Towards ecologically meaningful foundation models

Ross J. Gardiner[1†], Henry Cerbone[2†], Hammed A. Akande[3], Carly H. Batist[4], Amber Cowans[5], Stella M. Felsinger[2], Premdeep S. Gill[6, 7], Taniya Kapoor[8], Eliot T. Miller[9], Rodrigo Oyanedel[10, 11], Steven Reece[12], Yan Ying Tan[13], Jonas Antony[14], Carlos Rodriguez-Pardo[15, 16, 17], Amy Hinsley[2], Micah Bowles[18, 19], Sebastian A. Heilpern[20], Rachel H. Parkinson[2, 21*]

[1]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, United Kingdom.
[2]Department of Biology, University of Oxford, United Kingdom.
[3]Department of Biology, Concordia University, Canada.
[4]Conservation International, Moore Center for Science & Solutions, USA.
[5]Centre for Research into Environmental and Ecological Modelling, School of Mathematics and Statistics, University of St Andrews, United Kingdom.
[6]British Antarctic Survey, United Kingdom.
[7]Scott Polar Research Institute, Department of Geography, University of Cambridge, United Kingdom.
[8]Artificial Intelligence Group, Wageningen University & Research, Netherlands.
[9]American Bird Conservancy, USA.
[10]MAR FUTURA, Chile.
[11]Instituto Milenio en Socio-Ecología Costera (SECOS), Chile.
[12]School of Geography and the Environment, Oxford University Centre for the Environment, University of Oxford, United Kingdom.
[13]Michigan Institute for Data and AI in Society (MIDAS), University of Michigan, USA.
[14]Dian Fossey Gorilla Fund, Rwanda.
[15]Politecnico di Milano, Italy.
[16]Euro-Mediterranean Center on Climate Change (CMCC), Italy.

[17]RFF-CMCC European Institute on Economics and the Environment (EIEE), Italy.
[18]Ellison Institute of Technology, United Kingdom.
[19]Department of Physics, University of Oxford, United Kingdom.
[20]Department of Earth System Science, Stanford University, USA.
[21]School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom.

*Corresponding author(s). E-mail(s):
rachel.parkinson@biology.ox.ac.uk; r.parkinson@qmul.ac.uk;
†These authors contributed equally to this work.

Abstract

Ecology aims to explain and predict how organisms interact with each other and their environments across space and time. Yet both ecological data and theory are fragmented, leading to models that generalise poorly beyond specific systems or scales. Empirical evidence spans diverse modalities, resolutions and contexts, while theory is distributed across partially overlapping frameworks that are rarely integrated within a single predictive model. We argue that ecological foundation models (ecoFMs), trained on large, multimodal ecological datasets and informed by ecological theory, offer a route toward unifying data and theory within a common framework. Beyond their scientific value for ecology, ecoFMs present a challenging and consequential testbed for machine learning (ML), demanding advances in multimodal representation learning, theory-guided modelling, and uncertainty-aware inference. By learning shared representations of organisms, environments and interactions, ecoFMs could improve generalisation, link pattern to process, and enable synthesis across ecological sub-disciplines. We outline a roadmap for developing ecoFMs, including requirements for data infrastructure, model architectures, evaluation strategies and governance, and assess where current ML and ecological approaches fall short. If developed responsibly and collaboratively with ecological practitioners and other actors, ecoFMs could enable new modes of analysis and strengthen ecological forecasting, while simultaneously driving advances in ML for multimodal integration, theory-guided learning, and generalisation in complex, data-limited systems.

Keywords: ecology, foundation models, multimodal data, species interactions, ecological forecasting

1 Introduction

Ecology seeks to explain how organisms interact with each other and their environments [1]. Uncovering these relationships is difficult because ecological patterns are

often context-dependent and arise from the interaction of multiple processes. It is also impossible to talk about ecology without including and considering the related fields and processes of evolution and development. The field has struggled to disentangle the incredible complexity in nature and extract general principles that hold across scales, from genes to ecosystems, and from specific locations to global patterns [2, 3].

