Foundation Models in shaping the future of ecology

Albert Morera¹

¹University of Lleida

November 14, 2023

Abstract

In the field of ecology, we are facing urgent challenges related to biodiversity loss, global change and ecosystem sustainability. In this context, the application of Foundation Models emerges as a powerful tool. These models have the potential to reshape our understanding of natural systems, incorporating large volumes of data from different sources and generating results with a more holistic view of ecosystem functioning. However, the application of foundational models in ecology presents challenges that will need to be addressed, such as model interpretation, training efficiency, and the ethical considerations due their implementation.

Title: Foundation Models in shaping the future of ecology

Albert Morera¹,²*

¹Department of Agricultural and Forest Sciences and Engineering, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain

²Joint Research Unit CTFC – AGROTECNIO – CERCA, Ctra. Sant Llorenç de Morunys km 2, 25280 Solsona, Spain

* corresponding author: Albert Morera (albert.morera@udl.cat)

Statement of authorship: Albert Morera conceived the study idea and wrote the manuscript.

Data accessibility statement: Data sharing not applicable – no new data generated

Running title: Foundation Models shaping ecology’s future

Keywords: Artificial Intelligence; Machine Learning; Unsupervised learning; Modelling; Big Data

Article type: Viewpoint

Number of words in the abstract: 92 words

Number of words in the main text: 2021 words

Number of words in text box: no text box

Number of references: 25 references

Number of figures, tables and text box: 1 figure

Corresponding telephone number: (+34) 623173307

Corresponding e-mail: morera.marra@gmail.com

Mailing address: Campus de l’Escola Tècnica Superior d’Enginyeria Agrària de la Universitat de Lleida. Av. Rovira Roure, 191, 25198, Lleida
Title: Foundation Models in shaping the future of ecology

Albert Morera1,2*

1 Department of Agricultural and Forest Sciences and Engineering, University of Lleida, Av. Alcalde Rovira Roure 191, E-25198 Lleida, Spain
2 Joint Research Unit CTFC – AGROTECNIO – CERCA, Ctra. Sant Llorenç de Morunys km 2, 25280 Solsona, Spain

* corresponding author: Albert Morera (albert.morera@udl.cat)

Abstract

In the field of ecology, we are facing urgent challenges related to biodiversity loss, global change and ecosystem sustainability. In this context, the application of Foundation Models emerges as a powerful tool. These models have the potential to reshape our understanding of natural systems, incorporating large volumes of data from different sources and generating results with a more holistic view of ecosystem functioning. However, the application of foundational models in ecology presents challenges that will need to be addressed, such as model interpretation, training efficiency, and the ethical considerations due their implementation.

Keywords

Artificial Intelligence; Machine Learning; Unsupervised learning; Modelling; Big Data

Introduction

The field of artificial intelligence (AI), which focuses on creating intelligent systems and machine learning (ML), a subset of AI that involves systems learning from data, has progressed in the last decades (Jordan & Michell, 2015), and ecology research has benefited from it (Borowiec et al., 2022; Tuia et al., 2022; Christin et al., 2019; Wäldchen & Mäder, 2018). Among these advances, Foundation Models (FM) have emerged as a powerful tool for analyzing large volumes of data and for generating, understanding and manipulating human-like text. These models hold great promise for revolutionizing the field of ecology and improving the understanding of natural ecosystems. Even so, the implementation of these models in ecology will be a challenge for scientists and decision-makers.

To date, ML algorithms in ecology have been used for instance to study species distribution modeling (Liang et al., 2022; Ryo et al., 2021), land use change characterization (Gelabert et al., 2021), identification of conservable or restorable areas (Cheng et al., 2018; Duhart et al., 2019; Kwok, 2019; Moradi et al., 2019), ecosystem management support (Lauer et al., 2020), ecosystem services management (Dietterich et al., 2012; Scowen et al., 2021), biodiversity conservation (Tuia et al., 2022) and invasive species risk assessment (Barbet-Massin et al., 2018; Jensen et al., 2020). More recently, deep learning, a specialized form of ML using neural networks, have also started to be used for other types of tasks, such as species recognition through images or audios (Gray et al., 2019; Tabak et al., 2019; Mac Aodha et al., 2018; Guirado et al., 2018; Lasseck, 2018; Fritzler et al., 2017). Another example is the NCX model by Microsoft, which can predict the height and species of trees across the US landscape.

