – review & editing
Peter Kedron: Data curation, methodology, investigation, writing – review & editing
Eftychia Koukouraki: Data curation, investigation, writing – review & editing
Miguel Matey: Data curation, investigation, writing – review & editing
Rémy Decoupes: Data curation, investigation, writing – review & editing
Sergio Trilles: Data curation, investigation, writing – review & editing
Anita Graser: Data curation, investigation, writing – review & editing
Tom Niers: Data curation, investigation

Acknowledgments

We thank the Association of Geographic Information Laboratories in Europe (AGILE, https://agile-gi.eu/) community and, in particular, the AGILE Council, which promoted the introduction of the AGILE Guidelines and the reproducibility review process at the AGILE conference series. DN was in part supported by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, https://nfdi4earth.de/) within the German National Research Data Infrastructure (NFDI, https://www.nfdi.de/).

[8] https://doi.org/10.3789/ansi.niso.z39.104-2022

References

[1] *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C., 2019. doi:10.17226/25303.

[2] ACZEL, B., SZASZI, B., NILSONNE, G., VAN DEN AKKER, O. R., ALBERS, C. J., VAN ASSEN, M. A., BASTIAANSEN, J. A., BENJAMIN, D., BOEHM, U., BOTVINIK-NEZER, R., BRINGMANN, L. F., BUSCH, N. A., CARUYER, E., CATALDO, A. M., COWAN, N., DELIOS, A., VAN DONGEN, N. N., DONKIN, C., VAN DOORN, J. B., DREBER, A., DUTILH, G., EGAN, G. F., GERNSBACHER, M. A., HOEKSTRA, R., HOFFMANN, S., HOLZMEISTER, F., HUBER, J., JOHANNESSON, M., JONAS, K. J., KINDEL, A. T., KIRCHLER, M., KUNKELS, Y. K., LINDSAY, D. S., MANGIN, J.-F., MATZKE, D., MUNAFÒ, M. R., NEWELL, B. R., NOSEK, B. A., POLDRACK, R. A., VAN RAVENZWAAIJ, D., RIESKAMP, J., SALGANIK, M. J., SARAFOGLOU, A., SCHONBERG, T., SCHWEINSBERG, M., SHANKS, D., SILBERZAHN, R., SIMONS, D. J., SPELLMAN, B. A., ST-JEAN, S., STARNS, J. J., UHLMANN, E. L., WICHERTS, J., AND WAGENMAKERS, E.-J. Science forum: Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife 10* (2021), e72185. doi:10.7554/eLife.72185.

[3] BARBA, L. A. Terminologies for Reproducible Research, 2018. doi:10.48550/arXiv.1802.03311.