This challenge takes two forms. Firstly, the amount of multiscale data that has been generated in long-term projects is immense and has outpaced researchers' ability to fully interpret within a specific context, and critically what can be inferred and generalised about other systems. Efforts such as the Wytham Great Tit Project [4], the Isle Royale Wolf-Moose Project [5], the US Long Term Ecological Research Program [6], and the Park Grass Experiment [7] alone account for a combined 250 years of ecological data, spanning molecular samples and genomic sequencing through to population and demographic records. Secondly, while there are many theories in ecology, their unification suffers from the fact that they exist in separate sub-disciplines and are infrequently unified. Even though different theories explain similar phenomena, they do so using different language, variables, and models precluding straightforward transfer and comparison. For example, as dominant theories of modern ecology, both metabolic theory and co-existence theory centre food web dynamics, with one explaining consumer resource dynamics through the language of energy flows, and the other from the language of community interactions. Thirdly, the problem of effect size is crucial in ecology [8]. Despite substantial global data, data fragmentation often hampers scientists' ability to realise the full potential of the ecological data produced to date.

Looking forward, these challenges are intensifying. Ecological data are growing rapidly in both volume and modality, driven by the integration of long-standing field methods and emerging automated and sensing technologies [9]. Addressing the mismatch between data generation and integrative understanding is essential for realising the full value of ecological knowledge. This realisation is necessary to understand Earth system dynamics and for anticipating the impacts of human activity in the Anthropocene.

In the past decade, science has undergone a major shift, with many fields now leveraging artificial intelligence (AI) and machine learning (ML) models to tackle diverse research problems. Foundation models, coined in 2021, are large, unsupervised or self-supervised models pre-trained using substantial data and computation, and then adapted to specific tasks, requiring far fewer resources than training task-specific models from scratch [10]. Over the last three years, large language models (LLMs), the most widely recognised class of foundation models, have had a widespread impact across domains [11]. Their success has been enabled by advances in algorithms, large-scale computational infrastructure, and the construction of massive training datasets. Although LLMs have been applied in ecology [12, 13], they fall short of both being fundamental ecological tools and of being useful representations for ecological study [13, 14]. What fine-tuned LLMs lack is the ability to interpret and relate ecological phenomena in their native representation (the originally collected format) instead relying on secondary interpretation (published literature). Outside of language, the range of models released in the last decade in computer vision from CLIP [15] to SAM3 [16] to DINO [17] have ushered in an age of foundation models in vision. The success

of large, pre-trained models across input modalities has catalysed the proliferation of foundation models in disparate areas ranging across language [18, 19], computer vision [16, 20, 21], earth science [22–24], weather [25], fluid dynamics [26], genomics [27], proteins [28], astrophysics [29], and psychology [30] with calls for the development of more [31].

Despite major advances in these fields, no complete ecological foundation model yet exists despite many efforts [32]. Early conceptual discussions have begun to outline the potential role of foundation models in ecology, including proposals framing such systems as "ecoFMs" [33], but no previous framework has explicitly combined multimodal ecological representation learning, graph structure, uncertainty quantification, and ecological theory. With accelerating global change, ecological research must scale to meet increasingly complex and urgent challenges. These are new challenges not foreseen by past approaches, making methodological advancement necessary.

In this paper, we argue that an ecological foundation model that incorporates a multimodal representation backbone, graph-based ecological structures, and uncertainty quantification would be transformative for both ecological and computer sciences. We present the first blueprint for a complete ecoFM that, if implemented, would transform the field's ability to reveal connections within the Earth system and synthesise the extensive, albeit fractured, corpus of ecological data. In turn, meeting ecology's unique challenges would also advance foundation models and AI for science-driven challenges [34] by introducing domain-specific demands not found in standard benchmarks, motivating methods that can accommodate ecology's non-standard data, and improving generalisation and interpretability in settings characterised by mechanistic structure, long time horizons, and semi-supervised paradigms. These features create opportunities for machine-learning research which we believe will continue to motivate computer scientists to work on ecological applications. Drawing on an interdisciplinary team, we frame our proposal to be actionable for computer scientists, ecologists, and data curators. We begin by surveying related efforts to create foundation models in other fields and highlight points of distinction relevant to ecology. We then outline our proposed framework for what an ecoFM should be, from concept to capabilities. We describe its core requirements and, leveraging extensive discussions across ecology and AI, identify the ethical and societal risks that must be centred in its design. The proposed model would further both ecology and computer science through addressing shared challenges in prediction, generalisation and scaleability.

