Data-based, synthesis-driven: setting the agenda for computational ecology

Timothée Poisot *1,2,3* Richard LaBrie *1,2* Erin Larson *4* Anastasia Rahlin *5*

1: Université de Montréal, Département de Sciences Biologiques; **2**: Groupe de Recherche Interuniversitaire en Limnologie et environnement aquatique; **3**: Québec Centre for Biodiversity Sciences; **4**: Department of Ecology and Evolutionary Biology, Cornell University; **5**: Illinois Natural History Survey

Computational ecology, defined as the application of computational thinking to ecological problems, has the potential to transform the way ecologists think about the integration of data and models. As the practice is gaining prominence as a way to conduct ecological research, it is important to reflect on what its agenda could be, and how it fits within the broader landscape. In this contribution, we suggest areas in which empirical ecologists, modellers, and the emerging community of computational ecologists could engage in a constructive dialogue to build on one another expertise; specifically, about the need to make predictions from models actionable, about the best standards to represent ecological data, and about the proper ways to credit data collection and data reuse. We discuss how training can be amended to improve computational literacy.

Keywords: computational ecology - ecological synthesis - data sharing

Correspondence to Timothée Poisot – `timothee.poisot@umontreal.ca`

`2017-03-02T00:00:00.000Z`

Computational science happens when algorithms, software, data management practices, and advanced research computing are put in interaction with the explicit goal of solving "complex" problems. Typically, problems are considered complex when they cannot be solved appropriately with mathematical modelling (*i.e.* the application of mathematical models that are not explicitely grounded into empirical data) or data-collection only. Computational science is one of the ways to practice computational thinking (Papert 1996), *i.e.* the feedback loop of abstracting a problem to its core mechanisms, expressing a solution in a way that can be automated, and using interactions between simulations and data to refine the original problem or suggest new knowledge. Computational approaches are commonplace in most areas of biology, to the point where one would almost be confident that they represent a viable career path (Bourne 2011). Data usually collected in ecological studies have high variability, and are time-consuming, costly, and demanding to collect. In parallel, many problems lack appropriate formal mathematical formulations, which we need in order to construct strong, testable hypotheses. For these reasons, computational approaches hold great possibilities, notably to further ecological synthesis and help decision-making (Petrovskii & Petrovskaya 2012).

Levin (2012) suggested that ecology (and evolutionary biology) should continue their move towards a *marriage of theory and data*. In addition to the lack of adequately expressed models, this effort is hampered by the fact that data and models are often developed by different groups of scientists, and reconciling both can be difficult. This has been suggested as one of the reasons for which theoretical papers (defined as *papers with at least one equation in the main text*) experience a sharp deficit in numbers of citations (Fawcett & Higginson 2012); this is the tragic sign that empirical scientists do not see the value of theoretical work, which of course can be blamed on both parties. One of the leading textbooks for the mathematical models in ecology and evolution (Otto & Day 2007) is more focused with algebra and calculus, and not with the integration of models with data. Other manuals that cover the integration of models and data tend to lean more towards statistical models (Bolker 2008; Soetaert & Herman 2008). This paints a picture of ecology as a field in which dynamical models and empirical data do not interact much, and instead the literature develops in silos.

What is computational ecology? It is the application of computational thinking to ecological problems. This defines three core characteristics of computational ecology. First, it recognizes ecological systems as complex and adaptive; this places a great emphasis on mathematical tools that can handle, or even require, a certain degree of stochasticity (Zhang 2010, 2012). Second, it understands that data are the final arbiter of any simulation or model (Petrovskii & Petrovskaya 2012); this favours the use of data-driven approaches and analyses (Beaumont 2010). On this point, computational approaches differ greatly from modelling, that can often function on their own. Finally, it accepts that some ecological systems are too complex to be formulated in mathematical or programmatic terms (Pascual 2005); the use of conceptual, or "toy" models, as long as they can be confronted to empirical data, is preferable to "abusing" mathematics by describing the wrong mechanism well (May 2004). By contrast, modelling approaches are by construction limited to problems that can be expressed in mathematical terms.

Ecology as a whole (and community ecology in particular) circumvented the problem of model and data mismatch by investing in the development and refinement of statistical models (see Warton et al. 2014 for an excellent overview) and "numerical" approaches (Legendre & Legendre 1998) based on multivariate statistics. These models are able to *explain* data, but very rarely do they give rise to new predictions – despite it being a very clear priority even if we "simply" seek to further our understanding Houlahan et al. (2017). Computational ecology can fill this niche; at the cost of a higher degree of abstraction, its integration of data and generative models (*i.e.* models that, given rules, will generate new data) can be helpful to initiate the investigation of questions that have not received (or perhaps cannot receive) extensive empirical treatment, or for which usual statistical approaches fall short.

