Uncertainty at scale: how CS hinders climate research

Patrick Ferris, Michael Dales, Tom Swinfield, Sadiq Jaffer, Srinivasan Keshav, Anil Madhavapeddy
Departments of Computer Science & Technology and Zoology, University of Cambridge, UK

Computer science is a powerful tool for enabling data-driven advances in global ecology and conservation. However the amplification cuts two ways, as mechanisation can also compound problems inherent with just how uncertain [13] anything to do with natural ecosystems are! Species habitat datasets are uncertain, local observations are uncertain, the resulting inferences about species distributions are uncertain, side-effects from interventions are uncertain; conservation action has evolved to take this into account [17]. Computer science when applied without consideration of these factors can amplify the uncertainty by running ever-larger datasets through increasingly complex data pipelines and algorithms, all built upon wobbly foundations. What exact version of the dataset *is* being used? What exact version of the dataset *did* you use? What assumptions went into generating that dataset? What libraries, system dependencies and environment variables were used to calculate the results?

In this talk, we first segment sources of uncertainty across ecological data sources (§1) and computation over them (§2), and then reflect how these uncertainties impact ecological research and how we might cleanly bound the uncertainty for future conservation research (§3).

1 UNCERTAINTY IN DATA

As early as 2013, ecologists have been calling on their colleagues "...to treat data as an enduring product of research, not just a precursor to publication" [12]. This narrative has played out in computer science with version control systems (VCS) for software management, or for machine learning with *Weights and Biases* for AI experiment tracking and *HuggingFace* for AI model versioning [2, 3]. However, pivotal datasets within ecology are versioned ambiguously, leading to differing end results from seemingly identical inputs.

One example is the Tropical Moist Forest (TMF) dataset by the European Commission's Joint Research Centre [18], which calculates forest cover worldwide from satellite observations. TMF historic data is upgraded as new algorithms and analysis become available (which is good practise), but the means by which the data is published does not make these updates to historically available data obvious. Once updated, earlier versions on which other calculations were made are no longer easily available, hugely impacting research reproducibility [7]. Table 1 shows some differences partially caused by "Improvements and corrections of errors in the Annual Change collection in the sequence of values for deforestation of old regrowth forest..." [4]. Whilst they seem small, these proportions represent some 6.7 million squared kilometres area of the Amazon rainforest basin[1]! Without access to the original 2021 JRC release, these differences propagate silently through further downstream research.

Land Use Class Proportions (%) in 2020

	JRC 2021	JRC 2022	Difference
Undisturbed	74.83	74.71	-0.12
Degraded	5.07	5.16	0.09
Deforested	7.49	7.39	-0.10
Regrowth	0.83	1.74	0.91
Water	2.11	1.93	-0.18
Other	9.67	9.07	-0.60

Table 1: The proportions of difference Land Use Classes (LUC) in the Amazon basin as calculated from two JRC datasets for the same year.

2 UNCERTAINTY IN CODE

Once ingested into a data pipeline, the data's provenance is usually lost, as untracked inputs are combined into derived datasets whose origins only the operator may remember, as operating systems do no automatically track such things. Ecological data are often transformed using programming languages like Python and R, but the ecological community has yet to adopt a robust culture of openly publishing their code alongside their methodologies. One reason for this, that Mislan et al. recognise, is that ecologists "...may not be aware of the steps needed to archive code..." [16]. The tools for versioning and archiving code are not easily accessible for ecologists, and whilst the number of ecological journals requiring or encouraging code to be shared has increased, most authors still do not adhere to these requirements [8].

Even when care is taken, without reproducible artefacts (e.g., using a system like Nix [9]), the sheer multiplicity of factors that can change the final results is unmanageable within conventional operating systems. For example, consider the popular geospatial data manipulation library GDAL [1]. The image in figure 1 is the difference between data derived with *the same* command-line but with GDAL 3.2 and GDAL 3.3.

The low availability of source code only exacerbates the reproducibility story when combined with the data versioning issues (§1). Whilst some computer science research has offered accessible versioning of large datasets, like `DataHub`, it is not habitually used within ecological research [6].