Most of these ecological research is performed using uni-task models based on supervised machine learning (SML), which creates the first bottleneck for the full potential of ML. This happens because SML requires human-specified training labels (Pichler et al., 2023), i.e., where each example has information about the desired or expected result. For example, in order to train a model to detect tree crowns from satellite images, a human must first select individual tree crowns from a given number of satellite images under specific conditions. Ecology studies involve extremely complex and dynamic systems, where the interaction of multiple variables and factors often escapes our holistic understanding. Simply collecting labeled data is an arduous and costly task (Wäldchen et al., 2018), and labels established by humans can introduce inherent biases and limitations to our understanding and observation. In order to solve these limitations and overcome the specificity of the models, pre-trained and multi-task models were born.
The arrival of multi-tasking models has brought a new revolution to ML world. In the field of natural language processing (NLP) GPT-4 model by OpenAI is a prime example of this revolution. GPT-4 is not only efficient in understanding texts, but can also summarize them, translate them into different languages, among many other functionalities. Models such as Midjourney 5.1 by Midjourney, Inc. or Stable Diffusion by Stability AI are able to generate any type of image, whereas a few years ago the norm was to find models that could only generate faces, flowers or houses, but not all of them at the same time. This ability to generate images of any type opens up new possibilities in the field of graphic design and visual content creation. Another example is Gato by Google DeepMind, which demonstrates the versatility of multi-tasking models: Gato is able to caption images, chat and control a robotic arm.

In contrast to SML models, multi-task models are mostly based on self-supervised learning (SSL), and therefore use an inherent data structure and do not require user-specified training labels. This new approach has been fundamental in overcoming the bottleneck of data labeling. With SSL the model itself generates the task from which it can learn, masking data and challenging itself to rebuild or predict what it has hidden. Therefore, if human labeling is not required the models can be trained with a massive amount of available data and create much more powerful and generalizable AIs. This ability to handle large volumes of information and learn subtle patterns has led to the emergence of FM. The Center for Research on Foundation Models (CRFM) at Stanford University coined the term "Foundation Model" in August 2021, tentatively referring to "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al., 2021).

Foundation Models in Ecology

FM have not yet been widely applied in the ecology field. This may be partly due to the fact that FMs are an emerging tool that is still untapped in many fields. In later sections, the specific challenges of implementing FM in ecology will be addressed. The different nature of the multitude of ecological data together with existing literature on ecosystem functioning, represent a solid base on which to train FM to solve a wide variety of ecology tasks. For instance, numerous daily satellite data sources exist, including land cover, vegetation changes, deforestation, net primary production, and urban expansion. There are also large global biodiversity databases with species distribution, changes in flora and fauna, and information on threatened species. High-resolution environmental data covers temperature, precipitation, humidity, air and water quality, along with physical data like topography and hydrology. Socio-economic data encompasses populations, human activities, and socio-economic development. Real-time data comes from sensors and IoT devices, while citizen science and social networks also provide data. These abundant data sources offer the potential to unlock the capabilities of SSL-based FM.

Applicability Examples of Foundation Models in Ecology

Here are some possible applications of FM that could help solve hot topics in the field of ecology.

- Better understanding of natural ecosystems: In the field of ecology, it is easy to see studies on flies, fungi or forest cover, for example. But they are usually studied separately. In fact, the eco-systemic resources provided by fungi could be affected by the proliferation of dipteran larvae depending on the radiation that tree canopies let into the forest. The implementation of FM could give a more holistic view of ecosystem functioning by integrating large amounts of data from different sources.

- Prediction of environmental change: predict the impact of global change from an integral perspective on current and future ecosystems. For example, estimating the fire regime in the future is impossible only taking into account climate change, but it is mandatory to incorporate changes in land use, species migration, forest management and social development. Global FM that allow all these factors to be taken into account, will make it possible to implement preventive mitigation and adaptation measures in the coming years.

- Biodiversity conservation: FM themselves could suggest effective conservation strategies for threatened species or conservation areas based on the most exhaustive inventory of the actual state of ecosystems based on extensive databases and knowledge of the specific legislation of each region.
Remote sensing imagery: ML models, and specifically DL, have made great strides in image interpretation. FM could identify anomalies in satellite imagery to help diagnose species invasions and species diseases more accurately and efficiently. This would lead to earlier detection and intervention, improving ecosystem health.

The use of FM extends to ecology-related fields, such as sustainable natural resource management and environmental policy formulation, with the aim of promoting sustainability, conservation and interdisciplinary collaboration to address environmental challenges from a global perspective. In general, the application of FM in ecology encompasses all those challenges that need a more comprehensive and interdisciplinary view, and that would benefit from massive databases that are currently impossible to analyze as a whole.

