



Article

How to Improve the Reproducibility, Replicability, and Extensibility of Remote Sensing Research

Peter Kedron^{1,2,*} and Amy E. Frazier^{1,2,3}

- ¹ School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85283, USA
² Spatial Analysis Research Center (SPARC), Arizona State University, Tempe, AZ 85283, USA
³ Center for Global Discovery and Conservation Science, Arizona State University, Tempe, AZ 85283, USA
* Correspondence: peter.kedron@asu.edu

Abstract: The field of remote sensing has undergone a remarkable shift where vast amounts of imagery are now readily available to researchers. New technologies, such as uncrewed aircraft systems, make it possible for anyone with a moderate budget to gather their own remotely sensed data, and methodological innovations have added flexibility for processing and analyzing data. These changes create both the opportunity and need to reproduce, replicate, and compare remote sensing methods and results across spatial contexts, measurement systems, and computational infrastructures. Reproducing and replicating research is key to understanding the credibility of studies and extending recent advances into new discoveries. However, reproducibility and replicability (R&R) remain issues in remote sensing because many studies cannot be independently recreated and validated. Enhancing the R&R of remote sensing research will require significant time and effort by the research community. However, making remote sensing research reproducible and replicable does not need to be a burden. In this paper, we discuss R&R in the context of remote sensing and link the recent changes in the field to key barriers hindering R&R while discussing how researchers can overcome those barriers. We argue for the development of two research streams in the field: (1) the coordinated execution of organized sequences of forward-looking replications, and (2) the introduction of benchmark datasets that can be used to test the replicability of results and methods.

Keywords: reproducibility; replicability; repeatability; benchmark data; provenance



Citation: Kedron, P.; Frazier, A.E. How to Improve the Reproducibility, Replicability, and Extensibility of Remote Sensing Research. *Remote Sens.* **2022**, *14*, 5471. <https://doi.org/10.3390/rs14215471>

Academic Editor: Jason A. Tullis

Received: 10 October 2022

Accepted: 28 October 2022

Published: 31 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Over the last decade, the field of remote sensing has undergone a remarkable shift, with a considerable amount of data now easily accessible at higher spatial, temporal, spectral, and radiometric resolutions than ever before. Researchers can browse and immediately download data from vast catalogs like USGS EarthExplorer, Google Earth Engine (GEE), or the ESA Copernicus Open Access Hub, or request data from private sector firms like Planet that are imaging the earth every day. Cloud-based computing platforms like GEE have made the process even easier by allowing researchers to bundle image acquisition, processing, and analysis tasks altogether and rapidly conduct large-scale computations in the cloud [1]. Complementing the availability of satellite-based remote sensing data has been the advent of low cost, easy to use, close-range remote sensing technologies including uncrewed aircraft systems (UAS, or drones). It is now possible for anyone with a moderate budget to gather their own remotely sensed data by mounting sensors on a UAS, and that data can then be shared through open platforms online.

Researchers have gained similar flexibility in their data processing and analysis options. The continuous collection of multi-sensor, high resolution data that is too large to efficiently process using traditional techniques has motivated an explosion of artificial intelligence (AI) research within the remote sensing community [2,3]. New methods are released almost daily, and investments in large-scale infrastructure initiatives and high-performance

computing aim to make these methods readily accessible. As importantly, tools such as Binder and WholeTale that facilitate the documentation and sharing of the data, code, and software environment used to generate research findings with these emerging methods are also free and widely available [4–6].

The increasing data availability, methodological expansion, and tool proliferation in remote sensing creates both the need and opportunity to reproduce, replicate, and compare new research methods and findings across spatial contexts, measurement systems, and computational infrastructures. At the core of the scientific method is the idea that research methods and findings are progressively improved by repeating and expanding the work of others. This process is called reproduction and replication (R&R). By reproducing or replicating studies, and examining the evidence produced, scientists are able to identify and correct prior errors [7,8], verify prior findings, and improve scientific understanding. R&R is likewise closely linked with Popper's falsifiability criterion [9] because the falsification of scientific theories depends in part on the verification of auxiliary assumptions that link the hypothesis in question to something observable. When falsification attempts are made, reproductions and replications can help researchers identify, make explicit, and test the often unacknowledged auxiliary assumptions. As such, reproductions and replications can help researchers differentiate whether a research result is the product of some violation of the hypothesis or some auxiliary assumption. That knowledge can then be used to refine the specification of hypothesis and further develop theory [10]. Moreover, falsification is only informative if researchers discover a reproducible effect that refutes a theory [9,10].

There are innumerable opportunities for R&R in remote sensing due to the open availability of data and the consistency of repeat coverage from sensors like Landsat and Sentinel. As a complement, local data collected from UAS and other aerial platforms offer the opportunity to examine outcomes more closely and test new methods across measurement systems. As these new methods are developed, reproductions can be used to test their validity and credibility, and replications can be used to assess their robustness and reliability. These opportunities have not yet been fully explored though because basic reproducibility remains an issue in remote sensing research [11,12]. In a recent survey of 230 synthetic aperture radar (SAR) researchers, 74 percent responded that they experienced some difficulty when attempting to reproduce work in their field, while an additional 19 percent responded that they have never attempted a reproduction [10]. In a review of 200 UAS studies, Howe and Tullis [12] found that a mere 18.5 percent of authors shared their data, and only 8 percent shared their workflow or code. At the same time, the privatization of earth observation by companies like Planet and Maxar has created new challenges for reproducibility such as the need to rectify sensor discrepancies across rapidly progressing sensor generations and standardize radiometric and geometric corrections [13]. The growing use of UAS adds to the challenge of collecting and sharing data in dynamic, multi-user environments without recognized documentation or reporting standards [14,15].

As in other fields, remote sensing researchers also face a set of tradeoffs surrounding reproducibility when planning and implementing their studies. Ensuring that a remote sensing study can be accessed and recreated by others requires that researchers spend time and resources that might otherwise be dedicated to original research [16]. Recording workflow details, adequately commenting code, containerizing, preserving, and sharing research artifacts, and running internal reproducibility checks all take time, but are often met with little direct reward. While it is true that increasing the accessibility of work may increase its adoption by others, beyond adopting reproducibility badging systems, publishers have done relatively little to incentivize reproducibility. It also remains difficult across disciplines to publish reproductions and replications, even when this type of study may be particularly salient to resolving competing conceptual or policy claims [17–19]. Similarly, we are unaware of any criteria that reward the adoption of practices that facilitate R&R in the promotion and tenure process. Beyond these practical considerations, Wainwright [20] poses a range of ethical questions a researcher might consider when planning the reproducibility of their work. These questions include the ethics of sharing data that by

themselves may not violate ethical principles to do no harm, but could when combined with other data streams allow others to infer sensitive information resulting in a violation of privacy principles and cause harm. Wainwright similarly questions the for-profit motives of publishing companies seeking to expand their business models into data and metadata analytics. Bennett et al. [21] offer a still wider view of the challenge by situating remote sensing in the broader political economy. Adopting their framing, researchers could well ask whether making their research reproducible also reproduces existing power dynamics that disadvantage some communities relative to others.

