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

Markus Konkol, Christian Kray & Max Pfeiffer

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

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

Markus Konkol ⬥, Christian Kray ⬥ and Max Pfeiffer

Institute for Geoinformatics, University of Münster, Münster, Germany

ABSTRACT

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

1. Introduction

Reproducibility is an essential element of scientific work in general, as it enables researchers to re-run and re-use experiments reported by others. Further benefits of working and publishing reproducibly include increased transparency and more efficient review processes (Gil *et al*. 2016). Despite these advantages, publishing results in a reproducible way is still not common practice (Reichman *et al*. 2011), which is part of the reason why some have proclaimed a 'reproducibility crisis' (Baker 2016). A recent study in economics (Gertler *et al*. 2018) has shown that even when authors make the data and code publicly accessible, it is not guaranteed that readers can successfully reproduce the results published in the paper. On top of that, the inconsistent usage of the terms *reproducibility* and *replicability* within and across disciplines can cause further

confusion (Bollen *et al*. 2015). It is thus not surprising that the topic of reproducible research and how to realise it are discussed across many disciplines such as biology (Leek and Jager 2017) and computer science (Stodden 2010).

In this article, we focus on computational reproducibility in general and *open reproducible research* (*ORR*) in particular. Goodman *et al*. (2016) state that in *ORR* all used research components, e.g. data, software, and configuration are publicly accessible and produce the same results (i.e. numbers, tables, figures) compared to those reported in the paper. This is particularly relevant in the geosciences which encompasses all domains related to earth sciences, such as climatology and landscape ecology (see the list of geoscientific domains by Nature Geosciences (2018)). According to Goodchild (1992), three relevant topics in geographic information science are spatial statistics, algorithms that operate on geographic information, and the display of geographic information. Many papers published in the geosciences apply spatial statistics based on geographic information, and the results are often displayed as maps or time series. Thus, to achieve a minimum standard of credible research results, computational reproducibility is essential. However, compared to other disciplines, the field of computational geosciences (geoscientifc research based upon code and data) has given little attention to reproducibility (cf. Giraud and Lambert 2017). This paper aims to address this gap by investigating how geoscientists who conduct computational research understand reproducibility, whether and how it is practiced, and what obstacles hinder it. Hence, we carried out three studies with geoscientists (a survey, interviews, and a focus group), and we performed a reproduction study using previously published reports that apply spatial statistics in *R*.

Contributions. This article contributes the following insights. First, we report on what geoscientists understand *ORR* to mean. Second, we identify practical obstacles that stand in the way of authors publishing ORR, and we also identify some obstacles that readers face when reproducing others' work. Third, from our reproducibility study, we report on technical issues when attempting to execute the original code provided in the papers we aimed to reproduce. Next, we describe key differences that impeded the comparability of the original and our reproduced figures. Finally, we propose a set of guidelines for authors to address the identified issues.

Scope. Reproducibility is a complex concept involving different stakeholders (Nüst *et al*. 2017) across multiple disciplines. To keep the scope of the research manageable, we do not consider qualitative research and how to reproduce it, and amongst all stakeholders, we only focus on authors and readers.

In the following section, we first review related work on reproducible research in general and in particular in the geosciences. Then, we report on the four studies we conducted here, i.e. the survey, the interviews, the focus group, and the reproducibility study. We then discuss our findings and their limitations. We conclude by summarising key insights and providing a set of guidelines for authors wishing to publish reproducibly.

2. Related work

We first review work on the various definitions of reproducible research, the obstacles authors and readers face when producing and using reproducible research, the incentives for publishing reproducibly, and approaches to overcome the associated barriers.

2.1. *Reproducible research*

Different definitions of and perspectives on *reproducible research* have been proposed. According to Leek and Peng (2015), research results reported in a paper are reproducible if they can be re-computed based on the same data and 'knowledge of the data analysis pipeline'. Easterbrook (2014) considers research to be reproducible if it enables the re-creation of the results based on *given* code or an *own* program. Both definitions are flexible regarding the use of procedures and software. In contrast, Gentleman and Temple Lang (2007) require that for authors' research to be called reproducible, they must include the software they used to produce their results. Similarly, Bollen *et al.* (2015) equates reproducibility with being able to achieve the same results as reported using the same data and procedure. Goodman *et al.* (2016) linked reproducibility to specific purposes: they distinguish between *methods reproducibility*, referring to achieving the same results based on the same data and code, and *results reproducibility*, which corresponds to *replicability*, i.e. achieving consistent results by independent experiments with new data and code. Peng (2011) proposed a reproducibility spectrum ranging from *not reproducible* (if no research materials are provided), to *reproducible* (if code and data are available), to *fully replicable*. In contrast, Leek and Jager (2017) make a binary distinction based on the outcome of an attempt to reproduce results: research is reproducible if the results are the same; if they are not, the research is not reproducible. In summary, we can thus note that while the basic notion of re-use is consistent across different definitions of reproducibility, there are also substantial differences: some definitions simply require a detailed methodology section whereas others demand access to all used materials (e.g. data and software).

2.2. *Incentives for publishing reproducible research*

Independent of what definition of reproducibility is used; there are a number of reasons why it makes sense for authors to publish reproducible work and for readers to make use of such work. Most importantly, reproducible research facilitates the re-use of the results in the paper, including the methods and data that were used to produce the results (Collberg and Proebsting 2016, Gil *et al.* 2016). Furthermore, detecting errors is easier if research is reproducible (Gentleman and Temple Lang 2007), such as when there are differences between the reported and replicated results (Donoho *et al.* 2009). Readers or reviewers can then check if there was an error in the original analysis, e.g. by studying the data analysis and parameters that were used (Stodden *et al.* 2016). Moreover, it is likely that reproducibility will become a requirement for reputable publication outlets (Gil *et al.* 2016); working reproducibly from the beginning can make it easier to meet a journal's standards (Hillebrand and Gurevitch 2013). With the growing trend towards Open Science (open data, open code) (Gewin 2016), transparency can be increased and the credibility crisis can be tackled (Reichman *et al.* 2011).

Further benefits of reproducible research arise from new possibilities afforded by the approach, e.g. meta-analyses (Stodden *et al.* 2016), continuously evolving papers (Brunsdon 2016), and new cooperations (Costello 2009). Journals such as *Distill*[1] support publication of transparent research that can include, e.g. interactive figures. The ReScience[2] journal encourages replication of the computational steps in published

articles, ideally as open source implementations for future use. Finally, providing public access to code and data increases citation numbers (Piwowar *et al.* 2007, Vandewalle 2012), which have a direct impact on researchers' reputations. Despite these benefits, it is important to keep in mind that reproducible research cannot prevent flaws during data collection (Ostermann and Granell 2017). Nevertheless, it can help to establish a minimum standard for credible computational research (Bollen *et al.* 2015).

