Responses to reviewers

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

---

RECOMMENDATION BY THE SUBJECT EDITOR

We now have the comments of three independent reviewers, who provided very disparate perspectives on the manuscript.

On one hand, the reviewers and I have not identified any substantial conceptual flaw in the manuscript. However, on the other hand, two reviewers and I have not found any substantial insights or novel perspectives in the current version of the manuscript. Reviewer #2 does provide a potential route to substantially improve the manuscript.

Because the current version of the manuscript has potential to provide a contribution to the field, I would like to invite the authors to re-submit a substantially revised version of their manuscript.

Thiago Rangel

REVIEWER: 1

Comments on "Data-based, synthesis-driven: setting the agenda for computational ecology".

> RESPONSE: Regarding their dismissiveness towards prediction as being important: Houlahan et al. 2017 in Oikos have a great paper "The priority of prediction in ecological understanding", which deftly argues that prediction is the only way that ecologists can show understanding. However, those authors focus on statistical modeling approaches only as the path forward to prediction.

With much interest have I read this manuscript multiple times and I have to say that I find reviewing this paper quite difficult. The manuscript itself is well written, concise and reads quite well, however, I was left with a feeling of conflict or rather being challenged by its content and the presentation, where I expected to be inspired and having my agenda set. This does not need to be bad, being challenged can be a good thing. Having said that, I was and still am very excited by the topic, and impressed by the achievement of the authors in trying to address a hugely timely and

necessary task in such a good manner. But, and I guess here belongs a contra punctuation, I felt that this manuscript an even better job in what the authors have chose to be: setting the agenda for an entire field. Maybe this is asking for too much and I try to refrain from contradicting the authors solely based on differences in opinion, but I feel that specifically in a "forum" type of a manuscript there is hardly anything left to comment on that is not based on "opinions". So my suggestion overall to the authors (and editor) is to take my comments as an incentive to maybe make some of the claims that I had problems with more founded in numbers, facts or somehow better explained, such that I can better understand (albeit not necessarily agree to) their point of view. I tried to be nice but direct, but should I have failed to remain polite, or should I have overstepped my limited horizon, please apologize I did not mean to. I tried to be critical in the matter, and I understand my comments as what they are: suggestions to the authors from my specific perspective. As this is not a research paper, there much and more that we can discuss with no avail, so I leave it to the authors to decide what their "opinion" should be and express their view on a very exciting topic, and that can and probably should differ from others (including mine). I only tried to comment on things that caught my attention or where I thought the manuscript could benefit from by better or differently explained matters. It is my policy to always reveal my identity in non-double blind reviews and I am more than happy to explain myself better (or openly) if I should have failed to do so or learn about things in cases where I was mistaken and I hope that the authors can make good use of the comments.

With my best regards, Kamran Safi Max-Planck Institute for Ornithology.

First of all I did think that there is a strong discrepancy between what the title suggests will be in the manuscript and what actually is being provided. Making grand claims in a title is fine, but I felt that there was the promise to the community of computational-ecologists that there is somewhat an agenda that goes beyond what I in hind-sight would say is more or less what everyone sort of knows. This sounds harsh, but honestly having read the manuscript a few times, I have a hard time to say how what is in the MS fits to what is says on the label. This might be a problem that is routed in the next point that I felt needs addressing: audience.

One of the underlying problems, which is particularly stark in the mismatch between title and content, but imminent throughout the manuscript, is the ambiguity in the audience this paper wants to reach out to. The title suggests that it is the computational-ecologist being addressed, but then again large parts of the paper actually seem to be more reaching out to a general ecologist community. Some of the remarks then again are statements that I am sure quite a few non- or less computational ecologists among the ecology ecosystem of sub-disciplines would strongly disagree with.