These challenges hinder scientific progress in remote sensing research, but they can be overcome. This paper discusses R&R in the context of remote sensing and links the recent changes in the field to key barriers preventing R&R. We argue that foregrounding R&R during the design of individual studies and remote sensing research programs can position researchers to take advantage of recent developments and hasten progress. We organize this paper into three parts. First, we define the terms and connect R&R to some of the central objectives of remote sensing researchers. We then discuss the challenges that current developments in remote sensing present for R&R. We link those challenges to steps the remote sensing community can take to improve the reproducibility of research. Lastly, we look forward and discuss how planning for R&R can be incorporated into the research design and evaluation process. We aim to show how the challenges that presently exist in remote sensing research can be turned into opportunities, and we present actionable steps to facilitate that conversion.

2. Reproducible, Replicable, and Extensible Research

Different fields have adopted different definitions and uses of the terms ‘reproducible’ and ‘replicable’. We adopt the U.S. National Academy of Sciences, Engineering, and Medicine (NAEM) definitions of R&R [22] (Figure 1). A remote sensing study is reproducible if an independent research team can obtain results that are consistent with an original study using the same input data, computational steps, methods, code, and conditions of analysis as the original research team. Within the AI literature, Gunderson and Kjensmo [23] make a distinction between the conceptual approach to a problem (AI Method) and the implementation of that approach in the form of a computer program (AI Program), which is also a useful distinction in remote sensing. Separating methods from programs, the authors introduce what they call data reproducibility, which allows for an alternative implementation of the method used in the prior study. A test of data reproducibility isolates the effect of the implementation of a method as a program. This distinction is useful in remote sensing because researchers often wish to know how well a method, such as a classification algorithm, is performing independent of its computational implementation and, conversely, how consistent and relatively efficient different implementations of the programming are. Therefore, we discuss both methodological and programmatic reproducibility.

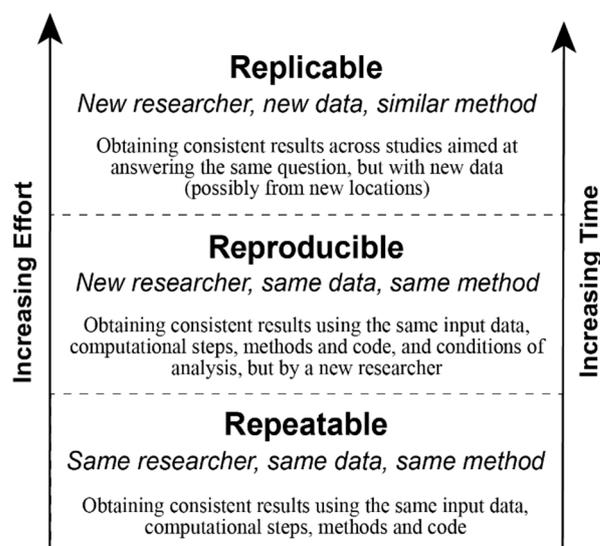


Figure 1. Typology of R&R concepts showing progression of effort and time. Adapted from Essawy et al. [24].

Using the NASEM definition, a remote sensing study is replicable if an independent research team addressing the same question can obtain consistent results when using different data. Replications provide information about the robustness and reliability of the claims of a study [25,26] because they provide evidence that the findings are not just characteristic of the original study. If the replication used all of the same methods and only changed the data, then the replication can support the inference that the remote sensing method, and the software(s) in which the analysis was implemented, are reliable. For example, a study replicating a new classification algorithm that was originally tested in a grassland environment using Landsat data may find that the algorithm produces consistent results using Sentinel data captured in the same grassland location. This replication would demonstrate the robustness of the classification method across the two sensors, but it does not provide any evidence that the method would work in a different location or ecosystem. Therefore, the method may be well suited for grasslands, but may perform poorly in urban systems. If the replication was instead performed using Landsat data but in an urban system, and the algorithm performed similarly well as in the grassland system, then the replication would demonstrate the robustness of the algorithm across study areas/environments. However, until the study is replicated successfully in different locations and with different sensors, it cannot be assumed that the algorithm is scientifically valid in all contexts.

Of the two concepts, R&R, reproductions are generally considered more straightforward and easier to implement as long as the researchers who conducted the prior study share the complete set of provenance information from their study. Provenance documentation records the full set of components and activities involved in producing a research outcome, which includes data-focused descriptive metadata and process-focused contextual metadata. With the complete provenance of a study, an independent researcher should be able to check the study for errors, or develop a research design that attempts to replicate the result with a new dataset. In this way reproducibility facilitates replication. Reproductions are also often easier to interpret than replications. Holding constant aspects of the research process, a researcher would expect a reproduction to produce the same results as a prior study. When a replication is pursued, the injection of new data requires that a researcher develop some criteria to make the comparison to the results of the prior study. Selecting those criteria can be complicated because different amounts of natural variation can occur across studies [27–29].

Reproducibility is also a key stepping stone for making remote sensing research extensible. Extensibility refers to the potential to extend existing research materials by adding

new functionality or expanding existing functionality [30]. There are many examples of extensibility in remote sensing, but one brief one is the evolution of spectral unmixing methods. Spectral mixture analysis was introduced to remote sensing in the 1980s [31,32] to extract multiple signals from a single pixel. Since then, the field has built on and extended those original methods to account for endmember variability [33] and incorporate spatial information [34], among other uses.

Extensibility depends on the research materials of a study having three characteristics: they must be available, reproducible, and understandable to the point that an independent research team can reuse them [30]. For example, remote sensing research material that is reproducible and understandable, but not available, can only be used as the basis for future development if an independent researcher takes the time to carefully reconstruct that material. Research materials may be reproducible and available, but if they are not intelligible or are not prepared in a transparent and understandable manner, then they have limited use as the foundation of further development. Of course, materials that are not reproducible cannot be fully vetted by independent researchers, which then requires a researcher to take on faith or reputation that the materials are reliable and worthy of further development.

The ability to build on prior research depends on the reproducibility of research materials in combination with their availability and transparency. Sharing data and code is essential, but it is also necessary to make the methods transparent and provide understandable documentation of the provenance of a study. A key challenge facing the remote sensing community is not just to create repository systems that store data and code, as many systems already exist. Rather the challenge is to make it easier to understand the provenance information that researchers share. This challenge is non-trivial as remote sensing workflows become more complex, integrate data from multiple platforms, require greater computational resources and customized software, and involve larger research teams.

3. Reproducibility and Replicability in Remote Sensing

While structures such as the wide availability of open data captured through calibrated sensors like Landsat that consistently re-image the same spots on Earth should help foster reproducibility, replicability and extensibility in remote sensing research, these goals remain elusive for several reasons. First, the top journals in the field including Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, and IEEE Transactions on Geoscience and Remote Sensing prioritize publishing new methods and algorithms rather than supporting reproductions or replications of existing methods. This preference by editors and publishers for new methods deters researchers from completing highly valuable, yet underappreciated, reproductions and replications. The result is that the field of remote sensing has become flooded with new methodologies, many of which are only ever used by the researcher or team that developed them. Researchers are drowning in a sea of methods and thirsting for evidence to help them decide which ones to use. At best, many methodological studies include pre-existing methods as a comparison case against the new method being introduced. These comparisons should not be mistaken for reproductions or replications as it is not the original study that is being scrutinized, it is the new method. Furthermore, these comparisons are only typically included when they support the development of the new method, and they are omitted if their performance disproves the new method. Thus, these comparisons are wrought with reporting biases.