In a thought-provoking essay, Markowetz (2017) suggests that *all biology is computational biology* – the rationale behind this bold statement being that integrating computational advances, novel mathematical tools, and the usual data from one field, has a high potential to deliver synthesis. A more reasonable statement would be that *all ecology can benefit from computational ecology*, as long as we can understand how it interacts with other approaches; in this paper, we attempt to situate the practice of computational ecology within the broader landscape of ecological research. In particular, we highlight the ways in which computational ecology differs from, and complements, ecological modelling. We finally move on to the currency of collaborations between different sub-disciplines of ecologists, and discuss the need to add more quantitative skills in ecological training.

1 A success story: Species Distribution Models

The practice known as "species distributions modelling" (and the species distribution models, henceforth SDMs, it generates) is a good example of computational practices generating novel ecological insights. At their core, SDMs seek to model the presence or absence of a species based on previous observations of its presence or absences, and knowledge of the environment in which the observation was made. More formally, SDMs can be interpreted as having the form $\mathrm{P}(S|E)$ (or $\mathrm{P}(S|E=1)$ for presence-only models), where S denotes the presence of a species, and E is an array of variables representing the local state of the environment at the point where the prediction is made (the location is represented, not by its spatial positions, but by a suite of environmental variables).

As Franklin (2010a) highlights, SDMs emerged at a time where access to computers *and* the ability to effectively program them became easier. Although ecological insights, statistical methods, and data already existed, the ability to turn these ingredients into something predictive required what is now called "computational literacy" – the ability to abstract, and automate, a system in order to generate predictions through computer simulations and their validation. One of the strengths of SDMs is that they can be used either for predictions or explanations of where a given species occur (Elith & Leathwick 2009) and can be corroborated with empirical data. To calculate $\mathrm{P}(S|E)$ is to make a prediction (what are the chances of observing species S at a given location), that can be refined, validated, or rejected based on cross-validation (Hijmans 2012) or *de novo* field samplig (West et al. 2016). To understand E, *i.e.* the environmental aspects that determine species presence, is to form an explanation of a distribution that relates to the natural history of a species.

SDMs originated as statistical and correlative models, and are now incorporating more ecological theory (Austin 2002) – being able to integrate (abstract) ideas and knowledge with (formal) statistical and numerical tools is a key feature of computational thinking. In fact, one of the most recent and most stimulating developments in the field of SDMs is to refine their predictions not through the addition of more data, but through the addition of more processes (Franklin 2010b). These SDMs rely on the usual statistical models, but also on dynamical models (for example simulations; *e.g.* Wisz et al. (2012) or Pellissier et al. (2013) for biotic interactions, and Miller & Holloway (2015) for movement and dispersal). What they lack in mathematical expressiveness (*i.e.* having a closed-form solution (Borwein & Crandall 2013), which is often ruled out by the use of stochastic models or agent-based simulations), they assume to gain in predictive ability through the explicit consideration of more realistic ecological mechanisms (D'Amen et al. 2017; Staniczenko et al. 2017). SDMs have been a success, but there are many other areas of ecology that could be improved by a marriage of computational ecology and empirical data.

2 Computational ecology in its broader landscape

2.1 The four quadrats of ecological research

In {fig. 1}, we propose a rough outline of four quadrats for ecological research. The horizontal axis is based on the degree of integration between data and models, ranging from disconnected (for purely data-based or model-based) to highly integrated. The vertical axis is based on the ability to *document* natural processes and their underlying mechanisms (through direct or indirect observation of natural systems) rather than *suggest* (through focus on a reduced number of mechanisms and their interactions). A classification this coarse is bound to be caricatural, but it serves as an illustration of where computational ecology exists in the overall research methodology. Because computational ecology relies on the integration of data (if possible *raw* data from observational and manipulative experiments) and models (either statistical or phenomenological), it can *suggest* general trends through an abstraction of the idiosyncracies of a particular system.

[Figure 1 about here.]

