# Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers

**Frank O. Ostermann** [1]
Faculty of Geo-Information Science and Earth Observation (ITC),
University of Twente, Enschede, The Netherlands

**Daniel Nüst**
Institute for Geoinformatics, University of Münster, Germany

**Carlos Granell**
Institute of New Imaging Technologies, Universitat Jaume I de Castellón, Spain

**Barbara Hofer**
Christian Doppler Laboratory GEOHUM and Department of Geoinformatics - Z_GIS,
University of Salzburg, Austria

**Markus Konkol**
Faculty of Geo-Information Science and Earth Observation (ITC),
University of Twente, Enschede, The Netherlands

## Abstract

GIScience conference authors and researchers face the same computational reproducibility challenges as authors and researchers from other disciplines who use computers to analyse data. Here, to assess the reproducibility of GIScience research, we apply a rubric for assessing the reproducibility of 75 conference papers published at the GIScience conference series in the years 2012-2018. Since the rubric and process were previously applied to the publications of the AGILE conference series, this paper itself is an attempt to replicate that analysis, however going beyond the previous work by evaluating and discussing proposed measures to improve reproducibility in the specific context of the GIScience conference series. The results of the GIScience paper assessment are in line with previous findings: although descriptions of workflows and the inclusion of the data and software suffice to explain the presented work, in most published papers they do not allow a third party to reproduce the results and findings with a reasonable effort. We summarise and adapt previous recommendations for improving this situation and propose the GIScience community to start a broad discussion on the reusability, quality, and openness of its research. Further, we critically reflect on the process of assessing paper reproducibility, and provide suggestions for improving future assessments. The code and data for this article are published at `https://doi.org/10.5281/zenodo.4032875`.

**2012 ACM Subject Classification** Information systems → Geographic information systems

**Keywords and phrases** reproducible research, open science, reproducibility, GIScience

**Digital Object Identifier** 10.4230/LIPIcs.GIScience.2021.II.2

**Related Version** *Previous Version*: `https://doi.org/10.31223/X5ZK5V`

**Supplementary Material** The input data for this work are the full texts of GIScience conference proceedings from the years 2012 to 2018 [35, 7, 20, 34]. The paper assessment results and source code of figures are published at `https://github.com/nuest/reproducible-research-at-giscience` and archived on Zenodo [27]. The used computing environment is containerised with Docker pinning the R version to 3.6.3 and R packages to the MRAN snapshot of July 5th 2019.

**Funding** *Daniel Nüst*: Project *o2r*, German Research Foundation, grant number PE 1632/17-1.
*Carlos Granell*: Ramón y Cajal Programme of the Spanish government, grant number RYC-2014-16913.
*Markus Konkol*: Project *o2r*, German Research Foundation, grant numbers KR 3930/8-1 and TR 864/12-1.

[1] Corresponding author

11th International Conference on Geographic Information Science (GIScience 2021) – Part II.
Editors: Krzysztof Janowicz and Judith A. Verstegen; Article No. 2; pp. 2:1–2:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

**Acknowledgements** Author contributions (see CRediT): *all authors* contributed to conceptualisation, investigation (number of assessed papers in brackets), and writing – original draft; FO (33): writing - review & editing, software; DN (33): software, writing - review & editing, visualisation; CG (30): writing - review & editing, software; BH (21): writing - review & editing; MK (30). We thank Celeste R. Brennecka from the Scientific Editing Service of the University of Münster for her editorial support and the anonymous reviewers for their constructive feedback.

## 1 Introduction

The past two decades have seen the imperative of Open Science gain momentum across scientific disciplines. The adoption of Open Science practices is partially prompted by the increasing costs of using proprietary software and subscribing to scientific journals, but more importantly because of the increased transparency and availability of data, methods, and results, which enable reproducibility [22]. This advantage is especially relevant for the computational and natural sciences, where sharing data and code is a prerequisite for reuse and collaboration. A large proportion of GIScience research today uses software to analyse data on computers, meaning that many articles published in the context of the GIScience conference series[2] fall into the categories of data science or computational research. Thereby, these articles face challenges of transparency and reproducibility in the sense of the Claerbout/Donoho/Peng terminology [2], where *reproduction* means a recreation of the same results using the same input data and methods, usually with the actual code created by the original authors. The related concept of replication, i.e., the confirmation of insights gained from a scientific study using the same method with new data, is of crucial importance to scientific progress, yet it is also frequently challenging to realise for interested readers of a published study. So far, despite the GIScience conference series' rigorous review process, reproducibility and replicability have not been a core concern in the contributions. With reproducibility now being a recognised topic in the call for papers, it is time to take stock and identify possible action. In previous work [26], we assessed the reproducibility of a selection of full and short papers from the AGILE conference series[3], a community conference organised by member labs of the Association of Geographic Information Laboratories in Europe (AGILE). Using systematic analysis based on a rubric for reproducible research, we found that the majority of AGILE papers neither provided sufficient information for a reviewer to evaluate the code and data and attempt a reproduction, nor enough material for readers to reuse or extend data or code from the analytical workflows. This is corroborated by research in related disciplines such as quantitative geography [3], qualitative GIS [21], geoscience [16], and e-Science [10]. The problems identified in these related research areas are transferable to the scientific discipline of GIScience, which operates at the intersections of aforementioned fields [11]. In any case, observations on the lack of reproducibility in all scientific fields contrast with the clear advantages and benefits of open and reproducible research both for individuals and for academia as a whole (cf. for example [6, 19, 17, 5]). As a consequence, we have initiated a process to support authors in increasing reproducibility for AGILE publications; as a main outcome, this initiative has produced author guidelines as well as strategies for the AGILE conference series[4].

[2] `https://www.giscience.org/`

[3] `https://agile-online.org/conference`

[4] See the initiative website at `https://reproducible-agile.github.io/`, the author guidelines at `https://doi.org/10.17605/OSF.IO/CB7Z8` [24] and the main OSF project with all materials `https://osf.io/phmce/` [25].

