Reproducibility and Replicability in SAR Remote Sensing

Timo Balz, *Senior Member, IEEE*, and Fabio Rocca

***Abstract*—Modern science is built on systematic experimentation and observation. Today's form of communicating scientific results with written articles is the foundation we build our work on. The reproducibility and replicability of the experiments and observations are central to this process of validation. However, reproducibility and replicability are not always guaranteed, sometimes referred to as "crisis of reproducibility." We believe that remote sensing, in general, suffers from this crisis. To analyze the extent of the crisis, we conducted a survey on the state of reproducibility in remote sensing. Based on this survey, we map the problem of reproducibility with a focus on synthetic aperture radar remote sensing, as this is our area of research. We also give advice on how to improve reproducibility in remote sensing.**

***Index Terms*—Remote sensing, replicability, reproducibility, survey.**

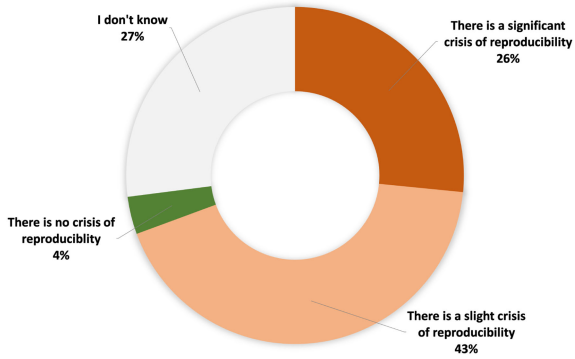

Fig. 1. Which of the following statements regarding a "crisis of reproducibility" in remote sensing do you agree with?

I. Introduction

REPRODUCIBILITY and replicability are pillars of science [1]. However, these terms are not clearly defined, and their meaning is different in different scientific communities [2]. Here, we use the definition of the National Academies of Science, Medicine, and Engineering [3]. Reproducibility is defined as the ability of researchers to obtain consistent results using the same input data; computational steps, methods, and code; and conditions of analysis [3]. Replicability is defined as a situation when researchers obtain consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data [3]. While replicability is a goal of our scientific endeavors, reproducibility is a basis for truth claims in every scientific work. These terms are sometimes interchanged or used as synonyms.

However, how many of the published results are actually reproducible? Macleod *et al.* [4] estimate that about 85% of biomedical research efforts are wasted. Ninety percent responded to a recent survey from Baker and Penny [5] that there is a "reproducibility crisis" [6]. The numbers are smaller in our survey, with about 70% believe there is a crisis (see Fig. 1), but still high. However, care has to be taken as the term reproducibility is not defined equally across disciplines and research cultures. Nevertheless, there are several reasons for this so-called crisis in reproducibility and replicability, including not only the increased career pressure on scientists, but also the increasing complexity and interdisciplinarity of the research, as well as the increase of the amounts of data [7].

Studies showing that scientific papers often leave out details that are essential for reproduction [8] may also indicate intent, as research groups hope to keep their advantage in the competition on research funding and career advancement. As understandable as that can be on an individual level, it is hurting science.

Remote sensing has replicability issues right at the core. The interpretation of images, even for basic elements such as lines, is interpreter dependent [9]. This has far reaching implications in terms of replicability, as, for example, the selection of training and test sets in classification can be determined by the interpreter's decisions. Even the selection of land use and land cover classes by subjective criteria can influence the significance of a study and hinder replicability [10]. Additionally, implicit assumptions on the relation between image features and ground attributes are not always true, even when using methods that should mitigate, e.g., atmospheric effects, such as band indices [11]. Temporary variation in reflectance even occurs on targets used for image calibration [12]. Other factors that can influence the reproducibility and replicability of research in remote sensing can be the preprocessing steps as done by the researchers, or even the computational back ends of data service providers [13].

With respect to these and other problems in reproducibility and replicability in remote sensing, the number of studies on the extent of reproducibility problems in remote sensing is rather

Manuscript received April 10, 2020; revised May 30, 2020; accepted June 22, 2020. Date of publication June 30, 2020; date of current version July 13, 2020. *(Corresponding author: Timo Balz.)*

Timo Balz is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: balz@whu.edu.cn).

Fabio Rocca is with the Department of Electronics, Information, and Bioengineering, Politecnico di Milano, 20133 Milano, Italy (e-mail: fabio.rocca@polimi.it).

Digital Object Identifier 10.1109/JSTARS.2020.3005912

limited. Ostermann and Granell [14] analyzed the reproducibility of studies using volunteered geographical information. In their findings, approximately one-third of the analyzed papers do not support reproducibility or replicability at all. The second group is considered replicable, as data and methods are properly documented, but only a very limited number are reproducible, as data or source code is not accessible. The largest group is that of limited replicability, as the papers describe and document the methods and tools used, but do not provide the details of the procedure and/or the data collection. None of the papers analyzed provided access to raw data used in the experiments, or analysis code and tools [14]. We assume a similar situation for papers in remote sensing.

In this article, we want to analyze the extent and the perception of the state of reproducibility in remote sensing. To this end, we decided to conduct our own survey. We designed an online survey targeting mostly experienced scientists in the field. This article reports the results of our survey. Based on these data, we analyze the current situation with respect to reproducibility and replicability in remote sensing, focusing on synthetic aperture radar (SAR) remote sensing, but believing that these issues also exist in remote sensing in general. We will then define some guidelines and suggestions to work toward a more rigorous remote sensing science.

The rest of this article is organized as follows. Section II introduces the problem of reproducibility in SAR remote sensing. Section III presents and discusses the results of our survey. Afterward, in Section IV, suggestions on how to improve reproducibility and replicability in remote sensing are given. In Section V, we discuss the issue with respect to modern publication forms. Finally, in Section VI, we draw our conclusions, hoping that the readers and the remote sensing community draw their conclusions as well.

II. Reproducibility in SAR Remote Sensing

In this section, we describe the current situation in SAR remote sensing from our perspective. We believe the situation to be similar in other fields of remote sensing. Our survey also does not show significant differences in the perception of reproducibility in SAR remote sensing compared to other fields. However, we decided to describe problems in SAR remote sensing, as we are most familiar with those, believing that similar issues may also exist more generally in remote sensing.