First paragraph of the introduction (pre 1) says that "Data usually collected in ecological studies have high variability, and are time-consuming, costly, and demanding to collect." Besides that data per-se cannot be time-consuming (minor suggestion: change to time-consuming to collect), I believe that together with the claim that "many problems lack appropriate formal mathematical formulations…" do not do the historic and recent developments in ecology justice. First, I strongly

believe that the whole field of computational-ecology emerges from an accelerated drop in cost of data generation and heterogeneity in ecology. I believe that the whole field only exists because data is being more openly shared (like macro-ecology mentioned in the text), which in turn is a consequence of the technological developments that decrease the costs of collection and lead to an accumulation that requires novel approaches, such as computational ecology. When the authors say that appropriate formal mathematical formulations are lacking, then I have to reply to that, that specifically in the examples they chose in the text as strongholds for the computational-ecology field, Species Distribution Modelling and predator-prey interactions, as also in many other conceptual fields on the ecological discipline, strong theoretical concepts existed and were formulated. The SDMs are based on the Grinnellian niche concept from the early 19-hundreds, with lots and lots of theory, and even equations. I would like to argue, instead, that computational-ecology emerges parallel to what we have witnessed in the molecular biology and biochemistry when computational-biology (in the narrow sense) was born: in ecology we are starting to have more data that we can reasonably process with the methods we have been using so far. Statistics were necessary when ecology suffered from scarcity and noise. In that era of data scarcity the whole concept of frequentist statistics is routed and still we are having a hard time to publish anything that lacks a "p-value". Now, however, we are progressing into an era where we need to learn how to distinguish signal from noise in an increasingly data rich environment, and where p-values have no meaning anymore, but code has. I do agree that the challenges in the field of computational-ecology are more complex and maybe paralyzing than in the "simple" field of molecular biology due to the inherent complexity and stochasticity of ecological systems. But therein exactly, in combination with data repositories and sharing initiatives, lies the "birth-right" of the discipline.

Narrow definition (although not explicit) of what the authors (I believe) think what computational-ecology is, limits in my view the scope of the agenda setting. Under the premise that technological developments are providing the modern ecologists with more and different kinds of data, for me, computational-ecology is a field that contains a wider range of sub-sub-disciplines, if you so want. I would have loved to read more about aspects like challenges in computational-ecology such as data integration, fusion, especially from heterogeneous sources. But maybe that would have been a different audience (again it is not clear to me who exactly the authors are talking to). Or more in-depth discussion about error-propagation, which the authors touch upon. Or challenges that come along with new sensors and kinds of data collected in ecology that we computational people could help with, or seek help by reaching out to existing other computational communities outside ecology. Visualisation of data is a big research area in computational-biology, and immersive data exploration an exploding field, likewise in ecology I think as a community of computational people, we should start to think about the kinds and ways we could make accessibility of data better to serve the purpose of generating knowledge, particularly understanding the ecological mechanisms and evolutionary pathways. But before I start writing my own "forum" contribution here in reply to a review request, I think that it would be great if the authors would define computational-ecology either more concretely than they did, or at least include some overview of the breadth that this field in my opinion has or soon will have beyond SDMs. And I truly believe that the four quadrats of ecological research is although quite simple, too simple.

On page 6 lines 137-138 the authors say that the rate at which ecological data are collected in not improving, to which I have to say that I must strongly disagree with. We are in fact in an era where ecological data are being gathered, stored, curated and shared with an exponential growth and speed. To some extent, to draw the parallels with molecular biology, PCR has been invented in ecology and we are fast going towards whole genoms, particularly with citizen-scientists reporting online to data bases, such as e-bird and ornitho. With the ease of use and reporting, the amount of data increases and costs per datum plummets. Movement data are being made available through global, free webservices such as movebank.org and the technology used produces data of animals in the wild at ever increasing spatial (sub metre error) and temporal (sub seconds) rates and streams everything to the data bases. The armchair ecologist is long real. At the same time, national and international space agencies have opened their data repositories providing ecologists access to an unprecedented amount of data that they can potentially use to relate their study species to fluctuations over time periods of several hours (weather data) to decades (NOAA data bases). And we are just facing the emergence of the consumer grade UAV remote sensing age. So, no I disagree. The challenges are how to bring together all these data.