The AGILE conference is related to GIScience conference in terms of scientific domain and contributing authors, but is different in organisational aspects. Two open questions are thus whether the GIScience conference series faces the same issues, and whether similar strategies could be applied successfully. To begin this investigation, we conducted a simple text analysis of GIScience conference proceedings[5] to evaluate the relevance of computational methods in the conference papers. The analysis searched for several word stems related to reproducibility: Generic words indicating a quantitative analysis, e.g., “data”, “software”, or “process”; specific platforms, e.g., “GitHub”; and concrete terms, e.g., words starting with “reproduc” or “replic”. Table 1 shows the results of the search for each year analysed. The take-away message from the text analysis is that algorithms, processing, and data play an essential role in GIScience publications, but few papers mentioned code repositories or reproduction materials. Therefore, an in-depth assessment of the reproducibility of these publications was deemed necessary.

The main contribution of this work addresses two objectives: First, it aims to investigate the state of reproducibility in the GIScience conference community. This investigation broadens our knowledge base about reproducibility in the GIScience discipline and informs us about the situation in the GIScience conference series specifically (details in section 4). Second, it aims to apply the assessment procedure used for AGILE conference papers (presented in section 3) to the papers of the GIScience conference, so that the broader suitability of this procedure is evaluated using a different dataset, and thereby providing evidence of its replicability. Such a transfer validates the developed methodology. We discuss these findings and present our conclusions in the final two sections (5 and 6). Together, these objectives yield important findings for the discussion of reproducibility within the GIScience conference community and the GIScience discipline at large. We believe that GIScience as a discipline would greatly benefit from more studies that reproduce and replicate other studies, similar to other disciplines that are recognising the value of replication for innovating theory [23], and argue that such a replication study is not lacking innovation but is a prerequisite for innovating community practice. Only then can a fruitful dialogue take place on whether and how to improve reproducibility for the GIScience conference series, and whether the recent steps taken at AGILE[6] could be an inspiration for GIScience conferences as well.

## 2 Related work

This work builds and expands on earlier work [26], which already provides an overview of reproducible research in general, including definitions, challenges, and shortcomings. In the following, we focus therefore on recently published works and briefly introduce related meta-studies.

Few groups have attempted practical reproduction of computational works related to GIScience. Konkol et al. [16] conducted an in-depth examination of the computational reproducibility of 41 geoscience papers with a focus on differences between the recreated figures. The set of papers was, similar to our work, drawn from a fixed group of two outlets

[5] The full text analysis and the results is available in this paper’s repository in the following files: `giscience-historic-text-analysis.Rmd` contains the analysis code; the result data are two tables with counts for occurrences of words respectively word stems per year in `results/text_analysis_topwordstems.csv` and `results/text_analysis_keywordstems.csv`; a wordcloud per year is in file `results/text_analysis_wordstemclouds.png`.

[6] See the initiative website at `https://reproducible-agile.github.io/`, the author guidelines at `https://doi.org/10.17605/OSF.IO/CB7Z8` [24] and the main OSF project with all materials `https://osf.io/phmce/` [25].

**Table 1** Reproducibility-related word stems in the corpus per year of proceedings.

| year | words | reproduc.. | replic.. | repeatab.. | code | software | algorithm(s) | (pre)process.. | data.* | result(s) | repository/ies | github/lab |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2002 | 23782 | 6 | 2 | 0 | 11 | 61 | 191 | 150 | 897 | 129 | 62 | 0 |
| 2004 | 26728 | 4 | 1 | 0 | 34 | 50 | 138 | 258 | 849 | 263 | 4 | 0 |
| 2006 | 32758 | 6 | 0 | 0 | 12 | 32 | 335 | 250 | 856 | 164 | 0 | 0 |
| 2008 | 27356 | 3 | 6 | 1 | 3 | 11 | 331 | 146 | 854 | 218 | 17 | 0 |
| 2010 | 23004 | 3 | 1 | 0 | 8 | 16 | 164 | 276 | 650 | 162 | 0 | 0 |
| 2012 | 28860 | 2 | 0 | 0 | 101 | 27 | 238 | 190 | 1048 | 311 | 3 | 0 |
| 2014 | 29534 | 3 | 4 | 1 | 12 | 18 | 255 | 159 | 1070 | 228 | 3 | 0 |
| 2016 | 24838 | 2 | 0 | 0 | 23 | 21 | 333 | 150 | 1007 | 202 | 4 | 1 |
| 2018 | 23318 | 3 | 10 | 0 | 15 | 15 | 201 | 160 | 891 | 294 | 6 | 6 |
| Total | 240178 | 32 | 24 | 2 | 219 | 251 | 2186 | 1739 | 8122 | 1971 | 99 | 7 |

*Note:* The very high value for ’code’ in 2012 is due to a single paper about land use, for which different “land use codes” are defined, discussed and used.

(journals), but it was further limited to recent papers providing code in the R language. The main issues raised by Konkol et al. [16] are similar to those identified in a recent report on the reproducibility review during the AGILE conference 2020[7], where the reproducibility committee summarised the process and documented relevant obstacles to reproducibility of accepted papers.

Within the geospatial domain, Kedron et al. [13] provide a recent review of opportunities and challenges for reproducibility and replicability. They transfer solutions from other domains but also discuss and conceptualise the specific nature of a reproducibility and replicability framework when working with geospatial data, e.g., handling context, uncertainty of spatial processes, or how to accommodate the inherent natural variability of geospatial systems. In a similar manner, Brunsdon and Comber [4] investigate reproducibility within spatial data science, with special attention to big spatial data. They support the need for open tools, knowledge about code, and reproducibility editors at domain journals and conferences, but they also introduce the perspective that spatial analysis is no longer conducted only by GI/geo-scientists or geographers and connect reproducibility with critical spatial understanding. The more conceptual work in those articles is complemented by the assessment of reproducibility conducted in this paper.

Two recent studies from distant disciplines, wildlife science [1] and hydrology [32], also relate to our work in this paper. Both studies investigate a random set of articles from selected journals and use a stepwise process of questions to determine the availability of materials and eventually reproduce workflows if possible. Archmiller et al. [1] use a final ranking of 1 to 5 to specify the degree to which a study’s conclusions were eventually reproduced. Similar to our classification scheme, their ranking models fit the general notion of a *“reproducibility spectrum”* [30].