2.3. *Reasons for irreproducible papers*

Given the long list of benefits and incentives for publishing reproducibly, it might seem surprising that not all research is published in this way. There are, however, a number of different reasons that explain why most papers are published in a non-reproducible way. These reasons include cultural (Reichman *et al.* 2011) and technical (Easterbrook 2014) barriers as well as authors who cannot reproduce their own results (Vandewalle *et al.* 2009). One key issue is that data is rarely available (Ioannidis *et al.* 2009). If not archived, data availability declines with article age, making it particularly challenging to reproduce older publications (Vines *et al.* 2014). A second issue is the source code is rarely accessible, and if it is, it is not always in the right version (Collberg and Proebsting 2016). This is largely because preparing code and data for publication requires considerable effort (Barnes 2010). Furthermore, authors are frequently not aware of the incentives that might be worth the extra effort (Nosek *et al.* 2015) and of the drawbacks of not publishing reproducibly, e.g. having to respond to questions about the code (Gewin 2016). Another problem is that many scientists worry about falling behind (Gewin 2016) if they spend time 'unwisely', as the credit system does not sufficiently reward scientists for fully disclosing their own work (McCullough *et al.* 2008). In addition, researchers fear that if they are fully transparent, others question their conclusions (Piwowar *et al.* 2007) and thereby tarnish their reputation, but this is in fact a key process in science (Benestad *et al.* 2016). Further barriers to reproducible research are legal aspects, sensitive data (Stodden *et al.* 2016), and ethical concerns (Darch and Knox 2017). As a result of these issues, some have proclaimed that science is suffering from a reproducibility crisis (Baker 2016). Examples that support this claim highlight the drawbacks of irreproducible papers (including a study with 100 replication attempts by Open Science Collaboration (2015)) and the flaws detected in published articles (cf. Benestad *et al.* 2016).

2.4. *Guidelines and recommendations*

Several authors have proposed ways to overcome the issues and barriers outlined in the previous subsection. Nosek *et al.* (2015) presented eight *Transparency and Openness Promotion* (TOP) guidelines addressing, for example, citation standards for materials, sharing of data and methods, and preregistration. Each guideline has four levels ranging from *standard not met* to *standard fully met*. Based on these guidelines, Stodden *et al.* (2016) suggested the *Reproducibility Enhancement Principles* (REP) for computational research, but they also highlighted that journals should demand all research components underlying the analysis. Ideally, these components should be shared via public repositories and archives (Gewin 2016). Another recommendation is to design an improved credit system to address citation issues such as being able to cite individual

research components (Gil *et al*. 2016). Consequently, *all* citable components will need a legal statement on reusability (Stodden *et al*. 2016). Scientists should choose open source software instead of proprietary tools, and they should add information on the computational environment (Fehr *et al*. 2016). Both practices facilitate reproduction by third parties, and Steiniger and Hay (2009) also showed that free tools can be as useful as proprietary software. Moreover, figures should be created by using scripts instead of ready-to-use toolboxes, which hide the computational steps (Sandve *et al*. 2013) and hinder reproduction. In this context, being able to reproduce figures is particularly important, as they are popular means to visualise computational results (cf. Claerbou and Karrenfach *et al*. 1992). Several technical solutions support scientists in publishing reproducible research. A popular approach is *literate programming*, which allows authors to combine text and code into a single document (Knuth 1984), e.g. using RMarkdown (Allaire *et al*. 2016) or Jupyter Notebooks (Kluyver *et al*. 2016). This approach can be extended to reproducible books such as the openly developed book *Geocomputation with R* by Lovelace *et al*. (2016), to which everyone can contribute. A useful tool for sharing code is *GitHub*,[3] a popular software development platform. However, one issue with publishing material on publicly accessible platforms is that it can interfere with a double-blind reviewing process. To counter this issue, the *Open Science Framework*[4] also enables sharing materials anonymously for peer review.

2.5. *Reproducible research in the geosciences*

Ostermann and Granell (2017) investigated if research on volunteered geographic information (VGI) is reproducible, i.e. by using the same data and methods, and is replicable, i.e. by conducting an independent experiment with new data and a similar method. According to their results, none of the investigated publications were reproducible and less than half of them were replicable. To facilitate reproducibility, Gil *et al*. (2016) proposed the *Geoscience Paper of the Future*, which provides public access to research components enriched by metadata. Giraud and Lambert (2017) argued that figures, such as maps, frequently depict key results of geographic research and thus should be reproducible. Brunsdon (2016) investigated the importance of code in quantitative geography. Key observations were that making code available facilitates tasks, such as comparing different implementations of an analysis. In addition, researchers can rerun the analysis with another or updated dataset. In this context, it is particularly important to provide the original code, as textual descriptions might be inaccurate. Increasingly, geoscientific journals such as *Nature Geosciences* encourage publishing code underlying the reported results. In the *Vadose Zone Journal* (Skaggs *et al*. 2015), authors can submit reproducible research articles by attaching data, code, and metadata. While this constitutes a big step towards making research fully reproducible, the problem of making the code executable on different machines and in different environments persists. In order to tackle this issue, Nüst *et al*. (2017) proposed the *Executable Research Compendium* (ERC), which encapsulates the runtime environment and all research components underlying the analysis in a Docker container. This approach can thus lead to improved reusability, accessibility, and transparency.

   In summary, we can observe that reproducible research has gained importance in the computational geosciences in recent years. However, it is largely unknown what

geoscientists who conduct computational research such as spatial statistics understand *ORR* to mean, what roadblocks they face in realising *ORR*, and what differences exist compared to other disciplines. In the next sections, we therefore report on a series of studies we conducted to shed light on these questions. While research on reproducibility mostly focused on the accessibility of materials, we took one further step and examined if the code attached to papers is actually executable. Then, we proceeded to compare the resulting figures to those in the original article.

3. Methods

In order to obtain an initial but comprehensive overview of reproducibility in the computational geosciences and to better understand obstacles impeding *ORR*, we ran four complementary studies: an online survey, semi-structured interviews, a focus group discussion with geoscientists who conduct computational research, and a reproducibility study. The combination of these methods enabled us to gather qualitative and quantitative data directly from researchers in the geosciences as well as to objectively assess of how reproducible recent papers in the computational geosciences are. By interrelating both types of data, we hoped to be able to gain deeper insights than either method could offer on its own. To keep the scope of the research manageable, we focused on authors and readers as participants and a subset of publication outlets with readily available material.

3.1. *Approaches*

Online survey. Online surveys are efficient means for collecting responses from a large number of (Lazar *et al*. 2017). Our goal was to examine key aspects in reproducible research, i.e. the accessibility of code and data, published by scientists with a geoscientific background. We analysed the data using descriptive statistics and diverging stacked bar charts as suggested by Heiberger and Robbins (2014).

Semi-structured interviews. To receive deeper insights into what geoscientists understand by the term *ORR* and what obstacles they face when publishing *ORR* or when reproducing others' work, we conducted semi-structured interviews. During interviews, participants can express their thoughts freely and the interviewer can ask more concrete or follow-up questions if required (Lazar *et al*. 2017). We applied the grounded theory approach for data analysis (Glaser and Strauss 1967).

Focus group. In focus group discussions, participants can interact with each other, because such conversations may elicit opinions and ideas different from those mentioned in interviews (Glaser and Strauss 1967). We thus organised a focus group session to complement the interviews and the other data we gathered. We used the same topics and the same grounded theory approach as we did with the interviews.

