RFI on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

Submission Deadline: Monday, December 16 at 5pm Eastern

Title - Cyberinfrastructure in support of iterative ecological forecasting

Abstract (200 words) - Summarize Response

Like weather, changes to ecological systems play an important role in human decisions. Ecological forecasting systems analogous to weather forecasting systems have the potential for not only large societal benefit, but also advances in ecology as a science. However, ecological forecasting faces many technical barriers that can be addressed by investment in cyberinfrastructure research. One major challenge is processing and analyzing ecological data that come in a wide range of sizes, scales, types, formats, and distribution methods. Other cyberinfrastructure challenges are related to forecasting models and algorithms, which must be able to accommodate these diverse data, propagate uncertainty, publish results to publicly available archives, and do all of this repeatedly (and therefore efficiently). Finally, aside from cyberinfrastructure, ecological forecasting will depend on novel funding streams to support sustained operation and maintenance, and on coordination across a wide range of ecological subdisciplines.

Question 1 (max 400 words) *- Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E (science and engineering) research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.*

Weather forecasting was a defining innovation of the 20th century; weather forecasts have become a part of our daily lives, and the science of weather forecasting has not only advanced the field of atmospheric science, but also led to breakthroughs in mathematics, computer science, and engineering that have had major societal benefits. Like weather, changes to ecological systems play an important role in human decisions, and the process of iteratively forecasting these changes can reveal gaps in our ecological understanding. Ecological forecasting offers valuable answers to tailored questions (e.g. Will crops survive a drought? Will algal blooms make this water undrinkable?) while also cutting to fundamental questions about the predictability of ecological processes (i.e. Are there coherent, understandable patterns to the predictability of ecological processes across different spatial and temporal scales, and what factors best explain these patterns?). Unfortunately, operational, near-term ecological forecasting systems comparable to their weather analogues are rare. In part, this rarity is related to general challenges of operational forecasting in any system, such as high data throughput and automation. However, ecological forecasting also faces a unique challenge due to the large diversity of its target study systems, datasets, and modeling approaches; while almost all weather forecasting is concerned with the same relatively small number of physical variables, ecology involves the study of hundreds of thousands of vastly different organisms and their interactions with each other and their environment. Importantly, the diversity of ecological forecasting cyberinfrastructure problems increases the likelihood that general solutions to these problems would be applicable to other disciplines as well.

Question 2 (maximum 600 words) *- Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s). Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of* ***data heterogeneity, data integration and interoperability****? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?*

Building the cyberinfrastructure required to support ecological forecasting supports boundary-pushing research across the fields of mathematics, statistics, ecology, library science, and software development. For one, ecology is becoming increasingly reliant on big data, such as those generated by remote sensing, automated sensor networks, and genomics. The sizes of these data push or exceed limits of hard disk storage and computer memory, making common processing tasks challenging; methods for remote processing (i.e. “moving compute to data”) are therefore needed. At the same time, analyzing spatial and temporal patterns in these data, particularly for multiple data sources simultaneously, demands novel multivariate analysis algorithms. Both of these challenges are amplified by the iterative nature of forecasting: these data processing and analysis steps need to be done in near-real time to allow forecasts to respond to new observations. Besides size, ecological data are highly heterogeneous in type (e.g. continuous, zero-bounded, categorical, count, presence-absence), spatial and temporal scale, format (e.g. tabular, unstructured, gridded), and distribution method (ranging from curated databases with APIs to hand-written field notebooks). This variability not only makes it difficult to develop generic tools for data processing, but also calls for novel mathematical and statistical approaches that can accommodate these data and their interrelationships.

Cyberinfrastructure challenges associated with ecological forecasting extend beyond data to modeling, software, and workflows. Compared to one-off analyses common in ecology, iterative forecasts have higher requirements in terms of reproducibility, robustness, computational efficiency, and automation. One challenge in particular is uncertainty estimation and propagation, an essential component of forecasting. Statistical modeling languages such as BUGS/JAGS/NIMBLE and Stan have made probabilistic analyses more accessible by allowing users to use complex statistical models without having to worry about implementation of complex optimization and sampling algorithms. However, these languages are currently limited in their ability to iteratively update models as new data become available (e.g. for state data assimilation), as is required by operational forecasting systems. There are also opportunities to make these approaches more efficient mathematically (i.e. how they explore solution; for instance, by leveraging likelihood function gradients) and computationally (e.g. taking advantage of parallel computing by making algorithms less Markovian and more naively parallelizable; leveraging graphics processing units for massive parallelization). The need for uncertainties in forecasts also means that forecast outputs take up significantly more disk storage. In addition to having many of the same processing challenges as ecological big data discussed above, forecasts also have to be frequently (and ideally, automatically) published and archived in a publicly-accessible, machine-readable, and searchable manner. The ability of existing open data archives (e.g. Zenodo, Open Science Framework) to accommodate iterative forecasts is unclear.

Ecology is a large and diverse discipline, and it is unlikely that every cyberinfrastructure solution will work equally well for every system. However, expecting each ecological subdiscipline to develop its own solutions to its particular cyberinfrastructure challenges is untenable, especially given resource constraints of researchers, agencies, and other organizations that may come to rely on ecological forecasts. Therefore, the general success of ecological forecasting will require solutions to the cyberinfrastructure challenges described above that are modular, interoperable, and accessible to users without advanced technical backgrounds. To be maximally effective, these tools will need to account for the diversity of programming languages and computing systems used by developers and users of forecasts, and may therefore depend on virtualization or containerization technologies for interoperability.

Question 3 (maximum 300 words) *- Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.*

Challenges associated with ecological forecasting extend beyond cyberinfrastructure. The underlying mathematical and scientific principles of forecasting are specialized, and even with well-designed tools, the next generation of researchers will require substantial training to make sure these tools are used correctly. Also, current models of scientific funding (and, correspondingly, the ways researchers do their work) are well-adapted to projects with discrete, one-off deliverables (e.g. papers, software), but are not effective for long-term maintenance of operational ecological forecasting systems once they have been built. Finally, an essential precondition to the successful development and widespread application of forecasting cyberinfrastructure is community engagement and consensus about which problems are most important and how they will be solved; this will require open communication channels and collaborative networks.

To respond to this RFI, please use the official submission form available at https://www.surveymonkey.com/r/NSFDataCIRFI.