A. Culture of Antireproducibility

Reproducibility is not very highly regarded in science, at least not in the publication process. Studies focusing on reproducibility have only a small chance of getting published [15]. Besides negligence, in part caused by the publish or perish culture, there is a tendency to even avoid reproducibility, which we believe to be particularly strong in SAR remote sensing.

This is, at least in the past, related to the data sparsity in the early years of SAR remote sensing [16]. With SAR data being sparse and rather difficult to get access to, reproducibility was a challenge. The access to data was an advantage in the scientific competition. Only half-jokingly, the first author of this article once described being able to get SAR data as the core competency of his research group. A certain amount of a "guanxi"[1] culture was and still is prevalent in the global SAR remote sensing community.

Certainly, that is the opposite of a culture of reproducibility. With the now wider availability of SAR data, especially with open data policies from the European Commission and their Sentinel mission, as well as other data providers, getting data is not a core competency anymore, and reproducibility is becoming more common. Providing test datasets, often used in research on classification, etc., also improved the reproducibility in many fields of remote sensing. These are common, e.g., in data fusion [17], but not so much in SAR interferometry, although, for example, the DLR is providing data of geologically active sites as part of the supersites program.

The availability of SAR data from scientific and commercial missions not only improves reproducibility, but also lowers the entry hurdle for new research groups. However, this development also leads to a certain backlash with established groups moving toward the use of proprietary data, e.g., from airborne systems, to keep the competition at arm's length. Hence, we speak of a culture of antireproducibility.

B. Remote Sensing Data: Legal and Licensing Issues

Data are at the core of remote sensing science. However, sharing the data is often made impossible by the licensing terms of many remote sensing data providers. Additionally, sharing of geographic information and precise ground-truth data is often prohibited by law or corporate interests. These restrictions make reproducibility difficult.

Standard test sites and publicly available test datasets with ground-truth information are the norms in some remote sensing fields. They are uncommon in SAR interferometry, although some attempts toward establishing such datasets have been undertaken, for example, in the context of the Terrafirma project [18]. The DLR supersites can also be seen as an approach to establish relevant test sites that could be used for standardized processing.

Now, large amounts of data are publicly available. However, the impossibility of data sharing from the past is still a driver of present behavior. It is still uncommon to strive toward reproducibility or even demand it during peer review and publication processes.

C. Proprietary Software and Reproducibility

Proprietary software is problematic in terms of reproducibility. The high costs for some of the remote sensing software packages can deny reviewers and readers the option of reproducing the results. In-house software developments are even worse, as it is impossible to reproduce the results independently.

Now, with some products being commonly used, the version of the software needs to be specified to allow for reproducibility. In recent times, however, some standard tools and operating systems do not have a (clear) version number anymore. With the

[1]Chinese for relationship. For western readers: imagine the Godfather.

model of continuous and silent updates, as it is, for example, practiced in Windows 10 and Microsoft Office 365, referencing the precise version of the software used to generate the data or the figures can become problematic.

In all fairness, similar problems can arise from open-source systems as well. With the complicated set of dependencies and the continuous update cycles, the underlying shared libraries of an open-source software tool can also change. The usage of such libraries can be opaque to the end user and can lead to similar reproducibility problems, if, for example, the output of figures changes due to such updates.

D. Consequences of the Lack of Reproducibility

The consequences are rather long term and influence our discipline in general, which is the reason why it can be beneficial for an individual to ignore reproducibility, as it can slow down the publication in a short-term view. The long-term consequences are severe though and are felt already throughout science. The main consequence is a lack of trust in publications, especially along scientists themselves. The lack of trust in scientific results in the general public has other reasons as well, where lack of reproducibility is not the main concern. However, for insiders, the lack of reproducibility is an issue that leads to general mistrust in everything published. Nevertheless, there is also a responsibility of the reader to read critically and carefully judge the validity of research results.

The lack of trust is often based on bad experiences. When trying to reproduce work found in other publications, it is common to encounter difficulties. This is also shown in our survey results, where 75% of the respondents have had issues reproducing work found in other papers. These problems also typically occur in the early stages of research careers, as young scientists often start with reproducing the results of others. In this early stage, this can severely influence their outlook for the rest of their career.

Now, in our opinion, this has direct consequences on our science as well as on the commercial market build around SAR remote sensing. Here, the need for constant revalidation of well-established methods is one of the direct consequences. In interferometry, many customers do not believe in the previously demonstrated results and demand a new validation. This is especially prevalent in low-trust societies with generally highly educated decision makers, such as, e.g., China. These revalidation efforts are costly and slow down the commercial development of interferometric SAR.

In our science, this also slows down the development. We have no way to even estimate the amount of time lost in trying to reproduce nonreproducible papers, the time that could have been spent on the further advancement of our knowledge.

Another consequence of this is the coping mechanisms of the readers. Assuming that many publications are not reproducible and, therefore, of low significance for them, readers tend to search for trust in other ways. For example, they trust publications from certain research groups or entities that have been shown to be trustworthy in the past and ignore publications from entities or countries that are perceived to be untrustworthy.

As a long-term strategy, it can, therefore, be beneficial for individuals, research entities, and journals, to be seen as trustworthy providers of reproducible and well-written results. This may be in conflict with a short-term tactic in publishing many results fast; however, we do believe that the long-term benefits are worth it.

III. Survey on Reproducibility in Remote Sensing

To learn more about the perception of the state of reproducibility in remote sensing science, we conducted an online survey [19]. We based our questionnaire on the survey of *Nature* [5], so we reproduced the reproducibility study. However, compared to the definition we use in this article based on [3], the *Nature* study uses the term reproducibility and replicability differently. As these are mostly semantics that many of the respondents were also not completely aware of, we suggest the term reproducibility used in the survey to be understood synonymously for reproducibility and replicability in the interpretation of the survey results.

Our survey was conducted as an online survey. We kept the number of questions small to keep the time to answer questionnaire below 5 min. To this end, we removed long Likert scale questions [20] and focused on questions that could be answered fast. Thanks to this, we achieved a 99.6% completion quota, i.e., all but one respondent completed the survey.