There are also intrinsic reasons why replications and reproductions are not frequently attempted by the remote sensing community. Within a remote sensing workflow, a myriad of small but important decisions are made by the analyst/s. These decisions include everything from how the data will be geometrically and radiometrically corrected, to any spatial and spectral resampling that might be needed to match the resolution of other data sources, to band or pixel transformations to make the images more interpretable. While each of these decisions may seem minor on its own, when taken together, they can produce thousands or even millions of different possible outcomes even when the basic

methodology described in the paper is followed. Exacerbating this issue is the fact that intermediate data that are produced during an analysis and result from these decisions are often not preserved, making it difficult to trace back these decisions, errors, and uncertainty. In the rare case that these intermediate products are included, they usually do not include accompanying metadata, making it difficult for researchers to fit them into the puzzle.

The uptick in UAS has also complicated the R&R of remote sensing studies. While these alternative data streams have put more options and control over data collection into the hands of users, they also decrease the consistency and stability of data sources, which was one of the advantages enhancing the potential for R&R in remote sensing. Specifically, when users collect their own data with UAS, an entirely new set of decisions is added into the workflow that impact the data and analyses. These decisions include the type of platform and sensor, altitude of the flight, ground sampling distance, flight pattern, flight speed, image/flight line spacing, and sensor trigger cadence, among many others. Since researchers rarely report a complete set of image collection parameters in their publications [14], it can be nearly impossible to replicate a data collection event in a new study area. Most concerning for UAS remote sensing is that the commercial, off-the-shelf (COTS) sensors being used to capture imagery from drones are almost never calibrated pre-flight, and the image reflectances are rarely validated against reflectance targets [14]. Since it is notoriously difficult to determine the reflectance from COTS sensors because manufacturers do not publish the spectral responses [35], researchers are using unknown reflectances to extract biophysical and/or biochemical properties, such as through NDVI. Not only is this problematic for individual remote sensing studies and calls into question any biophysical and biochemical findings based on COTS sensors, but it makes it nearly impossible to replicate UAS studies since mimicking the data collection when using a new platform, sensor, and having to recreate all the flight parameters is all but impossible.

Commercial smallsats introduce parallel, yet unique, issues for R&R in remote sensing. These 'cubesats' are much smaller than traditional satellite platforms (about the size of a loaf of bread) and do not require a dedicated launch vehicle, making them easy to build and quickly deploy into space. The sensor systems on cubesats are more sophisticated than the digital cameras being flown on drones, and there is more transparency in the band positions and spectral responses compared to COTS sensors. However, unlike traditional platforms like Landsat and Sentinel in which a single platform/sensor is capturing each image, cubesats rely on hundreds of platforms in a constellation to capture data. This multiplicity leads to substantial scene-to-scene variation in spectral reflectance [36] (Figure 2), and the scenes often require considerable preprocessing to correct radiometric effects prior to running any analyses [13]. Since repeat acquisitions for the same spot on earth may not be imaged by the same platform/sensor, it is not as straightforward to complete replications over time in the same place as with traditional platforms like Landsat or Sentinel.

From a methodological standpoint, the increasing use of machine learning and neural networks in remote sensing workflows make the field subject to data leakage concerns. Data leakage exists when there is a spurious relationship between the predictor and response variables of a model that is generated by the data collection, sampling, or processing strategy of a researcher [37,38]. Kapoor and Avind [38] identify three distinct forms of data leakage, which are all relevant to the performance and R&R of remote sensing models. First, data leakage can occur when features are used during training that would not otherwise be available to other researchers during the modeling process. For instance, higher resolution proprietary data are often used to train and test spectral unmixing techniques [39], but these data are often unavailable when the model is being run by an independent researcher.



Figure 2. Example of sensor spectral differences in PlanetScope sensors that impact potential replicability of studies.

Second, data leakage can occur when data used to train the model are drawn from a distribution that does not match the distribution of the features of interest or the environment in which the model will eventually be deployed. This form of leakage may be particularly challenging for remote sensing researchers because models trained with data from a spectral library or a single reference site are commonly used to make predictions about new locations at new times. Even when deployed in the same ecosystem, a model may not perform well if the ecosystem has undergone seasonal or other longer term changes. Both temporal and spatial leakages are of increasing concern in remote sensing as the impacts of climate change alter the environment and raise the variability of phenomena.

Finally, remote sensing researchers often touch on security issues and study vulnerable communities that may make it inappropriate to share the information needed to ensure reproducibility [40,41]. When studying such issues, researchers may simply not be allowed to share their data or their workflow information. In these instances, Shepard et al. [42] propose creating and releasing simulated datasets that match the important characteristics of the real data. Whether this approach would work well in remote sensing is debatable, as the spatial, temporal, and spectral structure of remote sensing data often carry indicators of from where and when observations were drawn. As Bennett et al. [21] note, remote sensing research may also reinforce entrenched representations of the world and the unequal hierarchies of power that create them. Replicating research that reinforces existing inequalities may serve to further amplify those inequalities and could be misconstrued as lending credibility to the social and economic structures that generate them.

4. Laying the Groundwork to Capitalize on Reproducibility and Replicability

Concern over the reproducibility of remote sensing research has motivated a range of responses from the remote sensing community. Many of these responses center on improving the sharing of data and code and the collection of provenance information—a record trail that describes the origin of the data (in a database, document or repository) and an explanation of how and why it got to its present place [12,43]. A popular solution to these challenges in computational remote sensing studies is the combined adoption of computational notebooks that implement literate programming practices and version control software that track and manage project code [44–46]. Jupyter and R Markdown are the most commonly used notebooks in many contexts [47] and have recently seen integrations into the increasingly popular Google Earth Engine for remote sensing applications [48]. An important feature of executable notebooks using either Jupyter or R Markdown is that they can be stored and rendered within Github, which allows a researcher to share a detailed

history of code development with an interactive rendering of their final analysis. Joint adoption of both technologies has the additional merit of being an already popular, user driven solution to workflow tracking and sharing, which allows remote sensing researchers to make rapid, if incremental, improvements in reproducibility. A systematic review of what remote sensing researchers have already done with these technologies could yield insights into what a set of best practices for the discipline might look like. However, we are of unaware of any such reviews existing at this time.

Another solution to the challenge of deciding what to share is to develop reproducibility checklists that set criteria for recording and reporting provenance information. Checklists have been developed in several fields that intersect with remote sensing research including machine learning and artificial intelligence [49,50] that can easily be applied to remote sensing studies that use these techniques. Tmusic et al. [51] provide a similar reference sheet, which lists technical and operational details to be recorded during UAS data collection, while Nust et al. [52] provide a list for geocomputational research.