The specific example of predator-prey interactions should be a familiar illustration of how the same problem can be addressed in different ways. The classical prey–predator equations of Lotka & Volterra are an instance of a "modelling" based perspective, wherein mathematical analysis reveals how selected parameters (rates of interactions and growth) affect an ecologically relevant quantity (population stability and coexistence). These models, although they have been formulated to explain data generated through empirical observations, are disconnected from the data themselves. In fact, this family of model lies at the basis of a branch of ecological modelling that now exists entirely outside of data (Ackland & Gallagher 2004; Gyllenberg et al. 2006; Coville & Frederic 2013). These purely mathematical models are often used to describe trends in time series. But not all of them hold up to scrutiny when explicitely compared to empirical data. Gilpin (1973) famously reports that based on the predictions of the Lotka-Volterra model, hares in the Hudson bay are feeding on Lynx.

By contrast Sallan et al. (2011) study the same issue (sustained persistence and fluctuations of predator–prey couples through time) using a paleo-ecological timeseries, and interpret their data in the context of predictions from the Lotka-Volterra family of models (namely, they find support for Lotka-Volterra-like oscillations in time). Although dynamical models and empirical data interact in this example, they do not do so directly ; that is, the analysis of empirical data is done within the context of a broad family of model, but not coupled to *e.g.* additional simulations. A number of other models have been shown to generate predictions that quantitatively match empirical data (Nicholson & Bailey 1935; Beverton & Holt 1957) – this represents, in our opinion, the sole test of whether a mathematical model is adapted to a particular problem and system. While models are undeniably useful to make mechanisms interact in a low-complexity setting, it is in our opinion a grave mistake to assume they will, in and of themselves, be relevant to empirical systems.

Meta-analyses, such as the one by Bolnick & Preisser (2005), are instead interested in collecting the outcome of observational and manipulative studies, and synthetizing the *effects* they report. These are often purely *statistical*, in that they aggregate significance, effect size, to measure how robust a result is across different systems. Meta-analyses most often require a *critical mass* of pre-existing papers (Lortie et al. 2013). Although they are irreplaceable as a tool to measure the strength of results, they are limited by their need for primary literature with experimental designs that are similar enough.

2.2 Computational ecology in context

In *Life on the Mississippi*, Mark Twain wrote that “There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact”. This is a good description of the purpose of computational ecology: in a data-limited context, merging phenomenological models with pre-existing datasets is a way to efficiently develop conjectures, or more appropriately, build on our knowledge of models and data to put forward testable, quantified hypotheses. Pascual (2005) outlines that computational ecology has a unique ability to go from the complex (natural systems) to the simple (representations and conceptual models), and back (testable predictions). Although the natural world is immensely complex, it is paradoxically the high degree of model abstraction in computational approaches that gives them generality. Because (with the exception of a still narrow family of problems that can be addressed by remote-sensing) there has been no regime shift in the rate at which ecological data are collected – observations from citizen science accumulate, but are highly biased by societal preferences rather than conservation priority (Donaldson et al. 2016; Troudet et al. 2017), by proximity to urban centers and infrastructure (Geldmann et al. 2016), *as well as* by the interaction between these factors Tiago et al. (2017). On the other hand, our needs for testable and actionable predictions increased dramatically. Refining the models and further integrating them with data is necessary.

In {tbl. 1}, the quadrats of ecological approaches are ranked in (again, approximate and arbitrary) order of cost and effort. Ecological models make, by definition, high accuracy predictions, but they tend to be difficult to test (Rykiel 1996) – models relying on precise mathematical expressions can be difficult to calibrate or parameterize. Observations (field sampling) or manipulative approaches (micro/meso/macro-cosms, field experiments) are highly accurate (but have also immense human and monetary costs that limit the scale at which they can be applied). There is simply too much nature around for us to observe, monitor, and manipulate it all.

Table 1 Overview of the properties of the quadrats delineated in {fig. 1}. Empirical observations are the most effort-intensive way of doing ecology. Computational approaches are ranked immediately below because the need to maintain a computational infrastructure is incurring immense (though often invisible) costs. Models are accurate in the limit of their definition, and meta-analysis are accurate in the limit of the empirical studies on which they are based.

Approach	accuracy	testability	suitability for prediction
Empirical observation	yes		
Computational models	unknown	yes	directly
Mathematical models	yes	variable	indirectly
Meta-analysis	yes	no	no

3 En route towards synthesis

The field of ecology as a whole needs to improve the ways in which it can improve synthesis in order to become policy-relevant. Most of the global policy challenges have an ecological or environmental component, and outside of the socio-ecological, socio-economical, socio-cultural, aspects, ecologists can contribute to the mitigation or resolution of these challenges by i) assessing our knowledge of natural systems, ii) developing methods to produce scenarios using state-of-the-art models and tools, and iii) communicating the output of these scenarios to impact policy-making. White et al. (2015) propose that this falls under the umbrella of *action ecology*, *i.e.* using fundamental knowledge and ecological theory to address pressing, real-world questions.