Our survey focused on experienced scientists that are active in the scientific publication process. We contacted editors and associate editors of the IEEE Geoscience and Remote Sensing Society journals (i.e., IEEE Geoscience and Remote Sensing Letters, IEEE Transactions on Geoscience and Remote Sensing, and IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing), editors and associate editors of *Remote Sensing*, and the commission and working group officers of the International Society for Photogrammetry and Remote Sensing. We got 230 responses to our online survey. This gives us a rather representative insight into the state of reproducibility in remote sensing science. Baker and Penny's study [5] had 1576 respondents, but of those, only 95 are from the field of Earth and Environmental Sciences and only 66 from Engineering, the two topics we deem most close to remote sensing.

As shown in Fig. 2, 85% of the respondents are active reviewers, and about half of them serve as associate editors and/or session chairs in conferences, we can consider the responders to be strongly engaged in the scientific publication process in remote sensing.

We look at the question at hand: Is there such a thing like a crisis of reproducibility? In Baker and Penny's study [5], 90% answered that there is such a crisis (52% a significant crisis and 38% a slight crisis). Only 3% said that there is no crisis and 7% answered that they did not know. In our survey, 70% answered that there is such a crisis (27% a significant crisis and 43% a slight crisis). Similar to the *Nature* study, only 3.5% said that there is no crisis, but 27% answered that they did not know (see Fig. 1). In general, the perception of the problem of

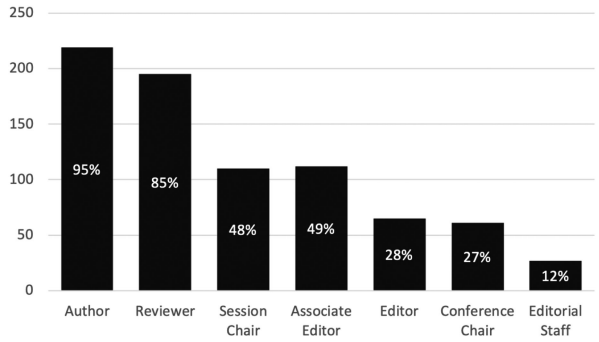

Fig. 2. Roles of respondents in the survey. Multiple answers are possible.

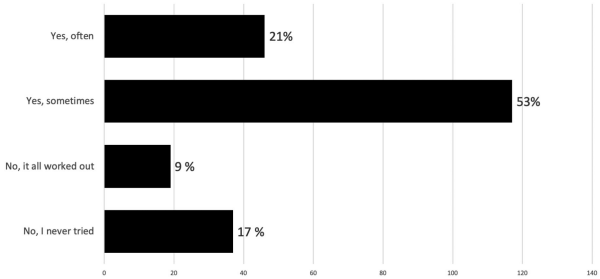

Fig. 3. Have you ever encountered difficulties in reproducing results from other researchers?

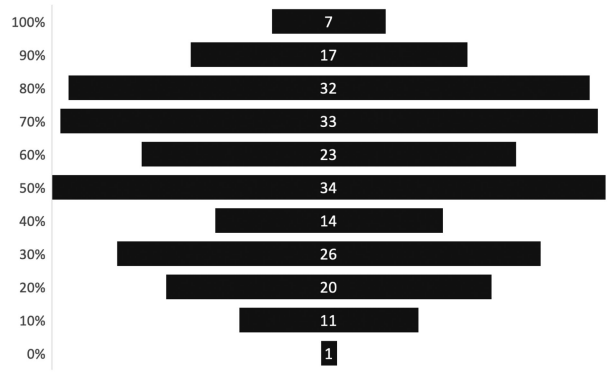

Fig. 4. What proportion of published work in remote sensing do you think is reproducible? In average, the respondents believe 55.6% of the papers to be reproducible.

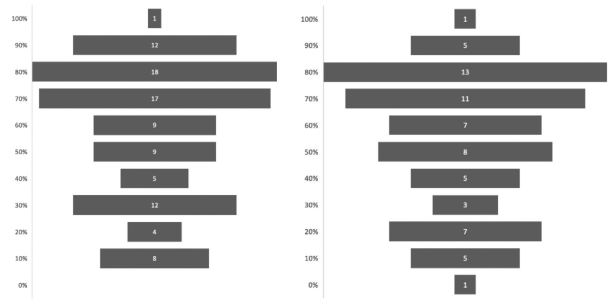

Fig. 5. Portion of reproducible work estimated by respondents of the *Nature* study [5]. Earth and Environmental Sciences (left) and Engineering (right). In average, the respondents believe 58.1% (left) and 55.5% (right) of the papers in their respective fields to be reproducible.

reproducibility in remote sensing seems to be less severe, but with 27% unsure, it may also indicate a lack of awareness.

To see the dimensions of the problem, we asked, "Have you ever encountered difficulties in reproducing results from other researchers?" As shown in Fig. 3, 75% of researchers encountered at least some problems when trying to reproduce results. This puts remote sensing in the top three of the research fields if compared to [5], below Chemistry and about the same as Biology. If we take only researchers who tried to reproduce results from others, an astonishing 90% encountered at least some problems when reproducing results. The source of the difficulties as well as the severity of these difficulties remains unclear.

In this regard, we also look at the question on the proportion of reproducible work. We asked, "What proportion of published work in remote sensing do you think is reproducible?" shown in Fig. 4. We compare that to Baker and Penny's result [5] shown in Fig. 5 with the results from Earth and Environmental Sciences left and Engineering on the right.

Among the respondents, 41% believe that 70% or more of the published works are reproducible in remote sensing science and 67% believe that 50% or more are reproducible. These results are similar to the *Nature* study, where in Earth and Environmental Science (69%) as well as in Engineering (68%) of the respondents believe that 50% or more of the published works are reproducible.

The belief in the amount of reproducible work is higher in Earth and Environmental sciences, where the majority (51%) believes that 70% or more of the published works are reproducible. In Engineering, this number is closer to remote sensing, where 45% believe this statement. However, this comparison is to be interpreted with care, as the number of respondents is low and not representative. Furthermore, the results are derived from different surveys in a different context. Nevertheless, these results seem to put remote sensing in the scientific fields with a comparably large perceived reproducibility issue.

