Plugging in to the environment: new technologies to address hard questions

Session summary - September 4, 2020

Organizer(s): Brett Dickson, Conservation Science Partners, brett@csp-inc.org; Ranjan Anantharaman, MIT, ranjanan@mit.edu; Vincent Landau, Conservation Science Partners, vincent@csp-inc.org; Tony Chang, Conservation Science Partners, tony@csp-inc.org; Kimberly Hall, The Nature Conservancy, kimberly.hall@tnc.org.

Session abstract: Solving today's complex conservation problems will require the creation of new paradigms and scalable, collaborative solutions that cut across geographic, ecological, social, organizational, and technological boundaries. Increasingly, advances in data access and computing technology offer tools and new partnership opportunities for the practice of conservation. On the data side, the conservation field is experiencing a rapid increase in the amount, variety, and quality of readily available data streams, including remotely sensed data or other high resolution ecological information. Tools like Google Earth Engine, cloud computing platforms, more efficient programming languages, advancements in machine learning libraries, and new applications of traditional inferential statistical approaches allow access to or analyses of vast archives of remotely-sensed and other datasets. To help us integrate data and analyses, insights from software design are informing creation of repeatable workflows that allow rapid scaling-up of the use of these datasets and analytical tools across multiple focal geographies and stakeholder communities. In this context, and in a friendly, diverse, and curiosity-driven environment, we hoped to achieve the following goals: (1) engage in cross-disciplinary dialogue on how advances in computing resources, data, and modeling processes/workflows can advance conservation; (2) help build the user base for key tools and programming languages; and (3) identify some exciting areas where the conservation community sees great opportunities to advance the pace and rigor of conservation science through stronger collaborations.

Session Format: This interactive session was held directly after the *Leveraging big data, new software, and high performance computing platforms to accelerate the co-production and impact of conservation applications* symposium. The intention was for the symposium to spark questions, ideas, and interest from attendees that can then be discussed during the interactive session, although we also welcomed people that did not attend the symposium. Our original plan was to use a world cafe format, and we did our best to simulate this approach with 4 break out rooms in Zoom, with participants staying in the same room, and moderators with a set of focal questions moving from one room to the next. The questions and a set of notes based on our discussions are included below.

Proposed next steps (timeline TBD).

- Reaching out to our session participants to help develop a perspective piece for Conservation Science and Practice;
- Developing a set of recommendations or 'best practices' that reflect the feedback we received across the core questions and issues covered;
- Working with the SCB Board(s) to develop materials that raise awareness around these issues at the organizational level.

The questions below were presented to the participants in the four breakouts (each repeated with three groups) and framed discussion. We also developed a set of ideas and subtopics represented by the bulleted lists below each question to spark more conversation as needed. Conversations in the different groups tended to overlap and carry over from one moderator & focal question to the next, so we summarized discussion topics and comments by theme, rather than by the specific questions. In general, we really only scratched the surface in terms of discussing these topics, and especially points under questions 3 and 4 below could be productive areas to focus on in similar sessions next year.

1. **Where are we seeing the greatest constraints in the computational tasks or infrastructure that support conservation science?**

- Do you have examples of where you were not able to run a particular analysis due to computational constraints?
- Do you have ideas on how to address these problems, or do you need training?
- Do you have access to computing power?
- If you are hitting a constraint, are there opportunities to simplify the problem? Can these issues be addressed in part through modeling philosophy/math/increasing abstraction?

2. What kind of training & support in compute/programming is needed to help you be more prepared to collaborate with computer scientists?

- Many options/roles possible in terms of goals for building expertise - i.e., knowing enough to effectively collaborate is at one end of the spectrum, vs. learning "all the things" and building complete workflows, programs, and user-friendly tools at the other.
- If you decide to learn more, what are the pathways?
- What is the time required to get the technical training on advanced computational methods and when is/isn't it worth the effort?
- What applications of high performance computing/big data are you most excited about in terms of potential for conservation impact?

3. Given the urgency of conservation need, how do we transition from one-off scientific analyses to scalable analyses without losing the end user along the way?

- Core concepts, key practices that should be fostered - how do we get there?
- What would we like to see everyone adopt in terms of reproducibility, data management & storage, finding efficiencies in their compute?
- How do we encourage people to publish "imperfect" scripts and take advantage of the open source community as peer reviewers & supporters
- How do we overcome the hurdles associated with sharing data / code. Tips for balancing sharing data & code and keeping your job (getting credit for your work). Proper credit for using other people's workflow etc. - how do we build from and acknowledge sources?

