Reviewer 1

Reviewed by anonymous reviewer, 31 Dec 2021 23:13

Here the authors outlined some important considerations when considering using high resolution remote sensing data. This type of data is becoming increasingly accessible and this is a worthwhile thought piece on how to approach this type of data.

Authors: Thank you.

I feel this is a very good manuscript. Researchers often default toward “more data is better” and thus are attracted toward high resolution remote sensing imagery. The points outlined here will help them reconsider the practical considerations of using that data, and either reconsider the decision or be better prepared for the process.

Authors: Thank you.

I have only minor suggestions, mostly adding useful citations which can help further inform readers. Note I reviewed this version 2 here https://doi.org/10.31219/osf.io/kehqz

Intro: Can you define what you feel high resolution data is currently (maybe just which sensors are considered hi-res in 2021), and how that might change in the future?

Authors: We agree about the importance of clarity of language. We specified “sub-meter, sub-hourly or hyper-spectral” in the abstract on line 20, and broadened the discussion with an added paragraph to the introduction on lines 37-50 to expand upon this topic.

L50: “know your question”. This is a great basic rule that researchers should follow for all studies. Recommend citing et al Betts et al 2021, along with Alon 2009, which has some very general points regarding this.

Betts, M.G., Hadley, A.S., Frey, D.W., Frey, S.J., Gannon, D., Harris, S.H., Kim, H., Kormann, U.G., Leimberger, K., Moriarty, K. and Northrup, J.M., 2021. When are hypotheses useful in ecology and evolution?. Ecology and evolution.

Authors: We have added this citation on line 84

Rule 1: This could use some examples of what can go wrong if you don’t focus on an overarching question/ hypothesis. For example, if someone focuses solely on using high resolution data, their analysis may not be able to explain ecological phenomena better than coarse resolution data, since it can be difficult to accurately model the fine resolution variability. (Hallet et al 2004). High resolution data may also have inflated accuracies due to autocorrelation (Ploton et al 2020).

Hallett, T B, T Coulson, J G Pilkington, T H Clutton-Brock, J M Pemberton, and B T Grenfell. 2004. “Why Large-Scale Climate Indices Seem to Predict Ecological Processes Better than Local Weather.” Nature 430 (6995): 71–75. https://doi.org/10.1038/nature02708.

Ploton, Pierre, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, et al. 2020. “Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models.” Nature Communications 11 (1): 4540. https://doi.org/10.1038/s41467-020-18321-y.

Authors: Thank you for these suggestions. We have included both of these references in the text for rule 3, on lines 166 and 172.

L87: Since this is in PCI Ecology some literature on scaling in ecology would provide great context here. I recommend including the following papers which provide great practical overviews of the topic:

Knegt, H. J. de, F. van Langevelde, M. B. Coughenour, A. K. Skidmore, W. F. de Boer, I.M. A. Heitkönig, N. M. Knox, R. Slotow, C. van der Waal, and H. H. T. Prins. 2010. “Spatial Autocorrelation and the Scaling of Species–Environment Relationships.” Ecology 91 (8): 2455–65. https://doi.org/10.1890/09-1359.1.

Sandel, Brody. 2015. “Towards a Taxonomy of Spatial Scale-Dependence.” Ecography 38 (4): 358–69. https://doi.org/10.1111/ecog.01034.

Authors: We added text on lines 117-119 to address this point.

L101: “knowing your data” should also consider the tradeoffs of high resolution data. Finer spatial scale usually means both a coarser temporal scale (ie. Daily MODIS to 16-day Landsat) and less robust radiometric quality (eg. Planet lab sensors are not as precise as Landsat or Modis, Houborg et al. 2018)

Houborg, Rasmus, and Matthew F. McCabe. 2018. “A Cubesat Enabled Spatio-Temporal Enhancement Method (CESTEM) Utilizing Planet, Landsat and MODIS Data.” Remote Sensing of Environment 209 (May): 211–26. https://doi.org/10.1016/j.rse.2018.02.067.

Authors: We agree that these tradeoffs are important to consider. We have added a sentence at the end of the “Understand the data” rule on lines 155-158, encouraging readers to consider the spatial and temporal resolution tradeoffs between selecting data sets. We also added this citation, as it highlights radiometric quality differences and other inconsistencies across spaceborne sensor systems.

L101: For how data processing is done you can cite the following guides for UAV and Landsat imagery.