These results are also slightly more negative than the estimation of the survey respondents. On the question to please complete the following sentence: "In my opinion, the level of reproducibility in my field is...," 17% suggested it to be "...better than for other scientific fields on average." Forty-four percent suggested it to be about the same and 12% believed remote sensing to be "... worse than for other scientific fields on average." Twenty-eight percent were unsure about this. These differences in the perception as well as the large number of unsure responses in several of our questions show, in our opinion, a certain lack of sensitivity to the problem of reproducibility in remote sensing. This is also shown in other answers. While on the statement of "I think that the failure to reproduce scientific studies is a major problem in my field," 44% agree and 24% disagree, on the statement "I think that the failure to reproduce scientific studies is a major problem for all fields," 54% agreed and only 10% disagreed. This clearly shows that for many of the respondents, reproducibility is mostly perceived as a problem

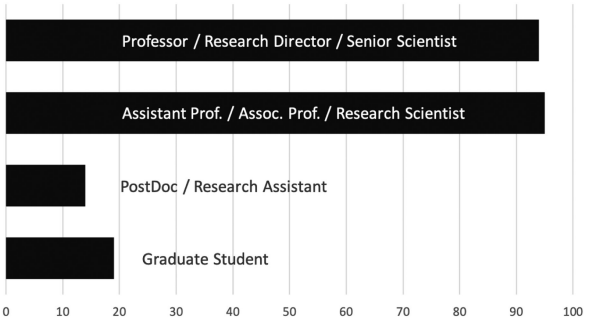

Fig. 6. Which of the following job titles best applies to you?

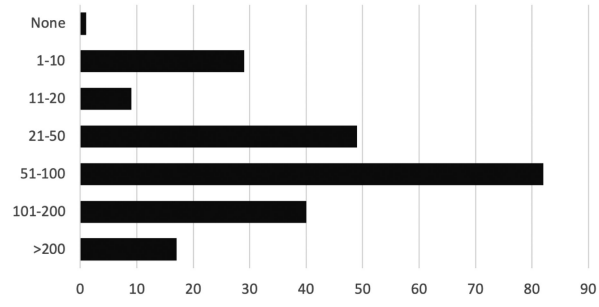

Fig. 7. How many scientific papers have you authored or coauthored?

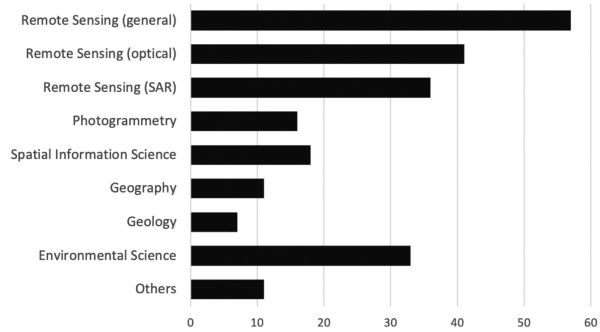

Fig. 8. How would you characterize your main research field?

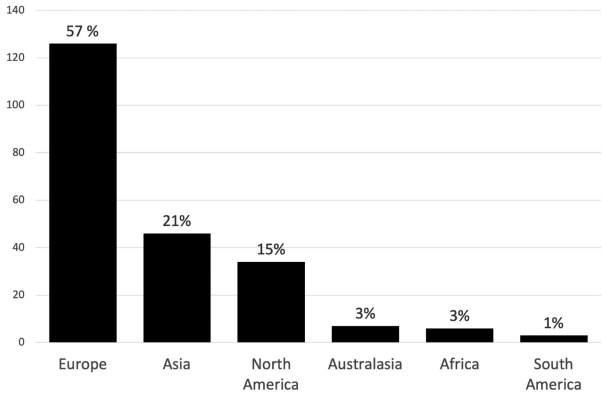

Fig. 9. In which continent do you (mainly) live?

of other sciences or science in general, not something that particularly affects remote sensing.

Compared to the rather low trust in the reproducibility of published articles in remote sensing and the 75% of researchers encountered at least some problems when trying to reproduce results, only 16% said that they have been contacted by colleagues that could not reproduce their results. This number is very similar to the 20% in the *Nature* study. This indicates a lack of communication in the community. Many reproduction attempts seem to fail, but this is also not openly discussed within the scientific community.

With our study having a strong focus on reviewers and editors, we also asked about reproducibility in the review process. Seventy-four percent of the respondents who reviewed a paper before said that they have raised questions or concerns with respect to the reproducibility of the results when reviewing a paper. Sixty percent of the authors answered that such concerns never have been raised in reviews, which might be related to the comparatively strong participation of experienced journal reviewers and editors in the survey. A large fraction of them may have encountered submissions of questionable reproducibility, even if only a minority of the papers are not reproducible.

That there is a need to improve reproducibility is shared by many in the survey. About 50% already implemented procedures ensuring reproducibility, and 56% believed that this should even be improved by themselves or their institution. However, only 28% encountered any efforts or directives from funding agencies to improve the reproducibility of their work, while 34% encountered such efforts from journal publishers. That is to say, there is still room for improvement.

Our survey had mostly rather experienced respondents, mainly because the call to participate in the survey was mostly distributed to editors, associate editors, and workshop and conference organizers. This led to an overrepresentation of more experienced researchers, as shown in Figs. 6 and 7. The vast majority of the respondents are on an Assistant Professor/Associate Professor/Full Professor level in their career, with an astonishing 42% being Professor, Research Director, or Senior Scientist. Similarly, 61% of the respondents had already published more than 50 scientific papers.

Fifty-eight percent of the respondents would characterize their field as directly related to remote sensing, as shown in Fig. 8. Of those, about 27% would characterize their research to be SAR related, a portion that is very similar to our bibliometric analysis in [16].

The respondents are not very geographically representative though. Fifty-seven percent of the participants were based in Europe and only 21% from Asia (see Fig. 9). This is certainly not representative of remote sensing domain. These numbers are a bit surprising, also because the survey was also advertised on Chinese social media channels, with very limited response rate.

