Peer Community In Ecology

Best practices for ecological analysis are required to act on concrete challenges

Timothée Poisot *based on peer reviews by* ***Nick Isaac*** *and 2 anonymous reviewers*

Coline Royaux, Jean-Baptiste Mihoub, Marie Jossé, Dominique Pelletier, Olivier Norvez, Yves Reecht, Anne Fouilloux, Helena Rasche, Saskia Hiltemann, Bérénice Batut, Marc Eléaume, Pauline Seguineau, Guillaume Massé, Alan Amossé, Claire Bissery, Romain Lorrilliere, Alexis Martin, Yves Bas, Thimothée Virgoulay, Valentin Chambon, Elie Arnaud, Elisa Michon, Clara Urfer, Eloïse Trigodet, Marie Delannoy, Gregoire Loïs, Romain Julliard, Björn Grüning, Yvan Le Bras (2024) Guidance framework to apply best practices in ecological data analysis: Lessons learned from building Galaxy-Ecology. EcoEvoRxiv, ver. 3, peer-reviewed and recommended by Peer Community in Ecology. https://doi.org/10.32942/X2G033

Submitted: 12 April 2024, Recommended: 07 October 2024

Cite this recommendation as:

Poisot, T. (2024) Best practices for ecological analysis are required to act on concrete challenges. *Peer Community in Ecology*, 100694. 10.24072/pci.ecology.100694

Published: 07 October 2024

A core challenge facing ecologists is to work through an ever-increasing amount of data. The accelerating decline in biodiversity worldwide, mounting pressure of anthropogenic impacts, and increasing demand for actionable indicators to guide effective policy means that monitoring will only intensify, and rely on tools that can generate even more information (Gonzalez et al., 2023). How, then, do we handle this new volume and diversity of data?

This is the question Royaux et al. (2024) are tackling with their contribution. By introducing both a conceptual ("How should we think about our work?") and an operational ("Here is a tool to do our work with") framework, they establish a series of best practices for the analysis of ecological data.

It is easy to think about best practices in ecological data analysis in its most proximal form: is it good statistical practice? Is the experimental design correct? These have formed the basis of many recommendations over the years (see e.g. Popovic et al., 2024, for a recent example). But the contribution of Royaux et al. focuses on a different part of the analysis pipeline: the computer science (and software engineering) aspect of it.

As data grows in volume and complexity, the code needed to handle it follows the same trend. It is not a surprise, therefore, to see that the demand for programming skills in ecologists has doubled recently (Feng et al., 2020), prompting calls to make computational literacy a core component of undergraduate education

(Farrell & Carrey, 2018). But beyond training, an obvious way to make computational analysis ecological data more reliable and effective is to build better tools. This is precisely what Royaux et al. have achieved.

They illustrate their approach through their experience building Galaxy-Ecology, a computing environment for ecological analysis: by introducing a clear taxonomy of computing concepts (data exploration, pre-processing, analysis, representation), with a hierarchy between them (formatting, data correction, anonymization), they show that we can think about the pipeline going from data to results in a way that is more systematized, and therefore more prone to generalization.

We may buckle at the idea of yet another ontology, or yet another framework, for our work, but I am convinced that the work of Royaux et al. is precisely what our field needs. Because their levels of atomization (their term for the splitting of complex pipelines into small, single-purpose tasks) are easy to understand, and map naturally onto tasks that we already perform, it is likely to see wide adoption. Solving the big, existential challenges of monitoring and managing biodiversity at the global scale requires the adoption of good practices, and a tool like Galaxy-Ecology goes a long way towards this goal.