I find the first sentence of section 3 highly interesting and again problematic. The authors say that "The field of ecology as a whole needs to improve the ways in which it can improve synthesis in order to become policy-relevant." Well, I would say that it quite daring to say something like that in the least, because it clearly assumes that hitherto ecology has just been toying around and ecologists weird storytellers hugging trees. Depending on what policies we are actually talking about (which would be nice to have defined or narrowed down in the paper), ecologists have always been making policy-relevant contributions, beginning with the theorem of limited growth. The contributions of ecologists have been and will be, no matter how improved we eventually will become, policy-relevant under the current limitations of data availability. Knowledge is not absolute, particularly not in ecology, no matter how much computation we do. And policy makers will have to make policies and more often than not the quality of the knowledge behind decision is not primarily relevant.

The concept of using predictions as the holy grail of computational-ecology bares a few questions for me. First of all, through-out the text, I wasn't sure what there was that computational-ecology should predict. In the narrow sense of ecology as a discipline that tries to estimate numbers of biological entities (abundance, diversities), I was somewhat disappointed by the narrow sense of the task. That might be owed to me not understanding the text correctly, or how specific the use of the example SDM as a potential field of research for computational-ecology was meant. Then again, I think that the emphasis on the importance of prediction for a measure of quality and excellence for computational-ecologists was too narrow for my sense of all the wide range of things that computational-ecology in my view could and should encompass, such as visualisation, method development, new inferrential deductions from data, error propagation, computational efficiency and algorithm optimisation, etc etc. What is more is that focusing on prediction only has such an utilitarian touch to it that I fear that the underlying purpose, the scientific raison d'etre for ecology, is being forgotten. Ecology, in the end of the day is a scientific discipline, that needs to integrate the steps of observation, hypothesizing, predicting and testing, to understanding, to

then go back and start again and so forth. Computers are great in predicting but particularly bad in understanding (at least up to now, which makes me really hear more about the future with AI and deep learning in generating hypothesis rather than making predictions in our field). To conclude, I agree that we can help with the predictions, among many other things that we can do to advance the field of ecology, but that ability requires that the way we approach predictions is founded in ecological theory, concepts and hypotheses. Maybe that is what the authors are saying, I wasn't sure though, or maybe in the course of the reading I tended to forget. My confusion with the concept of prediction is maybe best exemplified by the first sentence of the concluding remarks. There the authors say that "… direct observation and experimentation trumps all, and serve as the validation, rejection, or refinement of predictions derived in other ways …". This all sounds to me like it should say hypotheses instead of predictions. I mean how can one reject a prediction. A prediction is a prediction, can be good, bad, or terrible, but rejection is a subjective decision I take to say I do not trust a prediction (find it really bad).

I strongly welcome the section on the currencies of collaboration and think that this is the section that clearly addresses the computational-ecologists among the readers. But unlike the authors, I believe that we can be a bit bolder also in making it crystal clear that anyone should be entitled to ask for the data to be shared and accordingly documented with others for the benefit of the advancement of the field, society, funding bodies and ultimately hopefully also the natural world. Of course it would be great to see more inclusiveness and collaborations emerge, but then again, it not happening should not premit anyone to withhold data. Particularly, and here I believe that the idiosyncratic nature of data collections in ecology make it mandatory, for the sake of reproducibility and scientific validation, data, meta-data and code to be published along with the nice typeset story which appears in the papers published. At some stage, collaborations might be hindering alternative "challenging" views, even hindering progress, if the consent of all contributors has to be guaranteed. In order to remain scientifically acceptable, we have make data openly available, including meta data, and we are on our way there. I have seen a few cases where data owners refused to permit publication of results of multi-authored collaboration efforts, bringing down the entire paper, because the findings were not in line with what they expected to find. I think here, we have a moral responsibility as scientists to not let these kinds of evil alliances to happen. But as a community we do need to discuss the way we want to do this. I see computational-ecologists as a kind of data scavengers, or detritus ecosystem, where otherwise "dead" data gets another chance to turn into knowledge. Of course there are limits to data accessibility, for example in areas where poaching is using this knowledge, but then again other disciplines have these kinds of conflicts too. A minor comment here line 213: scale of individual research groups (remove the "s"). So in short: important to talk about it, define within the community a best practice procedure, but also not being shy about asking for them. There is in addition enough evidence that publishing data provides great synergistic benefits.