2 From Ecological Theories and Models to Machine Learning

In ecology, even simple systems can exhibit unexpectedly complex behaviour, as early non-linear population models demonstrated through oscillations, instabilities, and, later, chaotic dynamics [35]. These properties make ecology especially dependent on advances in quantitative modelling, both for inference and for understanding the mechanisms that underlie ecological patterns and dynamics. So, the development of ecological theory has long progressed in tandem with advances in statistical and mathematical methods [36, 37]. This trajectory continues today in more complex models

of ecological interactions, including graph-theoretic models of trophic structure, and community stability [38–40].

The complexity of ecological systems outlined above contributes to the inherent difficulty of modelling ecology. Ecology, most simply, the study of life's processes, must account for conscious agents and unconscious environmental factors operating across multiple spatial and temporal scales, all of which are highly context dependent [41]. Through its close connections with evolution and development [42], ecology presents a uniquely dynamic modelling challenge that differs from other more predictive scientific disciplines. Consequently, there remains ongoing debate over whether ecology is governed by universal causal laws or whether its primary insights are necessarily correlative [43]. These challenges have long been recognised. In their comparison of ecology to rocket science, Hilborn & Ludwig (1993) highlighted the difficulties posed by long timescales; limited opportunities for replication, randomisation and experimental control, and the fundamentally non-stationary nature of ecological systems – all factors that constrain prediction and generalisation [44].

Ecology and evolution possess a rich tradition of systems-level theories that seek to explain how biological organisation, abundance, and diversity emerge across scales. Over the past century, a number of conceptual frameworks have become central to the ecological canon, including neutral theory, metabolic theory, coexistence theory, and trait-based approaches. Although differing in their assumptions and key variables, these theories share a common goal: to identify the mechanisms that generate large-scale biological patterns from local processes. Each focuses on a particular set of measurable quantities, such as genetic variation, energy flux, species interactions, demographic rates, or functional traits, and formalises relationships among them through mathematical models.

For example, neutral theories demonstrate how large-scale patterns can emerge from stochastic processes such as mutation, drift, dispersal, and demographic chance, often without invoking strong selective explanations. In contrast, metabolic and trait-based theories focus on how organismal characteristics constrain ecological performance, while coexistence and network-based theories examine how interactions among species shape community structure and ecosystem dynamics. Advances in computation have further expanded these traditions, with ML increasingly used to infer population structure, identify selective pressures, and model ecological interactions from large datasets [45].

Collectively, these frameworks have generated powerful explanations for particular aspects of ecological systems, yet they remain difficult to unify because they differ in language, variables, and mathematical structure. Each captures a different subset of the processes thought to govern ecological systems, making it difficult to connect mechanisms operating across levels of biological organisation within a single modelling framework. Recent syntheses, including "eco-evo-devo" perspectives and comparative demographic approaches, have sought to bridge some of these divides by emphasising connections between development, evolution, ecology, and environmental change. Nevertheless, integrating diverse ecological processes across scales remains a central

problem in ecology, motivating the search for modelling approaches capable of combining heterogeneous data, interacting mechanisms, and emergent dynamics within a common representational framework.

In many ways, the culmination of these multi-scale modelling efforts were general ecosystem models [46] which sought to represent, "all of life on earth," [46, 47]. The first major example, the Madingley, was introduced in 2013 [48]. Despite the ambition of this approach, uptake among ecologists has remained limited [46], in part because such models are highly prescriptive rather than data-driven and therefore require explicit specification of vast numbers of ecological relationships and Earth-system processes. More broadly, ecological modelling relies heavily on abstraction for managing complexity, for example by using traits as proxies for species, treating species as aggregates of underlying genetic variation or using genetic changes as a proxy for spatial processes [49]. Yet despite these different levels of abstraction and scale, ecological processes remain fundamentally connected through the Earth system, and the data collected across scales are inherently linked.