Aasen, Helge, Eija Honkavaara, Arko Lucieer, and Pablo Zarco-Tejada. 2018. “Quantitative Remote Sensing at Ultra-High Resolution with UAV Spectroscopy: A Review of Sensor

Technology, Measurement Procedures, and Data Correction Workflows." Remote Sensing 10 (7): 1091. https://doi.org/10.3390/rs10071091.

Young, N.E., Anderson, R.S., Chignell, S.M., Vorster, A.G., Lawrence, R. and Evangelista, P.H., 2017. A survival guide to Landsat preprocessing. Ecology, 98(4), pp.920-932.

Vong, A., Matos-Carvalho, J.P., Toffanin, P., Pedro, D., Azevedo, F., Moutinho, F., Garcia, N.C., Mora, A., 2021. How to Build a 2D and 3D Aerial Multispectral Map?—All Steps Deeply Explained. Remote Sens. 13, 3227. https://doi.org/10.3390/rs13163227

Authors: These were added on lines 133-135.

L115: This seems like a good spot to emphasize the time costs of high-resolution images. For example MODIS data is relatively easy to acquire in an analysis ready form and analysis can be run on most desktop computers. On the other end UAV imagery has a large time cost and requires a user to perform all processing steps from acquisition onward.

Authors: We have expanded the paragraph on lines 148-159 to address this point.

L119: Rule #3 is probably the best advice in this manuscript, and I would suggest a more direct statement such as: "Do not use high-resolution unless there is a clear need which justifies the increased cost of acquisition, processing, storing, and analysis".

Authors: We added "Use high-resolution data when there is a clear need to justify the increased cost of acquisition, processing, storing, and analysis." on lines 163-164.

L159: bibtex citation issue here for weinsten2020cross

Authors: This is now fixed

L187: this is something to consider in literally any study and also something emphasized in Betts et al 2021.

Authors: Thanks for pointing this out. We have added citations for Betts et al. 2021 and Alon 2009 on line 247.

L198: A good point to make for #6: investing in training workshops (eg. Data Carpentry) is a worthwhile investment if one is planning to use high resolution imagery, since tools that make their analysis easier generally require scientific programming skills.

Authors: We extensively modified this whole section (lines 261-305) and in doing so addressed this point.

Reviewer 2

Review: Ten simple rules for working with high resolution remote sensing data

This paper offers some useful tips such as a new graduate student entering a field that uses remote sensing data might find helpful. These are offered in the form of 10 ‘rules’ under the assumption that using high resolution remote sensing data requires researchers in Earth and Environmental science to be equipped with skills that are seldom taught. The following critiques are offered with agreement that there would be value in this paper being published, however, it is in need of significant revision.

Authors: Thank you for taking the time to review our paper!

1. While the reviewer agrees there would be value in the paper concept (Ten simple rules for working with high resolution remote sensing data), the need for such has not been given here. It is simply stated that analysing particularly high resolution remote sensing data is now common yet requires specialised skill and researchers are not taught these skills. In my experience the later is not true. Many free and quality resources exist covering a wide range of the specialised skills needed, from the many software and data carpentry courses through the equally prolific range of short courses most Environmental and Geographical Science Departments and Computing centers run focused on remote sensing data processing, to formal graduate level courses that I know exist in many related departments. Finally NASA, NOAA, the ESA, QGIS, and FOSS4G among others all offer training resources on their data/tools/formats, and Esri offers paid commercial training courses.

Authors: We appreciate the comment that the reviewer agrees with the value of our paper concept, and we acknowledge the many training resources that exist. We have added text to note existing training resources (lines #), but highlight the components that are missing from the training landscape (lines #).

2. The paper would be better positioned as a science blog aimed at new graduate students than a peer reviewed journal. The authors do not state the target audience for this paper, however, the content would be of some value to a new graduate student but they are not indicated as a target and publishing under peer review would suggest this is not the intended primary audience. The majority of experienced researchers would not find any significant value in reading this paper. Additionally, the tone of the paper is more in line with what might be expected in a science blog than a peer reviewed paper. Examples of this are throughout the text but to point to two:
- “Show your work and create open workflows to ensure that your effort is also accessible to the community. Weigh the pros and cons of innovation for your particular project. Don’t reinvent the wheel”

- “But, consider finishing your plate (answering your science question) before eating dessert.”