These checklists and standards are also being integrated into the publication process. For example, James et al. [53] created reporting standards for geomorphic research employing structure from motion (SfM) photogrammetry that have been adopted by the journal *Earth Surface Processes and Landforms* as the benchmark for studies utilizing SfM. The *IPOL Journal* (Image Processing Online) is incentivizing reproducible research practices by publishing fully reproducible research articles that are accompanied by relevant data, algorithm details, and source code [54,55]. The journal has published 130 public experiments since 2018. Others have translated standards into badging systems to credit authors for reproducible research [56,57]. For example, Frery et al. [58] developed a badging system specific to remote sensing research and included a proposal for implementing this system through a reproducibility committee, which would oversee the four journals of the *IEEE Geoscience and Remote Sensing Society*. In alignment with this effort, the IEEE has also introduced the *Remote Sensing Code Library* [59], which collects and shares code used in remote sensing research. However, that library is not currently accepting new submissions, and Frery et al.'s badging system has not been adopted, which reflects ongoing hurdles to implementing and sustaining efforts even with interest from the community.

Scaling checklists and badging systems to cover the diversity of work within the remote sensing community presents a further challenge. While laudable, the IEEE effort to ensconce Frery et al.'s badging system covers only a small subset of the publication outlets where remote sensing researchers share their work. That work is also diverse. While the badging system may indeed work well for computationally intensive workflows, it will likely be difficult to apply the same system to mixed methods research in which remote sensing forms only a portion of the larger workflow, which might also include household surveys, stakeholder interviews, or archival analyses. For example, badging may not be a sufficient system for the type of remote sensing imagined in Liverman et al.'s [60] classic *People and Pixels*, or for the complicated and integrated workflows of many projects funded through the NSF Dynamics of Integrated Socio-Environmental Systems (DISES) program.

While checklists and badges facilitate standardized reporting of remote sensing research, they are commonly used only at the end of the research process as reminders of what to report, and they do not include specific recommendations about how to record provenance information in a standardized format. Linking these checklists and badges to provenance models and reporting standards already in place would increase the legibility of provenance information and interoperability across studies. Provenance models standardize the reporting about the products (e.g., data, code, reports) made during the research process, the researchers that were involved in creating those things (agents), and the actions those researchers took to make them (activities).

One option is for remote sensing researchers to adopt the W3C PROV model [61], which has a specific syntax and ontology for recording entity, agent, and activity information. The W3C PROV model has specific advantages over the use of a checklist. For example, if a data file of a land cover classification is one product of a remote sensing

research study, a checklist will remind the researcher to share that original imagery and code that went into creating that land cover product. While the code may provide some insight into how the classification was created, it is often difficult to interpret code due to poor commenting practices. Even when code is commented, details about who generated that code or how the code was used in combination with other research materials (e.g., hardware) to create the final product is often omitted. The W3C PROV model can overcome these challenges and be used to track these relationships and ascribe specific actions to individuals and times. In this way the development history of the same code could be viewed. Moreover, if sections of that code are borrowed from previous code, those relationships can also be tracked thereby marking the intellectual legacy of the new code. The provenance model could similarly record connection with data inputs of even dataset used to develop and test code performance, which would again provide greater insight into how and why the code was developed as it was.

Integrating the W3C PROV model with other computational resources would further improve consistency across remote sensing workflows and facilitate the comparison and extension of research. For example, remote sensing researchers could containerize their workflows and track progress with version control software. Containerization packages the data, code, software dependencies, and operating system information of an analysis together in a format that can run consistently on any computer. Sharing a container, rather than individual scripts, data, and software and hardware documentation facilitates reproductions and replications by making the computational aspects of a remote sensing analysis explicit. When a researcher couples a container with documentation following a widely adopted provenance model, they make their analysis understandable, available, and reproducible, and therefore extensible.

One challenge to containerization and sharing is the use of specialized infrastructure (e.g., institutional supercomputing resources) or sensitive/proprietary data in analyses. Limited or restricted availability of these resources make it difficult or impossible for other researchers to execute containers on their own hardware. However, even when a container cannot be run, containerization and provenance documentation still improve the legibility of research by linking data and code with provenance. Moreover, if the container, restricted data, and provenance record are stored on specialized infrastructure, the researcher that developed that analysis could allow curated access to that infrastructure and container. Developing protocols, review procedures, and standards for access that institutions can adopt for remote sensing research is an important task to facilitate the transition to containers and provenance documentation. A leading example of this type of system that the remote sensing community might emulate is the developing Geospatial Virtual Data Enclave [62,63] hosted by the Inter-university Consortium for Political and Social Research. That enclave not only makes data and code accessible through virtual machines, its creators are also developing a credentialing system that allows varied levels of access to sensitive information to different researchers.

While combining containerization and provenance models has clear benefits for computational analyses in remote sensing, it must be stressed that this pairing can also improve the capture and communication of the entire remote sensing research process, including projects that use remote sensing data streams from multiple platforms (e.g., satellites, cubesats, and UAS). Consider the example of the Central Arizona-Phoenix Long Term Ecological Research Project (CAP LTER)—a multi-decade, multidisciplinary research project monitoring and assessing the socio-ecological dynamics of the Phoenix metro-area and Sonoran Desert. The CAP LTER team uses satellite-based optical data, aerial imagery, and lidar data to support numerous ongoing analyses of the entire Phoenix region. CAP LTER researchers are now also using UAS to collect high-resolution data in specific locations, and are analyzing that data in combination with the satellite- and airborne-base data products for the wider region. Applied at the level of an individual research project within CAP LTER, a researcher could use the W3C PROV Model to track UAS mission details and data processing steps. If the researcher also used satellite-based products (e.g., land cover

classification) prepared by another researcher in the CAP LTER in their analysis, and that processing was also documented using the W3C PROV model, the two independently executed research components could be linked and shared. Moreover, if both processes were containerized, a researcher using the two data streams could create and share a container with their complete analysis. Scaled to the CAP LTER project as a whole, consistent provenance documentation and containerization of all remote sensing data collection and analysis within the project would create interoperable and extensible research packages that could be compared and recombined through new analyses.

5. Leveraging Recent Developments in Remote Sensing by Focusing on the Reproducibility and Replicability of Research

While developing and adopting systems that facilitate community-wide documentation of provenance information and data sharing are clearly critical to enhancing the reproducibility of remote sensing research, we have the opportunity to look beyond these tasks as well. If remote sensing researchers focus on the function that reproductions and replications play in scientific investigation, they will find opportunities to restructure existing research streams and open new avenues of investigation. Here, we discuss two such opportunities: (1) forward-looking replication sequences, and (2) benchmarking for multi-platform integration.

5.1. Forward-Looking Replication Sequences

Rather than looking at replication as a means of assessing the claims of prior studies, we can instead design sequences of replications to progressively and purposefully gather information about a topic of interest [64,65]. The simultaneous improvement of the resolution of remote sensing data and our capacity to process that data through high-performance or cloud-based computing creates this opportunity to design forward looking research programs around the error correcting function of replication. For example, there is presently debate around how to best approach ecological niche modeling [66,67] and conservation in terrestrial and aquatic environments [68]. If a set of competing niche models that use remote sensing data were specified, they could leverage the global scope of remote sensing data to fit those models across a sequence of locations. At each location, researchers could compare the performance of the niche models, and use a measure of that performance (e.g., Bayesian Information Criteria) to update a vector of model weights using Bayes theorem. Repeated across successive studies, that vector of model weights would reflect the information available about each model from the complete prior sequence of studies. Niche models that repeatedly fit the data and consequently receive greater weight would garner greater belief. This same approach could be applied to identify which implementation of a remote sensing method, such as a ML-based classification method, is the most robust to changes in location or data inputs.