Reproducibility study. In order to objectively assess the technical issues that make it difficult to reproduce others' work, we also conducted a reproducibility study. We systematically collected papers that had included source code written in *R*, and we then executed the analysis. During the study, we took note of any issues and how we were able to solve them. The resulting insights enabled us to derive recommendations for authors on how to avoid these issues.

In the following sections, we report on each study and its key results in detail.

3.2. *Online survey*

We conducted an online survey in order to assess whether geoscientists who conduct computational research publish *ORR*.

3.2.1. *Participants*

We recruited respondents during a poster presentation at the European Geosciences Union General Assembly (EGU) 2016.[5] In total, 13,650 scientists from 109 countries and several research areas[6] (e.g. biogeosciences, climatology) participated in the conference. In addition, we emailed 1,554 researchers who contributed to the conference with a poster or talk. To be included into the analysis, participants had to submit the survey actively . In total, 215 geoscientists started filling out the survey, of which 146 completed it (mean $\mu = 17$ years in research, standard deviation $\sigma = 9$ years). In our analysis, we only included the responses by participants who had submitted the survey and who had a research background in the geosciences.

3.2.2. *Materials*

After defining *ORR*, the survey collected background information about the respondents, i.e. whether they were authors, readers, or both, and their research field. We asked them how often (i) they published recomputable results, how frequently their papers linked to (ii) the data used and (iii) the code employed, and (iv) how often they tried to reproduce the results of other researchers. If they answered (ii) and (iii) in any way other than 'never', we also asked how frequently they included persistent identifiers (e.g. *Digital Object Identifier* (DOI)). Respondents answered the frequency questions using a five-item scale from 'never' to 'always'. We evaluated the data with the help of descriptive statistics. The survey was available for six months (April-September 2016). All the materials we used for the survey, i.e. the questions, data, and code (reproducible RMarkdown document) are included in the supplements.

3.2.3. *Results*

Of those who responded to the survey, 49% indicated they published their research often or always in a way that enables re-computation (Figure 1). However, only 33% included links to the data underlying the paper. Among those who included such links, 27% included persistent identifiers. Only 12% of all authors linked to the code used to produce the results, and among those who did, 12% included persistent identifiers. Of all survey respondents, 7% tried to reproduce other researchers' results often or always, and among those who answered this question other than never, 24% succeeded often or always. We can thus observe that there is a mismatch between those who said they published re-computable research and the frequency with which data and, particularly, code were shared. Figure 1 summarises the responses obtained from all participants.

3.3. *Semi-structured interviews*

We conducted semi-structured interviews with geoscientists who conducted computational research to investigate their understanding of *ORR* and identify barriers to the realisation of *ORR*. We recruited nine geoscientists (mean $\mu = 9$ years in research) from
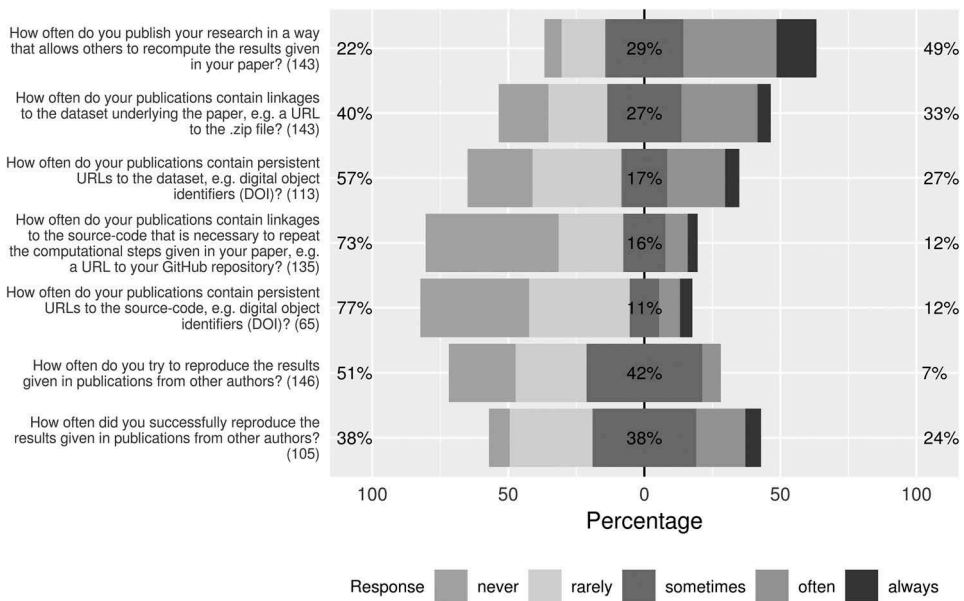
Figure 1. The diverging stacked bar chart (Heiberger and Robbins 2014) shows the percentage of respondents publishing reproducible research and reproducing the work of other researchers. Numbers in brackets show absolute numbers of respondents. Percentages are grouped into the categories never/rarely, sometimes, and often/always. Note: All except the last question were only shown to authors and not to readers.

geoinformatics, landscape ecology, geochemistry, and planetology (from within our faculty[7]) who had previously published papers that included geospatial figures based on computations, e.g. maps or time series.

3.3.1. *Materials*

The interview began with a brief introduction to the overall topic. We then asked participants to explain how they understand *ORR* in three consecutive steps. First, we asked what is meant by *reproducible research*, then *open research*, and finally *ORR*. Next, we presented our definition (see above) so that we could continue the interview with a common understanding. We then asked for obstacles participants perceive that hinder them from publishing *ORR* (author's perspective) and prevent the reproduction of other researchers' work (reader's perspective). Finally, participants had to fill out a brief survey to collect background information about them (research field, years in research).

3.3.2. *Procedure*

In order to ensure that the interview questions were understandable, we tested the interview with three Ph.D. students and revised the questions according to their feedback. All actual participants of the study received the final questions one day in advance. We recorded the interviews for later transcription (audio only). Before the interview started, participants were presented with a consent form that informed them about the audio recording, about their rights and about their statements being treated

anonymously. After they had signed it, we asked a series of questions during the actual interview and handed out a short questionnaire at the end. On average, the interview took 54 min (between 35 and 66 min). Seven interviews were conducted in German, two in English. We applied grounded theory (Glaser and Strauss 1967) to analyse the data. We captured key statements and assigned these statements to codes ('open coding'), which were then grouped into higher level themes and finally into categories. The supplemental files to this article include all materials we used for the interview, i.e. the questions we used, the questionnaire, the statements from the interviewees, the codes and categories we derived as well as the consent form.

3.3.3. Results

In line with existing literature, we found that geoscientists have a divergent understanding of *ORR*. For eight interviewees, *reproducible research* should describe the methods that were used to produce the results in sufficient detail for them to be repeated by others. They expected that reproducing such studies would achieve consistent results (7 mentions) and that reproducible research should make materials, e.g. code and data, accessible (3). Accessible materials were also relevant in *open research* (5), which should also be transparent (3) and free of charge (3). Interviewees combined these aspects to describe *ORR*, i.e. public access to data (5), code, and methods (4) to achieve the same results (3). Moreover, in *ORR* there should be an explanation of how results were produced (3) and the research components should have non-restrictive licenses. Three associated the term with *replicability*, i.e. confirming results with independent experiments (Bollen *et al*. 2015).