Modelling in practice is shaped by a trade-off between complexity and effort [50, 51]. Adding more processes can increase realism, but often with diminishing returns as implementation cost grows. Additionally, the "bias-variance" trade-off is realised in ecology in the questions of how many and what kind of interactions to include [52]. As a result, models are typically designed to be just complex enough to address the scientific question, leading to a wide landscape of fragmented and specialised ecological models tailored to specific scenarios [51]. A similar pattern historically occurred in ML: excessive or hand-crafted features often caused over-fitting and poor generalisation [53, 54], resulting in highly specialised models tailored to particular datasets and tasks.

This paradigm shifted with the combination of scalable self-supervised learning objectives and transformer architectures built around attention mechanisms, which laid the groundwork for foundation models. Self-supervised learning enabled models to be trained on vast quantities of unlabelled or weakly labelled data, allowing representations to emerge directly from the structure of the data itself. Attention mechanisms then provided a scalable way to allocate representational capacity dynamically, focusing computation on the most informative components of the input rather than relying on predefined task-specific structures [55]. Deep learning architectures can serve multiple roles, including learning latent representations that support downstream tasks such as classification or clustering, and modelling data distributions for generative applications. Modern foundation models frequently combine both capabilities, learning representations that are useful for analysis while also supporting large-scale generation. These architectures also scale incredibly well with increased data and compute, enabling models such as GPT to function as generalised text predictors across the majority of written human language [56].

Beyond language, foundation models have emerged across many domains. A defining feature is their ability to integrate multiple modalities, such as text, images, audio, and time series, within a unified architecture. This is achieved through a variety of mechanisms, including cross-attention, which allows representations from one modality to draw information from another; in other cases, modalities are fused through joint embeddings or by applying self-attention over combined input data [57–60]. These

approaches allow models to relate heterogeneous signals and learn shared representations without requiring bespoke architectures for each combination of inputs. As a result, foundation models can learn from diverse data sources and develop transferable representations that support a wide range of downstream tasks.

The recent advancement in foundation models presents a monumental opportunity for ecology. A unified modelling framework could support more complex, multi-purpose, scalable, and generalisable ecological models, capable of representing cross-scale, multi-modal data in ways that are currently not available. This was simply not possible before the advent of sufficiently large and complex ML paradigms. To understand how such a unified framework might be built, we next examine the growing landscape of existing models and ecology-adjacent approaches.

3 Ecology-Adjacent Deep Learning Landscape

Recent advances in ML have produced a growing set of deep learning-based models relevant to understanding ecological phenomena. These systems span three major domains: (i) perception, (ii) environmental feature mapping, and (iii) species distribution modelling. Understanding what these models already achieve and where they fall short clarifies the opportunities and constraints for developing an ecoFM.

3.1 Perception

We define perceptive models as those that retrieve relevant biological information from raw ecological data sources. Early work in this area largely centred on fine-grained visual classification benchmarks aimed at identifying visually similar taxa in images [61, 62]. This area has subsequently expanded with the emergence of foundation models, such as BioCLIP2, which learns joint image-text embeddings [63], and vision backbones such as DINO, which learn broadly useful visual representations that transfer remarkably well to species identification when fine-tuned or used as feature extractors [64, 65].

Scaling has also extended beyond purely visual inputs. Models such as TaxaBind integrate multiple information streams, including acoustic recordings, satellite imagery, and environmental covariates, to learn a unified embedding of a given observation with each of these data types [66]. In parallel, advances in bioacoustic processing enable robust automated recognition of species calls from noisy field deployments [67, 68].

The primary utility of these perceptive models in ecology is currently as data-curation tools: they filter camera-trap data [69–71], classify citizen-science images [64], and detect acoustic events [72], and are highly flexible through zero-shot adaptation, which allows for the identification of species not seen in training [73]. It is clear that this area will continue to improve as more advancements are made. Recent benchmarks such as VLM4Bio [74] further suggest that pre-trained vision-language models may support biological trait discovery directly from imagery, indicating that ecological foundation models may learn latent ecological and functional properties from observation records. These feature extraction techniques can form representational layers that can be repurposed for ecoFMs.