Raghavan et al. (2016) suggest that this approach can also accommodate stakeholder knowledge and engagement. By building models that rely on ecological concepts, empirical data, and stakeholder feedback, they propose a *computational agroecology* program, to use computational tools in the optimization of sustainable agricultural practices. This example suggests that not only can computational approaches yield fundamental research results in a short time frame, they can also be leveraged as a tool for applied research and knowledge transfer now. The definition of "a short time" is highly sensitive to the context – some predictions can be generated using routine tools (in a matter of weeks), whereas some require to develop novel methodologies, and may require years. Accelerating the time to prediction will, in large part, require the development of software that can be deployed and run more rapidly. Overall, computational ecology is nevertheless nimble enough that it can be used to iterate rapidly over a range of scenarios, to inform interactions with policy makers or stakeholders in near real time.

3.1 Mapping the domains of collaboration

Understanding how computational ecology will fit within the broader research landscape requires us to answer three questions: what can computational ecology bring to the table, what are the needs of computational ecologists, and what are the current limitations of computational approaches that could limit their immediate applicability. It seems, at this point, important to minimize neither the importance nor the efficiency of sampling and collection of additional data. Sampling is important because ecological questions, no matter how fundamental, ought to be grounded in phenomena happening in nature, and these are revealed by observation or manipulation of natural systems. Sampling is efficient because it is the final arbiter: how good any prediction is at explaining aspects of a particular empirical system is determined by observations of this system, compared to the predictions.

Relying heavily on external information implies that computational research is dependant on standards for data representation. The Ecological Metadata Language (Fegraus et al. 2005) is an attempt at standardizing the way meta-data are represented for ecological data; adherence to this standard, although it has been shown to improve the ease of assembling large datasets from single studies (Gil et al. 2011), is done on a voluntary basis (and is therefore abysmal). An alternative approach is to rely on community efforts to pre-curate and pre-catalog ecological data, such as with the flagship effort *EcoDataRetriever* (Morris & White 2013). Yet even this approach is ultimately limited, because of the human factor involved — when the upstream data change, they have to be re-worked into the software. A community consensus on data representation, although unlikely, would actually solve several problems at once. First, it would make the integration of multiple data sources trivial. Second, it would provide clear guidelines about the input and storage of data, thus maybe improving their currently limited longevity (Vines et al. 2014). Finally, it would facilitate the integration of data and models with minimum efforts and risk of mis-communication, since the format would be the same for all. To this extent, a recent proposal by Ovaskainen et al. (2017) is particularly interesting: rather than deciding on formats based on knowledge of eco-informatics or data management best practices, why not start from the ecological concepts, and translate them in digital representation? This task requires a strong collaboration between ecologists with topic expertise, ecologists with field expertise, and those of us leaning closest to the computational part of the field.

With or without a common data format, the problem remains that we have very limited insights into how error in predictions made on synthetic datasets will propagate from an analysis to the other (Poisot et al. 2016); in a succesion of predictive steps, do errors at each step amplify, or cancel one another? Biases exist in the underlying data, in the models used to generate the predictions, and this can turn out in three possible ways. First, predictions from these datasets accumulate bias and cannot be used. Second, because the scale at which these predictions are expressed is large, errors are (quantitatively) small enough to be over-ridden by the magnitude of actual variation. Finally, in the best-case but low-realism scenario, errors end up cancelling each other out. The best possible way to understand how errors propagate is to validate predictions *de novo*. Model-validation methods can be used, as they are with SDMs (Hijmans 2012), but *de novo* sampling carries the additional weight of being an independent attempt at testing the prediction. Improved collaborations on this aspect will provide estimates of the robustness of the predictions, in addition to highlighting the steps of the process in which uncertainty is high — these steps are natural candidates for additional methodological development.

Finally, there is a need to assess how the predictions made by purely computational approaches will be fed back into other types of research. This is notably true when presenting these approaches to stakeholders. One possible way to make this knowledge transfer process easier is to be transparent about the way predictions were derived: which data were used (with citations for credits and unique identifiers for reproductibility), which software was used (with versions numbers and code), and what the model / simulations do (White et al. 2013). In short, the onus is on practitioners of computational research to make sure we provide all the information needed to communicate how predictions came to be.