Furthermore, interviewees named several *obstacles* hindering the publication of open and reproducible results and preventing the reproduction of other researchers' results (Table 1): Insufficiently described methods were seen as impeding the understanding of how results were produced. Frequently, materials needed for reproducing results were inaccessible due to scientists' fears or concerns (e.g. regarding legal issues). Another problem was the use of proprietary tools, which can encapsulate essential processes such as how results are computed and thus can decrease transparency. In addition, several interviewees considered reproducible research as not being relevant for them yet. They also argued that their code was developed for a specific use case and thus not worth publishing, as others would not be able to reuse it. Finally, working reproducibly was seen as being too time consuming and not sufficiently supported by tools.

3.4. Focus group

In order to complement the insights from the interviews regarding researchers' understanding of *ORR* and the *obstacles* hindering *ORR*, we also conducted a focus group discussion. We recruited five additional geoscientists (mean $\mu = 5$ years in research) from our faculty with backgrounds in landscape ecology and geoinformatics based on the same criteria we used for recruiting interviewees.

3.4.1. Procedure

The focus group discussion comprised the same three parts as the interviews. Participants received the guiding questions one day in advance. On the day of the focus group session,

Table 1. Obstacles while publishing reproducible research (left) and while reproducing other researchers' work (right). Numbers in brackets show how many interviewees mentioned the obstacle. Aspects marked with (*) were also mentioned in the focus group discussion.

Obstacles while publishing reproducible research (authors)	Obstacles while reproducing others' work (readers)
Describe methodology sufficiently (5)	Missing details in methodology (7)*
Losing competitive advantages (4)*	Inaccessible materials (3)*
Prepare code and data (4)	Not yet relevant (3)*
Not yet relevant (4)	Proprietary software (2)
Proprietary software (4)	Time consuming (2)
Missing supporting tools (3)	Lack of expertise (1)
Licensing (3)*	Individual interpretations lead to other conclusions*
Code not worth publishing (2)	
Making it understandable for non-experts (1)	

we used the same questionnaire and consent form. We briefly introduced the topic, asked participants to introduce themselves, and then asked the questions. The focus group took 86 min in total and was conducted in German. Statements were analysed using the same grounded theory approach we applied for the interviews. All materials used for the focus group, i.e. the statements, codes and categories, the questionnaire, as well as the consent form are available in the supplements.

3.4.2. Results

Participants of the focus group described their understanding of *ORR* one after another and also referred to each other. They collaboratively achieved the following definition: *ORR* contains a detailed description of the methodology (3 mentions) and provides access to data (2), model, and code (3). Readers can achieve consistent results (5) by repeating the analysis. This then leads to research being more transparent (5), because readers better understand how results were achieved and which limitations they have.

The participants of the focus group discussion made many statements that were similar to those mentioned in the interviews (cf. Table 1 for a summary of repeated statements). In addition to those statements, there were also a number of points that did not come up during the interviews. One participant pointed out that reproduction might fail due to individual interpretations that could differ from one researcher to another. Using different versions of the same software could result in some required functionality not being included or lead to deviating results due to the functionality having changed from one version to another. Thus, reproducible research is not necessarily achieved even when the used materials are accessible. To better understand the relevance of this observation, in our reproducibility study (below) we decided to reproduce results from papers that made available all their original materials, namely the spatial statistics and code.

3.5. Reproducing geoscientific results

A key benefit of *ORR* is that other scientists can reuse existing materials such as code. However, this is only practical if the code is executable and produces the same results as those reported in the paper. In order to further examine obstacles for readers while reproducing published work, we tried to execute the code attached to papers and compared the figures depicting results in the reproduced and original versions.

3.5.1. Materials

We selected a paper for our study if it met the following criteria: (i) it was licensed as open access; (ii) it provided links to code written in R core (2018), a programming language that is used frequently in the geosciences (Giraud and Lambert 2017), and linked to data that was used (if applicable); and (iii) it was published between January 2016 and August 2017. The latter criterion was used in order to capture current practices of researchers and to keep the scope of the work manageable. We began our search by scanning the journals published by *Copernicus Publications'*,[8] which are all open access and many of which fall into one of the aforementioned geoscientific domains. In a first run, we searched for papers using the keyword *R Core Team*, as this term is used frequently to cite R as the programming language underlying the developed software. In order to find cases where the authors did not cite the programming language but rather linked to the externally hosted code, we searched a second time using the term *GitHub*, which is a popular platform for storing and sharing source code. This two-step search yielded 31 research articles from the geoscientific domain that met our three criteria. Because computational analyses used in articles are often based upon other software libraries, we broadened our analysis and also considered the 10 most cited papers of the *Journal of Statistical Software*[9] (state June 2016) that describe frequently used R packages based on code written in R. Although these papers are not specific to the geosciences, they describe important features for spatial statistics, e.g. handling spatial and temporal data.

3.5.2. Procedure

In order to execute the code included in the papers, we set up *RStudio* (RStudio Team 2015) on an Ubuntu system (Version: 16.04) using *rocker/geospatial* (Version 3.4.2), a Docker image tailored to the geoscientific domain (Boettiger and Eddelbuettel 2017). If we encountered issues while running the scripts that we were unable to resolve by ourselves, we searched the Web for solutions.[10] If this was unsuccessful, we contacted the corresponding author. If they did not reply within four weeks, we considered reproduction to have failed for this paper. Once the scripts compiled without issues, we did not further inspect the code or make any changes.

We documented all technical issues and how we solved them. Each issue we encountered was categorised into one of the four categories: *minor, substantial, severe, and system-dependent issues*. All scripts that successfully compiled were then executed, and we saved all figures that they generated during execution. Figures consisting of several sub-figures (e.g. Figure a, b) were considered as one figure. The generated figures were then visually compared to the original ones in a side-by-side manner. Since figures (such as maps or time series) are frequently used to relay key results in academic papers, comparing those produced during reproduction to the figures in the original paper is one way to confirm that the reproduced results are identical to the reported ones. During this comparison, we recorded any differences that we found. Each type of difference, such as a label being different in the reproduced and original figure, was counted only once. This approach was chosen to avoid issues with how to count different types of differences and to prevent over emphasising consistent but repeated differences (e.g. numeric labels next to an axis all changing due to the depicted range being different in the reproduced and the original figure).
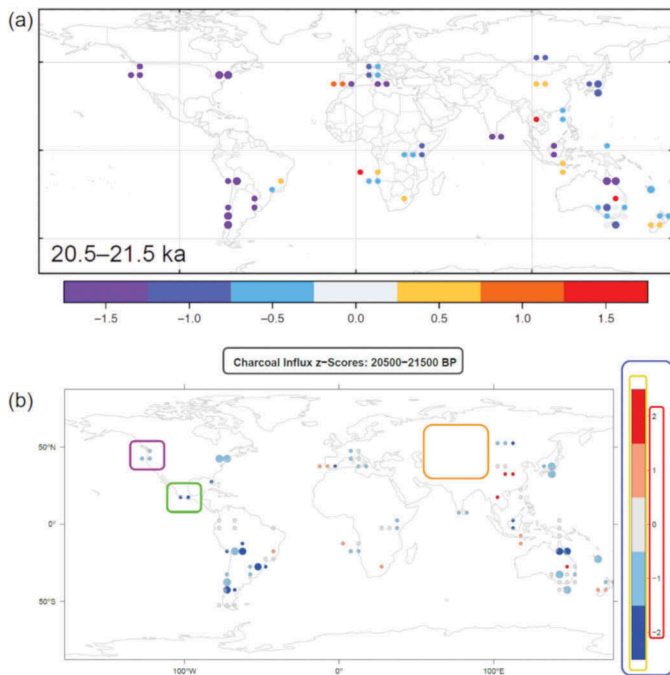
Figure 2. Differences between the original (a) and reproduced figure (b) (Marlon *et al*. 2016). Design: Placement of legend (blue box), aspect ratio, colour of results (purple), data type of legend numbers (red). Content: labelling (grey), background map (orange), number of classes (yellow), deviating results (green).