There is also a crucial link between forward-looking replication studies in remote sensing and the need to address climate change. The IPCC 6th Assessment Report [69] makes clear not only that climate change is happening, but that we have technologies and interventions to mitigate its impacts. A pressing question then is how to efficiently scale those technologies and interventions across the planet. Framed in this way, scaling is a question of whether an intervention will replicate in another location. Adapting the approach above, we might design intervention sequences across places that have been closely matched but are also known to vary in key dimensions from remote sensing studies. Such sequences could test what environmental limits constrain the effectiveness of those interventions.

The data and computational foundations needed to begin implementing replication-based approaches are already in place. The niche modeling example presented above could be implemented using the datasets, APIs, and computational resources available through Google Earth Engine [70] or the Microsoft Planetary Computer [71]. However, realizing the potential of this approach remains constrained by a set of challenges succinctly summarized

by the Trillion Pixel Challenge [72]. While researchers may be able to replicate and compare models across contexts, delving deeper into the theories behind those models and the processes they represent will depend on expanding the variety of remote sensing data available, integrating localized data sampled from non-orbital platforms, improving our ability to learn from labeled and unlabeled data, and developing a better understanding of the effects of the sequencing of replication series.

Ongoing research taking place inside and outside of remote sensing could be used to address these challenges. Researchers at Oak Ridge National Laboratory have recently developed the RESflow system that uses AI to identify, characterize, and match regions in remotely sensed data [73]. Understanding what characteristics regions do or do not share is important for designing replication sequences to maximize information about the applicability of a model, or the scalability of a policy intervention. Al-Ubaydil et al. [74–76] have explored this problem in depth within the experimental economics literature and have introduced a model and framework that outlines the threats to scaling policy interventions. The authors highlight four factors that shape the chances of an intervention scaling: the quality of the initial inference, situational confounds, population characteristics, and general equilibrium effect. Using this framework, they suggest designing study sequences that first evaluate a model or interventions where it is most likely to work, and then test the model in edge situations that vary on some dimension that could lead to non-replication. While not a perfect match for all types of remote sensing research, the lessons of this approach could be applied to the investigation or monitoring of interventions using remotely sensed data such as the climate change adaptation/mitigation strategies discussed above. For example, remotely sensed data could be used to identify environmental factors or population dynamics that define edge locations that provide information about why interventions might fail.

Another provocative avenue forward is to link remote sensing research sequences with randomized control trials (RCT) of landscape-scale interventions. In the language of scaling, RCT can provide clear answers about the efficacy of an intervention (what works here), while study sequences address the question of effectiveness (does it, or will it work somewhere else). Facilitated by the capabilities of remote sensing technologies, landscape-scale RCTs that cover multiple large geographic extents have the potential to answer both questions simultaneously. We are aware of few such studies in the literature. Wiik et al. [77] use remote sensing to monitor participant compliance in an RCT testing the impact of an incentive-based conservation program in the Bolivian Andes. Weigel et al. [78] similarly use satellite image-derived estimates of crop cover to measure adoption of an agricultural conservation program in the Mississippi River Basin. Both studies demonstrate how remote sensing can be integrated into RCTs but also open potentially insightful avenues for reproduction and reanalysis. For example, a researcher using the data shared by Weigel et al. could identify geographic variation in program adoption and then create imagined intervention sequences to test how rapidly different sequence structures converged to the known result. That knowledge could be used to plan subsequent interventions, which could again be monitored remotely.

Shifting the remote sensing research agenda in the direction of replication sequences will also require a shift in researcher incentives from ‘one-off’ examinations to holistic programs of study. As Nichols et al. [65] note, academic institutions and funding agencies appear to value standalone contributions with the potential for high citation more than repeated reproductions or replications. One solution could be for journal editors to commission articles or letter series that give updates about progress on a specific issue being studied by a replication sequence. Commissioning these series would be in the interest of the publisher and enhance the standing of a publication if that journal became the source of reference for the repeated release of evidence about key issues in remote sensing. Just as the release of datasets such as the National Landcover Dataset (NLCD) are anticipated every five years, a well-designed sequence of replication testing competing ecological theories or

assessing the relative performance of new methods could garner similar attention. Acting as the lead researcher of such an effort would have clear career advantages.

Agencies like the NSF and NASA that offer external funding opportunities can also start to change researcher incentives to perform replications and plan replication sequences by releasing calls for proposals for this specific type of work. To begin a transition, agencies can start with small targeted calls. For example, it would make sense to target funding for replication sequences at graduate students or early to mid-career scholars who are still forming their research programs. Funding for replication sequences could also be attached to existing data gathering missions or monitoring programs that are already designed to repeatedly produce the consistent quality data needed to compare models of methods through replication. This approach has the additional advantage of lowering barriers to adoption because research teams already using these datasets are already familiar with accessing and analyzing repeatedly collected data. Funding could also be directed to educational programming or as an additional educational component to already funded research. While emphasizing education would help improve future research practices by training future remote sensors in reproducibility standards and practices, one hindrance to this approach would be the limited capacity to teach these practices in the current research community.

5.2. Benchmarking Datasets for Method Comparisons and Multi-Platform Integration

When a researcher introduces a new remote sensing procedure, algorithm, or program and shares the provenance information and materials used in their development, another researcher is able to understand and assess what they have done and build on their work. Through reproduction, assessments can be made in relation to the materials used in the original work. Through replication, assessments can be made in relation to a new dataset, population, or context. However, researchers can only directly assess the method and materials used in a study in relation to other competing methods when both sets of methods are executed on a common dataset.

Benchmark datasets are openly accessible datasets that a community uses for testing and comparing new methods. By controlling for variation in data, benchmark datasets can be the basis for the fair comparison and validation of new remote sensing methods [79,80]. Good benchmark datasets have at least four characteristics. First, they are discoverable, accessible, and understandable. Prominent publication outlets such as *Scientific Data* already exist where benchmark remote sensing data could be shared. Remote sensing can also leverage its connection to well-established government data providers as sites to host benchmarks related to those products. Ideally, these datasets would also carry open source licenses that permit re-use with proper attribution (see Stodden [81,82] for a discussion of licensing and reproducibility). Second, good benchmark data are linked to a specific remote sensing problem (e.g., classification, change detection) and are complex enough to be interesting. For example, if the benchmark data will be used to compare the accuracy of classification algorithms, then a heterogeneous mixture of land covers that includes features with different minimum mapping units is preferable over a homogeneous scene with little variation.

Third, good benchmark data should be ground-truthed and labeled in a way that is relevant to the remote sensing problem it is being used to assess. A recent issue of *Remote Sensing* outlines some of the challenges of selecting remote sensing datasets for specified tasks and the need to understand what information datasets may contain [83]. Returning to the classification example above, the land cover classes must be able to be labeled correctly otherwise it is impossible to assess the performance of the algorithm. Remote sensing has established performance measures that can be used as the basis for straightforward comparisons, such as differences in overall classification accuracy or user/producer accuracy. However, it may be more difficult to compare methods that vary in complexity. For example, if a newly developed classification method is developed and tested using a bundle of datasets in a benchmark dataset but is compared to an existing approach that was originally

trained on a single dataset, it may not compare well because its performance threshold is now higher. Should the second approach be penalized for being more rigorous than the first? The field must grapple with these types of questions as algorithms become more complex and standards are simultaneously rising. Lastly, good benchmark data should be well documented with metadata that is written in a way that is accessible to a non-expert. The metadata should follow a widely adopted provenance model and contain information about how the data was constructed and for what purpose. Procedures that were used to clean and prepare the data should also be documented, and a container implementing those processes should be shared. To promote use and further ease access, an explanation of how to load and inspect the benchmark data would be shared.