Analyzing the results, we did not find many significant differences in the perception of reproducibility from Asian researchers. If anything, we found that the problem is perceived to be more significant there. Thirty-seven percent of Asian researchers said that there is a significant problem in reproducibility, 48% believe that there is a slight problem of reproducibility, 2% do not believe that there is a reproducibility crisis, and

13% do not know. On the question "Have you ever encountered difficulties in reproducing results from other researchers?" 85% of researchers based in Asia already encountered difficulties, while 15% never tried. That is to say, 100% of the respondents based in Asia trying to reproduce results from other researchers had difficulties doing so.

IV. Improving Reproducibility and Replicability in Remote Sensing

Improving reproducibility and replicability is important in science. However, not every study may be reproducible and problems in replicating studies can actually lead to a greater understanding of the methods and the underlying limitations. Advanced and new, leading edge content can be especially difficult to reproduce or replicate. However, in the majority of the cases, problems in reproducibility and replicability can be avoided or at least reduced when appropriate scientific workflows are followed.

A. Possible Reasons for Reproducibility and Replicability Issues

We, the authors, believe that reproducibility and replicability are in a crisis mode, especially in remote sensing. We explained our reasons for this in Section I. After looking at the perception of the state of reproducibility in remote sensing through an online survey, as discussed in Section II, we want to show ways to improve reproducibility and replicability in remote sensing.

The first step is the analysis of reasons for a lack of reproducibility. According to [3], five main reasons can be identified as:

1) inadequate record keeping;
2) nontransparent reporting;
3) obsolescence of digital artifact;
4) flawed attempts of reproduction;
5) barriers in the culture of research.

As we discuss in Section I, we believe that the barriers in the culture of SAR remote sensing are especially high. As we will discuss below, the first issue of record keeping can be relatively easily fixed by each researcher and research group themselves. Similarly, the nontransparent reporting is in the responsibility of each author, but stricter enforcement of this in the review process can also be helpful. However, the obsolescence of digital artifacts is a problem that deserves a more detailed discussion.

In terms of replicability, some of the issues are more systematic and require more systematic solutions from the scientific community. According to [3], six main sources for replicability problems can be identified as:

1) publication bias;
2) misaligned incentives;
3) inappropriate use of statistics;
4) poor study design;
5) errors;
6) incomplete reporting of studies.

We will discuss these issues and possible solutions below, focusing on these issues that can be addressed by the individual researcher and research groups as well as by our remote sensing community, whereas, for example, the misaligned incentives are to be addressed by funding agencies, universities, or governments and are, so to say, above our pay grade.

B. Improving Reproducibility

The minimum requirement for reproducibility is the ability to reproduce the published results by the authors themselves. We, the authors, currently do assume that all authors strive to be able to reproduce their own results, but even this minimum requirement is often not reached.

One of the main issues for this is the lack of appropriate record keeping. This includes clear, detailed, and complete information about the methods and steps taken to reach the results. This includes not only the methods reported, but also preprocessing steps and visualization details. As detailed record keeping can be quite time consuming, this is often ignored. Automated log tools and scientific workflows can help [21]. Modern forms of laboratory notebooks can directly include code, equations, visualizations, and text allowing for a direct link among the research work, the visualization, and the publication. One example for this would be the Jupyter Notebook or JupyterLab and the well-developed infrastructure around it [22].

Version control systems can help to keep track of source code changes and can be essential in reproducing the version of the software used in a publication. With the often ongoing changes in the software used, it can otherwise be very difficult to make sure that the correct version of the software is used in reproducing previous results. Besides the source code, scripts and other workflow details can also be kept in a version control system.

This is also important for the problem of obsolescence of digital artifacts, such as, e.g., source code, but also remote sensing data used in the study. In this respect, it is in actuality not as trivial to be able to reproduce the results even by the authors. This requires access to the materials, code, software, etc., used to generate the results. One question arises though: How long should they aim to make such things available so that the results are reproducible? For a published journal article, we would suggest a minimum period of five years. The difficulties lie, as so often, in the details. First, this requires archiving a snapshot of the software version used for processing. Each update may possibly destroy the reproducibility. To make things more complicated, it may be necessary to keep a snapshot of the complete operating system being used as well, as updates in the underlying frameworks and libraries can also disturb the reproducibility. This can, for some of the commercial operating systems, be a significant challenge. However, also additional tools for data pre- or postprocessing can change, which may change the output of figures and charts. With some tools having only internal version numbers and no way to reinstall outdated versions, this can also be a significant challenge in practice, especially considering long archive periods, covering several years.

Containerization can help mitigate these problems. A software container packages up code and all its dependencies, allowing an application to reliably run in different computer environments. The most well-known approach to software

containers is Docker [23]. This not only allows running applications on different environments, but also supports archiving strategies that can ensure reproducibility and replicability [24].

With respect to research work dealing with industry partners, patented code, or the authors aiming at commercialization of their work, the publication of source code may not be possible. As an alternative, an online system where a user can submit data to be processed by the algorithm on a dedicated server is proposed by [25].

The data also need to be archived. Depending on the size of the project and the data used, this can be a challenge. It is strongly recommended to archive the original data, which, under some circumstances, can be a problem with respect to licenses from commercial data providers may allow only a limited time for data usage. Furthermore, the original data should be archived, as a later reprocessing of the data, with a new version of the processing software of the data provider, can have differences, e.g., with respect to the orbit position or improvements in the calibration, etc. We also recommend archiving of intermediate results, allowing reproduction of the most important processing steps on the way to the final results.

This can require a significant amount of storage. Additionally, it requires a long-time archive strategy. Five years is over the expected lifetime of a standard hard disk. An archive strategy that goes beyond saving the data on an external disk and locking it away is required, although, in most cases, saving all the data on an external disk would already be a huge progress with respect to the mostly prevalent strategy of doing nothing about reproducibility.

However, this short list of requirements also demonstrates the difficulty in having the author alone taking care of it. A secure archiving strategy is rather an endeavor to be undertaken by the underlying research organization. This leads us to the next stage of ensuring reproducibility, which should be the responsibility of the research organization, such as the institute, university, or other relevant research bodies supporting the work.

In our opinion, research organizations also have to take responsibility to ensure reproducibility. There are two main reasons for this. First, research organizations have the means to organize a long-time archiving strategy for software, data, and if necessary hardware, allowing reproducibility of published results for a predefined period. Second, research organizations themselves benefit most from a well-established culture of reproducibility.