[4] BREZNAU, N., RINKE, E. M., WUTTKE, A., ADEM, M., ADRIAANS, J., AKDENIZ, E., ALVAREZ-BENJUMEA, A., ANDERSEN, H. K., AUER, D., AZEVEDO, F., BAHNSEN, O., BAI, L., BALZER, D., BAUER, P. C., BAUER, G., BAUMANN, M., BAUTE, S., BENOIT, V., BERNAUER, J., BERNING, C., BERTHOLD, A., BETHKE, F. S., BIEGERT, T., BLINZLER, K., BLUMENBERG, J. N., BOBZIEN, L., BOHMAN, A., BOL, T., BOSTIC, A., BRZOZOWSKA, Z., BURGDORF, K., BURGER, K., BUSCH, K., CASTILLO, J.-C., CHAN, N., CHRISTMANN, P., CONNELLY, R., CZYMARA, C. S., DAMIAN, E., DE ROOIJ, E. A., ECKER, A., EDELMANN, A., EDER, C., EGER, M. A., ELLERBROCK, S., FORKE, A., FORSTER, A., FREIRE, D., GAASENDAM, C., GAVRAS, K., GAYLE, V., GESSLER, T., GNAMBS, T., GODEFROIDT, A., GRÖMPING, M., GROSS, M., GRUBER, S., GUMMER, T., HADJAR, A., HALBHERR, V., HEISIG, J. P., HELLMEIER, S., HEYNE, S., HIRSCH, M., HJERM, M., HOCHMAN, O., HÖFFLER, J. H., HÖVERMANN, A., HUNGER, S., HUNKLER, C., HUTH-STÖCKLE, N., IGNÁCZ, Z. S., ISRAEL, S., JACOBS, L., JACOBSEN, J., JAEGER, B., JUNGKUNZ, S., JUNGMANN, N., KANJANA, J., KAUFF, M., KHAN, S., KHATUA, S., KLEINERT, M., KLINGER, J., KOLB, J.-P., KOŁCZYŃSKA, M., KUK, J., KUNISSEN, K., KURTI SINATRA, D., LANGENKAMP, A., LEE, R. C., LERSCH, P. M., LIU, D., LÖBEL, L.-M., LUTSCHER, P., MADER, M., MADIA, J. E., MALANCU, N., MALDONADO, L., MARAHRENS, H., MARTIN, N., MARTINEZ, P., MAYERL, J., MAYORGA, O. J., MCDONNELL, R., MCMANUS, P., MCWAGNER, K., MEEUSEN, C., MEIERRIEKS, D., MELLON, J., MERHOUT, F., MERK, S., MEYER, D., MICHELI, L., MIJS, J., MOYA, C., NEUNHOEFFER, M., NÜST, D., NYGÅRD, O., OCHSENFELD, F., OTTE, G., PECHENKINA, A., PICKUP, M., PROSSER, C., RAES, L., RALSTON, K., RAMOS, M., REICHERT, F., ROETS, A., ROGERS, J., ROPERS, G., SAMUEL, R., SAND, G., SANHUEZA PETRARCA, C., SCHACHTER, A., SCHAEFFER, M., SCHIEFERDECKER, D., SCHLUETER, E., SCHMIDT, K., SCHMIDT, R., SCHMIDT-CATRAN, A., SCHMIEDEBERG, C., SCHNEIDER, J., SCHOONVELDE, M., SCHULTE-CLOOS, J., SCHUMANN,

S., SCHUNCK, R., SEURING, J., SILBER, H., SLEEGERS, W., SONNTAG, N., STAUDT, A., STEIBER, N., STEINER, N. D., STERNBERG, S., STIERS, D., STOJMENOVSKA, D., STORZ, N., STRIESSNIG, E., STROPPE, A.-K., SUCHOW, J. W., TELTEMANN, J., TIBAJEV, A., TUNG, B., VAGNI, G., VAN ASSCHE, J., VAN DER LINDEN, M., VAN DER NOLL, J., VAN HOOTEGEM, A., VOGTENHUBER, S., VOICU, B., WAGEMANS, F., WEHL, N., WERNER, H., WIERNIK, B. M., WINTER, F., WOLF, C., WU, C., YAMADA, Y., ZAKULA, B., ZHANG, N., ZILLER, C., ZINS, S., ŻÓŁTAK, T., AND NGUYEN, H. H. The reliability of replications: a study in computational reproductions. *Royal Society Open Science 12*, 3 (2025). doi:10.1098/rsos.241038.

[5] BRODEUR, A., MIKOLA, D., AND COOK, N. Mass Reproducibility and Replicability: A New Hope. SSRN Scholarly Paper 16912, IZA Discussion Paper, Rochester, NY, 2024. Available at SSRN: https://ssrn.com/abstract=4790780.

[6] BRUNSDON, C., AND COMBER, A. Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems 23*, 4 (Aug. 2021), 477–496. 10.1007/s10109-020-00334-2.

[7] CANNON, M., KELLY, A., AND FREEMAN, C. Implementing an open & FAIR data sharing policy – a case study in the earth and environmental sciences. *Learned Publishing 35*, 1 (2022), 56–66. doi:10.1002/leap.1442.

[8] CHANG, J. R., AND NORDLING, T. E. M. ChatGPT struggles to recognize reproducible science. *Knowledge and Information Systems 67*, 8 (2025), 6825–6843. doi:10.1007/s10115-025-02428-z.

[9] CHIARELLI, A., LOFFREDA, L., AND JOHNSON, R. *The Art of Publishing Reproducible Research Outputs: Supporting emerging practices through cultural and technological innovation.* 2021. doi:10.5281/ZENODO.5521077.