References:

Farrell, K.J., and Carey, C.C. (2018). Power, pitfalls, and potential for integrating computational literacy into undergraduate ecology courses. Ecol. Evol. 8, 7744-7751.
https://doi.org/10.1002/ece3.4363

Feng, X., Qiao, H., and Enquist, B. (2020). Doubling demands in programming skills call for ecoinformatics education. Frontiers in Ecology and the Environment 18, 123-124.
https://doi.org/10.1002/fee.2179

Gonzalez, A., Vihervaara, P., Balvanera, P., Bates, A.E., Bayraktarov, E., Bellingham, P.J., Bruder, A., Campbell, J., Catchen, M.D., Cavender-Bares, J., et al. (2023). A global biodiversity observing system to unite monitoring and guide action. Nat. Ecol. Evol., 1-5.
https://doi.org/10.1038/s41559-023-02171-0

Popovic, G., Mason, T.J., Drobniak, S.M., Marques, T.A., Potts, J., Joo, R., Altwegg, R., Burns, C.C.I., McCarthy, M.A., Johnston, A., et al. (2024). Four principles for improved statistical ecology. Methods Ecol. Evol. 15, 266-281.
https://doi.org/10.1111/2041-210X.14270

Coline Royaux, Jean-Baptiste Mihoub, Marie Jossé, Dominique Pelletier, Olivier Norvez, Yves Reecht, Anne Fouilloux, Helena Rasche, Saskia Hiltemann, Bérénice Batut, Marc Eléaume, Pauline Seguineau, Guillaume Massé, Alan Amossé, Claire Bissery, Romain Lorrilliere, Alexis Martin, Yves Bas, Thimothée Virgoulay, Valentin Chambon, Elie Arnaud, Elisa Michon, Clara Urfer, Eloïse Trigodet, Marie Delannoy, Gregoire Loïs, Romain Julliard, Björn Grüning, Yvan Le Bras (2024) Guidance framework to apply best practices in ecological data analysis: Lessons learned from building Galaxy-Ecology. EcoEvoRxiv, ver.3 peer-reviewed and recommended by PCI Ecology.
https://doi.org/10.32942/X2G033

Reviews

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.32942/X2G033
Version of the preprint: 2

Authors' reply, 05 September 2024

Download author's reply
Download tracked changes file

Decision by Timothée Poisot, posted 12 August 2024, validated 12 August 2024

Both reviewers are satisfied with the improvements, and are only recommending (very) minor wording / clarification changes. I will not send this preprint for review again when they are done.

Reviewed by Nick Isaac, 01 July 2024

This revised ms is much clearer and I am pleased to see that many suggestions from the first round of review have been adopted. I would be happy to see this work published: in addition to providing an advertisement for the Galaxy platform, it makes a number of really good points about computational practices in ecological research. Most of my recommended changes (below) are aimed at clarifying the message and simplifying the language:

Some of the new text appears to have been rather hastily inserted and could be improved. For example, on line 162: "propositions have been delimited within several thematic communities in ecology to evaluate and enhance best practices application, notably the SDM communities" should be simplified to "Individual subdisciplines have taken initiative to assert community best practices, e.g. in SDMs" (or similar)

Line 138: "although data accessibility": this paragraph is not just a single sentence. Please combine with previous or next para.

Line 181: replace "the expertise and research" with "the scientific process"

Line 182: replace "FAIR research data" with "the adoption of FAIR practices"

Line 244: the heading "frameworks towards best practice" is a bit confusing. Perhaps replace with "Principles for best practice"

Line 247: would benefit from a clearer definition of atomisation, perhaps with an example. Something like: "In a maximally-atomised workflow, each small step would be conducted by its own bespoke function"

Line 195-7: replace "mechanically reduces the number of potential users ... fragilizing .... " with "creates barriers to uptake and challenges for peer-review."

Line 220: replace "single" with "distinct". Also insert "each" between "steps" and "constituting".

Line 224-6: Around here it would be appropriate to have a sentence explaining how/why atomisation is part of the solution. (I realise much of this is covered in the later section from line 245, so overlap between sections should be minimised). My own recent experience of working on a colleague's code has been that atomisation makes it much easier to alter the code, to make it do something different from what the original coder intended. Altering how atomised code is used is, of course, the first step towards generalisation. However, as written the two steps appear to be quite distinct.

Ine 227: replace "this framework .... The Galaxy-Ecology" with "Atomisation and Generalisability are central organising principles in the design of the Galaxy-Ecology"

Line 229: I don't think that "sharing and processing data" does justice to what Galaxy is aiming for. How about "analysing data and sharing outputs".