3.2 Establishing the currencies of collaboration

An important question to further the integration of computational approaches to the workflow of ecological research is to establish *currencies* for collaborations. Both at the scale of individuals researchers, research group, and larger research communities, it is important to understand what each can contribute to the research effort. As ecological research is expected to be increasingly predictive and policy-relevant, and as fundamental research tends to tackle increasingly refined and complex questions, it is expected that research problems will become more difficult to resolve. This is an incentive for collaborations that build on the skills that are specific to different approaches.

In an editorial to the *New England Journal of Medicine*, Longo & Drazen (2016) characterized scientists using previously published data as "research parasites" (backlash by a large part of the scientific community caused one of the authors to later retract the statement – Drazen (2016)). Although community ecologists would have, anyways, realized that the presence of parasites indicates a healthy ecosystem (Marcogliese 2005; Hudson et al. 2006), this feeling on unfair benefit for ecological data re-analysis (Mills et al. 2015) has to be addressed. It has no empirical support: Evans (2016) shows that the rate of data re-use in ecology is low and has a large delay – he found no instances of re-analysing the same data for the same (or similar) purpose. There is a necessary delay between the moment data are available, and the moment where they are aggregated and re-purposed (especially considering that data are, at the earliest, published at the same time as the paper). This delay is introduced by the need to understand the data, see how they can be combined, develop a research hypothesis, etc..

On the other hand, there are multiple instances of combining multiple datasets collected at different scales, to address an entirely different question (see GBIF 2016 for an excellent showcase) – it is more likely than data re-use is done with the intent of exploring different questions. It is also worth remembering that ecology as a whole, and macroecology and biogeography in particular, already benefit immensely from data re-use. For example, data collected by citizen scientists are used to generate estimates of biodiversity distribution, but also set and refine conservation target (Devictor et al. 2010); an overwhelming majority of our knowledge of bird richness and distribution comes from the *eBird* project (Sullivan et al. 2009, 2014), which is essentially fed by the unpaid work of citizen scientists.

With this is mind, there is no tip-toeing around the fact that computational ecologists will be *data consumers*, and this data will have to come from ecologists with active field programs (in addition to government, industry, and citizens). Recognizing that computational ecology *needs* this data as a condition for its continued existence and relevance should motivate the establishment of a way to credit and recognize the role of *data producers* (which is discussed in Poisot et al. 2016, in particular in the context of massive dataset aggregation). Data re-users must be extremely pro-active in the establishment of crediting mechanisms for data producers; as the availability of these data is crucial to computational approaches, and as we do not share any of the cost of collecting these data, it behooves us to make sure that our research practices do not accrue a cost for our colleagues with field or lab programs. Encouraging conversations between data producers and data consumers about what data will be shared, when, and how databases will be maintained will improve both collaborations and research quality. In parallel, data producers can benefit from the new analytical avenues opened by advances in computational ecology. Research funders should develop financial incentives to these collaborations, specifically by dedicating a part of the money to developing and implementing sound data archival and re-use strategies, and by encouraging researchers to re-use existing data when they exist.

3.3 Training data-minded ecologists

The fact that data re-use is not instantaneously convenient reveals another piece of information about computational ecology: it relies on different skills, and different tools than those typically used by field ecologists. One of the most fruitful avenue for collaboration lies in recognizing the strengths of different domains: the skills required to assemble a dataset (taxonomic expertise, natural history knowledge, field know-how) and the skills required to develop robust computational studies (programming, applied mathematics) are different. Because these skills are so transversal to any form of ecological research, we are confident that they can be incorporated in any curriculum. If anything, this calls for increased collaboration, where these approaches are put to work in complementarity.

Barraquand et al. (2014) highlighted the fact that professional ecologists received *less* quantitative and computational thinking that they think should be necessary. Increasing the amount of such training does not necessarily imply that natural history or field practice will be sacrificed on the altar of mathematics: rather, ecology would benefit from introducing more quantitative skills and reasoning across all courses, and introductory ones in particular (Hoffman et al. 2016). Instead of dividing the field further between empirically and theoretically minded scientists, this would showcase quantitative skills as being transversal to all questions that ecology can address. What to teach, and how to integrate it to the existing curriculum, does of course requires discussion and consensus building by the community.