[7] `https://osf.io/7rjpe/`

# 3 Reproducibility assessment method

## 3.1 Criteria

The assessment criteria used for the current study were originally defined in previous work, so we provide only a short introduction here and refer to Nüst et al. [26] for details. The three assessment criteria are *Input Data*, *Methods*, and *Results*. *Input Data* comprises all datasets that the computational analysis uses. *Methods* encompasses the entire computational analysis that generates the results. Since *Methods* is difficult to evaluate as a whole, we split this criterion into three subcriteria: *Preprocessing* includes the steps to prepare the *Input Data* before the main analysis; *Methods, Analysis, Processing* is the main analysis; *Computational Environment* addresses the description of hard- and software. Finally, the criterion *Results* refers to the output of analysis, e.g., figures, tables, and numbers.

For each of these (sub)criteria, we assigned one of four levels unless the criterion was not applicable (*NA*). *Unavailable* (level 0) means that it was not possible to access the paper's data, methods, or results, and that it was impossible to recreate them based on the description in the paper. *Documented* (level 1) indicates that the paper still did not provide direct access to datasets, methods, or results, but that there was sufficient description or metadata to potentially recreate them closely enough for an evaluation; yet, often a recreation was unlikely due to the huge amount of effort needed. For example, with regard to the methods criteria, *Documented* means that pseudo code or a textual workflow description was provided. *Available* (level 2) was assigned if the paper provided direct access to the materials (e.g., through a link to a personal or institutional website), but not in the form of an open and permanent identifier, such as a digital object identifier (DOI). The indication of a DOI does not apply to the methods criteria, as it is not yet common practice to make a permanent reference to code, libraries, and system environments with a single identifier. The gold standard, *Available and Open* (level 3), requires open and permanent access to the materials (e.g., through public online repositories) and open licenses to allow use and extension.

Note that levels are ordinal numbers that can be compared (3 is higher than 2), but absolute differences between numbers must not be interpreted as equals: Moving one level up from 0 to 1 is not the same as from level 1 to level 2. While reaching level 1 is fairly straightforward, moving to level 2 means one must create a fully reproducible paper.

## 3.2 Process

The overall approach to assessing the reproducibility of GIScience papers followed the previous assessment of AGILE papers [26], and was conducted by the same persons. Contrary to the AGILE investigation, all full papers in the GIScience conference series (from the 2012 to 2018 editions) were assessed. This is partly because no obvious subset exists, such as the nominees for best papers as in the case of the AGILE conference series, but also because we aimed to work with a larger dataset for potentially more informative results. Each GIScience conference paper was randomly assigned to two assessors who evaluated it qualitatively according to the reproducibility criteria. The assessors were free in the way they approached the assigned evaluations, depending on the structure of the paper and the assessor's familiarity with the topic. An evaluation could range from browsing the paper to identify relevant statements in case of high familiarity to a thorough reading of the full text. The identification of relevant content could be supported to some extent by a PDF reader with multiple highlights, using keywords like e.g., "data, software, code, download, contribution, script, workflow". The results of the individual assessments were joined in a collaborative Google Spreadsheet. This spreadsheet also had a comments column for assessors to record relevant sources and decisions. In case of disagreement between assessors, arguments for and against a certain reproducibility level were discussed in the entire group of five assessors until a consensus was reached. Only then were the assessments merged into a single value. A snapshot of both the unmerged and merged values was stored as a CSV file in the collaboration repository for transparency and provenance[8]. Two independent assessors per paper increased the objectivity of the final assessment. Disagreements and conducting the assessment one year at a time, going backwards from the most recent year, were found helpful in aligning the interpretation of criteria and, in rare cases, led to an adjustment of similar cases in other papers.

The discussion about the correct assignment of levels led to a reflection on how to apply the rubric for special situations. For the *Input Data* criterion, some papers had input data "available" at the time of writing/publication that was not available anymore at the time of evaluation, due to broken links, changes in the URL structure of a website, or projects and/or personal websites that were down or moved. In such cases, we gave the authors the benefit of the doubt and assumed the data were accessible some time after the publication of the conference proceedings. We did not give those papers an arbitrary score and discussed internally the best level per case; yet, such papers never earned a `3`, which would require permanent resolving of the link. Related to this criterion, simulation data, like the specification or configuration of agents in an agent-based system, was not treated as input data (resulting in *NA* if no other data was used), but as parameters of the main analysis, i.e., as part of the *Methods, Analysis, Processing.*

*Preprocessing* covers preparatory work for the actual analysis involving various tasks such as data selection, cleaning, aggregation, and integration. However, the dividing line between data preprocessing and processing (i.e., the main analysis) proved to be often vague, and occasionally assessors disagreed whether the preprocessing criterion should be assigned *NA*, *Unavailable*, or *Documented* (`0` or `1`, respectively). Therefore, we decided eventually to apply the *Preprocessing* criterion only in cases where papers specifically mentioned a preprocessing task independent of the actual analysis or method, e.g., when clearly stated in a separate sub-section of the paper.

Lastly, human subject tests and surveys were also a special case. Human-related research activities were rated as `1` in the methods/analysis/processing criterion if sufficiently documented; nonetheless, a sufficient documentation in these cases did not mean that original sources were available or could be exactly recreated.

## 3.3 Paper corpus

In total, 87 papers from the GIScience conferences in 2012, 2014, 2016, and 2018 were assessed. A table in the reproducibility package shows the full results of the assessment and the included raw data provides details on assigned assessors, authors, etc. [27]. 12 papers (14%) across all years were identified as conceptual papers[9] and were not included in the corpus. The number of conceptual papers in GIScience conferences was low over the analysed

[8] The assessment results are in the file `results/paper_assessment.csv`. As an example, commit `464e630` and `2e8b1be` are the pre-merge and post-merge commit after completing the assessment of the papers from 2014. The pre-merge commit contains the assessments including the assessors' initials, e.g. "CG: 1, MK: 1".

[9] See [26] for a definition of "conceptual".

**Table 2** Statistics of reproducibility levels per criterion (rounded to one decimal place).

| | input data | preproc. | method/analysis/proc. | comp. env. | results |
|---|---|---|---|---|---|
| Min. | 0.0 | 0.0 | 0 | 0.0 | 0.0 |
| Median | 1.0 | 1.0 | 1 | 0.0 | 1.0 |
| Mean | 0.7 | 0.8 | 1 | 0.3 | 1.1 |
| Max. | 2.0 | 2.0 | 2 | 1.0 | 2.0 |
| NA's | 1.0 | 24.0 | 0 | 0.0 | 0.0 |

years (2012: 4; 2014: 5; 2016: 3), and none in 2018. This might suggest an increasingly predominant and ubiquitous role of analytical datasets and computational workflows in the generation of the final published results in the field.