[10] DAVIS-STOBER, C. P., SARAFOGLOU, A., ACZEL, B., CHANDRAMOULI, S. H., ERRINGTON, T. M., FIELD, S. M., FISHBACH, A., FREIRE, J., IOANNIDIS, J. P. A., OBERAUER, K., PESTILLI, F., RESSL, S., SCHAD, D. J., TER SCHURE, J., TENTORI, K., VAN RAVENZWAAIJ, D., VANDEKERCKHOVE, J., AND GUNDERSEN, O. E. How can we make sound replication decisions? *Proceedings of the National Academy of Sciences 122*, 5 (2025), e2401236121. doi:10.1073/pnas.2401236121.

[11] DI COSMO, R., GRANGER, S., HINSEN, K., JULLIEN, N., LE BERRE, D., LOUVET, V., MAUMET, C., MAURICE, C., MONAT, R., AND ROUGIER, N. P. Stop treating code like an afterthought: record, share and value it. *Nature 646*, 8084 (Oct. 2025), 284–286. doi:10.1038/d41586-025-03196-0.

[12] DUDDA, L., KORMANN, E., KOZULA, M., DEVITO, N. J., KLEBEL, T., DEWI, A. P. M., SPIJKER, R., STEGEMAN, I., VAN DEN EYNDEN, V., ROSS-HELLAUER, T., AND LEEFLANG, M. M. G. Open science interventions to improve reproducibility and replicability of research: a scoping review. *Royal Society Open Science 12*, 4 (2025), 242057. doi:10.1098/rsos.242057.

[13] FIŠAR, M., GREINER, B., HUBER, C., KATOK, E., AND OZKES, A. I. Reproducibility in management science. *Management Science 70*, 3 (2024), 1343–1356. doi:10.1287/mnsc.2023.03556.

[14] GOODCHILD, M. F., AND LI, W. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences 118*, 35 (2021), e2015759118. doi:10.1073/pnas.2015759118.

[15] GRANELL, C., OSTERMANN, F. O., NÜST, D., KEDRON, P., KOUKOURAKI, E., MATEY-SANZ, M., DECOUPES, R., TRILLES, S., GRASER, A., AND NIERS, T. Longitudinal assessment of research in GIScience domain shows a positive impact of reproducible research practices. *EarthArXiv* (2025). doi:10.31223/X5RJ3W.

[16] GRANELL, C., OSTERMANN, F. O., NÜST, D., KEDRON, P., KOUKOURAKI, E., MATEY-SANZ, M., DECOUPES, R., TRILLES, S., GRASER, A., AND NIERS, T. Reproducibility Package for 'Improving reproducibility of GIScience publications through novel reproducibility guidelines and revised review procedures', 2026. doi:10.5281/zenodo.21097308.

[17] GRANELL, C., OSTERMANN, F. O., NÜST, D., KEDRON, P., KOUKOURAKI, E., MATEY-SANZ, M., DECOUPES, R., TRILLES, S., GRASER, A., AND NIERS, T. Software Heritage Deposition for repository 'nuest/reproducible-research-giscience-longitudinal-study', 2026. https://archive.softwareheritage.org/swh:1:dir:d91b644451add768e90efbcce40976ccb525590b.

[18] HARDWICKE, T. E., MATHUR, M. B., MACDONALD, K., NILSONNE, G., BANKS, G. C., KIDWELL, M. C., HOFELICH MOHR, A., CLAYTON, E., YOON, E. J., HENRY TESSLER, M., LENNE, R. L., ALTMAN, S., LONG, B., AND FRANK, M. C. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science 5*, 8 (2018), 180448. doi:10.1098/rsos.180448.

[19] HARDWICKE, T. E., WALLACH, J. D., KIDWELL, M. C., BENDIXEN, T., CRÜWELL, S., AND IOANNIDIS, J. P. A. An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *Royal Society Open Science 7*, 2 (2020), 190806. doi:10.1098/rsos.190806.

[20] HILDEMANN, M. J., MURRAY, A. T., AND VERSTEGEN, J. A. Genetic programming for computationally efficient land use allocation optimization. In *12th International Conference on Geographic Information Science (GIScience 2023)* (2023), Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPICS.GISCIENCE.2023.4.

[21] JANOWICZ, K., ZHU, R., VERSTEGEN, J., MCKENZIE, G., MARTINS, B., AND CAI, L. Six GIScience ideas that must die. *AGILE: GIScience Series 3* (2022), 7. doi:10.5194/agile-giss-3-7-2022.