4. How might the field of conservation work to increase diversity in the data science and analytics disciplines?

- How do equity and diversity challenges manifest in this space?
- What about diversity related to demographics (challenges for established scientists & mentors trying to keep up with changes in technology) & areas of expertise?
- What are the most exciting areas for connecting across models and topics (integration across disciplines)?
- How can we encourage adoption of new technologies and computation methods by the larger conservation biology community?

Report-out on discussion areas

Computational constraints. In general, while people recognized the problem of having programs crash or take a really long time to run, many had not run into these issues in their own work (lots of early-stage students in the group). One participant noted there have been "plenty of times where I've had to wait for days" for models to run - examples include work with LIDAR data, environmental DNA, & work in population genomics. Computational scale-up is often data driven - for example by use of finer-scale data sets. Sometimes model complexity drives up compute - examples included multi-factor land use planning, wetland planning. Participants also identified challenges with scaling models, and finding collaborators that can help them create these complex computational pipelines. In terms of collaborating with computer scientists, participants suggested they may not have the bandwidth, or interest (other fields perceived as more interesting or lucrative). We may need to create incentives for computer scientists to join us.

One participant identified 2 main challenges: (1) having a powerful enough computer set-up (especially at home when access to university resources are limited/not an option, i.e., in a pandemic). They noted that being able to run an analysis in the cloud is very empowering. (2) the learning curve for powerful analytical tools and cloud computing and related tools can be steep. However, others observed that reliable internet can make it sketchy to use the cloud, to which another responded that Docker containers can be used (and are great…) but things get complicated, because there is too much to learn! Cost of cloud computing also came up as a constraint.

Several people pointed out that easy to use software with good documentation and video tutorials are needed to help people use the computational power and tools they can access. Many packages are powerful, but without enough training material. Another constraint mentioned was that some powerful software packages don't have support for datasets that we might be interested in. For example, GEE is powerful, but does not support historical NOAA IDS/ISH data - but R does, so that determined how this participant did their analysis.

Participants appreciated that faster software/massive computing power allows people to run connectivity analysis for lots and lots of species to look for patterns, instead of having to narrow down pick ahead of time (commenting on Tim Poissot's talk).

Specific requests to computer scientists:

- Exposing easy-to-use and powerful APIs for parsing commonly used datasets would be good."
- "I'm dealing with NOAA data for example, and doing data cleanup and processing is not easy. You can help with creating high level languages with good documentation. Also, something that can run in the cloud and does not require a powerful home computer would be helpful."

In the conversations on computation, conversations often shifted to constraints related to data access and usability. In users' experience, a key constraint can be that spatial datasets are just not user friendly and/or have poor metadata, and are difficult to download, assemble, and integrate. Also, people sharing the data don't care that you can't use it (described as "poor customer service"). Open access is great, but can be a time trap because it does not mean it's useful/usable, and it can take a long time to get to the point with the data where you can even tell if it will be useful or not. Similarly, many participants noted that data integration is a big challenge. One was working with Google lands free data, high resolution purchased data & ground truthing data, and noted their main constraint was figuring out how to make inference from a stack of different datasets.

Lack of data is also a major issue, especially outside of North America. One participant noted a real difference in the availability of spatial data/imagery between the US and field sites in Central America. In conversations related to connectivity modeling (building from the symposium), participants highlighted the lack of validation data, and noted there is a lot of variation in terms of how this issue is addressed -- and pointed out that many applications really don't address the lack of validation at all.

Ability to store giant datasets is also a constraint for those trying to apply new technology, for example the need to store large amounts of acoustic data collected at remote field sites - but also a general issue as datasets can be really large, and every modification takes up more space. Another participant asked "Is there guidance to scales of ways to store/manage data for projects/people with different capacities (with or without internet access, with or without access to programs, with or without database creation knowledge, etc.) that maintains security?"

Additional points on data:

- It would be great if spatial data were offered at different extents rather than just one, so you don't have to aggregate tons of files if you are working at larger scales.
- We need more incentives/support for making data accessible, and more help with use. Potentially also a review system or way for others to report back on their experience with a data source so that new users have more information to go on before investing time.
- Data security is also an issue - some work with sensitive data and it could be poached. How do I share with others without loss in security?

- Data discovery is a problem - what if we could create an AI app that can crawl the literature and find data -- what if the right filter could be designed to help you more efficiently compare datasets of a specific type/geographic extent?
- Methods are always changing. I have data from my dissertation but can't access it because it required mini-tab. I'm at USGS, and we are required to push all our data to open access. This has created a cultural shift. So I freely publish my data now. When you make a software open source, you get more users. NSF funding should require data to be open.