Benchmark datasets that have these four characteristics can be used to address a range of remote sensing challenges. Perhaps the most obvious application is the relative comparison of new algorithms and programs. Several researchers have already introduced highly cited benchmark datasets with their remote sensing papers [84,85]. Less obviously, a collection of related benchmark datasets can be used to address the challenge of remote sensing data gathered from multiple platforms. For example, if a collection of benchmark datasets included traditional satellite-based (e.g., Landsat), personal data (e.g., 4-band imagery collected with a UAS), and terrestrial (e.g., ground-based Lidar) remote sensing products, researchers could use that data to compare data integration and analysis workflow linked to a specific research problem. To test the sensitivity of the resulting workflows to features like coverage or resolution, the packages could be altered and re-analyzed. For instance, the spatial coverage of the UAS data included in the original package may be changed to only partially overlap with the traditional data products.

6. Conclusions

Recent changes in the technologies, practices, and methodologies of remote sensing have created opportunities to observe and understand the world in new ways. To capitalize on these opportunities, the remote sensing community should bring R&R questions forward on the disciplinary research agenda. Activities to improve reproducibility should include prioritizing the collection and sharing of provenance information in standardized and easily understood formats. However, we argue that researchers should give equal, if not greater attention, to the functions reproduction and replication play in scientific research and act to build individual and community-wide research programs around those functions.

Well-designed sequences of replications focused on a topic could help researchers distinguish between competing explanations and methods, which could translate into better monitoring and intervention programs. Establishing benchmark datasets and incentives to use those resources in methods comparisons and data integration challenges could simultaneously create a culture of reproducible research practices and facilitate the assessment and extension of new methodologies. Funding agencies and academic publications can facilitate both activities by creating calls for proposals, or papers or organizing Special Issues around benchmark datasets or replication sequences on a particular topic or method. Such calls could also be organized in the form of competitions in which reproducible research that makes the most progress on a specific task receives a cash award and prominent publication.

Author Contributions: P.K. and A.E.F. contributed equally to all aspects of this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant to P.K. from the U.S. National Science Foundation, grant number 2049837. A.E.F. is supported by U.S. National Science Foundation grants number 1934759 and 2225079.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, L.; Driscoll, J.; Sarigai, S.; Wu, Q.; Chen, H.; Lippitt, C.D. Google Earth Engine and Artificial Intelligence (AI): A Comprehensive Review. *Remote Sens.* **2022**, *14*, 3253. [CrossRef]
2. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
3. Wang, L.; Yan, J.; Mu, L.; Huang, L. Knowledge discovery from remote sensing images: A review. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1371. [CrossRef]
4. Jupyter, P.; Bussonnier, M.; Forde, J.; Freeman, J.; Granger, B.; Head, T.; Holdgraf, C.; Kelley, K.; Nalvarte, G.; Osherooff, A.; et al. Binder 2.0—Reproducible, interactive, sharable environments for science at scale. In Proceedings of the 17th Python in Science Conference, Austin, TX, USA, 9–15 July 2018; pp. 75–82. [CrossRef]
5. Nüst, D. Reproducibility Service for Executable Research Compendia: Technical Specifications and Reference Implementation. Available online: <https://zenodo.org/record/2203844#.Y17s6oTMluV> (accessed on 10 October 2022).
6. Brinckman, A.; Chard, K.; Gaffney, N.; Hategan, M.; Jones, M.B.; Kowalik, K.; Kulasekaran, S.; Ludäscher, B.; Mecum, B.D.; Nabrzycki, J.; et al. Computing environments for reproducibility: Capturing the “Whole Tale”. *Futur. Gener. Comput. Syst.* **2019**, *94*, 854–867. [CrossRef]
7. Woodward, J.F. Data and phenomena: A restatement and defense. *Synthese* **2011**, *182*, 165–179. [CrossRef]
8. Haig, B.D. Understanding Replication in a Way That Is True to Science. *Rev. Gen. Psychol.* **2022**, *26*, 224–240. [CrossRef]
9. Popper, K.R. *The Logic of Scientific Discovery*, 2nd ed.; Routledge: London, UK, 2002; ISBN 978-0-415-27843-0.
10. Earp, B.D. Falsification: How Does It Relate to Reproducibility? In *Research Methods in the Social Sciences: An A–Z of Key Concepts*; Oxford University Press: Oxford, UK, 2021; pp. 119–123. ISBN 978-0-19-885029-8.
11. Balz, T.; Rocca, F. Reproducibility and Replicability in SAR Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3834–3843. [CrossRef]
12. Howe, C.; Tullis, J.A. Context for Reproducibility and Replicability in Geospatial Unmanned Aircraft Systems. *Remote Sens.* **2022**, *14*, 4304. [CrossRef]
13. Frazier, A.E.; Hemingway, B.L. A Technical Review of Planet Smallsat Data: Practical Considerations for Processing and Using PlanetScope Imagery. *Remote Sens.* **2021**, *13*, 3930. [CrossRef]
14. Singh, K.K.; Frazier, A.E. A meta-analysis and review of unmanned aircraft system (UAS) imagery for terrestrial applications. *Int. J. Remote Sens.* **2018**, *39*, 5078–5098. [CrossRef]
15. Tullis, J.A.; Corcoran, K.; Ham, R.; Kar, B.; Williamson, M. Multiuser Concepts and Workflow Replicability in SUAS Applications. In *Applications in Small Unmanned Aircraft Systems*; CRC Press/Taylor & Francis Group: Boca Raton, FL, USA, 2019; pp. 34–55.
16. Kedron, P.; Li, W.; Fotheringham, S.; Goodchild, M. Reproducibility and replicability: Opportunities and challenges for geospatial research. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 427–445. [CrossRef]
17. Waters, N. Motivations and Methods for Replication in Geography: Working with Data Streams. *Ann. Am. Assoc. Geogr.* **2021**, *111*, 1291–1299. [CrossRef]
18. Gertler, P.; Galiani, S.; Romero, M. How to make replication the norm. *Nature* **2018**, *554*, 417–419. [CrossRef]
19. Neuliep, J.W. Editorial Bias against Replication Research. *J. Soc. Behav. Personal.* **1990**, *5*, 85.
20. Wainwright, J. Is Critical Human Geography Research Replicable? *Ann. Am. Assoc. Geogr.* **2021**, *111*, 1284–1290. [CrossRef]
21. Bennett, M.M.; Chen, J.K.; León, L.F.A.; Gleason, C.J. The politics of pixels: A review and agenda for critical remote sensing. *Prog. Hum. Geogr.* **2022**, *46*, 729–752. [CrossRef]
22. Committee on Reproducibility and Replicability in Science; Board on Behavioral, Cognitive, and Sensory Sciences; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; Nuclear and Radiation Studies Board; Division on Earth and Life Studies; Board on Mathematical Sciences and Analytics; Committee on Applied and Theoretical Statistics; Division on Engineering and Physical Sciences; Board on Research Data and Information; et al. *Reproducibility and Replicability in Science*; National Academies Press: Washington, DC, USA, 2019; p. 25303. ISBN 978-0-309-48616-3.
23. Gundersen, O.E.; Kjensmo, S. State of the Art: Reproducibility in Artificial Intelligence. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
24. Essawy, B.T.; Goodall, J.L.; Voce, D.; Morsy, M.M.; Sadler, J.M.; Choi, Y.D.; Tarboton, D.G.; Malik, T. A taxonomy for reproducible and replicable research in environmental modelling. *Environ. Model. Softw.* **2020**, *134*, 104753. [CrossRef]
25. Nosek, B.A.; Errington, T.M. What is replication? *PLoS Biol.* **2020**, *18*, e3000691. [CrossRef]
26. Kedron, P.; Frazier, A.E.; Trgovac, A.; Nelson, T.; Fotheringham, A.S. Reproducibility and Replicability in Geographical Analysis. *Geogr. Anal.* **2021**, *53*, 135–147. [CrossRef]
27. Gelman, A.; Stern, H. The Difference between “Significant” and “Not Significant” is not Itself Statistically Significant. *Am. Stat.* **2006**, *60*, 328–331. [CrossRef]
28. Jilke, S.; Petrovsky, N.; Meuleman, B.; James, O. Measurement equivalence in replications of experiments: When and why it matters and guidance on how to determine equivalence. *Public Manag. Rev.* **2017**, *19*, 1293–1310. [CrossRef]
29. Hoepfner, S. A note on replication analysis. *Int. Rev. Law Econ.* **2019**, *59*, 98–102. [CrossRef]
30. Goeva, A.; Stoudt, S.; Trisovic, A. Toward Reproducible and Extensible Research: From Values to Action. *Harv. Data Sci. Rev.* **2020**, *2*. [CrossRef]