[1]Source code: https://github.com/carboncredits/jrc-diff

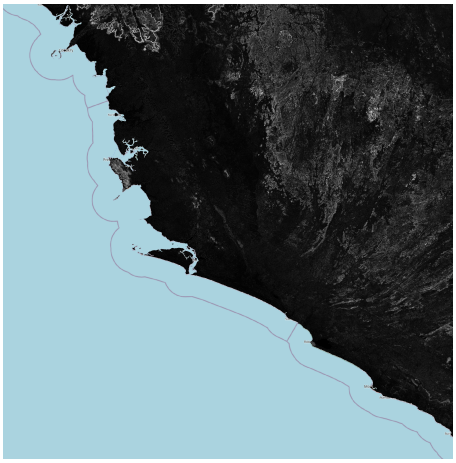

Figure 1: A map showing the difference in terrain ruggedness index (TRI) as calculated by the `gdaldem tri` command between GDAL versions 3.2 and 3.3. The difference would ideally be zero (all black) but includes pixels as different as 372 metres.

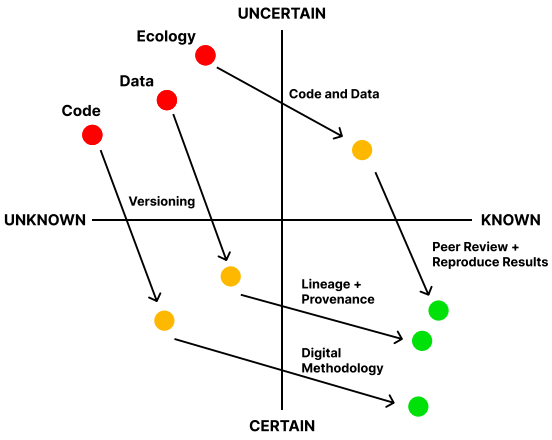

Figure 2: Shifting unknown code and data certainties in order to allow computation ecological research to be meaningfully peer-reviewed and reproduced.

One platform has cornered the remote sensing, ecological research market and that is Google Earth Engine (GEE) [11]. GEE has enabled many ecologists to produce useful analysis and datasets that may have been difficult to achieve otherwise. For example, the global map of travel time to healthcare facilities which can be used as a proxy for "remoteness" in ecological analysis [19]. However, the longevity and endurance of the platform is tied to the whims of a large corporation that has a record of closing down non-core products [10]. Also, many datasets in GEE have been processed to make them easier to consume, but the methodology by which they've been processed is not published, preventing easy migration away from GEE once datasets are adapted.

3 UNCERTAINTY IN ECOLOGY

The lack of software engineering skills and tools impacts the rigour and quality of ecological research. The analysis can explode in both space and time: even a relatively simple species area of habitat analysis requires hundreds of gigabytes of raster data to be analysed per species, with typically many thousands of species being processed [14], potentially requiring days of compute on high-end hardware.

Unoptimised code limits the analysis that can be performed due to the time or resources it would require. This is important for sensitivity analysis for methodologies with a large number of inputs and possible configurations. In the context of Merton's notion of "specified ignorance" (the knowledge that is not yet known but ought to be) these issues due to computer science should be flagged as sources of ignorance and therefore fruitful research directions for others [15]. This idea is explored in terms of known and unknown certainties in Figure 2. Code and data are bounded via versioning, provenance and lineage tracking and shifting to digital methodologies. This is not dissimilar to the use of *continuous integration* systems for managing production-level code. Only once code and data are bounded can we hope to return to applying the peer review process to reproduce results and have known, certain ecological research.

The computer science aspects must be dealt with first. In recent years, we have seen the rise of the *Research Software Engineer* [20]; the primary role of the research software engineer is to provide software skills to another discipline. Computer science and software can be thought of as a new instrument for experimental scientists. In their paper on the research software engineer, Baxter et al. argue a key difference between more traditional scientific methods and the newer computational ones are that "...no-one with the same casual attitude to experimental instrumentation as many researchers have to code would be allowed anywhere near a lab" [5]. They later cite one reason for the short supply of research-focused software engineers is the lack of "institutional homes and career progression paths for their work" which could explain why software is treated differently to more traditional instruments used in experiments.