> RESPONSE: I liked the analogy of computational ecologists as data scavengers (although the way he structured his metaphor made it clear he does not have a strong understanding of food webs – there is no such thing as a "detritus ecosystem", just a

detrital resource loop – i.e. brown vs. green foodwebs.

What I missed is a section on what we computational-ecologists should do in order to make our contributions to ecology more accessible to the entirety of the field. Here, I believe that there are many things we can improve. Besides, citing the relevant resources from where the data was taken. I believe that computational-ecologists often do not go the extra mile it takes to make their efforts useful for the community of the ecologists. Even though investing in computer literacy is important (not only in ecology), there will be always a wide range of differently skilled (individually specialised, or individual niche seggregation), people that we ought to reach out to. This includes the way we decide to communicate, but also in terms of tools that the community develops requires additional (often academically not-rewarding nor rewarded) activities such as making code actually work and documenting functions etc. Even small obstacles for computer savvy people mark often the difference between using or dumping a method, no matter how wise its use would be. Here, I think this contribution could make a better job in including a wider range of putting task items on the list (see also next point).

Training ecologists in data management. I theoretically agree, but as a practicing field ecologist and yet a computational-ecologist, I have to say that there is a wide gap between theory and practice particularly in ecology also for people knowing data management, in what data management should be and what it ends up being. In addition many ecologists are happy to include some procedures, if it was not too costly and if only someone told them what they should make to document. We could agree on the necessity of publishing guidelines for that matter. I don't think we can eliminate this issue, surely improve on it, but no way we can make this problem go away. One thing I do, which might be a suggestion to make, is assist (or force) people willing to collaborate to go through data management and storage tools to standardise their data. The use of collaborative data storages is very helpful. At the same time, as a community in need of standardised data, maybe we should think about how we ask for infrastructure that could possibly serve the entire community by bringing together the different needs, such as data standardisation, but also correct citation and tracking of data (re)utilisation.

RESPONSE: I'm wondering about rearranging section 3 at the end specifically targeting the three groups – empirical ecologists, modellers and computational ecologists, and what they need to do to enable synthesis. I do think that a strength of the paper could be putting those three groups in conversation with each other about computational ecology, if that's something you'd like to do. I'd be happy to contribute thoughts about the agenda for empirical ecologists, since that's closer to my background.

Line 293: check the beginning of the sentence.

REVIEWER: 2

RESPONSE: This reviewer seemed to have largely mild comments that could be fixed by reframing some of this in terms of the designated audience. Both this reviewer and reviewer 1 argue that citizen science is increasing data availability, which is true, but I think it's still important to note that citizen science is great for answering some questions and not others in ecology. It also is limited to certain taxonomic groups currently, which restricts its scope.

Poisot et al provide a clear overview of considerations surrounding computational ecology. I can't really find anything to disagree with; it seems like it represents a pretty common perspective among modern ecologists. However I think it lacks any particular insights or perspective; it almost seems written so as not to offend anyone (which can be good). I think that it needs some sort of synthesis, prediction, or call to action that is not too vague to be useful. A casual reader of this article should come away with the perspective that computational ecology is critical, there are some great successes, and inspired about much remaining to do. Rather I'd predict they'd think, as I did, 'Sure, that seems fine.'

72: Technically, presence-only models, which are of course more common than presence/absence give P(E|S=1).

85: Rather than 'SDMs exist at the interface between ecological theory and statistical models' I'd say that its possible that they could but the theory papers are limited and usually ignored, in my opinion.

93: As strictly correlative models, I don't think it's accurate to say that they explicitly include mechanisms.

137: I'd say citizen science is improving the rate at which data are collected. And the new Icarus project is tracking individuals at an unprecedented rate.

REVIEWER: 3

The manuscript deals with an extremely important topic of computational ecology. Inability of existing ecological techniques to make accurate predictions remains one of the main factors that impedes further progress in ecological science. Further development of computational ecology should help to rectify the current situation and any comprehensive and consistent discussion of this topic must be encouraged and supported.

RESPONSE: I think Comment 1 is in line with Reviewer 1's concern about clearly

defining our audience. I think some of this reviewer's concern in comment 2 and comment 4 is the distinction between computational methods and dynamic modeling, which we could make clearer in the manuscript.