Line 284: again, an example would help to illustrate this point more clearly: "This means trying to avoid hard-coding anything that is specific to the structure of the original dataset (e.g. number of years)."

Line 290 the word "step" appears twice in this sentence. Replace the second with "element" or similar word.

Line 301: the section title is quite clunky: how about "Practical steps towards atomised and generalised coding".

Line 319: the figure legend needs more detail here. Explain that the different colours refer to different scripts/categories (1-3) and that the boxes refer to functions/scripts.

Line 321: re “code-writing habits”: I think it would be useful to make some kind of comment about how ecologists learn to code. Most of them learn by analysing their own data, and/or from a statistical ecology module that focusses on specific applications, in which the statistics and the ecology are given primacy but the computation. Few are taught formal programming skills in the way that computer science students would be. This means that most of us have generated a large number of “bad habits”!

Line 345: can you explain why someone should feel less embarrassed or fearful out sharing code if it has been atomised and generalised? Is it because they will feel confident about having followed best practice (I doubt it) or because they will feel confident that someone else will be able to actually read and implement it (more likely)?

Table 2 legend: would be more informative to write as “illustration of how Galaxy implements and confirms to best practice”. The second sentence about “limitations …” makes no sense to me. Perhaps it is a reference to the fact that Galaxy is a work-in-progress that the table is a snapshot of current status.

Line 444: can you elaborate on the figure 3 legend to relate it back to the concepts of atomisation and generalisability? In particular, it is appropriate to describe the named items with checkboxes as atoms in the workflow?

Line 459: this is very unclear. Does “eventually” mean “when the user becomes expert” or “there is an aspiration for Galaxy to have this new functionality”. If the former then perhaps replace with something like “In addition to using existing tools, users may develop and upload entirely new tools to the Galaxy server”.

Line 462: “utterly” is superfluous

Line 464: “notably” is superfluous

Line 515: there is perhaps another level in this hierarchy: I have authored papers that were fully executable at the time of publication. However, we did not user Docker or other tools to account for changes in the underlying software, so the code no longer works and the work is therefore not reproducible. Distinguishing between “reproducible now vs reproducible forever” might be helpful.

Line 531: “additionally”

Line 550: replace “correctly” with “appropriately”

Line 552: “heavier” is not clear. I get that the “heaviness” refers to the amount of time investment required to realise the advantages of using Galaxy (but this could be clearer – please use a different word). What I don’t understand is the comparator: is it heavier for experienced than non-experienced, heavier to learn Galaxy than the principles of atimosation and generalisation, and are you referring to absolute or relative terms (i.e. cost vs cost:benefit ratio).

Reviewed by anonymous reviewer 2, 09 August 2024

This draft is greatly improved in structure from the previous version and I found it significantly easier to follow. The presentation of reproducibility as the goal and the Galaxy workflow as a solution is strong and clear, and generally the message is more concise. The section describing Galaxy-E and the discussion however have some issues with redundancy of some topics, and the absence of other topics, that make them harder to follow.

The “Entering a new dimension” section lists many of the ways Galaxy meets the criteria for reproducibility, or the benefits it might bring, but it doesn’t describe what it *is*. That piece is critical for this type of introductory paper – where does it live? How does the user engage? What are the key pieces? This part might include existing parts of the section such as the description of the community, or how a user uploads data, but should be more comprehensive and systematic. I think in the previous version many of those ideas lived in the

"methods" section, which has rightly been moved to a different venue, but a paragraph or two of description is still necessary. Starting with a clear description also gives grounding for the platform's benefits, as the reader already has clear evidence of how the tool might achieve those things. One possible structure for the "Entering a new dimension" section would be:

- What Galaxy-E is (how users engage with it)
- How Galaxy-E follows the atomisation/generalisation and reproducibility/fairness philosophy
- Other benefits
- Examples of its success

I think improving the structure will also reduce some of the redundancy of ideas and language throughout the section. For example, the paragraph starting at line 403 is a repetition of the general philosophy of Galaxy-E that echoes a similar sentiment given many times throughout the paper, I don't think It's necessary here.