Having the processing framework of published results available in a way that allows reproducibility by the organization, instead of only by the author, allows for a better continuity of the research progress within the organization. In addition, increasing trust by reproducible results is a long-term strategy. Organizations tend to benefit more from long-term approaches, whereas individual authors may benefit less.

C. Improving Replicability

Replicability is a goal for our science. While reproducibility is a sign of good scientific practice, it is not necessarily a prerequisite for replicability. Studies may be replicable with new data, but not reproducible, for example, due to changes in software, loss of notices/preprocessing steps/data through the mentioned obfuscation of digital artifacts. In contrast, studies can be reproducible with the same data, but cannot be replicated in another experiment, test site, etc.

Failures in replicability can be of great scientific significance, as they may show limitations of previously published approaches. It is very useful for replicability, if the authors of a study discuss limitations and uncertainties openly and directly.

In this context, we also have to consider the so-called publication bias. Journals prefer positive and statistically significant results. This can lead to a reluctance of the authors to explore and discuss limitations and uncertainties, as this may disturb the "perfect" picture authors may like to paint for the reviewers. Writing for reviewers instead of for readers is part of this problem. This is related to a tendency of seeing getting published as the goal of the scientific endeavor, so that the review has to be passed or overcome. We see this in contrast to an approach, where the publication is a means of communicating science with the communication and the scientific discussion as a value in itself. In such an approach, the review is not seen as something blocking the way to success, but as a help in communicating better. In such an approach, where the scientific communication is the goal, a more self-critical discussion on the results would be the norm.

Several journals and publishers are already working on improving the reproducibility. For example, *Science* requires researchers to make data and code available on request [26].

Other journals offer badges for good practices, e.g., for preregistered studies, open data, and open materials. Such badges can help increase the rate of data sharing [27]. The IEEE Xplore Digital Library also assigns reproducibility badges for some articles. With the introduction of the IEEE DataPort, data sharing and linking to published article is also becoming possible.

Other forms of publication and scientific discussion can also encourage reproducibility. *Nature* supports comments on articles on their website, which can encourage scientific discussion and also may nudge authors toward more reproducible and replicable research. Similar preprint archives, such as, e.g., medRxiv, allow comments to articles, as does *Remote Sensing*.

As will be discussed in Section IV, new forms of publication can offer space for more details and an extended discussion on limitations. This could dramatically change scientific publications, as, e.g., the computational workflow as well as data, code, etc., forming the basis of a paper can be published together [28]. In contrast, with an increase in overall publications, getting an overview of the previous research is also getting more time consuming and difficult.

A more frequent publication of negative results would also be useful in exploring the limitations of known approaches and as starting points on scientific discussion on the replicability of other results. However, with the publication bias toward novel and positive results, negative results are difficult to publish. Specialized journals on negative results can help [29], but it would be expected that publications in such journals would be not very prestigious, and given the misaligned incentives of today's scientific practices, in terms of career advancement and bonus systems, would discourage such publications as well.

However, besides the lack of incentives to publish negative results, it is also important to state that writing a good paper with negative results is much more difficult. Writing about a negative result, it is necessary to prove the correctness of the implementation and correct data handling. Furthermore, reasons for the negative result need to be given. In contrast, a positive result stands on its own feet and is much easier to write convincingly, even if not every step is done correctly along the way.

Although these problems can hamper replicability, they are not per se signs that the original findings of a study are wrong. However, these issues for sure also exist. Many common issues are related to the inappropriate use of statistics. These schemes include "p-hacking" and "cherry picking" as well as "HARKing." As in remote sensing the p-value is less used, "cherry picking" is more common. Cherry picking describes a practice of selective reporting of results that meet a certain criterion. Certainly, it is quite common to select the data or the subset of a data that best fits the results. This may be deliberately or unconscious.

HARKing describes a confirmatory research approach that retrospectively develops hypotheses to fit the data and then validates the hypothesis with that data. The extent of HARKing in remote sensing and other disciplines is unknown, but it is estimated to be a rather common practice. Clearly, such statistical errors and poor study designs can also be traced back to the publication bias and publication pressure, as journals require clear positive results, where some cherry picking can help to make the results even more convincing.

The respondents from Baker and Penny [5] stated that "more robust experimental design," "better statistics" and "better mentorship" would help. The training and education of young scientists with respect to the issues of reproducibility and replicability should be improved, but also the mentorship. Experienced researchers should be aware of their function as role models. These would be comparably slow but very sustainable measures in improving the replicability and quality of our science.

Although we do not like to admit it, intent and misconduct also play a role. It is not in the scope of our article to address issues of criminal behavior and deliberate scientific misconduct, but we can encourage the scientific community to reduce misaligned incentives wherever possible.

V. Discussion With Respect to Modern Forms of Publication

The inception of computer-based, open-access, journals and the substantial end of the page limit created an entirely new type of scientific articles, impossible before. The big change from papyrus to the cloud had a large impact on the way science is discussed and diffused.

In the days of paper journals, space was indeed limited, and the justification for its use was mainly the presence of innovation that only could motivate the expensive archiving. Nowadays, archives are free, and their access in the form of Journals has to be construed in a new and different way. Articles did change. Their quality is not in their existence as a publication, now close to being valueless, but in the number of followers that read and refer that article. This is the teaching of Facebook.

Which were the most significant changes induced in the article manufacturing by the infinite archives? The introductory part of any new article, in the past limited as much as possible and often just a pure list of references, now can be transformed in a tutorial part, a long and often well-written synopsis of the available literature. At times, this discussion is already a motivation to the readers to learn how the authors cataloge and coordinate the existing literature on any given topic. This could be already quite a help.

Another big change is in the examples; rather than being a cramped and short chapter at the end of the article, the examples can now be expounded with leisure, allowing the reader to check, in really difficult cases, the efficacy of the proposed recipe.