Authors: We appreciate this comment. We believe that this paper has potential value to anyone working with high-resolution data, from new graduate students to experienced researchers, particularly those with expertise in a different subfield but who are starting to use high-resolution data or are beginning collaborations for which a basic familiarity with high-resolution data is vital. Our paper is potentially especially useful as a broad overview of best practices for working with these data, given the gaps in training that we highlight in the paper (please see response to comment 1 above). Further, we agree that the tone was too informal. We extensively edited throughout the paper to achieve a more formal tone.

3. Multiple of the “rules” presented and elaborations given are guidance that should be covered in any introduction to research course and are not specific to remote sensing data. Examples of such include the recommendations to:

- Carry out a thorough literature review and develop a robust research plan early on.
- Clearly define a research question
- Keep in mind your research goal so as to not be diverted by interesting aspects emerging from the data during analysis, and to rather save such for future work.
- Begin with a small exploration
- Allow for/anticipate the unexpected

Authors: We appreciate this comment. We feel that it can be the case that often the allure of high resolution data can cause researchers to forget their research fundamentals, and new and exciting high-res data can lead to unique distractions and rabbit holes which we discuss. We included text to ground these broad recommendations in the specific context of high-resolution data. For example…

4. For a formal paper, **there is a lack of term definition.** This is very important generally in all academic literature, but particularly when discussing open science, data, software and reproducibility currently, as these topics are still becoming normalised in the global scientific community and therefore very open to misinterpretation. **For instance the terms “open” and “reproducible” and “high resolution” are casually used but not defined.**

Authors: We agree that clarity of language is important. We define high-resolution in the abstract (line 20) and discuss it at more length in the introduction on lines 37-50:

“Current understanding of high-resolution may include sub-meter, sub-hourly or hyper-spectral, but this is constantly changing, and what is considered high-resolution has to be considered in the context of the spatial and temporal coverage. We may even be reaching the useful limits of resolution with some products, but at limited coverage, or high-resolution in one aspect but low in others. For example, the Geostationary Operational Environmental Satellites [GOES, @schmidt_goes_2003] have sub-hourly

resolution for most of the western hemisphere, but low (1.5 km) spatial resolution. Future advances may center around increasing the resolution of all facets of a single product. For example, Landsat and Sentinel are considered moderate resolution in all facets, but with global coverage, and have been progressing towards higher resolution in all facets since the first Landsat satellite was launched in 1972. Landsat 8 has higher spatial and spectral resolution than previous Landsat products [@roy2014landsat]. Now, with the launch of Landsat 9 [@masek2020landsat], the temporal resolution is doubled. Furthermore, the Landsat products have since been harmonized with Sentinel 2 for a unified product with even higher temporal resolution [@claverie2018harmonized].“

We now define open and reproducible on lines 405-408.

“Software is open source when “the source code is available for anyone to view, use, change, and then share” (Open Source Initiative 2007). Science can be considered open and reproducible when it is conducted in such a way that scientific methods, data, and outcomes are available to everyone (Gezelter 2009).”

5. Rule 3 is “3. Don’t use high resolution data”. However, this is not justified beyond a statement that doing so can be costly. It is debatable whether or not this is actually advisable given the ready access researchers in the developed world at least have to extensive computing resources. While there are some conditions under which this would be reasonable these are not given.
Further, it might lead new researchers to use outdated lower resolution data when higher resolution but/and more up to date data is available, this in turn is likely to lead to incorrect conclusions.

Authors: We changed the name of the rule to “Use high resolution data when resolution matters” to clarify that our intention in this rule is to consider *under what circumstances* to use high resolution data. We added an additional example on lines 178-184 to indicate how a consideration of scale can inform data choices:

“The natural scale at which a climatic variable like temperature responds to atmospheric circulation is relatively coarse, and so a typical resolution for climate data may be between 800m to a full degree (cite prism, gridmet, worldclim). But the temperature that might be experienced by an individual organism can depend on extremely fine-scale variations in topography. Thus, in vegetation ecology, climate data are often downscaled with high-resolution topographic data to identify areas where climatic trends will lead to suitable microclimates for seedling survival, for example (Rodman 2019).”

Finally, we disagree that access to extensive computing resources and/or the knowledge of how to use them is a given, even in the developed world; we suggest that researchers consider the trade-offs in using high resolution data for this and other reasons.