## 4 Reproducibility of GIScience conference papers

Table 2 shows aggregated values for the assessed reproducibility levels. If we look at the median values of the five criteria (Table 2), a typical GIScience paper scores `1 1 1 0 1`. This score translates in practical terms into a paper that is sufficiently documented to claim that reproduction could be attempted within a reasonable time frame after publication. While such a level of reproducibility is typically accepted by journals and conferences today, it does not guarantee that a reproduction would be possible and practical. A reproduction of such a paper would require considerable effort, namely technical skills, communication with authors, and time not only to both gather, recreate, and/or analyse all the necessary resources (data, code, etc.) but also to recreate the specific computational environment of the paper. Especially the latter is very unlikely, as the computational environment is generally not specified at all, as demonstrated by the median value of `0` (*Unavailable*) for this sub-criterion.

Figure 1 shows the distribution of the reproducibility levels for each criterion. None of the papers reached the highest reproducibility level of `3` (*Available and Open*) on any criterion. Only 12 papers reached level `2` (*Available*) in the *Input Data* criterion. Similar to previous results [26], the number of papers with level `0` for *Input Data* was especially high (33, corresponding to 44%), which is a significant barrier to reproduction since input data is not only unavailable but also cannot be recreated from the information provided in the paper.

*Preprocessing* applied to only 51 publications. For 24 papers, the *Preprocessing* criterion was not applicable (*NA*). This large number is a result of our decision to assess *Preprocessing* only if papers explicitly stated or described a preprocessing step in their analysis, which few did. This does not mean the assessment ignored missing information on preprocessing step, only that such missing information would then reduce the level of the *Methods* criterion instead. Obviously, if data preprocessing is required but it is either not indicated in the paper or is not provided as an additional (computational) step or resource, the ability to reproduce the paper will be limited. The achieved levels for *Preprocessing* remained low: 37 papers reach level `1` (*Documented*), about half of the papers with level `1` in the *Methods* criterion. For the other half, it was not clear whether data preprocessing tasks existed at all, or whether these tasks were part of the main analysis.

*Methods* and *Results* criteria show a similar distribution (see Figure 1). Indeed, 65 publications had level `1` in both criteria, which represents 87% of the papers assessed. In this sense, most of the assessed papers fall below the minimum standard for reproduction in the methods and results criteria. All papers except one reached level `1` for the *Results* criterion, which shows that the peer review worked as expected for almost all articles. In other words,

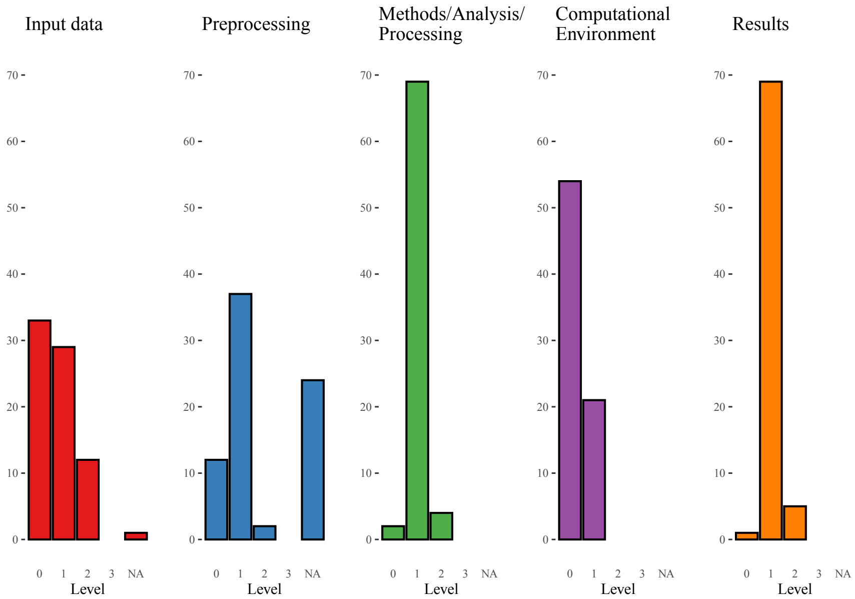

**Figure 1** Barplots of reproducibility assessment results; levels range from 0 (leftmost bar) to 'not applicable' (rightmost bar).

authors are concerned with making the results understandable to the reviewers, which is not always the case for the other criteria. More generally, this aspect raises the question of whether peer review should stop in the absence of minimal evidence of the input data, analysis, and computational environment used in a paper.

Finally, papers scored worse on the *Computational Environment* criterion. Overall, 54 publications (72%) remained at level 0, which means that no information was provided in the paper about the computing environment, tools, or libraries used in the reported analysis. The *Computational Environment* criterion and the *Input Data* criterion accounted for a significant number of 0 values, which clearly signals an impediment to reproduction. It also shows a rather low recognition of data and software as academic outputs, because both data and software should be properly cited to give credit to their creators [18, 12].

Figure 2 shows an alluvial diagram of all scores, i.e., combinations of criteria values of those 49 papers without any *NA* criterion. Most of the excluded papers have *NA* for *Preprocessing*, therefore this criterion is not included in the figure. The diagram confirms overall patterns seen before. The vast majority of papers have level 1 in *Methods/Analysis/Processing* and *Results*. *Input data* is most diverse, with a surprisingly large number of papers with level 0 but also the largest fraction of papers reaching level 2. Many papers show low levels in *Computational Environment*.

The diagram illustrates how groups of papers with similar properties "flow" through the different criteria Three major groups, which represent 34 of the papers (69%) included in the figure, become visible as broad bands. Two groups with 10 papers each start with level 0 for *Input Data* and 1 for *Methods/Analysis/Processing* and reach a 1 for *Results*, while they are divided equally between level 0 and 1 for *Computational Environment*. These two groups seem to indicate that the authors and reviewers alike follow the established pattern that results outweigh concerns for transparency and reproducibility, since computational papers with *Unavailable* input data are irreproducible The third and largest group matches the overall mean values for the typical GIScience paper with level 1 for all criteria except for *Computational Environment*.