A related problem is that most practising ecologists are terrible role models when it comes to showcasing good practices of data management (because there are no incentives to do this); and data management is a crucial step towards easier computational approaches. Even in the minority of cases where ecologists do share their data on public platforms, there are so few metadata that not being able to reproduce the original study is the rule (Roche et al. 2014, 2015). This is a worrying trend, because data management affects how easily research is done, regardless of whether the data are ultimately archived. Because the volume and variety of data we can collect tends to increase over time, and because we expect higher standard of analysis (therefore requiring more programmatic approaches), data management has already became a core skill for ecologists to acquire.

This view is echoed in recent proposals. Mislan et al. (2016) suggested that highlighting the importance of code in most ecological studies would be a way to bring the community to adopt higher standards, all the while de-mystifying the process of producing code. This also requires teaching ecologists how to evaluate the quality of the software they use (Poisot 2015). Finally, Hampton et al. (2015) proposed that the "Tao of Open Science" would be particularly beneficial to the entire field of ecology; as part of the important changes in attitude, they identified the solicitation and integration of productive feedback throughout the research process. Regardless of the technical solution, this emphasizes the need to foster, in ecologists in training, a culture of discussion across disciplinary boundaries.

All of these points can be distilled into practical training recommendations for different groups in the community of ecologists. Classes based around lab or field experience should emphasize practical data management skills, and introduce tools that would make the maintenance of data easier. Modelling classes, especially when concerned about purely mathematical models, should add modules on the way these models can be integrated with empirical data. Finally, computational classes should emphasize communication skills: what do these new tools do, and how can they be used by other fields in ecology; but also, how do we properyl track citations to data, and give credit to data producers? Building this practices into training would ensure that the next generation of ecologists will be able to engage in a meaningful dialogue across methodological boundaries.

4 Concluding remarks

None of these approaches to ecological research have any intrinsic superiority – in the end, direct observation and experimentation trumps all, and serve as the validation, rejection, or refinement of predictions derived in other ways, but lacks the scaling power to be the only viable solution. The growing computational power, growing amount of data, and increasing computational literacy in ecology means that producing theory and predictions is becoming cheaper and faster (regardless of the quality of these products). Yet the time needed to test any prediction is not decreasing (or at least not as fast). Computational science has resulted in the development of many tools and approaches that can be useful to ecology, since they allow ecologists of all kinds to wade through these predictions and data. Confronting theoretical predictions to data is a requirement, if not the core, of ecological synthesis; this is only possible under the conditions that ecologists engage in meaningful dialogue across disciplines, and recognize the currencies of their collaborations.

Discussing the place of computational ecology within the broader context of the ecological sciences will highlight areas of collaborations with other areas of science. Thessen (2016) makes the point that long-standing ecological problems would benefit from being examined through a variety of machine learning techniques – We fully concur, because these techniques usually make the most of existing data (Halevy et al. 2009). Reaching a point where these methods are routinely used by ecologists will require a shift in our culture: quantitative training is currently perceived as inadequate (Barraquand et al. 2014), and most graduate programs do not train ecology students in contemporary statistics (Touchon & McCoy 2016).

Ultimately, any additional data collection has its scope limited by financial, human, and temporal constraints — or in other words, we need to chose what to sample, because we can't afford to sample it all. Computational approaches, because they can work through large amounts of data, and integrate them with models that can generate predictions, might allow answering an all important question: what do we sample, and where?

Some rely on their ecological intuition to answer; although computational ecologists may be deprived of such intuitions, they have the know-how to couple data and models, and can meaningfully contribute to this answer. Computational ecology is also remarkably cost-effective. Although the reliance on advanced research computing incurs immense costs (including hardware maintenance, electrical power, and training of highly qualified personnel; these are often absorbed by local or national consortia), it allows to generate predictions that are highly testable. Although the accuracy of these predictions is currently unknown (and will vary on a model/study/question basis), any additional empirical effort to *validate* predictions will improve their quality, reinforcing the need for dialogue and collaborations.

Acknowledgements: TP thanks Dr. Allison Barner and Dr. Andrew McDonald for stimulating discussions, and the Station de Biologie des Laurentides de l'Université de Montréal for hosting him during part of the writing process. We thank the volunteers of Software Carpentry and Data Carpentry, whose work contribute to improving the skills of ecologists.

References

Ackland & Gallagher. (2004). Stabilization of Large Generalized Lotka-Volterra Foodwebs By Evolutionary Feedback. *Phys Rev Lett.* 93.

Austin. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling.* 157:101–18.

Barraquand et al. (2014). Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions. *PeerJ.* 2:e285.

Beaumont. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics.* 41:379–406.