3.2 Environmental Feature Mapping

Earth-observation (EO) foundation models constitute the second major group. These systems learn general-purpose representations of environmental states from large volumes of satellite imagery. Models such as TESSERA [23] and AlphaEarth [22] provide dense, globally consistent embeddings, which give insight into ecological processes and patterns such as land-use change, vegetation structure, and ecosystem disturbance. Such representations offer an efficient route to integrating remotely sensed information into ecological analyses. By leveraging already existing models, the need to develop specialised EO models is bypassed. However, EO data have inherent limitations: spatial resolution remains coarse relative to fine-scale habitat variation, microhabitats are often unresolved, and performance is not alone sufficient for aquatic environments and beneath dense canopies [75, 76]. EO models therefore provide valuable inputs for ecoFMs but, on their own, do not represent ecological processes.

3.3 Species Distribution Modelling

Species distribution modelling (SDM) has traditionally been grounded in statistical approaches, including hierarchical Bayesian models, occupancy–detection frameworks, point process models, and joint species distribution models (JSDMs) [77]. These methods explicitly model uncertainty, spatial structure, and species co-occurrence, and continue to form the backbone of ecological inference.

More recently, deep learning methods have been introduced into this landscape [78, 79], largely as a response to the growth of global-scale biodiversity datasets and high-resolution environmental feature mapping data. Neural networks trained on occurrence data, environmental rasters, or remote-sensing imagery can achieve predictive performance comparable to, and in some cases exceeding, classical SDMs such as MaxEnt, particularly at large spatial extents. However, these gains are context-dependent and often diminish for rare or range-restricted species, under spatially structured validation, or when models are required to extrapolate beyond the training domain [80].

Progress in deep SDMs is expected to continue as deep learning methods mature, particularly through improved architectures, uncertainty estimation, and the integration of further data sources. Multimodality, in particular, offers a promising avenue for incorporating complementary information from remote sensing, citizen science, and other ecological data streams [81–83]. Nevertheless, even recent extensions largely treat prediction tasks in isolation, focusing on mapping species–environment relationships rather than representing the underlying ecology as a series of interconnected interactions and systems.

This limitation motivates the development of ecological foundation models. Rather than modelling species–environment relationships independently, ecoFMs aim to learn shared representations that integrate perception, environmental context, and biological interactions within a unified framework. This aligns with recent syntheses, which have highlighted the potential for AI systems to address major global shortfalls in biodiversity knowledge, particularly through integrating heterogeneous ecological observations across scales and regions [84]. By enabling transfer across tasks, scales,

and taxa, ecoFMs extend beyond the scope of SDMs toward a more general and integrative representation of ecological systems.

4 Proposed Framework

Our proposed framework for an ecoFM is a unified intelligent system for ecology that can integrate global data across genes, organisms, communities, environments, space, and time within a single coherent modelling framework. Such a model would learn representations that capture how species interact with each other and with their environments, support tasks that range from species identification to system forecasting and mechanistic inference, and enable synthesis across the currently fragmented landscape of ecological data, models, and theory. Below, we outline the datasets needed for training, our proposed architecture and implementation details, and fully define the capabilities of an ecoFM.

4.1 Data

Large-scale, curated datasets are the core of any foundation model [56, 85]. Ecological data span many biological, spatial and temporal scales and occur in heterogeneous forms. In ML, a modality refers to a distinct type of data with its own structure, statistical properties and representation. In ecology, these modalities include images of species or their habitats, acoustic recordings, genetic sequences, trait measurements, environmental time series, satellite imagery, species occurrence records and more (Figure 1, Table A1). Each modality captures different aspects of ecological systems, requires different processing steps and has no standardised format that links it to other modalities.