31. Dozier, J. A method for satellite identification of surface temperature fields of subpixel resolution. *Remote Sens. Environ.* **1981**, *11*, 221–229. [[CrossRef](#)]
32. Adams, J.B.; Smith, M.O.; Johnson, P.E. Spectral mixture modeling: A new analysis of rock and soil types at the Viking Lander 1 Site. *J. Geophys. Res. Earth Surf.* **1986**, *91*, 8098–8112. [[CrossRef](#)]
33. Somers, B.; Asner, G.P.; Tits, L.; Coppin, P. Endmember variability in Spectral Mixture Analysis: A review. *Remote Sens. Environ.* **2011**, *115*, 1603–1616. [[CrossRef](#)]
34. Shi, C.; Wang, L. Incorporating spatial information in spectral unmixing: A review. *Remote Sens. Environ.* **2014**, *149*, 70–87. [[CrossRef](#)]
35. Mathews, A.J. A Practical UAV Remote Sensing Methodology to Generate Multispectral Orthophotos for Vineyards: Estimation of Spectral Reflectance Using Compact Digital Cameras. *Int. J. Appl. Geospat. Res.* **2015**, *6*, 65–87. [[CrossRef](#)]
36. Csillik, O.; Asner, G.P. Near-real time aboveground carbon emissions in Peru. *PLoS ONE* **2020**, *15*, e0241418. [[CrossRef](#)]
37. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1–21. [[CrossRef](#)]
38. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in ML-Based Science. *arXiv* **2022**, arXiv:2207.07048.
39. Frazier, A.E. Accuracy assessment technique for testing multiple sub-pixel mapping downscaling factors. *Remote Sens. Lett.* **2018**, *9*, 992–1001. [[CrossRef](#)]
40. Fisher, M.; Fradley, M.; Flohr, P.; Rouhani, B.; Simi, F. Ethical considerations for remote sensing and open data in relation to the endangered archaeology in the Middle East and North Africa project. *Archaeol. Prospect.* **2021**, *28*, 279–292. [[CrossRef](#)]
41. Mahabir, R.; Croitoru, A.; Crooks, A.T.; Agouris, P.; Stefanidis, A. A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. *Urban Sci.* **2018**, *2*, 8. [[CrossRef](#)]
42. Shepherd, B.E.; Peratikos, M.B.; Rebeiro, P.F.; Duda, S.N.; McGowan, C.C. A Pragmatic Approach for Reproducible Research with Sensitive Data. *Am. J. Epidemiol.* **2017**, *186*, 387–392. [[CrossRef](#)] [[PubMed](#)]
43. Tullis, J.A.; Kar, B. Where Is the Provenance? Ethical Replicability and Reproducibility in GIScience and Its Critical Applications. *Ann. Am. Assoc. Geogr.* **2020**, *111*, 1318–1328. [[CrossRef](#)]
44. Rapiński, J.; Bednarczyk, M.; Zinkiewicz, D. JupyterTEP IDE as an Online Tool for Earth Observation Data Processing. *Remote Sens.* **2019**, *11*, 1973. [[CrossRef](#)]
45. Wagemann, J.; Fierli, F.; Mantovani, S.; Siemen, S.; Seeger, B.; Bendix, J. Five Guiding Principles to Make Jupyter Notebooks Fit for Earth Observation Data Education. *Remote Sens.* **2022**, *14*, 3359. [[CrossRef](#)]
46. Hogenson, K.; Meyer, F.; Logan, T.; Lewandowski, A.; Stern, T.; Lundell, E.; Miller, R. The ASF OpenSARLab A Cloud-Based (SAR) Remote Sensing Data Analysis Platform. In Proceedings of the AGU Fall Meeting 2021, New Orleans, LA, USA, 13–17 December 2021.
47. Nüst, D.; Pebesma, E. Practical Reproducibility in Geography and Geosciences. *Ann. Am. Assoc. Geogr.* **2020**, *111*, 1300–1310. [[CrossRef](#)]
48. Owusu, C.; Snigdha, N.J.; Martin, M.T.; Kalyanapu, A.J. PyGEE-SWToolbox: A Python Jupyter Notebook Toolbox for Interactive Surface Water Mapping and Analysis Using Google Earth Engine. *Sustainability* **2022**, *14*, 2557. [[CrossRef](#)]
49. Gundersen, O.E.; Gil, Y.; Aha, D.W. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Mag.* **2018**, *39*, 56–68. [[CrossRef](#)]
50. Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv* **2020**, arXiv:2003.12206.
51. Tmušić, G.; Manfreda, S.; Aasen, H.; James, M.R.; Gonçalves, G.; Ben-Dor, E.; Brook, A.; Polinova, M.; Arranz, J.J.; Mészáros, J.; et al. Current Practices in UAS-based Environmental Monitoring. *Remote Sens.* **2020**, *12*, 1001. [[CrossRef](#)]
52. Nüst, D.; Ostermann, F.O.; Sileryte, R.; Hofer, B.; Granell, C.; Teperek, M.; Graser, A.; Broman, K.W.; Hettne, K.M.; Clare, C.; et al. AGILE Reproducible Paper Guidelines. Available online: osf.io/cb7z8 (accessed on 10 October 2022).
53. James, M.R.; Chandler, J.H.; Eltner, A.; Fraser, C.; Miller, P.E.; Mills, J.; Noble, T.; Robson, S.; Lane, S.N. Guidelines on the use of structure-from-motion photogrammetry in geomorphic research. *Earth Surf. Process. Landforms* **2019**, *44*, 2081–2084. [[CrossRef](#)]
54. Colom, M.; Kerautret, B.; Limare, N.; Monasse, P.; Morel, J.-M. IPOL: A New Journal for Fully Reproducible Research; Analysis of Four Years Development. In Proceedings of the 2015 7th International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, 27–29 July 2015; pp. 1–5.
55. Colom, M.; Dagobert, T.; de Franchis, C.; von Gioi, R.G.; Hessel, C.; Morel, J.-M. Using the IPOL Journal for Online Reproducible Research in Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6384–6390. [[CrossRef](#)]
56. Nüst, D.; Lohoff, L.; Einfeldt, L.; Gavish, N.; Götz, M.; Jaswal, S.; Khalid, S.; Meierkort, L.; Mohr, M.; Rendel, C. Guerrilla Badges for Reproducible Geospatial Data Science (AGILE 2019 Short Paper); Physical Sciences and Mathematics. 2019. Available online: <https://eartharxiv.org/repository/view/839/> (accessed on 15 October 2022).
57. Wilson, J.P.; Butler, K.; Gao, S.; Hu, Y.; Li, W.; Wright, D.J. A Five-Star Guide for Achieving Replicability and Reproducibility When Working with GIS Software and Algorithms. *Ann. Am. Assoc. Geogr.* **2021**, *111*, 1311–1317. [[CrossRef](#)]
58. Frery, A.C.; Gomez, L.; Medeiros, A.C. A Badging System for Reproducibility and Replicability in Remote Sensing Research. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4988–4995. [[CrossRef](#)]