The central part, the presentation and discussion of the innovation, stays the same. However, here there is a new danger: maybe the innovation is just a minuscule detail, as the unlimited archive removes the motivation for any and all lower limits[2] [30]. Reproducibility and replicability are helped by the fact that recipes can now be described accurately, as there is no motivation to be short. However, we should remember the final goal: to be referred, not to be published. Now, a well-described recipe can be easily replicated. Then, it will be easy for the competitors to add an irrelevant additional new detail to claim innovation. However, as this is just an infinitesimal contribution to an infinite archive, the danger of uselessness is not so relevant for the general readers.

The real danger comes to the generous authors who in detail explained their work, if the minuscule innovators stop referring to the original article but instead refer to their own, deviating from the followers. The originator is ruined, and the not so honest competitors strive.

Does limiting or preventing replicability protect from that? In the case of having only limited replicability, the newcomers are pushed, if not obliged, to refer to the original article, first because they are not sure that that they replicated it perfectly, and second because they need it to compare with their own work, as they are not sure that that article was superseded or not.

Then, are replicability and article survival in the references at odds? How to avoid that? This is the main task of the reviewer, who should indicate, which real data examples should be used to demonstrate the validity of any new concept. The dangerous synthetics, acceptable just for explanation purposes, should always be accompanied by significant real data cases, or publicly shared whenever relevant. Then, there should always be a clear

[2]From those incontrovertible premises, the librarian deduced that the Library is "total"-perfect, complete, and whole-and that its bookshelves contain all possible combinations of the twenty-two orthographic symbols (a number which, though unimaginably vast, is not infinite)-that is, all that is able to be expressed, in every language. All: the detailed history of the future, the autobiographies of the archangels, the faithful catalog of the Library, thousands and thousands of false catalogs, the proof of the falsity of those false catalogs, a proof of the falsity of the true catalog, the gnostic gospel of Basilides, the commentary upon that gospel, the commentary on the commentary on that gospel, the true story of your death, the translation of every book into every language, the interpolations of every book into all books, the treatise Bede could have written (but did not) on the mythology of the Saxon people, the lost books of Tacitus. (Translated from Spanish. Online. [Available]: http://biblio3.url.edu.gt/Libros/borges/babel.pdf)

comparison with the prior art, to show the advantages that should be significant to justify publication. Thus, referring to the prior art becomes necessary, the innovation gets stimulated, and the innovators rewarded.

Otherwise, a new recipe, another acronym, and we could have a new paper, ready to deviate references, forgetting the past. It is the reviewer's task, however, to guarantee that significant real data examples be always present, proposing also more difficult situations, where the new algorithm could fail or behave as the many others already available, either outright killing the paper or giving to it the proper perspective.

These points might push some authors toward being not completely open on the details of the manufacturing recipe. In contrast, still to learn that something is doable, that there is a light that could be reached on the opposite side of the valley will push newcomers to try to cross the valley avoiding to drowning in the river and looking for the not so evident bridge.

In other words, innovation has to have the foremost place, because it is the tool for the advancement of science. At times, very nice examples, if necessary on sites proposed by the reviewers, could be enough to justify publication, even if not every detail of the research is plainly spelled out. However, again, the role of the reviewers is paramount in addressing this point and enhancing reproducibility as much as possible, while safeguarding priority. They should insist that the original authors be present in the coveted references, avoiding the exaggerated self-referencing that too often pushes authors to be closed and not open and reproducible.

There is another additional point that should be made: the impact of "Publish or Perish." This is most true for Academia and much less for researchers belonging to groups, often in well-established research groups for which reputation of solidity is more significant than a new paper being published. Unfortunately, the time axis is extremely relevant for academics, who need a paper in time for a selection and maybe they are ready to make a mistake. Here, not only the reviewer's savvy but the readers too should remember "caveat emptor" (buyers beware). However, if the readers should beware, maybe the reviewers should be laxer, to avoid boredom and to push for innovation. A last useful check is if there is a patent pending: no one throws money away.

Finally, maybe the unlimited archives are pushing toward a life without archives at all, where we live just in the present. Then, rather than striving to be remembered, authors should go ahead and produce innovation, proving their resilience with continuous creation.

VI. Conclusion

Reproducibility and replicability are in a crisis. It remains open if this crisis is similar to other scientific fields, as the majority of the respondents of our study believe, or even bigger.

With 75% of the respondents of our survey encountering at least some problems when reproducing or replicating studies, it is clear that there is an issue. The extent and the severity of the issue cannot be directly inferred from this finding. Nevertheless, the first step in solving the problem is openly addressing that there might be an issue, which is what we intend to do here. We believe that there is an issue with reproducibility and replicability in remote sensing science, and we need to improve ourselves.

There are simple measures that can significantly improve reproducibility, while also effectively improving reliability and productivity in computational science:

1) record keeping, ideally supported by automatic logs;
2) use of version control systems (e.g., GIT);
3) developing an archive strategy;
4) containerization;
5) proper study design.

With the inception of computer-based journals and the end of page limitations as well as the availability of extended online appendices, more details on the experiments can be given and reproducibility can be enhanced. Reviewers play an important role in ensuring the research to be valid and acknowledging the original contributors and innovators. However, readers should also be responsible. In times of prevalent discussions on "fake news," the ability and willingness to read critically and carefully should be emphasized.

Paying attention to reproducibility and replicability in writing, reviewing, teaching, mentoring, and everyday work is a responsibility that, if taken seriously, will have a profound impact. We encourage everybody working in remote sensing to increase awareness of reproducibility issues and working on improving workflows to achieve better reproducibility and replicability throughout our field.

Acknowledgment

The authors would like to thank Prof. Alejandro Frery for the fruitful discussions on this study and his endeavors in promoting a more reproducible science. They would also like to thank the anonymous reviewers for their support in improving this article.