In terms of data volume, several streams are now large enough to plausibly support foundation model development. Earth-observation archives (e.g. Sentinel) operate at petabyte scale [86]. Occurrence databases are also large: GBIF currently aggregates over 3.5 billion records, interaction resources such as GloBI combine open datasets and include over 6 million species interaction records [87]. While it is extremely difficult to reason about the general quantity of data required, existing foundation models have pre-trained using a number of data points ranging from hundreds of millions [29, 63] to billions [88, 89]. Considering approximately 30 large-scale datasets across ecology, there are more than 10 billion ecological data products available. In addition, while there are many large datasets in ecology, the volume of poorly standardised, inaccessible, uncurated, or otherwise 'dark' ecological data is likely substantially larger [90–92]. The relevant question is therefore not only whether there are enough data, but whether the data we have can be harmonised, cleaned, and engineered around biases and limitations to support useful ecological representations [92].

To give a concrete example, consider the challenge of integrating species occurrence records with remotely sensed environmental data and species interaction records from GloBI. These sources all clearly describe overlapping ecological systems but in very different forms: occurrence records are irregular, often presence-only observations of taxa in space and time; remotely sensed data are more regularly sampled and spatially continuous, but may be collected at resolutions or over time periods that do not

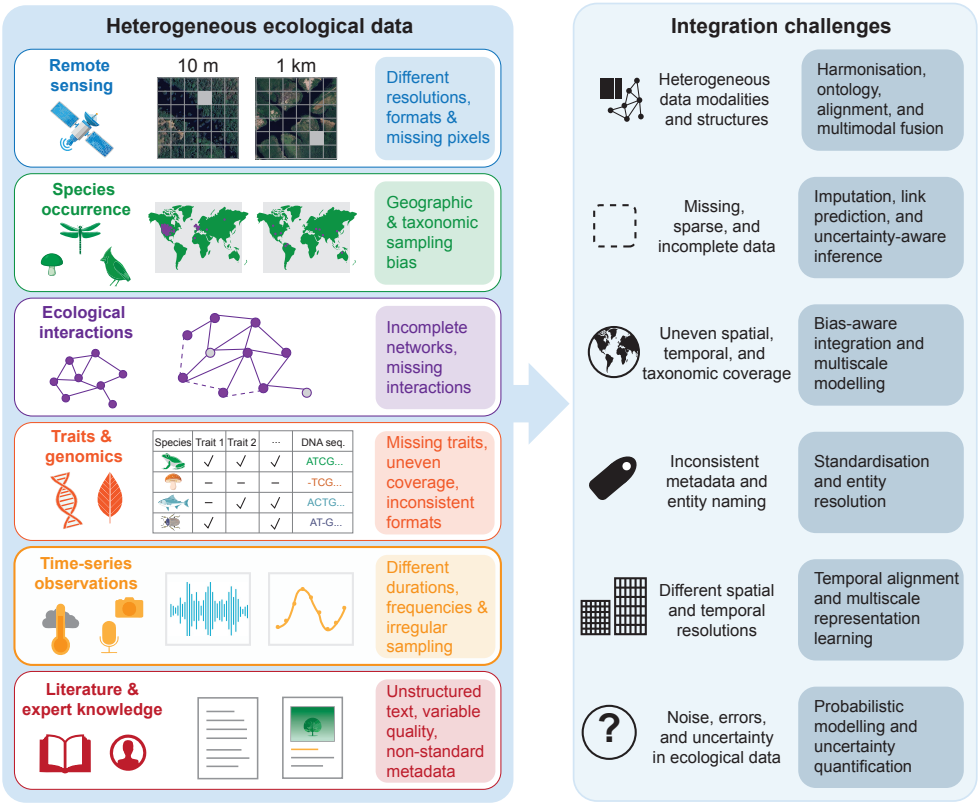

Fig. 1 Challenges in integrating heterogeneous ecological data for ecological foundation models. Ecological data span diverse modalities and structures, including remote sensing, species occurrences, interaction networks, traits, genomics, time-series observations, and literature-derived knowledge. These datasets differ in scale, completeness, coverage, metadata standards, and uncertainty, creating major challenges for integration into unified ecological foundation models. Identified large-scale ecological datasets are listed in Table A1.