The supplemental material attached to this article includes a list of all papers that we examined as well as a reproducible RMarkdown document for Figure 2.

3.5.3. *Results*

Below, we first report on the technical issues we encountered and then summarise the differences between the original and reproduced figures that we observed.

Technical issues: The code of two papers ran without any issues, 33 had resolvable issues, and two were partially executable, i.e. the code produced output but also had issues that we could not resolve. We classified four papers as being irreproducible, as we could not solve all issues. The code of 15 papers contained issues that required contacting the corresponding author. Eleven authors helped us to find solutions, e.g. by pointing out code changes or solutions to the problems. Five of them sent additional code and data. One author helped with some but not all issues; three authors did not reply within the four-week time limit. In total, we encountered 173 issues in 39 papers (mean $\mu = 4.4$ issues per paper), which we categorised as follows (see Table 2).

Minor issues were defined as being resolvable without any manipulation of the code that was provided by the authors. They mainly resulted from code calling a library that was not installed but could be found in public repositories, i.e. *CRAN* for *R*. This issue emerged 49 times in 24 papers (49/24). *Minor issues* also comprised negligible issues (4/

Table 2. Issues we encountered during code execution. Numbers in brackets show how often and in how many papers they occurred (overall occurrence/number of papers). In total, we encountered technical issues in 39 papers.

Minor (53, 25)	Substantial (73, 25)	Severe (41, 22)	Sys.-dependent (6, 5)
Library not found but available in repository (49, 24)	Wrong directory (34, 13)	Flawed functionality (13, 9)	Insufficient RAM (2, 2)
Faulty variable call (4, 3)	Deprecated function (10, 4)	Missing data or code (11, 9)	Function behaves differently across OSes (3, 3)
	Output not storable in local folder (10, 2)	Flawed data integration (11, 8)	Installing libraries on different OSes (1, 1)
	Function not found or missing library (9, 4)	Code in PDF (6, 6)	
	Library not found and not in repository (8, 7)		
	Broken link (2, 2)		

3), e.g. a faulty function irrelevant for further computations. In total, we encountered 53 *minor issues* in 25 papers. Six papers only had *minor issues*.

Substantial issues (73/25) required manipulating the code in order to resolve them, e.g. by adjusting file directories (34/13). We encountered deprecated functionalities (10/4) that had to be resolved by installing an archived version of the corresponding library or by using the current version with potentially deviating results. *Substantial issues* also resulted from functionalities which saved outputs locally but did not execute properly (10/2). These issues could be addressed by plotting the results within the programming environment. Further issues in this category were libraries that did not exist in the *CRAN* repository (8/7). scripts called functions without explicitly importing the library that provided these functions (9/4). Again, we had to find and embed the right library. The search was more demanding when the name of the library was unknown, as different libraries can provide the same function name but with different implementations. We considered one reproduction as failed since we were unable to find an outdated library that provided the functions called in the script. Finally, in two papers the links to the required materials were broken, which meant that we had to contact the author and look for other repositories from the same author in order to access those materials.

Severe issues (41/22) required a deeper understanding of the source code and the programming language in order to be resolved. These issues arose when it was necessary to adapt functions or parameters (13/9) or when data or code segments were missing (11/9); while we were able to resolve the issues of one paper by ourselves, we had to contact the authors of the other eight papers to tackle the issues that emerged. Five authors sent the required material and one updated the library that had caused issues. Two authors did not reply within the four-week period, which meant that we classified the reproduction as failed. Further causes for severe issues included data not loading correctly (11/8) and having to extract code from a PDF (6/6), which entailed a number of copy-and-paste issues.

System-dependent issues (6/5) relate to problems resulting from the computational environment in which the code was run. Two analyses required more random access memory (RAM) than was available on the machine we used for our study. Some scripts (partially) failed when run inside a Docker container but worked fine when we ran them

Table 3. Number of design and content differences between original and reproduced figures. Bolded numbers in brackets show how many figures had the corresponding figure component. Italic numbers in brackets show in how many papers the differences occurred.

Comparability	Total	Legend (38)	Labelling (93)	Results	Aspect Ratio	Axes (90)	Placement	Background (50)
Design	315 *(28)*	18 *(12)*	60 *(20)*	44 *(19)*	78 *(25)*	64 *(22)*	18 *(11)*	33 *(9)*
Content	105 *(27)*	15 *(11)*	21 *(13)*	46 *(20)*	N/A	13 *(9)*	N/A	10 *(4)*

on Windows (3/3). Finally, the installation of libraries might be different in a Docker container and on Windows (1/1).

Differences between original and reproduced figures. The code of 28 out of 41 papers produced 97 figures, which we compared to the ones contained in the original paper. We observed the following differences which impeded the comparability of figures and deviations between original and reproduced figures (Table 3).

Cosmetic differences: 78 out of 97 reproduced figures had a different *aspect ratio* than the original one. We encountered 44 cases where the visualisation of the *results* differed regarding line widths or colours of barcharts and data points. In total, 90 figures consisted of diagrams that included *axes*, and 64 of those had a different font, interval, or data type (e.g. '2' instead of '2.0'). The *axes* were missing or had a different scale unit in a further 13 cases. We counted 50 figures that had a *background* (e.g. maps, grids). In 33 of these, the line widths of boundaries or grid structures differed, and in 10 cases the level of detail was different or the grid was missing entirely. A *legend* was present in 38 figures, and in 18 cases, it differed in terms of colours, font, or data types. Frequently, the *legend* was completely absent or incomplete (15 cases). Further differences we spotted relate to the *placement* of figure components (e.g. subfigures), which differed in 18 cases. *Labels* were present in 93 figures; we counted 60 cases with different fonts and 21 cases with a different or missing text. Overall, we counted 374 cosmetic differences that can affect the comparability of original and reproduced figures: 315 of those differences were related to the design of the figures and 59 differences were related to the actual content of the figures.

Deep differences: The results of 46 out of 97 reproduced figures deviated from the original figures on a deeper level, e.g. graphs had different curves, and key numbers were missing or different. These differences make it harder for readers to determine whether or not the reproduced figures depict the same results as those shown in the original figures. We did not find systematic correlations between the specific issues, cosmetic differences discussed, and deeper differences. In the case of deeper differences, confirming successful reproduction will thus most likely require a deeper inspection, e.g. using raw and/or intermediate results produced by the scripts.