However, while the manuscript contains some interesting thoughts and suggestions on how to incorporate computational ecology into the mainstream of ecological research, my overall impression about the manuscript is that the authors need careful re-consideration of basic concepts they try to discuss. Several major concerns can be found below.

Abstract, page 1. The authors' claim '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' is not fully justified in the manuscript. It would be helpful if the authors could give a clear list of those areas (probably with bullet points) in the conclusions or anywhere in the text.

Introduction, page 3, lines 63-64. If the authors want to discuss a difference between computational ecology and ecological modelling, they have to give their definition of ecological modelling first. That definition should include 'core characteristics' of ecological modelling like it has been done by the authors for computational ecology on page 2, lines 41-46.

Section 1, pages 3-4. (i) The authors' conclusion about success of SDM is not convincing because their discussion of SDM is unclear. As someone who is not familiar with the SDM technique I wonder what the predictive power of SDM is based on. Statistical methods are unable of making predictions and the authors do not give details of any predictive model.

Section 1, pages 3-4. (ii) The last paragraph in Section 1 has to be carefully revised. Why do the authors believe that dynamic models and simulations are the same (page 4, lines 89-90)? Also, why is a closed-form solution ruled out by the use of stochastic simulations? Is the latter true just for stochastic simulations or for deterministic simulations as well?

Section 2. I am puzzled by the diagram in Figure 1 (page 5) and correspondingly by Table 1 (page 7). Why are 'Modelling' and 'Computational approaches' considered by the authors as separate entities? What is then computed by 'Computation' and how are 'Models' handled to get any result? Please explain.

Section 2, page 5. I find the example with the Lotka-Volterra (L-V) model mentioned in sub-section 2.1 to be very misleading and I require careful revision of this sub-section. The authors place the L-V model in wrong context and that substantially diminishes its importance. The conclusion prospective readers will make from lines 108-121 on page 5 is that sometimes the L-V model is 'good' and sometimes it is 'bad'. Meanwhile the power of the L-V model is based on its capacity to predict trends in the population dynamics as opposite to predicting some quantitative data-based answers. Hence in the example provided by the authors the L-V model did exactly what it was supposed to do: it predicted and confirmed a population cycle. Expressions like 'exists

entirely outside of data…' and '…by contrast…' are highly inappropriate in the context of this discussion.

Subsection 3.1, page 9, lines 193-199. What do the authors mean by 'error propagation'? Is it error accumulation when temporal dynamics is considered or is it propagation of errors in input data as data processing is made or is it error propagation in a computer network or anything else? Please explain.

Subsection 3.2, page 10, lines 242-246. The statement '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…' contradicts, in my opinion, to all the arguments made by the authors for computational ecology. It is not just that data are crucial to computational approaches, it also is that computational approaches should be crucial to data producers as a powerful and in most cases the only means of data processing. The requirement to data producers to admit importance of computational approaches must remain one of the cornerstones of computational ecology (and this is actually mentioned by the authors…).

> **RESPONSE:** "The requirement to data producers to admit importance of computational approaches must remain one of the cornerstones of computational ecology" – this could be part of the agenda for empirical ecologists.

Sub-section 3.3, page 11. The discussion of training trends is not quite relevant as there are a plenty of educational programmes in ecology across the world. Hence any conclusion/suggestion about training ecologists has to be linked to the higher education curriculum in a specific country and the authors should therefore come with more specific suggestions (or at least give references) on how the curriculum can be modified. Also, the title of the section is ambiguous - what landscape do the authors talk about?

> **RESPONSE:** "The discussion of training trends is not quite relevant as there are a plenty of educational programmes in ecology across the world." – I think the discussion of training is actually really relevant. As someone who is training in a top ecology program in the U.S., I've been really amazed how few resources we have here on campus in our department for folks who are interested in computational approaches. I think we can highlight Data and Software Carpentry as great resources, and perhaps talk about training considerations for graduate level students, especially?

Overall, I recommend major revision of the manuscript where the above-mentioned issues must be taken into account. There also are some minor concerns and questions (for example, I am intrigued by the star on page 7, line 150…), but there is no need to discuss them at this stage, as the manuscript should be reviewed again after the major revision has been made.