[22] KAMBOURIS, S., WILKINSON, D. P., SMITH, E. T., AND FIDLER, F. Computationally reproducing results from meta-analyses in ecology and evolutionary biology using shared code and data. *PLOS ONE 19*, 3 (2024), 1–22. doi:10.1371/journal.pone.0300333.

[23] KEDRON, P., BARDIN, S., HOFFMAN, T. D., SACHDEVA, M., QUICK, M., AND HOLLER, J. A replication of DiMaggio et al. (2020) in Phoenix, AZ. *Annals of Epidemiology 74* (2022), 8–14. doi:10.1016/j.annepidem.2022.05.005.

[24] KEDRON, P., BARDIN, S., HOLLER, J., GILMAN, J., GRADY, B., SEELEY, M., WANG, X., AND YANG, W. A framework for moving beyond computational reproducibility: Lessons from three reproductions of geographical analyses of COVID-19. *Geographical Analysis 56*, 1 (2024), 163–184. doi:10.1111/gean.12370.

[25] KEDRON, P., FRAZIER, A. E., TRGOVAC, A. B., NELSON, T., AND FOTHERINGHAM, A. S. Reproducibility and replicability in geographical analysis. *Geographical Analysis 53*, 1 (2021), 135–147. doi:10.1111/gean.12221.

[26] KEDRON, P., AND HOLLER, J. Replication and the search for the laws in the geographic sciences. *Annals of GIS 28*, 1 (2022), 45–56. doi:10.1080/19475683.2022.2027011.

[27] KEDRON, P., HOLLER, J., AND BARDIN, S. Reproducible research practices and barriers to reproducible research in geography: Insights from a survey. *Annals of the American Association of Geographers 114*, 2 (2024), 369–386. doi:10.1080/24694452.2023.2276115.

[28] KEDRON, P., HOLLER, J., AND BARDIN, S. A survey of researcher perceptions of replication in geography. *Annals of the American Association of Geographers 115*, 1 (2025), 184–204. doi:10.1080/24694452.2024.2415695.

[29] KEDRON, P., LI, W., FOTHERINGHAM, S., AND GOODCHILD, M. Reproducibility and replicability: opportunities and challenges for geospatial research. *International Journal of Geographical Information Science 35*, 3 (2021), 427–445. doi:10.1080/13658816.2020.1802032.

[30] KONKOL, M., KRAY, C., AND PFEIFFER, M. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science 33*, 2 (2019), 408–429. doi:10.1080/13658816.2018.1508687.

[31] KOUKOURAKI, E., DEGBELO, A., AND KRAY, C. Assessing Map Reproducibility with Visual Question-Answering: An Empirical Evaluation. In *13th International Conference on Geographic Information Science (GIScience 2025)* (Dagstuhl, Germany, 2025), K. Sila-Nowicka, A. Moore, D. O'Sullivan, B. Adams, and M. Gahegan, Eds., vol. 346 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 13:1–13:12. doi:10.4230/LIPIcs.GIScience.2025.13.

[32] KOUKOURAKI, E., AND KRAY, C. Map Reproducibility in Geoscientific Publications: An Exploratory Study. In *12th International Conference on Geographic Information Science (GIScience 2023)* (Dagstuhl, Germany, 2023), R. Beecham, J. A. Long, D. Smith, Q. Zhao, and S. Wise, Eds., vol. 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 6:1–6:16. doi:10.4230/LIPIcs.GIScience.2023.6.

[33] KOUKOURAKI, E., AND KRAY, C. A systematic approach for assessing the importance of visual differences in reproduced maps. *Cartography and Geographic Information Science 53*, 3 (2026), 304–319. doi:10.1080/15230406.2024.2409920.

[34] LI, W., HSU, C.-Y., WANG, S., AND KEDRON, P. GeoAI reproducibility and replicability: A computational and spatial perspective. *Annals of the American Association of Geographers 114*, 9 (2024), 2085–2103. doi:10.1080/24694452.2024.2373787.

[35] MUNAFÒ, M., NOSEK, B., BISHOP, D., ET AL. A manifesto for reproducible science. *Nature Human Behaviour 1* (2017), 21. doi:10.1038/s41562-016-0021.

[36] NÜST, D., AND EGLEN, S. J. CODECHECK: An Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. *F1000Research 10* (2021). doi:10.12688/f1000research.51738.2.