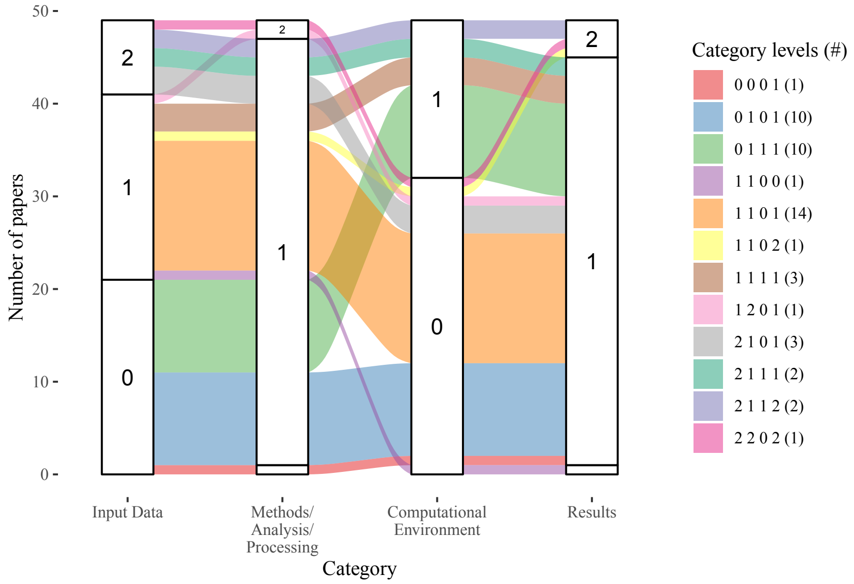

**Figure 2** Alluvial diagram of common groups of papers throughout 4 of 5 categories including only papers without any "not applicable" *(Level* `NA`*)* value; category *Preprocessing* was dropped because difficulty to clearly assess it lead to many "not applicable" values.

The diagram also shows additional interesting patterns for a few papers. The papers with the lowest level of `0` in *Results*, i.e., according to the assessors the results are documented insufficiently and thus difficult or impossible to fully understand, actually have better values in previous criteria. Only few papers that start with level `2` in *Input Data* can keep this level for *Methods/Analysis/Processing*, and even those who do later drop to level `0` in *Computational Environment*. Only one paper each shows the following surprising paths: Starting with level `1` for *Input Data* , then moving up to level `2` in *Methods*, before reaching level `2` in *Results* despite having only values of `1` or `0` in other criteria. In summary, not a single paper can reach the required levels for an immediate reproduction by ensuring that all required pieces are *Available* (level `2`), not even considering the further challenges for reproductions, such as incomplete documentation [28]. An investigation of yearly scores to track developments over time does not show any trend, i.e., there is little change in reproducibility over the study period[10]. The overall low values for *Computational Environment* are one signal that confirms the growing concerns for reproducibility and reusability of computational research are not misplaced.

# 5 Discussion

## 5.1 State of reproducibility in the GIScience conference series

Our first research objective was to assess the state of reproducibility in the GIScience conference series. A recurrent issue found in the analysis was the inability to access input data based on the information provided in the paper. Most of the links and pointers to

[10] See the additional analysis and plots published at `https://nuest.github.io/reproducible-research-at-giscience/giscience-reproducibility-assessment.html` or in the paper's reproducibility package [27].

datasets reported at the time of publication were either broken (e.g., non-existing resource, HTTP 404 error, invalid URL syntax) or not available anymore (URL works but redirects to a different generic page; specific resource from the paper no longer exists). In these cases, a level 2 in the *Input Data* criterion was deserved at the time of publication; however, when evaluating the level of reproducibility some time later, as was done in this work, level 2 is no longer suitable for those papers. From a reproducibility point of view, the input data was therefore not accessible, although contacting the authors could still be attempted. However, according to the meaning of the criterion and in practical terms, this is equivalent to including the statement "available upon request" in the paper and thereby level 0. An important part of reproducibility is that access to material should not degrade over time, which is best achieved by depositing data in repositories, including sensitive data (using the appropriate mechanisms), and properly citing it. In this assessment of reproducibility, we decided to give the authors the benefit of the doubt and awarded a value of 2 for *Input Data* even if we could not conclusively determine, e.g., by using the Internet Archive's Wayback Machine[11], whether the original website ever existed.

Regarding the common situation of a paper with *Documented* (level 1) for all criteria, our interpretation is that this is indeed a regular paper that is up to current scientific standards. Does this imply that a paper with *Unavailable* (level 0) in any criterion should not have been accepted? We believe that this requires differentiation between past and future papers. The criteria used in this paper were not included in the previous call for papers or in the reviewer guidelines, and therefore received less attention from authors or reviewers. Thus, we have analysed work in a historical context when there were few concrete incentives to push these aspects, beyond the general concerns for good scientific practice. Nowadays, with awareness about reproducibility being raised through initiatives, projects, and publications about it, we would expect that reproducibility levels increase, and argue that papers with *Unavailable* in one more criteria should not be accepted anymore without a clear and explicit justification (e.g., sensitive data on human subjects). This does not imply that it is always necessary to achieve the gold standard of *Available and Open*. The overall objective should be to make a paper as reproducible as possible before publication. We argue that, for most currently published works at the GIScience conference, *Available* would have been achievable and feasible with reasonable efforts.

However, such a change in standards for paper acceptance would also mean that researchers, editors, and publishers might have to reevaluate their focus on publishing novel and purportedly groundbreaking results in science, and give as much weight to publishing the full process and collection of parts that would allow readers to try to fully understand the research. Clearly, *Unavailable* for *Input Data* is the most problematic, because without sufficient knowledge about the characteristics of the input data, all attempts at reproducing results are bound to fail, even when the textual documentation of the data would potentially allow for an time-intensive recreation of the computational workflow.