In order to illustrate the differences and the resulting difficulties when comparing an original and a reproduced figure, consider the following example from Marlon *et al.* (2016). The different *aspect ratio* leads to a different appearance that might be interpreted as the results being different although the actual numbers from the original and the reproduced analysis are identical. Figure 2 shows a real example in which we highlighted typical differences in the reproduced figure. We are very grateful to the corresponding author (Marlon *et al.* 2016) for giving us permission to include their figure as an example to illustrate typical differences that can occur during reproduction.

4. Discussion

The main goal of the work presented in this article was to shed light on how reproducibility is perceived and practiced in the computational geosciences. In our studies, geoscientists had divergent perceptions of what ORR means. Half of the survey participants said they published re-computable research results, but they rarely linked to source code and data underlying the figures and numbers. Interviewees already conducted computational research such as spatial statistics but focused mainly on the methodology and associated it with *replicability* (Bollen *et al*. 2015). As a detailed description of the methodology can be sufficient to ensure replicability, this might explain the high number of respondents claiming that they publish reproducible research though many did not attach code and data. Three interviewees initially struggled when asked to define the term, which might indicate that the topic is not fully present in their daily work. The low number of geoscientists who regularly reproduce other researchers' work (7%) confirms this impression. The fact that only few of them succeeded confirms the findings reported by Baker (2016).

The *obstacles* hindering *ORR* mentioned by interviewed geoscientists also confirm what has been reported in the literature for other disciplines (cf. Baker 2016). It seems some of the obstacles impede both authors and readers, such as the use of proprietary software preventing code publication and reproduction. The results from the questionnaire and the interviews indicate that realising *openness* is one key issue in reproducible research, which provides support for the current trend towards openness in science in general (open access, open data, open code).

While executing the code of the papers, we encountered a substantial number of technical issues and observed various differences between original and reproduced figures. The code of only two papers was executable without any issues. Thirty-three out of 41 papers were executable after we resolved different types of issues that varied in terms of severity and effort required to address them. We also came across a number of *system-dependent issues* that resulted from implicit assumptions about the underlying system (e.g. the operating system). These issues might highlight the importance of describing the computational environment in reproducible research. Reproducibility in general hinges on being able to achieve the same results. The figures showing results (e.g. maps) provide an effective way to quickly compare original and reproduced results. In our study, we identified numerous cosmetic and deeper differences between the figures in the original publication and the figures that were generated during reproduction (see Figure 2 for an example). It was not straightforward to determine whether results deviate because reproduced figures had a different aspect ratio or because the computational steps produced a different outcome. Finding the right configuration (e.g. parameters) to produce identical figures usually requires effort and knowledge of the code. These figure-related issues as well as the technical ones mentioned above provide initial evidence that 'just' making code and data publicly available oftentimes does not guarantee that others can execute the analysis and produce identical results.

Besides code, it is still important to consider data underlying a geospatial analysis. Ideally, authors should provide their geographic data following the *FAIR* principles (Wilkinson *et al*. 2016): Data should be findable, i.e. by persistent identifiers; accessible, i.e. for free; interoperable, i.e. by using open formats (e.g. GeoJSON instead of shapefile);

and reusable, i.e. by using open licenses. Several other approaches might counteract the obstacles in *ORR*. Educating graduate students, offering workshops for scientists, or conducting hands-on seminars addressing the technical difficulties while publishing reproducible results might be promising solutions (cf. Leek and Jager 2017). Such educational initiatives could increase the awareness of *ORR* not only for one's own research but also while reviewing others' submissions.

Making available tools that integrate well with geoscientists' existing workflows and that provide support for open reproducible work might be another key element for boosting reproducibility in the geosciences. Some initial proposals have been made in this field (e.g. Nüst *et al*. 2017), and the common domain of space and time holds great potential in this respect. Geographic information systems, satellite imagery, and geospatial analysis systems (such as libraries for geospatial statistics in R) are widely used throughout the geosciences. Integrating those systems and datasets into reproducibility tools may drastically lower the effort needed to work reproducibly while keeping in line with geoscientists' existing workflows.

While it is highly desirable and necessary to address technical issues, provide supportive tools, and further educate geoscientists about *ORR*, these steps alone are probably not sufficient to eliminate the practice of publishing irreproducible research. The overall culture, processes, and reward systems around scientific work need to be adjusted, too. Scientists (including but not limited to those who conduct computational research) need to shift towards ORR in their working methods (Markowetz 2015). In addition, fears and worries of (geo-)scientists need to be addressed. For example, we observed that some authors were reluctant to fully disclose their own work since they feared that others would either 'steal' it or find issues that might damage their reputation. More fine-grained citation systems (e.g. for data, code) and 'evolving' publications (e.g. where credits are given to improvements proposed by researchers other than the authors) might be ways to deal with such fears. In addition, there is the issue of a lack of incentives to reproduce other researchers' work. If reviewers would receive rewards for their effort to reproduce submissions (Stodden *et al*. 2016), some issues could already be resolved during the review process and reproduction could become more widely practiced. Furthermore, publication outlets might consider desk-rejecting submissions that are not reproducible.

Limitations. The work presented in this article aims to shed light on how ORR is perceived and practiced in the computational geosciences. In order to keep the scope of the research manageable, we made several assumptions and decisions which limit the generalisability of our results. One key limitation pertains to the study participants. The number of participants and their research areas do not represent the diversity of the geoscientific domain. Moreover, interviewees and focus group participants were recruited from the same faculty. Researchers from other institutions or countries might hold different perceptions of *ORR* or may be able to identify other obstacles. Although we contacted scientists from several geoscientific domains to complete the survey, the total number of respondents (146) still does not represent all geoscientists. Consequently, some research domains within the geosciences might not have been included in our studies. In addition, it is also likely that the survey was completed by people who are inherently interested in reproducible research, which thus could have

introduced some bias. A truly representative selection from all geoscientists and all geoscientific domains is difficult to achieve.

Another limitation pertains to the selection criteria for papers we used in the reproducibility study. Due to practical reasons (e.g. the authors' familiarity with certain programming languages and tools) as well as time constraints, we focused on very recent papers from few outlets published by *Copernicus* and the *Journal of Statistical Software*. While this decision enabled us to systematically evaluate the reproducibility of the selected papers, it also reduces representativeness of the findings. However, in recent years, new technologies have emerged that assist scientists in sharing code and data (e.g. GitHub and Zenodo). Hence, recent papers likely do show the best current practice, and older papers might show even worse results. More papers from other outlets are needed to draw informed conclusions. In addition, we only focused on *computational* reproducibility rather than *general* reproducibility. While this limits the scope of the article, we argue that computational reproducibility is particularly relevant in the geosciences since many papers include results that are produced by source code. Although the identified technical issues were specific to the *R* programming language, it seems very likely that similar issues might also emerge when using other programming languages (such as Python, which is also very popular in the geosciences). In order to confirm this, we plan to run another study with articles based on Python to see if similar or different issues emerge and with what frequency.