6. While rule 6 is “6. Survey the computing landscape”. The discussion of this point is lacking. The need for a thorough literature review (which should include such) is already made earlier in rule 4 (and 2 slightly). **However, it would be useful to point a reader to the many research compute infrastructure facilities and their training materials available globally that might not appear in a domain specific focused literature review.** Furthermore, while python, Xarray and Dask are briefly mentioned in the context of explaining data size challenges, is software tooling for remote sensing data analysis is to be discussed there are many other tools that are far more relevant to a wide audience such as QGIS, the plethora of formal published and coded processing procedures and software in many disciplines, Numpy, and Esri’s tools.

Authors: We heavily revised rule 6 to address this point (lines 262-305).

“high-resolution data processing is time- and resource-intensive. Thus, before conducting an analysis, survey the software landscape to identify existing tools that can be part of an efficient, open workflow. Consider the computing environment that will be used to process the data and search for training resources that may serve as a guide through building efficient workflows, such as The Carpentries, https://earthdatascience.org, or the Pangeo community documentation. Foundational data processing and analysis tools include programmatic free and open-source tools such as Python and R, as well as graphical user interface-based tools such as the free QGIS and the proprietary ArcGIS software. The choice of which tools are used depends on the researcher’s familiarity, preference for graphical software versus coding, resources to support licenses, and the availability of add-ons specific to the analysis being conducted. For example, R may be best for statistical modeling with its many robust statistical packages while Python may be preferable for processing large arrays with the powerful Dask and xarray modules. It may be worthwhile to invest time and resources into learning a new tool that is better suited for the task rather than trying to replicate its functionality in the software language or package with which you are already familiar.

Understanding the hardware, memory, and CPU requirements will speed up the iterative process of writing code, troubleshooting bugs, and developing analyses. Understand which computing platforms meet the requirements for the analysis, whether it be in the cloud, a high performance computing cluster, or a local workstation.

Often, the data used define the software needed. For example, National Ecological Observatory Network (NEON) aerial hyperspectral imagery have 426 spectral bands spanning the visible to shortwave infrared wavelengths of the electromagnetic spectrum [@kampe2010neon]. One file may cover 7.5 km^2 and can be on the order of 2.5 GB compressed in the HDF5 (hierarchical data) format. This type of data may be too big and the HDF format too complex to open in a graphical tool such as QGIS or ArcGIS. Further, when loaded into memory as a numerical array it can require close to 26 GB of memory (e.g., a 6307x1239x426 floating point array). Many personal computers can not load the data in memory. However, the file format of the data supports both compression and

slicing operations with open source Python tools such as Xarray and Dask to scale computing tasks, allowing the data to be referenced and loaded only when computation is required, and distributing computations across multiple processors [@rocklin2015dask, @hoyer2017xarray]. These tools can enable analyses that would be challenging using graphical interface based tools otherwise.

Research whether there are existing software tools that have already been created and optimized to load and process the data. For instance, the neonHS R package enables efficient opening and processing of NEON hyperspectral imagery [@max_joseph_2021_4641288]. This process can begin with a domain-specific literature review, but does not end there. Packages that are stable, follow community software standards and are actively maintained and/or supported by rOpenSci and pyOpenSci can provide a good starting point [@boettiger2015building]. Seek tools from other disciplines that might prove useful (see Know when to innovate). For instance, the cloth simulator filter algorithm for classifying "ground" versus "not ground" in lidar or SfM photogrammetry point clouds is both accurate and efficient for this purpose, though it was originally developed for efficiently mimicking the movement of fabric in video games [@zhang2016easy].

Invest time early in a project to understand which tools will help achieve project goals."

7. The citations of Wyngaard et al 2019 and Wilkinson et al 2016 are familiar to the reviewer and both have been poorly used.
a. Wyngaard et al 2019 is used as part of the problem statement. Specifically: "The resulting bespoke approaches that applied researchers develop can be inconsistent, inefficient, and challenging to implement, reproduce, or extend (Wyngaard et al. 2019)" **Wyngaard et al 2019 is only discussing these challenges as relates to data captured with UAVs** as these problems are most prevalent with such data due to the immaturity of the tooling supporting UAV data management. These challenges do not exist to the same extent with for instance Satellite or manned aircraft data which are both very mature fields.
b. Wilkinson et al 2016 is used briefly in the discussion of Rule 10 (10. Show your work) to justify some form (**this is not discussed**) of publishing the method used. **However, Wilkinson et al 2016 is the primary source on the FAIR principles which concern the management and publication of data not methods or software.**