## 5.2 Transferability of method

Concerning our second research objective, we can state that the overall process and the application of the reproducibility rubric was successfully replicated with a different data set. This is not entirely surprising given that AGILE and GIScience conference series share similarities in target audience, review process, and publication of proceedings (more on

[11] https://web.archive.org/.

that in the following section). More importantly, the process faced similar challenges as we recalled from its earlier application. This is crucial information, because the successful replication of the process, including its challenges, enables us and others to ground any changes in solid evidence. In particular the *Preprocessing* criterion caused many discussions among the reproducibility reviewers during the assessment. It is often not clear or a matter of interpretation if a particular processing step belongs to a minor basic transformation of input data, if it is already part of the main analysis, and when it is a truly distinct step in the process. The borders are vague and hence scores should be interpreted with caution. Likewise, the *Computational environment* is also difficult to distinguish from analysis, and technology and practices for the effective management of the computing environment have reached mature states relatively recently. Future reproducibility assessments of papers could provide a more precise definition for *pre*-processing, e.g., only use it if the authors use the term, or might consider to drop the category, and benefit from rules to deal with the specific issues of older workflows, similar as discussed for input data above. Furthermore, it is important to remember that the levels of reproducibility are not equidistant in the sense that a level of `2` would be twice as good as a level of `1`, or that the effort needed is twice as high. A level of `1` should be the standard for current and future peer-reviewed papers. Reaching level `2` requires several additional steps, while reaching the gold standard of `3` is again a comparatively small step from level `2` in terms of effort - the main difference is to use public repositories with a DOI - yet with a high positive impact on permanent accessibility.

Although the replication was successful, the process was again labour-intensive, making it problematic to scale it up to assess multiple years of several popular journals, for example. Further, despite our best efforts for transparency and the four-eyes principle in the assessment, the process is inherently subjective. A different group of investigators might score papers differently. While natural language processing techniques have made great progress in the past decades, an automated assessment of a paper's reproducibility still seems out-of-reach. Including important information as machine-readable metadata could allow to come closer to automation.

## 5.3 Comparison of conferences

Given that we followed the same process as in [26] and demonstrated the transferability of the method, comparing the two conference series seems appropriate. It is important to remember that we do not attempt such a comparison with the objective of declaring a "winner". The published work and contributing community of the two conferences are similar enough for a comparison, yet their organisation (setup, process, geographic focus) differ too much for a simplistic ranking. However, a comparison is required to sensibly discuss whether the guidelines developed for AGILE might also be promising for GIScience: Are they transferable? If not, what adaptations seem necessary?

Concerning the contributing and participating academic communities, Egenhofer et al. [8] and Kemp et al. [14] both include both conferences series as outlets for GIScience research. Further, Keßler et al. [15] investigate the bibliographies of four GIScience conference series, including GIScience and AGILE for the year 2012, and identify 15 authors who have published in both conference series. We conducted a cursory investigation of the body of authors for full papers, revealing significant overlap[12]: Out of 571 unique AGILE and 405 unique GIScience full paper authors, 86 published in both conferences, and this includes all 15 authors mentioned by Keßler et al. [15]. Therefore, the strong relation between the AGILE and GIScience

[12] The data and code for the brief exploration into the authorship across the conferences considered in this work can be found in the directory `author_analysis` of this paper's reproducibility package [27].

**Table 3** Mean values per criterion for both conferences (rounded to two decimal places).

| Criterion | AGILE full papers | GIScience papers |
|---|---:|---:|
| input data | 0.67 | 0.72 |
| method/analysis/processing | 1.00 | 1.03 |
| computational environment | 0.62 | 0.28 |
| results | 0.88 | 1.05 |

conference series confirms our approach to apply the same methodology to GIScience that has been developed for AGILE conference publications, and it might lead to similar implications for improving reproducibility.

Nevertheless, before discussing any strategies to improve reproducibility, it is important to identify and consider the differences between the two conference series. GIScience is a biannual conference series whereas AGILE is annual, and they feature different pre-publication review processes and review management systems: In AGILE both authors and reviewers are anonymous, while in GIScience only the reviewers are. Furthermore, the AGILE conference series has the AGILE association[13] as an institutional supporter, which means a more stable organisational and financial framework for activities spanning more than one or between conferences. However, like GIScience, local conference organisers for AGILE have the main financial burden and experiences are informally handed over between organising committees. Geographic focus is also different: GIScience has a global target audience, and the individual conferences are likely to be different in their contributor communities because of the moving conference location, which often means lowered accessibility for authors from other parts of the world. AGILE, by comparison, has a European focus and accessibility is more homogeneous, although the conference location moves every year,. This likely translates into a less fluctuating and less geographically diverse audience at AGILE. Clearly, these observations will need a reassessment in several years to evaluate the impact of both conferences going full online in 2020/21 because of the travel and activity restrictions due to the COVID-19 pandemic.

Concerning the paper corpora, the publication years considered here (2012-2018) are similar to the assessment of AGILE papers (2010-2017), which makes the results comparable in the sense of what methods and tools would have been available for authors. Furthermore, we note that both conferences have a similar ratio of conceptual papers which were not assessed for reproducibility: In the AGILE corpus we identified 5 of 32 conceptual papers (15.6%), in the GIScience corpus there were 12 of 87 (13.8%). This indicates that both conferences have similar share of papers that used, at least in part, computational methods. On the content of the papers, our overall impression was that a larger share of GIScience papers included theoretical, conceptual, or methodological aspects, while AGILE papers seemed to feature more empirical and/or applied geoinformation science research.

Regarding the results of the reproducibility assessments as summarised in Table 3, the nature of the data and sample size does not support statistical analyses on significant differences. Nevertheless, looking at the *Input Data* criterion, GIScience has a slightly higher mean value compared to AGILE full papers (0.72 as opposed to 0.67) and a median of 1. These values indicate that the GIScience contributions had a slightly better, but by no means optimal, availability of input data. The pattern of reproducibility of the papers' workflows (category *Method, Analysis, Processing*) was very similar for the two conference

[13] `https://agile-online.org/`.

series: The majority of papers achieved a level of `1`, resulting in a mean of 1.03 for GIScience and 1 for AGILE full papers. The *Computational Environment* category shows the largest difference (although at overall low levels): AGILE scored better with a mean of 0.62 vs. 0.28 for GIScience. The *Results* category scores were again slightly higher for GIScience, with a mean of 1.05 vs. a mean of 0.88 for AGILE. Several papers in AGILE received a level of `0` here, indicating that crucial information is missing to connect analysis outputs and presented results. We refrain from comparing the *Preprocessing* category for the reasons stated earlier.