References

[1] J. Crocker and M. L. Cooper, "Addressing scientific fraud," *Science*, vol. 334, no. 6060, pp. 1182–1182, 2011.
[2] L. A. Barba, "Terminologies for reproducible research," 2018, *arXiv:1802.03311v1*.
[3] E. National Academies of Sciences and Medicine, *Reproducibility and Replicability in Science*. Washington, DC, USA: National Academies Press, 2019.
[4] M. R. Macleod *et al.*, "Biomedical research: Increasing value, reducing waste," *The Lancet*, vol. 383, no. 9912, pp. 101–104, Jan. 2014.
[5] M. Baker and D. Penny, "Is there a reproducibility crisis?" *Nature*, vol. 533, no. 7604, pp. 452–454, 2016.
[6] M. R. Munafò *et al.*, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, Jan. 2017, Art. no. 0021.
[7] B. R. Jasny, G. Chin, L. Chong, and S. Vignieri, "Again, and again, and again …," *Science*, vol. 334, no. 6060, pp. 1225–1225, 2011.
[8] A. Nekrutenko and J. Taylor, "Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility," *Nature Rev. Genetics*, vol. 13, no. 9, pp. 667–672, 2012.
[9] K. Burns, J. Shepherd, and M. Berman, "Reproducibility of geological lineaments and other discrete features interpreted from imagery: Measurement by a coefficient of association," *Remote Sens. Environ.*, vol. 5, pp. 267–301, 1976.

[10] M. Reis, M. Escada, L. Dutra, S. Sant'Anna, and N. Vogt, "Towards a reproducible LULC hierarchical class legend for use in the southwest of Pará State, Brazil: A comparison with remote sensing data-driven hierarchies," *Land*, vol. 7, no. 2, May 2018, Art. no. 65.
[11] M. J. Duggin and C. J. Robinove, "Assumptions implicit in remote sensing data acquisition and analysis," *Int. J. Remote Sens.*, vol. 11, no. 10, pp. 1669–1694, 1990.
[12] K. Anderson and E. J. Milton, "On the temporal stability of ground calibration targets: Implications for the reproducibility of remote sensing methodologies," *Int. J. Remote Sens.*, vol. 27, no. 16, pp. 3365–3374, 2006.
[13] B. Gwein, T. Miksa, A. Rauber, and W. Wagner, "Data identification and process monitoring for reproducible earth observation research," in *Proc. 15th Int. Conf. eSci.*, 2019, pp. 28–38.
[14] F. O. Ostermann and C. Granell, "Advancing science with VGI: Reproducibility and replicability of recent studies using VGI," *Trans. GIS*, vol. 21, no. 2, pp. 224–237, 2017.
[15] G. N. Martin and R. M. Clarke, "Are psychology journals anti-replication? A snapshot of editorial practices," *Frontiers Psychol.*, vol. 8, 2017, Art. no. 523.
[16] M. Liao, T. Balz, F. Rocca, and D. Li, "Paradigm changes in surface-motion estimation from SAR," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 1, pp. 8–21, Mar. 2020.
[17] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," vol. IV-2/W7, 2019, pp. 153–160.
[18] *Terrafirma. Qlty. Ctrl. Protoc. for Lvl. 1 Produc.*, Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany, 2006.
[19] T. Balz, "Survey on reproducibility and replicability in remote sensing," 2020. [Online]. Available: http://dx.doi.org/10.21227/hkgh-9y21
[20] R. Likert, "A technique for the measurement of attitudes," *Arch. Psychol.*, vol. 140, pp. 1–55, 1932.
[21] L. Wang, Z. Lu, P. van Buren, and D. Ware, "Terminologies for reproducible research," *BioInformatics*, vol. 34, no. 22, pp. 3917–3920, 2018.
[22] Jupyter, Project jupyter,2020. [Online]. Available: https://jupyter.org
[23] Docker, Empowering app development for developers, 2020. [Online]. Available: https://www.docker.com
[24] C. Knoth and D. Nust, "Enabling reproducible OBIA with open-source software in docker containers," Sep. 2016. [Online]. Available: http://proceedings.utwente.nl/456/
[25] P. Vandewalle, G. Barrenetxea, I. Jovanovic, A. Ridolfi, and M. Vetterli, "Experiences with reproducible research in various facets of signal processing research," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. IV-1253–IV-1256.
[26] AAAS, Science journals: Editorial policies, 2020. [Online]. Available: https://www.sciencemag.org/authors/science-journals-editorial-policies
[27] M. C. Kidwell *et al.*, "Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency," *PLOS Biol.*, vol. 14, 2016, Art. no. e1002456.
[28] Y. Gil *et al.*, "Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance," *Earth Space Sci.*, vol. 3, no. 10, pp. 388–415, 2016.
[29] T. Balz and E. Christophe, "Failing as chance—Negative results in geoscience and remote sensing," *IEEE Geosci. Remote Sens. Soc. Newslett.*, vol. 2011, pp. 15–16, Jun. 2011.
[30] J. L. Borges, *La Biblioteca de Babel*, 1941. [Online]. Available: http://biblio3.url.edu.gt/Libros/borges/babel.pdf

Timo Balz (Senior Member, IEEE) received the Diploma degree in geography and the Ph.D. degree in aerospace engineering and geodesy from the University of Stuttgart, Stuttgart, Germany, in 2000 and 2007, respectively.

He has been a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, since 2015. His research interests include synthetic aperture radar (SAR) remote sensing, SAR interferometry, and the use of SAR data for social dynamics, including the usage of SAR to support archaeological prospections.

Dr. Balz is a member of the editorial boards of *Remote Sensing*, *Geospatial-Information Science*, and the *Journal of Geodesy*. He is the Chair of the International Society for Photogrammetry and Remote Sensing Working Group I/3 on SAR and Microwave Sensing.

Fabio Rocca received the Graduate degree in electronic engineering from the Politecnico di Milano, in 1962, and the Doctorate degree (*Honoris Causa*) in geophysics from the Institut Polytechnique de Lorraine, Nancy, France, in 2001.

He is a Professor Emeritus of Telecommunications with the Politecnico di Milano, Milan, Italy. Since 2004, in cooperation with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, he has participated with the Chinese European space research programs Dragon (1–4), sponsored by the European Space Agency and the National Remote Sensing Center of China.

Dr. Rocca received the Award from Honeywell, in 1979, an Award from the Italgas for Telecommunications, in 1995, an Award from the Rhein Foundation for Technologies for Motion-Compensated Television Coders, in 1999, the Desiderius Erasmus Award from the European Association of Geoscientists and Engineers, in 2009, an Eni Award, in 2012, a Chinese Government International Science and Technology Cooperation Award, in 2013, and a Commemorative Medal for the 70th Anniversary of the People's Republic of China, in 2019.