Despite us defining *ORR* explicitly in the studies, participants might still have had replicability in mind while answering the questions, which might constitute another limitation. Since we conducted the studies in German and English, we took great care to provide identical definitions of *ORR* in both languages. Though in principle translation issues might constitute another limitation, we consider this unlikely. Finally, we did not investigate the authors' intentions behind attaching materials, i.e. whether they did it to facilitate reproduction of the results or just as supplements. This aspect could explain some of the issues, differences, and deviating results. If authors anticipate that readers want to reproduce their work, they might take greater care when preparing data and code than when authors only include material for completeness or to comply with a journal's requirements.

5. Conclusion

Reproducibility is essential to computational research in general and consequently to papers that apply spatial statistics based on code and data. In order to shed light on *ORR* in the computational geosciences, we reported on a series of studies that examined issues, perceptions, and practices related to *ORR*. We conducted an online survey (146 responses), semi-structured interviews ($n = 9$), a focus group ($n = 5$), and a reproducibility study with 31 articles including spatial statistics published by *Copernicus Publications* and 10 papers from the *Journal of Statistical Software*. Our main contributions are the initial identification of (1) geoscientists' understanding of the term *ORR* and (2) practical obstacles which might hinder reproducibility in the computational geosciences. Moreover, we report on (3) issues arising when reproducing papers that apply spatial statistics based on code and data, and (4) differences in the reproduced results. Our final contribution is the provision of (5) a set of guidelines for authors wishing to publish reproducible work.

The results from our studies indicate that geoscientists might have a divergent understanding what reproducibility means, and that reproducibility might be hindered by a lack of openness regarding data, code, and proprietary software. In order to gain direct insight into current practice, we tried to execute the source code attached to 41 scientific publications and analysed the figures that were generated during reproduction. In total, we identified 173 issues, which we classified into four categories: *minor, substantial, severe*, and *system-dependent* issues. We compared 97 original and reproduced figures and detected 420 differences, which were either cosmetic or of a deeper nature. It appears that publishing code and data with a paper does not guarantee that the reported results can be easily reproduced and that the figures generated during reproduction are identical to those in the original paper. To overcome the issues, we propose the following guidelines for publishing computational research reproducibly:

(1) Embed and install libraries within the source code and include their version numbers to facilitate finding the right libraries.
(2) Make directories relative to a top directory (see https://www.r-bloggers.com/making-an-r-package-to-use-the-here-geocode-api/here package for *R*) instead of the author's computer.
(3) Do not modify the source code once the results are copied into the paper, as later changes of the code might affect already extracted results.
(4) Execute the code in a clean programming environment after completing the analysis, e.g. by using *rocker/geospatial*, to spot issues other readers might have.
(5) Publish input data and processed data of figures to enable readers to assess whether deviating figures result from differences in the data analysis or in the settings of the systems used to generate the original and the reproduced figures.
(6) Encapsulate code and data in a project folder to facilitate execution (Fehr *et al*. 2016), e.g. as an *Executable Research Compendium* (Nüst *et al*. 2017).
(7) Provide code and data in original files instead of PDFs to avoid cut-and-paste issues and to lower the burden for readers to rerun and reuse the analysis.
(8) Use code to produce and design figures (Sandve *et al*. 2013), ideally embed- ded in executable documents, e.g. RMarkdown (Allaire *et al*. 2016) or Jupyter notebooks (cf. Kluyver *et al*. 2016), to avoid scaling issues.

Our recommendations confirm and complement suggestions regarding reproducible computational analyses in general Bailey 2016 facilitating but were specifically designed to address the issues we encountered when reproducing the 41 selected papers. The recommendations could also be used as author guidelines in conferences and journals. In addition, they might inform the design of tools that support working reproducibly.

Notes

1. Distill: https://distill.pub/journal/ Accessed 30.07.2018.
2. ReScience: http://rescience.github.io/ Accessed 30.07.2018.
3. GitHub: https://github.com/ Accessed 30.07.2018.
4. Open Science Framework: https://osf.io/ Accessed 30.07.2018.
5. EGU: https://www.egu2016.eu/. Accessed 30.07.2018.

6. See https://meetingorganizer.copernicus.org/EGU2016/sessions-of-special-interest/ECS for a full list of research domains. Accessed 30.07.2018.
7. Institute for Geoinformatics, University of Münster.
8. Copernicus Publications: https://publications.copernicus.org/ Accessed 30.07.2018.
9. Journal of Statistical Software: https://www.jstatsoft.org/index Accessed 30.07.2018.
10. see e.g. StackOverflow: https://stackoverflow.com/ Accessed 30.07.2018.

Acknowledgement

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Notes on contributors

Markus Konkol is a research associate in the DFG-project 'Opening Reproducible Research' at the Institute for Geoinformatics, University of Münster. He is currently doing his Ph.D. in the DFG-funded project Opening Reproducible Research that aims at facilitating access to and interaction with research results. His main topic is about interconnecting the textual publication, data sets, source code, and UI elements in order to assist scientists in exploring and creating dynamic publications.

Chris Kray is a professor of Geoinformatics at the Institute for Geoinformatics (ifgi) at the University of Münster. His research interests include location-based services, smart cities and human-computer interaction (in particular: interaction with spatial information). Chris is the scientific coordinator of the ITN 'GEO-C: enabling open cities' at ifgi, where he works on realising transparency, accessibility and privacy protection in the context of smart cities.

Max Pfeiffer is a postdoc at the Institute for Geoinformatics (ifgi) at the University of Münster. His research interests include human-computer interaction, electrical muscle stimulation, and user centered design.

ORCID

Markus Konkol http://orcid.org/0000-0001-6651-0976
Christian Kray http://orcid.org/0000-0002-4199-8976

References

Allaire, J., *et al.*, 2016. rmarkdown: dynamic documents for R. *R Package Version*, 1, 9010.

Bailey, D.H., Borwein, J.M., and Stodden, V., 2016. Facilitating reproducibility in scientific comput- ing: principles and practice. *In*: H. Atmanspacher and S. Maasen, eds. *Reproducibility: principles, problems, practices*. New York: John Wiley and Sons.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533 (7604), 452–454. doi:10.1038/533452a

Barnes, N., 2010. Publish your computer code: it is good enough. *Nature News*, 467 (7317), 753. doi:10.1038/467753a

Benestad, R.E., *et al.*, 2016. Learning from mistakes in climate research. *Theoretical and Applied Climatology*, 126 (3–4), 699–703. doi:10.1007/s00704-015-1597-5

Boettiger, C. and Eddelbuettel, D., 2017. *An introduction to rocker: docker containers for r*. arXiv preprint. Ithaca, NY, arXiv:1710.03675.

Bollen, K., *et al.*, 2015. *Social, behavioral, and economic sciences perspectives on robust and reliable science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, 3. Arlington, VA: National Science Foundation.

Brunsdon, C., 2016. Quantitative methods i: reproducible research and quantitative geography. *Progress in Human Geography*, 40 (5), 687–696. doi:10.1177/0309132515599625

Claerbou, J.F., *et al.*, 1992. Electronic documents give reproducible research a new meaning. *In*: *1992 SEG Annual Meeting*. Society of Exploration Geophysicists. New Orleans: Society of Exploration Geophysicists.