This comparison lets us draw two main conclusions: First, we conclude that both the target audience and the content of the two conference series are similar enough to be afflicted with similar shortcomings in terms of reproducibility, and thus, they both likely respond to similar solutions. Second, we conclude that the AGILE conference series seems structurally better positioned to support changing culture, because of a more stable audience and institutional support. The introduction of the AGILE reproducibility guidelines was achieved within a short time frame and with financial support in the form of an "AGILE initiative", including travel funding for an in-person workshop. For GIScience, the task of changing the review process to foster better reproducibility falls squarely on the shoulders of the changing program committees. However, the initial results of AGILE's new guidelines show that even small changes can lead to a significantly improved outcome.

## 6 Conclusions and outlook

In this work we investigated the reproducibility of several years of GIScience conference publications. The paper corpus is large enough for a representative sample and comparable to that used for the AGILE assessment study due to largely overlapping time window. However, this study does not intend to make judgements on AGILE vs. GIScience conference quality, nor to question the papers' scientific soundness or relevance, since they were accepted for publication at a reputable conference. Instead, we investigated the papers along a single desirable quality dimension, reproducibility, which implies requirements on openness and transparency.

Using a similarly high bar for reproducibility as in the earlier assessment study, the results show room for improvement, as none of the presented articles were readily reproducible. The majority of articles provided some information, but not to the degree required to facilitate transparent and reusable research based on data and software. Overall, this is very similar to the outcomes of our earlier study on AGILE papers. As part of the AGILE assessment, we described concrete recommendations for individuals and organisations to improve paper reproducibility [26]. We have argued that AGILE and GIScience share a sufficiently common domain/discipline characteristics, audience, and author community, such that for both communities the strategies to improve the situation should be similar. Therefore, the previously identified recommendations are transferable to the GIScience conference series, with the most important recommendations being (1) promoting outstanding reproducible work, e.g., with awards or badges, (2) recognizing researchers' efforts to achieve reproducibility, e.g., with a special track for reproducible papers, implementing a reproducibility review, open educational resources, and helpful author guidelines including data and software citation requirements and a specific data/software repository, and (3) making an institutional commitment to a policy shift that goes beyond mere accessibility [33]. These changes require a clear roadmap with a target year, e.g., 2024, when GIScience starts to only accept computationally reproducible submissions and to check reproducibility before papers are accepted. The concluding statement of Archmiller et al. [1] is directly transferable to GIScience: The challenges are not insurmountable, and increased reproducibility will ensure scientific integrity. The AGILE reproducible paper guidelines [24] and the associated reproducibility review processes as well as other community code review systems such as CODECHECK [9] are open and "ready to use". They can also be adopted for GIScience conferences, e.g., to suit the peer review process goals and scheduling. Kedron et al. [13] stressed the need for a comprehensive balanced approach to technical, conceptual, and practical issues. They further pointed out that simple availability does not automatically lead to adoption. Therefore, a broad discourse around these recommendations, tools, and concepts would be beneficial for all members of the community, whether their work is more towards conceptual, computational, or applied GIScience. A survey for authors, as conducted for AGILE [26], could help identify special requirements and specific circumstances, beyond the findings presented here and in related work.

Future work may replicate the reproducibility assessment at other major events and outlets for GIScience research, such as GeoComputation or COSIT conferences and domain journals (cf. [8] for an extensive list), but we would not expect significantly differing results. Practical reproductions of papers, and even more so replications of fundamental works, are promising projects to convincingly underpin a call for a culture change [29]. A successful *reproducibility turn* would not mean that every reproducible paper would be fully reproduced, nor would this be necessary. But at least for influential, e.g., highly cited papers, a validation of their applicability and transferability to other study areas should be possible – reproducibility is a prerequisite for that. For example, Egenhofer et al. [8] provide for a list of the most frequently cited articles as potential candidates. Such a project would ideally be supported with proper funding. There is currently growing activity in the GIScience discipline to address reproducibility and replicability of geospatial research. The GIScience conference community has the opportunity to play a leading and shaping role in this process, thereby ensuring its continuing attractiveness for authors to submit their work, and in consequence its high relevance for the wider GIScience discipline. A timely adoption of the technological and procedural solutions may allow GIScience researchers, together with the entirety of academia, to level up and approach the challenges of the *"second phase of reproducible research"* by tackling long-term funding for maintenance of code and data and building supporting infrastructure for reproducible research [31].

## References

1 Althea A. Archmiller, Andrew D. Johnson, Jane Nolan, Margaret Edwards, Lisa H. Elliott, Jake M. Ferguson, Fabiola Iannarilli, Juliana Vélez, Kelsey Vitense, Douglas H. Johnson, and John Fieberg. Computational Reproducibility in The Wildlife Society's Flagship Journals. *The Journal of Wildlife Management*, 84(5):1012–1017, 2020. doi:10.1002/jwmg.21855.

2 Lorena A. Barba. Terminologies for Reproducible Research. *arXiv:1802.03311 [cs]*, 2018. arXiv: 1802.03311. URL: https://arxiv.org/abs/1802.03311.

3 Chris Brunsdon. Quantitative methods I: Reproducible research and quantitative geography. *Progress in Human Geography*, 40(5):687–696, 2016. doi:10.1177/0309132515599625.

4 Chris Brunsdon and Alexis Comber. Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems*, August 2020. doi:10.1007/s10109-020-00334-2.

5 Giovanni Colavizza, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker, and Barbara McGillivray. The citation advantage of linking publications to research data. *PLOS ONE*, 15(4):e0230416, 2020. doi:10.1371/journal.pone.0230416.

6 David L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, July 2010. doi:10.1093/biostatistics/kxq028.

7 Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank, editors. *Geographic Information Science*. Springer International Publishing, 2014. doi:10.1007/978-3-319-11593-1.

8 M. Egenhofer, K. Clarke, S. Gao, Teriitutea Quesnot, W. Franklin, M. Yuan, and David Coleman. Contributions of GIScience over the past twenty years. In Harlan Onsrud and Werner Kuhn, editors, *Advancing Geographic InformationScience: The Past and Next Twenty Years*. GSDI Association Press, Needham, MA, 2016. URL: http://www.gsdiassociation.org/images/publications/AdvancingGIScience.pdf.