The scientific method in the computational age has become vastly more complicated. Computer science has thus far failed to deliver the tools to scientists to easily bound this complexity and reduce uncertainty in their analysis. Uncertainty within methodologies is difficult to correct. However, ensuring that the computing foundations upon which ecologists build their work minimises and tracks uncertainty is achievable, and this must be a priority for our field.

REFERENCES

[1] GDAL. https://gdal.org/.

[2] Experiment tracking with weights and biases. https://www.wandb.com/, 2020. Software available from wandb.com.

[3] Huggingface: We are on a mission to democratize good machine learning, one commit at a time. https://huggingface.co, 2023.

[4] List of updates integrated in the TMF 2022 products. https://forobs.jrc.ec.europa.eu/TMF/data#update, 2023.

[5] Baxter, R., Chue Hong, N., Gorissen, D., Hetherington, J., and Todorov, I. The research software engineer. Digital Research 2012 ; Conference date: 10-09-2012 Through 12-09-2012.

[6] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A. J., Madden, S., and Parameswaran, A. G. Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798* (2014).

[7] Clark, B., Deshane, T., Dow, E., Evanchik, S., Finlayson, M., Herne, J., and Matthews, J. Xen and the art of repeated research. In *Proceedings of the FREENIX* (2004), pp. 135–144.

[8] Culina, A., van den Berg, I., Evans, S., and Sánchez-Tójar, A. Low availability of code in ecology: A call for urgent action. *PLOS Biology 18*, 7 (07 2020), 1–9.

[9] Dolstra, E., and de Jonge, M. Nix: A safe and Policy-Free system for software deployment. In *18th Large Installation System Administration Conference (LISA 04)* (Atlanta, GA, Nov. 2004), USENIX Association.

[10] Gerken, T. Gamers say goodbye to google's stadia as platform shuts, 2023. Accessed on October 23, 2023.

[11] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. re-mote sensing of environment, 202, 18–27, 2017.

[12] Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H. Big data and the future of ecology. *Frontiers in Ecology and the Environment 11*, 3 (2013), 156–162.

[13] Lewandowsky, S., Ballard, T., and Pancost, R. D. Uncertainty as knowledge. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 373*, 2055 (2015), 20140462.

[14] Lumbierres, M., Dahal, P., Soria, C., Di Marco, M., Butchart, S., Donald, P., and Rondinini, C. Area of habitat maps for the world's terrestrial birds and mammals. *Sci Data* (2022).

[15] Merton, R. K. Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology 13*, 1 (1987), 1–29.

[16] Mislan, K., Heer, J. M., and White, E. P. Elevating the status of code in ecology. *Trends in Ecology & Evolution 31*, 1 (2016), 4–7.

[17] Sutherland, W. J., Ed. *Transforming Conservation: A Practical Guide to Evidence and Decision Making*. Open Book Publishers, Dec. 2022.

[18] Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragão, L. E. O. C., and Nasi, R. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. *Science Advances 7*, 10 (2021), eabe1603.

[19] Weiss, D. J., Nelson, A., Vargas-Ruiz, C. A., Gligorić, K., Bavadekar, S., Gabrilovich, E., Bertozzi-Villa, A., Rozier, J., Gibson, H. S., Shekel, T., Kamath, C., Lieber, A., Schulman, K., Shao, Y., Qarkaxhija, V., Nandi, A. K., Keddie, S. H., Rumisha, S., Amratia, P., Arambepola, R., Chestnutt, E. G., Millar, J. J., Symons, T. L., Cameron, E., Battle, K. E., Bhatt, S., and Gething, P. W. Global maps of travel time to healthcare facilities. *Nature Medicine 26*, 12 (Dec. 2020), 1835–1838.

[20] Woolston, C. Why science needs more research software engineers, 2022.