Collberg, C. and Proebsting, T.A., 2016. Repeatability in computer systems research. *Communications of the ACM*, 59 (3), 62–69. doi:10.1145/2897191

Costello, M.J., 2009. Motivating online publication of data. *BioScience*, 59 (5), 418–427. doi:10.1525/bio.2009.59.5.9

Darch, P.T. and Knox, E.J., 2017. Ethical perspectives on data and software sharing in the sciences: a research agenda. *Library & Information Science Research*, 39 (4), 295–302. doi:10.1016/j.lisr.2017.11.008

Donoho, D.L., *et al.*, 2009. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering*, 11 (1), 8–18. doi:10.1109/MCSE.2009.15

Easterbrook, S.M., 2014. Open code for open science? *Nature Geoscience*, 7 (11), 779–781. doi:10.1038/ngeo2283

Fehr, J., *et al.*, 2016. Best practices for replicability, reproducibility and reusability of computer- based experiments exemplified by model reduction software. *AIMS Mathematics*, 1 (3), 261–281. doi:10.3934/Math.2016.3.261.

Gentleman, R. and Temple Lang, D., 2007. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16 (1), 1–23. doi:10.1198/106186007X178663

Gertler, P., Galiani, S., and Romero, M., 2018. How to make replication the norm (vol 554, pg 417, 2018). *Nature*, 555 (7698), 580.

Gewin, V., 2016. Data sharing: an open mind on open data. *Nature*, 529 (7584), 117–119. doi:10.1038/nj7584-117a

Gil, Y., *et al.*, 2016. Toward the geoscience paper of the future: best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, 3 (10), 388–415. doi:10.1002/2015EA000136

Giraud, T. and Lambert, N., 2017. Reproducible cartography. *In*: *International Cartographic Conference*. Washington: Springer, 173–183.

Glaser, B. and Strauss, A., 1967. Grounded theory: the discovery of grounded theory. *Sociology the Journal of the British Sociological Association*, 12, 27–49.

Goodchild, M.F., 1992. Geographical information science. *International Journal of Geographical Information Systems*, 6 (1), 31–45. doi:10.1080/02693799208901893

Goodman, S.N., Fanelli, D., and Ioannidis, J.P., 2016. What does research reproducibility mean? *Science Translational Medicine*, 8 (341), 341ps12–341ps12. doi:10.1126/scitranslmed.aaf0746

Heiberger, R.M. and Robbins, N.B., 2014. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software*, 57 (5), 1–32. doi:10.18637/jss.v057.i05

Hillebrand, H. and Gurevitch, J., 2013. Reporting standards in experimental studies. *Ecology Letters*, 16 (12), 1419–1420. doi:10.1111/ele.12190

Ioannidis, J.P., *et al.*, 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41 (2), 149–155. doi:10.1038/ng.295

Kluyver, T., *et al.*, 2016. Jupyter notebooks-a publishing format for reproducible computational workflows. *In*: *ELPUB*. 87–90.

Knuth, D.E., 1984. Literate programming. *The Computer Journal*, 27 (2), 97–111. doi:10.1093/comjnl/27.2.97

Lazar, J., Feng, J.H., and Hochheiser, H., 2017. *Research methods in human-computer interaction*. Cambridge, MA: Morgan Kaufmann.

Leek, J.T. and Jager, L.R., 2017. Is most published research really false? *Annual Review of Statistics and Its Application*, 4, 109–122.

Leek, J.T. and Peng, R.D., 2015. Opinion: reproducible research can still be wrong: adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112 (6), 1645–1646. doi:10.1073/pnas.1421412111

Lovelace, R., Nowosad, J., and Muenchow, J., 2016. *Geocomputation with R*. Boca Raton, FL: CRC Press.

Markowetz, F., 2015. Five selfish reasons to work reproducibly. *Genome Biology*, 16 (1), 274. doi:10.1186/s13059-015-0667-4

Marlon, J.R., *et al.*, 2016. Reconstructions of biomass burning from sediment charcoal records to improve data-model comparisons. *Biogeosciences (BG)*, 13, 3225–3244. doi:10.5194/bg-13-3225-2016

McCullough, B.D., McGeary, K.A., and Harrison, T.D., 2008. Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne D'économique*, 41 (4), 1406–1420. doi:10.1111/j.1540-5982.2008.00509.x

Nature Geosciences, 2018. *Aims and scope*. Available from: https://www.nature.com/ngeo/about/aims

Nosek, B.A., *et al.*, 2015. Promoting an open research culture. *Science*, 348 (6242), 1422–1425. doi:10.1126/science.aaa5139

Nüst, D., *et al.*, 2017. Opening the publication process with executable research compendia. *D-Lib Magazine*, 23 (1–2). doi:10.1045/january2017-nuest

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716. doi:10.1126/science.aac4716

Ostermann, F.O. and Granell, C., 2017. Advancing science with vgi: reproducibility and replicability of recent studies using vgi. *Transactions in GIS*, 21 (2), 224–237. doi:10.1111/tgis.2017.21.issue-2

Peng, R.D., 2011. Reproducible research in computational science. *Science*, 334 (6060), 1226–1227. doi:10.1126/science.1213847

Piwowar, H.A., Day, R.S., and Fridsma, D.B., 2007. Sharing detailed research data is associated with increased citation rate. *PLoS One*, 2 (3), e308. doi:10.1371/journal.pone.0000308

R Core Team, 2018. *R: a language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing. Available from: http://www.R-project.org/

Reichman, O.J., Jones, M.B., and Schildhauer, M.P., 2011. Challenges and opportunities of open data in ecology. *Science*, 331 (6018), 703–705. doi:10.1126/science.1197962

RStudio Team, 2015. *RStudio: integrated development for R* [online]. Boston, MA: RStudio, Inc. Available from: http://www.rstudio.com/

Sandve, G.K., *et al.*, 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9 (10), e1003285. doi:10.1371/journal.pcbi.1003285

Skaggs, T., Young, M., and Vrugt, J., 2015. Reproducible research in vadose zone sciences. *Vadose Zone Journal*, 14 (10). doi:10.2136/vzj2015.06.0088

Steiniger, S. and Hay, G.J., 2009. Free and open source geographic information tools for landscape ecology. *Ecological Informatics*, 4 (4), 183–195. doi:10.1016/j.ecoinf.2009.07.004

Stodden, V., 2010. *The scientific method in practice: reproducibility in the computational sciences*. MIT Sloan Research Paper No. 4773-10.

Stodden, V., *et al.*, 2016. Enhancing reproducibility for computational methods. *Science*, 354 (6317), 1240–1241. doi:10.1126/science.aal1794

Vandewalle, P., 2012. Code sharing is associated with research impact in image processing. *Computing in Science & Engineering*, 14 (4), 42–47. doi:10.1109/MCSE.2012.63

Vandewalle, P., Kovacevic, J., and Vetterli, M., 2009. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26 (3), 37–47. doi:10.1109/MSP.2009.932122

Vines, T.H., *et al.*, 2014. The availability of research data declines rapidly with article age. *Current Biology*, 24 (1), 94–97. doi:10.1016/j.cub.2013.11.014

Wilkinson, M.D., *et al.*, 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.