Authors: We removed the Wyngaard reference. We rewrote the paragraph about FAIR to be more explicit about expanding the FAIR data principles to software development. The paragraph now on lines 427-434, has been changed to the following:

"The open data principles of findability, accessibility, interoperability, and reusability (FAIR, Wilkinson et al. 2016) can be extended to software and workflows as well. These principles can be translated to a variety of specific actions such as providing open access to your original and derived data products following community created standards (e.g. Research Data Alliance, Fair Data Maturity Model Working Group 2020),

documenting and releasing software (e.g. pyOpenSci and rOpenSci), recording and reporting metadata, releasing end-to-end workflows or data pipelines, and building research compendia around publications (Gray and Marwick 2019).”

8. Finally, as indicated the reviewer agrees there would be value in a significantly refactored version of this paper. Recommendations for such would include:

- Point to and summarise existing specific remote sensing related data ethics policies under the discussion of doing no harm.

Authors: Thank you for this suggestion. Lines 388-395 now read:

”If it could do harm, consider whether to proceed and how to mitigate harm. UNICEF’s Office of Research - Innocenti has published guidelines for ethical use of geospatial technologies, many of which apply to the use of high resolution data, including de-identifying visual information, conducting a risk assessment before proceeding with data collection, and engaging with stakeholder communities before, during, and after the research (Berman 2018). The American Association for the Advancement of Science also published a set of guidelines for using location-based data, specifically during crisis situations, including detailed decision trees and risk assessment tools (Hoy 2019).”

- As suggested in various above items: point to and discuss existing known best practices regarding remote sensing tooling, infrastructure, methods, and publication of method practices.

- Discuss the complexities of publishing and managing FAIR remote sensing data specifically given that applying FAIR to remote sensing data as geospatial data is notably more complexity than other data. For instance, multiple endorsed recommendations on some of these complexities exist from the Research Data Alliance’s (RDA) work.

- The same RDA body has also published best practices on software publication which would improve the value of the point on “showing your work”. Discussing these and the many other articles and tools available for making software dependent research reproducible would also be a uniquely valuable contribution.

Authors: Thank you for this suggestion. We added more detail on existing guidelines, policies, and discussions around publishing under FAIR guidelines on lines 412-441:

“In some applications, there is tension between accessible open research and the practical reality of working with high-resolution data which may involve expensive commercial software, proprietary data, or ethical concerns (see Do no harm). For

example, Agisoft provides robust software to create 3D models from 2D imagery (e.g., from drones) using structure from motion (SfM) photogrammetry, but the software is closed source with the actual algorithms employed being hidden from the end user. For many researchers, however, commercial software may be cheaper and more accessible than developing an open source alternative (Li et al. 2016). Google Earth Engine similarly is proprietary but provides unprecedented access to many high-resolution data products that would otherwise be out of reach for many researchers. These trade-offs can also arise with data, e.g., commercial satellite imagery may be expensive but necessary for a particular study (McGlinchy et al. 2019). In these cases, reproducibility can be increased if not fully realized by approaching it modularly (Nosek et al. 2015). For instance, reproducibility can be increased by: 1) disclosing all data and steps used in a workflow, 2) reporting all algorithms (with citations) and settings used in a data pipeline, and 3) if possible, modularizing workflow so that other tools and/or data can be substituted in the future. The Transparency and Openness Promotion Guidelines provide additional steps that can be taken to "show your work" (Nosek et al. 2015).

The open data principles of findability, accessibility, interoperability, and reusability (FAIR, Wilkinson et al. 2016) can be extended to software and workflows as well. These principles can be translated to a variety of specific actions such as providing open access to your original and derived data products following community created standards (Group et al. 2020), documenting and releasing software, e.g. pyOpenSci (Trizna, Wasser, and Nicholson 2021) and rOpenSci (Boettiger et al. 2015), recording and reporting metadata, releasing end-to-end workflows or data pipelines, and building research compendia around publications (Gray and Marwick 2019).

The volume and complexity of high-resolution remote sensing data can readily lead to complicated analyses, which makes showing the work particularly challenging. For the same reasons, it is also critical to show your work in order to produce high-quality, reproducible, usable science. Publishing the code used in analysis also serves to ease the barriers of using high-resolution data."