Similarly, the discussion repeats many of the basic concepts of reproducibility or atomisation/generalisation without linking them back to the platform. The first two paragraphs in particular could either be cut, or should be edited to be directly relevant to the platform and its strengths. The discussion could also benefit from a more detailed comparison to other existing platforms.

Detailed feedback:

Line 227-243: This description feels out of place here, as it's followed by a more detailed description of the atomisation/generalisation framework rather than the Galaxy-E approach, I'd move it to the "Entering a new dimension" section

Line 342: How or why might that be true?

Line 346–360: This feels redundant, it could be integrated more concisely into the earlier paragraphs.

Line 376: This feels redundant after the previous paragraph

Line 501: I don't think this bullet list is appropriate for the discussion, maybe a summarized version in the reproducibility section, but as it stands it's relevance to the Galaxy-E tool is not clear.

Line 527-535: These things are all true, but what are their relevance for Galaxy-E? How does it help a user achieve this?

Line 530: I don't think arborescence is the right word here, are you trying to describe the relationships of analysis pieces to one another?

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.32942/X2G033
Version of the preprint: 1

Authors' reply, 20 June 2024

Download author's reply
Download tracked changes file

Decision by Timothée Poisot, posted 07 June 2024, validated 07 June 2024

Revision needed to improve readability

Both reviewers see the value of the work (and I agree), but make important comments that would improve the readability of the manuscript. I think the separation between philosophy, implementation/architecture, and anticipated use-cases needs to be clarified. This will help readers navigate a very dense manuscript, especially readers with different types of background/expertise.

Reviewed by anonymous reviewer 1, 23 May 2024

This is a paper in three parts. The first part discusses issues around analytical practices in ecology, and the principles by which these can become more reproducible. The second half is a description of the Galaxy Ecology platform, and it's potential to realise better practice among the community. The third part described the technical details of implementation.

The visual abstract is really great, presenting a clear, cohesive message about how atomisation and generalisation can improve reproducibility and FAIR principles. However, the main text does not reflect this clarity of message.

Overall there is a lot of good material in here, but the article does not feel mature in how it is structured and presented. I do not get a clear idea who the article is aimed at, nor who is the target user of Galaxy Ecology. It's hard to see how the present version would lead to material change in how ecologists go about their work. Another challenge is that much of the material about Galaxy Ecology is likely to become quickly outdated. For these reasons, my overall suggestion would be to greatly reduce the amount of text in the section on Galaxy and in the Methods. Rewrite these sections in a way that uses Galaxy to illustrate the general principles in the first part of the manuscript.

The paper needs to make more comparisons to current workflow platforms. The examples provided are not current: Taverna retired in 2020 and the latest version of Kepler (version 2.5) was 2015. For example, ecologically specific tools such as Bon-in-a-box https://boninabox.geobon.org/ or more general tools such as the R package Targets: https://books.ropensci.org/targets/ .

Additionally, the methods section could be better as online supplementary materials (or zenodo repository) of simply part of the galaxy user guide. It is not relevant to the core message of the manuscript. As the galaxy-E platform develops, the user guide may change. Hosting this content online rather than in the manuscript allows that updating to ensure parity with the Galaxy-E platform.

The introduction needs to be more concise, a lot of text is used to set up the wider concerns about reproducibility in ecology, which is important, but the main contribution is not to review ecological reproducibility to provide the solution and at current it takes too long to get to the solution. Large parts of the text are not necessary for the delivery of the core message eg. lines 106-131, 151-185, 211-229.In the detailed comments below, I've made a few suggestions for how the Introductory sections could help to contextualise the issues being discussed. For the section on Galaxy, I think the authors need to give a much clearer exposition of what it is, who it is for and how it can help to deliver the principles outlined in the first section. At present, this text assumes too much knowledge of the system for a naïve reader to properly engage. The Methods section is even more difficult to follow: it's half-way between a user-manual and a conceptual overview but doesn't quite deliver to either of those goals. Perhaps it would help if these issues were illustrated via the use of one or more case studies.