[37] NÜST, D., GRANELL, C., HOFER, B., KONKOL, M., OSTERMANN, F. O., SILERYTE, R., AND CERUTTI, V. Reproducible research and GIScience: an evaluation using AGILE conference papers. *PeerJ 6* (2018), e5072. doi:10.7717/peerj.5072.

[38] NÜST, D., GRANELL, C., AND OSTERMANN, F. O. Impact of reproducible paper guidelines on computational papers: A longitudinal study on the agile and giscience conference series, 2023. doi:10.17605/OSF.IO/XZJCH.

[39] NÜST, D., MOMIN, A., EGLEN, S., DAVIES, I., AND GUIMARÃES, J. Codecheckers/register: CODECHECK Register Deposit July 2025, 2025.

[40] NÜST, D., OSTERMANN, F. O., GRANELL, C., AND KMOCH, A. Improving reproducibility of geospatial conference papers – lessons learned from a first implementation of reproducibility reviews. *Septentrio Conference Series*, 4 (2020). doi:10.7557/5.5601.

[41] NÜST, D., OSTERMANN, F. O., SILERYTE, R., HOFER, B., GRANELL, C., TEPEREK, M., GRASER, A., BROMAN, K., HETTNE, K., CLARE, C., BELLIARD, F., AND WANG, Y. AGILE Reproducible Paper Guidelines, 2020. doi:10.17605/OSF.IO/CB7Z8.

[42] OBELS, P., LAKENS, D., COLES, N. A., GOTTFRIED, J., AND GREEN, S. A. Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science 3*, 2 (2020), 229–237. doi:10.1177/2515245920918872.

[43] OSTERMANN, F. O., NÜST, D., GRANELL, C., HOFER, B., AND KONKOL, M. Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers. In *11th International Conference on Geographic Information Science (GIScience 2021) - Part II* (Dagstuhl, Germany, 2021), K. Janowicz and J. A. Verstegen, Eds., vol. 208 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 2:1–2:16. doi:10.4230/LIPIcs.GIScience.2021.II.2.

[44] OSTERMANN, F. O., NÜST, D., AND GRANELL, C. Assessment Protocol for the assessment of papers for the longitudinal study on the potential reproducibility of research in GIScience, 2026. doi:10.17605/OSF.IO/ZYM6Q.

[45] PAEZ, A. Reproducibility of research during COVID-19: Examining the case of population density and the basic reproductive rate from the perspective of spatial analysis. *Geographical Analysis 54*, 4 (2022), 860–880. doi:10.1111/gean.12307.

[46] SHEHZAD, F., BREUER, T., MAISTRO, M., AND JANNACH, D. "We Share Our Code Online": Why this is not enough to ensure reproducibility and progress in recommender systems research. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems* (New York, NY, USA, 2025), RecSys '25, Association for Computing Machinery, p. 884–893. doi:10.1145/3705328.3748157.

[47] SHI, M., CURRIER, K., LIU, Z., JANOWICZ, K., WIEDEMANN, N., VERSTEGEN, J., MCKENZIE, G., GRASER, A., ZHU, R., AND MAI, G. Thinking geographically about AI sustainability. *AGILE: GIScience Series 4* (2023), 42. doi:10.5194/agile-giss-4-42-2023.

[48] STARK, P. B. Before reproducibility must come preproducibility. *Nature 557*, 7706 (2018), 613–614. doi:10.1038/d41586-018-05256-0.

[49] STODDEN, V., SEILER, J., AND MA, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences 115*, 11 (2018), 2584–2589. doi:10.1073/pnas.1708290115.

[50] SUI, D., AND KEDRON, P. Reproducibility and replicability in the context of the contested identities of geography. *Annals of the American Association of Geographers 111*, 5 (2021), 1275–1283. doi:10.1080/24694452.2020.1806024.

[51] VAN DE WEGHE, N., DE SLOOVER, L., COHN, A., HUANG, H., SCHEIDER, S., SIEBER, R., TIMPF, S., AND CLARAMUNT, C. Opportunities and challenges of integrating geographic information science and large language models. *Journal of Spatial Information Science*, 30 (2025), 93–116. doi:10.5311/josis.2025.30.389.

[52] WAINWRIGHT, J. Is critical human geography research replicable? *Annals of the American Association of Geographers 111*, 5 (2021), 1284–1290. doi:10.1080/24694452.2020.1806025.