Beverton & Holt. (1957). On the dynamics of exploited fish populations. Springer Science & Business Media;

Bolker. (2008). Ecological models and data in R. Princeton University Press;

Bolnick & Preisser. (2005). RESOURCE COMPETITION MODIFIES THE STRENGTH OF TRAIT-

MEDIATED PREDATOR–PREY INTERACTIONS: A META-ANALYSIS. *Ecology.* 86:2771–9.

Borwein & Crandall. (2013). Closed Forms: What They Are and Why We Care. *Notices of the American Mathematical Society.* 60:50.

Bourne. (2011). Ten Simple Rules for Getting Ahead as a Computational Biologist in Academia. *PLoS Comput Biol.* 7:e1002001.

Coville & Frederic. (2013). Convergence To The Equilibrium In A Lotka-Volterra Ode Competition System With Mutations. *arXiv.*

Devictor et al. (2010). Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and distributions.* 16:354–62.

Donaldson et al. (2016). Taxonomic bias and international biodiversity conservation research. *FACETS.*

Drazen. (2016). Data Sharing and the Journal. *New England Journal of Medicine.* 374:e24.

D'Amen et al. (2017). Improving spatial predictions of taxonomic, functional and phylogenetic diversity. *Journal of Ecology.*

Elith & Leathwick. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu Rev Ecol Evol Syst.* 40:677–97.

Evans. (2016). Gauging the Purported Costs of Public Data Archiving for Long-Term Population Studies. *PLOS Biology.* 14:e1002432.

Fawcett & Higginson. (2012). Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences.* 109:11735–9.

Fegraus et al. (2005). Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America.* 86:158–68.

Franklin. (2010a). Mapping species distributions: spatial inference and prediction. Cambridge University

Press;

Franklin. (2010b). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions.* 16:321–30.

GBIF. 2016 Oct. GBIF Science Review 2016.

Geldmann et al. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions.* 22:1139–49.

Gil et al. (2011). Examples of ecological data synthesis driven by rich metadata, and practical guidelines to use the Ecological Metadata Language specification to this end. *International Journal of Metadata, Semantics and Ontologies.* 6:46.

Gilpin. (1973). Do Hares Eat Lynx? *The American Naturalist.* 107:727–30.

Gyllenberg et al. (2006). Limit cycles for competitor–competitor–mutualist Lotka–Volterra systems. *Physica D: Nonlinear Phenomena.* 221:135–45.

Halevy et al. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems.* 24:8–12.

Hampton et al. (2015). The Tao of open science for ecology. *Ecosphere.* 6:1–13.

Hijmans. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology.* 93:679–88.

Hoffman et al. (2016). Development and Assessment of Modules to Integrate Quantitative Skills in Introductory Biology Courses. *Cell Biology Education.* 15:ar14–4.

Houlahan et al. (2017). The priority of prediction in ecological understanding. *Oikos.* 126:1–7.

Hudson et al. (2006). Is a healthy ecosystem one that is rich in parasites? *Trends in ecology & evolution.*

21:381–5.

Legendre & Legendre. (1998). Numerical ecology. Oxford, UK: Elsevier;

Levin. (2012). Towards the marriage of theory and data. *Interface Focus.* 2:141–3.

Longo & Drazen. (2016). Data Sharing. *New England Journal of Medicine.* 374:276–7.

Lortie et al. 2013 Jun. Practical interpretation of ecological meta-analyses. PeerJ PrePrints; Report No.: e38v1.

Marcogliese. (2005). Parasites of the superorganism: Are they indicators of ecosystem health? *International journal for parasitology.* 35:705–16.

Markowetz. (2017). All biology is computational biology. *PLOS Biology.* 15:e2002050.

May. (2004). Uses and Abuses of Mathematics in Biology. *Science.* 303:790–3.

Miller & Holloway. (2015). Incorporating movement in species distribution models. *Progress in Physical Geography.* 39:837–49.

Mills et al. (2015). Archiving Primary Data: Solutions for Long-Term Studies. *Trends in Ecology & Evolution.* 30:581–9.

Mislan et al. (2016). Elevating The Status of Code in Ecology. *Trends in Ecology & Evolution.* 31:4–7.

Morris & White. (2013). The EcoData Retriever: Improving Access to Existing Ecological Data. *PLoS ONE.* 8:e65848.

Nicholson & Bailey. (1935). The Balance of Animal Populations.—Part I. *Proceedings of the Zoological Society of London.* 105:551–98.