Detailed comments:

· Abstract Line 75: perhaps add a few words to explain that the principles described here are applicable across all levels of ecological analysis, e.g. "from individual research projects to production-level analytical pipelines"

· Abstract line 77: Perhaps explain why "atomisation" is the right word. I thought this was a typo and that the authors meant to write "automation" instead. The authors need to provide rationale behind using the term 'atomisation' to describe the process of making a script more modular, there is no previous usage of this term in the software development literature. Whereas modular programming is a widely used and understood term (https://en.wikipedia.org/wiki/Modular_programming). If the authors want to use the term 'atomisation' it must be made clear it is a new, non-standard term introduced in this manuscript that may not be widely understood across other sectors.

· Introduction paragraph 2: a useful citation here would be Cassey & Blackburn, who distinguished between "reproducibility" and "repeatability", and discussed the relative merits of each. Also, with reference to later discussion of reproducibility around lines 205, it would be good to acknowledge that reproducibility is not

an absolute but rather a relative concept (who needs the results to be identical to the 9th significant figure?).

· Lines 117-119: complicated sentence. I suggest to simplify to "Given the increasing complexity of ecological analyses, there is a clear need for tools that facilitate greater reproducibility."

· The same paragraph (lines 119-130, but also the next section, starting at line 140) would really benefit from a big more context for what is the problem and why atomised workflows are needed. My perspective is that ecology has, until now, been a discipline in which most analysis happened on a single computer, but increasingly we are seeing papers derived from big collaborations involving code developed in different labs. This means we are moving into an era where analytical pipelines are becoming so complex that no individual researcher can understand all the details at a granular level. Other disciplines (e.g. meteorology, particle physics) have already passed through this phase. There is plenty of literature that could help illustrate these issues and clarify why this paper is novel. First, it would be good to include a citation to support the assertion that analyses are becoming more complex. One option would be to find some data on the average length of supplementary information on ecological papers (I know of no such data!). An alternative would be to cite papers that describe highly complex workflows, e.g. Boyd et al (2023) or Jetz et al (2019). Second, it might be worth acknowledging that individual branches of ecology have developed principles to enhance reproducibility within those sub-domains. The SDM community is perhaps the best example: citing papers such as Araujo et al (2019) and Golding et al (2017) might provide a way for the authors to explain what is different about the proposals in this paper.

· Line 156: try to avoid directly quoting from another article

· line 170: I would question whether long-term public archiving of code is as valuable as the authors assert. The most popular coding language among ecologists, R, is in a continuous state of evolution. Most R code written 10 years ago would not execute today. I'm not saying that we should not archive code: I just think it is important to be clear about what we are trying to achieve as a community and make decision about where to invest resources accordingly.

· Line 195: in the previous paragraph you made the case that code should be considered as data. So, for clarity, insert the word "observational" before "data" in this sentence.

· Line 298: "Atomisation refers to dividing …"

· Line 298-366: this text on atomisation and generalisation is absolutely fine, but I can't help thinking that these must be fundamental principle of computer science. If so, it may be worth mentioning here, perhaps with a citation.

· Line 369: "et" -> "at"

· Line 386: missing word "A" at beginning of sentence

· Line 412: at the beginning of this section, it would be useful to explain who is the target user. Are you recommending that everyone in ecology use it for all of their analysis? Or is it better suited to large collaborative projects?