59. Remote Sensing Code Library|Home. Available online: <https://tools.grss-ieee.org/rscl1/index.html> (accessed on 5 October 2022).
60. Liverman, D.M. *People and Pixels: Linking Remote Sensing and Social Science*; National Academy Press: Washington, DC, USA, 1998; ISBN 978-0-585-02788-3.
61. Missier, P.; Belhajjame, K.; Cheney, J. The W3C PROV Family of Specifications for Modelling Provenance Metadata. In Proceedings of the 16th International Conference on Extending Database Technology—EDBT '13, Genoa, Italy, 18–22 March 2013; p. 773.
62. Richardson, D.B.; Kwan, M.-P.; Alter, G.; McKendry, J.E. Replication of scientific research: Addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Ann. GIS* **2015**, *21*, 101–110. [[CrossRef](#)]
63. Richardson, D. Dealing with Geoprivacy and Confidential Geospatial Data. *ARC News* **2019**, *41*, 30.
64. Nichols, J.D.; Kendall, W.L.; Boomer, G.S. Accumulating evidence in ecology: Once is not enough. *Ecol. Evol.* **2019**, *9*, 13991–14004. [[CrossRef](#)]
65. Nichols, J.D.; Oli, M.K.; Kendall, W.L.; Boomer, G.S. A better approach for dealing with reproducibility and replicability in science. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2100769118. [[CrossRef](#)]
66. Feng, X.; Park, D.S.; Walker, C.; Peterson, A.T.; Merow, C.; Papeş, M. A checklist for maximizing reproducibility of ecological niche models. *Nat. Ecol. Evol.* **2019**, *3*, 1382–1395. [[CrossRef](#)] [[PubMed](#)]
67. Leitão, P.J.; Santos, M.J. Improving Models of Species Ecological Niches: A Remote Sensing Overview. *Front. Ecol. Evol.* **2019**, *7*, 9. [[CrossRef](#)]
68. Leidner, A.K.; Buchanan, G.M. (Eds.) *Satellite Remote Sensing for Conservation Action: Case Studies from Aquatic and Terrestrial Ecosystems*, 1st ed.; Cambridge University Press: Cambridge, UK, 2018; ISBN 978-1-108-63112-9.
69. Climate Change 2022: Mitigation of Climate Change. In *Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: New York, NY, USA, 2022.
70. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [[CrossRef](#)]
71. Microsoft Planetary Computer. Available online: <https://planetarycomputer.microsoft.com/> (accessed on 22 September 2022).
72. Yang, L.; Lunga, D.; Bhaduri, B.; Begoli, E.; Lieberman, J.; Doster, T.; Kerner, H.; Casterline, M.; Shook, E.; Ramachandran, R.; et al. *2021 GeoAI Workshop Report: The Trillion Pixel Challenge* (No. ORNL/LTR-2021/2326); Oak Ridge National Lab. (ORNL): Oak Ridge, TN, USA, 2021; p. 1883938.
73. Lunga, D.; Gerrand, J.; Yang, L.; Layton, C.; Stewart, R. Apache Spark Accelerated Deep Learning Inference for Large Scale Satellite Image Analytics. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 271–283. [[CrossRef](#)]
74. Al-Ubaydli, O.; List, J.A.; Suskind, D. 2017 Klein Lecture: The Science of Using Science: Toward an Understanding of the Threats to Scalability. *Int. Econ. Rev.* **2020**, *61*, 1387–1409. [[CrossRef](#)]
75. Al-Ubaydli, O.; Lee, M.S.; List, J.A.; Suskind, D. The Science of Using Science. In *The Scale-Up Effect in Early Childhood and Public Policy*; List, J.A., Suskind, D., Supplee, L.H., Eds.; Routledge: London, UK, 2021; pp. 104–125. ISBN 978-0-367-82297-2.
76. Kane, M.C.; Sablich, L.; Gupta, S.; Supplee, L.H.; Suskind, D.; List, J.A. Recommendations for Mitigating Threats to Scaling. In *The Scale-Up Effect in Early Childhood and Public Policy*; List, J.A., Suskind, D., Supplee, L.H., Eds.; Routledge: London, UK, 2021; pp. 421–425. ISBN 978-0-367-82297-2.
77. Wiik, E.; Jones, J.P.G.; Pynegar, E.; Bottazzi, P.; Asquith, N.; Gibbons, J.; Kontoleon, A. Mechanisms and impacts of an incentive-based conservation program with evidence from a randomized control trial. *Conserv. Biol.* **2020**, *34*, 1076–1088. [[CrossRef](#)]
78. Weigel, C.; Harden, S.; Masuda, Y.J.; Ranjan, P.; Wardropper, C.B.; Ferraro, P.J.; Prokopy, L.; Reddy, S. Using a randomized controlled trial to develop conservation strategies on rented farmlands. *Conserv. Lett.* **2021**, *14*. [[CrossRef](#)]
79. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]
80. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [[CrossRef](#)]
81. Stodden, V. Enabling Reproducible Research: Open Licensing for Scientific Innovation. *Int. J. Commun. Law Policy* **2009**, *13*, forthcoming.
82. Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. (Eds.) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*; Cambridge University Press: New York, NY, USA, 2014; ISBN 978-1-107-06735-6.
83. Vasquez, J.; Kokhanovsky, A. Special Issue “Remote Sensing Datasets” 2022. Available online: https://www.mdpi.com/journal/remotesensing/special_issues/datasets (accessed on 20 October 2022).
84. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [[CrossRef](#)]
85. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [[CrossRef](#)]