Otto & Day. (2007). A biologist's guide to mathematical modeling in ecology and evolution. Princeton University Press;

Ovaskainen et al. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters.*:n/a–a.

Papert. (1996). An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning.* 1.

Pascual. (2005). Computational Ecology: From the Complex to the Simple and Back. *PLoS Comp Biol.* 1:e18.

Pellissier et al. (2013). Combining food web and species distribution models for improved community projections. *Ecol Evol.* 3:4572–83.

Petrovskii & Petrovskaya. (2012). Computational ecology as an emerging science. *Interface Focus.* 2:241–54.

Poisot. (2015). Best publishing practices to improve user confidence in scientific software. *Ideas in Ecology and Evolution.* 8.

Poisot et al. (2016). Synthetic datasets and community tools for the rapid testing of ecological hypotheses. *Ecography.* 39:402–8.

Raghavan et al. (2016). Computational Agroecology. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16.* Association for Computing Machinery (ACM);

Roche et al. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology.* 13:e1002295.

Roche et al. (2014). Troubleshooting Public Data Archiving: Suggestions to Increase Participation. Eisen, ed. *PLoS Biology.* 12:e1001779.

Rykiel. (1996). Testing ecological models: the meaning of validation. *Ecological Modelling.* 90:229–44.

Sallan et al. (2011). Persistent predator-prey dynamics revealed by mass extinction. *Proceedings of the National Academy of Sciences.* 108:8335–8.

Soetaert & Herman. (2008). A Practical Guide to Ecological Modelling: Using R as a Simulation Platform.

Springer Verlag;

Staniczenko et al. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecology Letters.*:n/a–a.

Sullivan et al. (2014). The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation.* 169:31–40.

Sullivan et al. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation.* 142:2282–92.

Thessen. (2016). Adoption of Machine Learning Techniques in Ecology and Earth Science. *One Ecosystem.* 1:e8621.

Tiago et al. (2017). Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Scientific Reports.* 7:12832.

Touchon & McCoy. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere.* 7:e01394.

Troudet et al. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports.* 7:9132.

Vines et al. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology.* 24:94–7.

Warton et al. (2014). Model-based thinking for community ecology. *Plant Ecol.* 216:669–82.

West et al. (2016). Field validation of an invasive species Maxent model. *Ecological Informatics.* 36:126–34.

White et al. (2013). Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution.* 6.

White et al. (2015). The next generation of action ecology: novel approaches towards global ecological research. *Ecosphere.* 6:1–16.

Wisz et al. (2012). The role of biotic interactions in shaping distributions and realised assemblages of species:

implications for species distribution modelling. *Biological Reviews.* 88:15–30.

Zhang. (2010). Computational ecology: artificial neural networks and their applications. Singapore: World Scientific Publ;

Zhang. (2012). Computational ecology: graphs, networks and agent-based modeling. New Jersey: World Scientific;

List of Figures

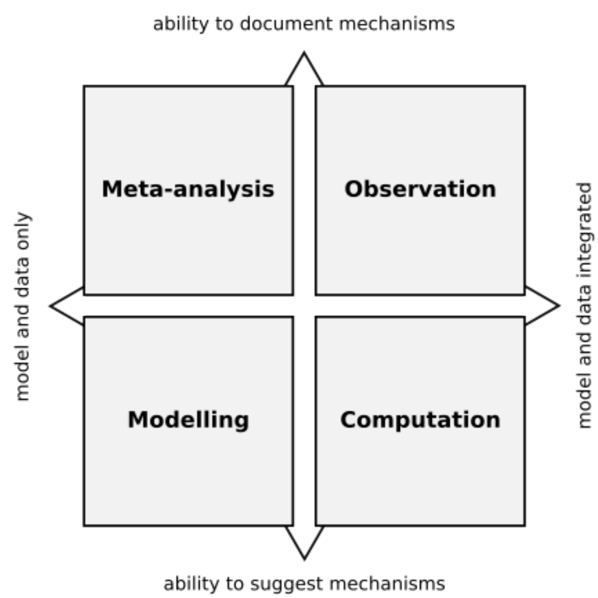

Figure 1 An overview of four quadrats of ecological research. The vertical axis differentiates the ability to document (by observation) or suggest (by simulation and inference) the action of ecological mechanisms. The horizontal axis indicates whether data and models are connected, or not. Computational ecology constitutes one of these quadrats, as it can bridge dynamical models with observations to further suggest mechanisms.