· Line 435: "tools" -> "tool"

REFERENCES

Araujo et al (2019) Standards for distribution models in biodiversity assessments. Science Advances https://www.science.org/doi/full/10.1126/sciadv.aat4858

Boyd et al (2023) Biological Reviews https://onlinelibrary.wiley.com/doi/full/10.1111/brv.12961

Cassey & Blackburn (2006) Reproducibility and Repeatability in Ecology. BioScicence https://academic.oup.com/bioscience/article/56/12/958/221622

Golding et al (2017) the Zoon package … Methods in ecology & Evolution. http://doi.wiley.com/10.1111/2041-210X.12858

Jetz et al (2019) Nature Ecology & Evolution https://doi.org/10.1038/s41559-019-0826-1

Reviewed by anonymous reviewer 2, 09 May 2024

This manuscript introduces and describes the Galaxy-Ecology tool, laying out different modes of engagement and the ways the approach addresses reproducibility issues in ecology. The Galaxy-Ecology project and associated community seem like a powerful framework to build and share ecological analyses and this paper includes all the essential introductory pieces for a new user. I particularly appreciate the way the paper discusses different types of users and the value it adds for each. I do however think there is a fair bit of related but ultimately superfluous content that, if paired down, would greatly improve the readability and clarity of the paper.

While the initial discussion of reproducibility is clearly motivating for development of the tool, the current level of depth is unnecessary and even a little misleading for the reader. For example, Galaxy-Ecology isn't mentioned until the fourth section of the introduction and is introduced in a way that makes the reader unsure if it's just a nice example to illustrate the reproducibility point or the main message of the paper in and of itself. Given that a history of reproducibility in ecology is not the goal of the paper, I would recommend editing everything prior to "Framework towards good practices" down to a few introductory paragraphs that also immediately introduce Galaxy-Ecology as a solution. If I have misunderstood the purpose of the paper and the goal of the paper is to first give a detailed context for the reproducibility crisis and second discuss Galaxy-Ecology, that structure can be better set up in the abstract and first couple paragraphs of the paper.

In general I found the structure of the paper hard to follow, as technical details about engaging with the tool or development are interspersed with motivation and philosophy. One possible approach for addressing that confusion is to lay out the technical details in an initial description of the tool, then describe which of those pieces different kinds of users might engage with, rather than introducing new information in the user sections. In a related formatting issue, currently the three guidelines sections read as multi-paragraph lists which I found fairly jarring. Rather than referencing different steps of the workflow as a starting bullet point, I would reference steps in the body of the paragraph (for example as "(step A)") to aid flow of the writing.

As a small linguistic note, I would suggest the phrase "good practices" be replaced with "best practices", which is a more common way to reference standards and will be immediately recognizable to a reader.

A few line comments:

Line 461: Unnecessary reference, either remove sentence or integrate into a paragraph.

Line 560: This heading is unnecessary.

Line 632: The colon here is confusing to me, I can see from the workflow what you're implying but the list within a list is quite hard to follow in the writing. This would be much easier to express in a paragraph rather than list format.

Line 644: This is a good example of the kind of concept that should be introduced outside the user descriptions. Is the "Galaxy history" an internal versioning system? At what level is it tracking, just the kinds of modules that are used in the pipeline?

Title and abstract

Does the title clearly reflect the content of the article? [x] Yes, [ ] No (please explain), [ ] I don't know

Does the abstract present the main findings of the study? [x] Yes, [ ] No (please explain), [ ] I don't know

Introduction

Are the research questions/hypotheses/predictions clearly presented? [ ] Yes, [x] No (please explain), [ ] I don't know

- As discussed above, it is difficult for the reader to initially figure out that Galaxy-Ecology is the focus of the paper.

Does the introduction build on relevant research in the field? [x] Yes, [ ] No (please explain), [ ] I don't know

Materials and methods

Are the methods and analyses sufficiently detailed to allow replication by other researchers? [x] Yes, [ ] No (please explain), [ ] I don't know

Are the methods and statistical analyses appropriate and well described? [ ] Yes, [x] No (please explain), [ ] I don't know

The paper is missing a detailed overview of the moving pieces of the tool.

Results

In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes, [ ] No (please explain), [ ] I don't know N/A

Are the results described and interpreted correctly? [x] Yes, [ ] No (please explain), [ ] I don't know

Discussion

Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [ ] Yes, [x] No (please explain), [ ] I don't know

There could be a little more comparison to other similar efforts to improve reproducibility and the limitations of what Galaxy-Ecology does.

Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [x] Yes, [ ] No (please explain), [ ] I don't know