Increasing use of coding & repeatable workflows:
Repeatable workflows was the main topic of Tony Chang's talk, and in general, it seemed that people were familiar with the idea, but most did not have direct experience, and it's likely that many did not feel like they had the programming skills to do what Tony described. This seems like a great potential area for promotion by the society - encouraging & promoting papers with accessible workflows, and highlighting the value of these investments.

Comments on repeatable workflows:

- I find it fascinating that creating consistency is such a potential big thing. We need to change the incentive structure from having to provide only novel ideas, to value validation and use of standardized methods. That could help us share more and support consistency in our data standards.
- Conservation applications are so specialized - how transferable are workflows really? As soon as you generalize, you lose some performance for those original applications.
- How do you find workflows -- where are they stored and is there a way to categorize them?

Equity & inclusion. One participant suggested that to help address equity and access issues, start off conversations asking about constraints. Access to Wi-Fi, phones, and tech varies, and we need to anticipate that and have a plan to address gaps. We also need clear "beginner places" -- bootcamps, centers on campus that are similar to statistics help centers would be a good place to start. Another observed that social barriers are a key aspect to leveraging big data. Many feel intimidated by the space -- this participant suggested seeking progress not perfection, and just keep moving forward. Often people went through the same path to try to figure out similar challenges and can be helpful, but it's not always easy to find those people.

Training. The main message we heard on the topic of training is some version of "there is so much to learn! How do I decide what aspect of technology and coding to focus on to support my interests?" For the conservation community, the need to consider access and the need for training come up multiple times, starting in academia, and again once people are working for NGOs, governments, and even still in academia. Tools that students and academics have access to may not be affordable once you are away from the university. Also, many academic programs are emphasizing coding skills now as part of the conservation toolkit, but this wasn't the case for everyone, so older people may be left behind, including academic mentors -- so the mentors may not be very helpful in this area. Differences in training and tools across fields can be a barrier, especially for groups trying to collaborate across disciplines, even before we bring in the computer scientists. One participant noted that it's frustrating to not be a native in this landscape -- either because these skills (coding, GIS) were not part of training, or because of the way students in different tracks are sorted. We need more entry points.

Comments on trying to keep up:

- (From a student) What do we learn first? Am I focused on programming? Just started with R from a straight novice level. It feels like a really steep curve especially with all the other commitments (qualifiers/classes). It feels like I'm always catching up with the new algorithms. What's a Shiny app? How do I learn this & is it worth it? Noted that Rmarkdown is super helpful to start to learn these techniques.
- We have seen such rapid technological advancements - for example visual basic to python with GIS. One question: Is there a point where we have a consistent framework? The different number of packages is huge, what do I learn now so I'm not having to do it again in 2 years?

Communication and Collaboration across disciplines. Many participants observed that language is a challenge: Different audiences use the same terms differently, or just don't understand each other. We need to improve understanding across conservation fields & computing/software. A co-created glossary may be helpful here. SImilarly, training/experience can be a barrier - in one participant's experience, a regional government doesn't like to use scientific models because they don't understand them. They are reliant on basic statistical methods.

Bias and bad behavior can also be an issue inhibiting multi-disciplinary collaborations. One participant noted that in the social science symposium during this year's meeting, there were comments about "it's not a real science" (or maybe "not a hard science") -- one participant noted we need to be more welcoming to social scientists, and also be more welcoming and intentional with our language. This will be helpful for collaboration and encouraging diversity. Another participant noted that it's not just language that differs -- methods and approaches vary. Working together requires respect from all sides, there's not one correct way to do things. In fact, these differences can make the science better!

One subset of the group noted that there are really important roles for translators/ connectors with bridging skills, and skills in app design and usability testing that can help us work more effectively with communities, stakeholders & partners. However, those skills don't seem to be appropriately valued (in academia). How do we strengthen the pipeline for this - and in the end strengthen boundary organizations that work in this space?

On the topic of building models collaboratively (context was social & ecological scientists working together), a social scientist noted that it's common to collect data without thinking about how the data will be used in the modeling world --there may be cases where design of data collection would really help improve the process. Another participant agreed with this point, and note that many of us are trained in this from the statistical analysis side (experimental design), but that this kind of training has not kept up with new issues related to advances in technology - issues of scale, data resolution & uncertainty.

Another participant noted there can be barriers to working together that need to be addressed. For example, when you connect from conservation to computer science, the technology in use differs - things like file format vary, which causes an additional challenge to language and other issues. To promote more collaboration, we need to cross pollinate more and cross train -- ecologists should learn the basic comp sci, and comp sci needs basic training in conservation science. A computer scientist in the group agreed (Ranjan), and emphasized that the need for training, crash-courses, and tutorials goes both ways.