9 Stephen Eglen and Daniel Nüst. CODECHECK: An open-science initiative to facilitate sharing of computer programs and results presented in scientific publications. *Septentrio Conference Series*, (1), 2019. doi:10.7557/5.4910.

10 Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041). *Dagstuhl Reports*, 6(1):108–159, 2016. doi:10.4230/DagRep.6.1.108.

11 Michael F. Goodchild. Geographical information science. *International journal of geographical information systems*, 6(1):31–45, 1992. doi:10.1080/02693799208901893.

12 Daniel S. Katz, Neil P. Chue Hong, Tim Clark, August Muench, Shelley Stall, Daina Bouquin, Matthew Cannon, Scott Edmunds, Telli Faez, Patricia Feeney, Martin Fenner, Michael Friedman, Gerry Grenier, Melissa Harrison, Joerg Heber, Adam Leary, Catriona MacCallum, Hollydawn Murray, Erika Pastrana, Katherine Perry, Douglas Schuster, Martina Stockhause, and Jake Yeston. Recognizing the value of software: a software citation guide. *F1000Research*, 9:1257, January 2021. doi:10.12688/f1000research.26932.2.

13 Peter Kedron, Wenwen Li, Stewart Fotheringham, and Michael Goodchild. Reproducibility and replicability: opportunities and challenges for geospatial research. *International Journal of Geographical Information Science*, 0(0):1–19, 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2020.1802032. doi:10.1080/13658816.2020.1802032.

14 Karen Kemp, Werner Kuhn, and Christoph Brox. Results of a survey to rate GIScience publication outlets. Technical report, AGILE Initiative - GIScience Publication Rating, 2013. URL: https://agile-online.org/conference_paper/images/initiatives/results_of_a_survey_to_rate_giscience_publications.pdf.

15 Carsten Keßler, Krzysztof Janowicz, and Tomi Kauppinen. spatial@linkedscience – Exploring the Research Field of GIScience with Linked Data. In Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, and Shashi Shekhar, editors, *Geographic Information Science*, Lecture Notes in Computer Science, pages 102–115, Berlin, Heidelberg, 2012. Springer. doi:10.1007/978-3-642-33024-7_8.

16 Markus Konkol, Christian Kray, and Max Pfeiffer. 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):408–429, February 2019. doi:10.1080/13658816.2018.1508687.

17 Christian Kray, Edzer Pebesma, Markus Konkol, and Daniel Nüst. Reproducible Research in Geoinformatics: Concepts, Challenges and Benefits (Vision Paper). In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *COSIT 2019*, volume 142 of *LIPIcs*, pages 8:1–8:13. Schloss Dagstuhl Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.COSIT.2019.8.

18 Lawrence, Bryan, Jones, Catherine, Matthews, Brian, Pepler, Sam, and Callaghan, Sarah. Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, 6(2), 2011.

19 Florian Markowetz. Five selfish reasons to work reproducibly. *Genome Biology*, 16:274, 2015. doi:10.1186/s13059-015-0850-7.

20 Jennifer A. Miller, David O'Sullivan, and Nancy Wiegand, editors. *Geographic Information Science*. Springer International Publishing, 2016. doi:10.1007/978-3-319-45738-3.

21 Jannes Muenchow, Susann Schäfer, and Eric Krüger. Reviewing qualitative GIS research—Toward a wider usage of open-source GIS and reproducible research practices. *Geography Compass*, 13(6):e12441, 2019. doi:10.1111/gec3.12441.

22 Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021, January 2017. doi:10.1038/s41562-016-0021.

23 Brian A. Nosek and Timothy M. Errington. What is replication? *PLOS Biology*, 18(3):e3000691, March 2020. doi:10.1371/journal.pbio.3000691.

24 Daniel Nüst, Frank Ostermann, Rusne Sileryte, Barbara Hofer, Carlos Granell, Marta Teperek, Anita Graser, Karl Broman, and Kristina Hettne. AGILE Reproducible Paper Guidelines, 2019. doi:10.17605/OSF.IO/CB7Z8.

25 Daniel Nüst, Frank Ostermann, Rusne Sileryte, Barbara Hofer, Carlos Granell, Marta Teperek, Anita Graser, Karl Broman, and Kristina Hettne. Reproducible Publications at AGILE Conferences, 2019. doi:10.17605/OSF.IO/PHMCE.

26 Daniel Nüst, Carlos Granell, Barbara Hofer, Markus Konkol, Frank O. Ostermann, Rusne Sileryte, and Valentina Cerutti. Reproducible research and GIScience: an evaluation using AGILE conference papers. *PeerJ*, 6:e5072, 2018. doi:10.7717/peerj.5072.

27 Daniel Nüst, Frank Ostermann, Carlos Granell, and Barbara Hofer. Reproducibility package for "Reproducible Research and GIScience: an evaluation using GIScience conference papers", September 2020. doi:10.5281/zenodo.4032875.

28 Daniel Nüst, Frank Ostermann, Carlos Granell, and Alexander Kmoch. Improving reproducibility of geospatial conference papers – lessons learned from a first implementation of reproducibility reviews. *Septentrio Conference Series*, (4), September 2020. doi:10.7557/5.5601.

29 Frank O. Ostermann. Linking Geosocial Sensing with the Socio-Demographic Fabric of Smart Cities. *ISPRS International Journal of Geo-Information*, 10(2):52, January 2021. doi:10.3390/ijgi10020052.

30 Roger D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, December 2011. doi:10.1126/science.1213847.

31 Roger D. Peng and Stephanie C. Hicks. Reproducible Research: A Retrospective. *arXiv:2007.12210 [stat]*, July 2020. arXiv: 2007.12210. URL: http://arxiv.org/abs/2007.12210.

32 James H. Stagge, David E. Rosenberg, Adel M. Abdallah, Hadia Akbar, Nour A. Attallah, and Ryan James. Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*, 6(1):190030, February 2019. Number: 1 Publisher: Nature Publishing Group. doi:10.1038/sdata.2019.30.

33 Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, 2018. doi:10.1073/pnas.1708290115.

34 S. Winter, A. Griffin, and M. Sester, editors. *Proceedings 10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114. LIPIcs, 2018. URL: http://www.dagstuhl.de/dagpub/978-3-95977-083-5.

35 Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, and Shashi Shekhar, editors. *Geographic Information Science*. Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-33024-7.