For the FM development we must consider the data, the model architecture, and the systems used to train and optimize them, along with all the theory that must be developed to understand this new paradigm. FM in ecology must fulfill five properties, namely expressivity (to capture and flexibly represent information), scalability (to efficiently consume large amounts of data), multimodality (to connect different modalities and domains), memory capacity (to store accumulated knowledge), and compositionality (to generalize to new contexts, tasks and environments). These properties allow distilling and accumulating knowledge from different sources and domains, organizing it into an efficient and scalable representation, and generalizing it flexibly to novel contexts (Bommasani et al., 2021).

Open Challenges

Because the development and use of FM is still in its early stages, there are still many challenges and open problems that need to be addressed.

- Quality of the training data: The different sources and input data formats, e.g. satellite images from different platforms or raw databases from different institutions, become a challenges in aligning them for their use. Therefore, standardize input data would enhance data aggregation efficiency by FM.

- Interpretability: Improving the interpretability of the ML model is a crucial step toward understanding the internal algorithm logic of FM, and this has been a hot topic in ecology (Ryo et al., 2021). For example, studying which data inputs are most relevant to model predictions or graphically representing the relationships between variables and model outputs can improve model interpretability to some extent (Lucas, 2020).

- Robustness: It is necessary to be sure that the data cover the whole distribution of possible scenes. Therefore, for example, data sharing initiatives such as those sharing global forest data becomes importance to cover the full range of scenes (Liang & Gamarra, 2020).

- Causal inference: To establish cause-effect relationships prior knowledge of theory, information flow, and methodological issues of the study design are required. These aspects cannot be derived from the data, which precludes FM per se to reach appropriate causal conclusions. Jing et al. (2023) revealed a critical shortcoming of FMs for causal inference. However, certain aspects of causal inference, i.e., those related to model specification and statistical inference, could benefit from machine learning (Lin & Ikram, 2020).

- Knowledge-based models: In addition to data itself, in the ecology field, there is also a large amount of scientific literature providing detailed research information. The generation of large linguistic models encompassing all this existing literature could strengthen the information in the databases analyzed by the FM, improving their predictive accuracy (Deng et al., 2023).

- Efficiency: As the quantity of data and FM parameters increases, there is a tendency for higher training overhead, i.e., more computational capacity, storage and processing time is required. Thus, the development of efficient fine-tuning algorithms will require further insights.

- Security: The world’s first Artificial Intelligence Act proposed by the European Commission, that will come into force in the next few years (Draft AI Act), may have significant direct implications on the application of FM in ecology. This law requires that IA systems used in the EU be secure, transparent, traceable, non-discriminatory and environmentally friendly. This need to regulate IA systems suggests that countries around the world will introduce their own laws in the coming years.
User-Friendliness: FM adoption requires solid technical knowledge in areas such as programming, mathematics and machine learning. Collaboration between ecologists and data scientists will be necessary. At the same time, to promote widespread, it is crucial to design simple and user-friendly systems.

Conclusions and Future Directions

As FM become established as a tool for recurrent use in different fields, it is expected that their performance, interpretation and application in the ecological research will improve. This will make it possible to address ecological problems that require a more holistic view of ecosystems, by integrating large databases of different sources. Even so, it will be necessary to establish regulations and ethical considerations that guarantee data privacy and security, professional responsibility in decision making, and reduce biases towards fair and equitable results for the entire population. In short, the future of FM in ecology is promising, with direct implications for understanding of natural systems and for addressing urgent challenges related to biodiversity loss, global change, and ecosystem sustainability. This overview of FM in ecology hopes to foster a more informed and collaborative future among ecologists, decision makers, and other disciplines, with the goal of improving our natural ecosystems.

Conflict of interest statement

The author declare that he have no conflict of interest.

References


Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., ... Jones, K. E. (2018). Bat detective—Deep learning tools for bat acoustic signal detection. PLOS Computational Biology, 14(3), e1005995. https://doi.org/10.1371/journal.pcbi.1005995


Figure 1. Representation generated by DALL-E 3, using ChatGPT (GPT-4), of the application of Foundation Models (shown as a box in the figure) in ecology. Figure shows the data used by the model (i.e., satellite information, biophysical data, scientific literature, citizen science) to analyze and understand the functioning of natural ecosystems and assist in responding to current ecological challenges. DALL-E 3 is a text-to-image conversion model that is based on OpenAI’s Foundation Model GPT-3.