match occurrence observations; and GloBI is an evolving interaction graph, updated as interactions are documented, but does not provide a uniformly spatial and temporal observation structure that can be directly aligned with occurrence records or remote-sensing products. So, these data cannot be integrated through simple means; each source must be transformed in ways that preserve its own structure, uncertainty, incompleteness, and biases while also allowing a larger model to learn where data sources are complementary. This integration forms a demanding problem in representation learning. Multimodal ML already identifies representation, alignment and fusion across heterogeneous data sources as core challenges [93]; ecoFMs extend these challenges to modalities which are irregularly sampled, weakly aligned, uncertain and linked by graph-structured relationships.

To clarify these opportunities and challenges, we identified representative datasets across key ecological modalities and scales (Table A1), visually illustrated by Figure 1. These include remote sensing products that capture environmental structure and dynamics, organismal observations from imagery and acoustics, genetic and trait datasets that characterise biological variation, taxonomic and phylogenetic knowledge graphs, and spatiotemporal environmental records used to contextualise ecological processes. This landscape motivates the need for unified tokenisation and multimodal representation learning methods capable of bridging these heterogeneous sources [29, 94]. In the following section, we detail how such methods integrate within the proposed ecoFM architecture.

4.2 Architecture

It is critical to define a realistic model architecture which can capture the relational structure of ecological systems and produce relevant, interpretable predictions. Our framework proceeds through several key stages: modality-specific encoders to embed diverse data types; multimodal masked modelling to learn shared structure across modalities; downstream task adaptation and interpretability tools to assess and probe the learned representation; uncertainty modelling to quantify confidence; and graph-based representations to express ecological interactions. These components are summarised in Figure 2.

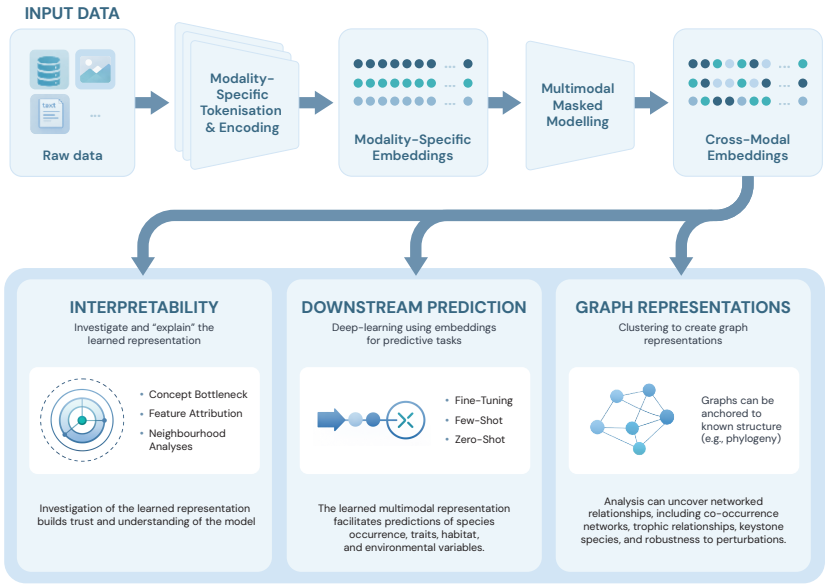

Fig. 2 Workflow illustrating how an ecological foundation model (ecoFM) integrates multimodal ecological data. Raw data from multiple modalities are tokenised and encoded into modality-specific representations, which are jointly encoded using a multimodal masked-modelling objective to produce a shared embedding space. This representation can then be adapted for downstream tasks, including predictive modelling (e.g., species distributions, occupancy, habitat inference). Interpretability methods (e.g., concept bottlenecks, global sensitivity analysis, or attribution analyses) can be applied to interrogate the learned representation. Graph-based tasks may additionally construct ecological networks from the shared embedding using clustering approaches, enabling network-based analyses (e.g., interaction networks, trophic structure, competitive hierarchies)

4.2.1 Modality Encoding

As previously discussed, ecological data span diverse modalities but are often sparse, imbalanced and incomplete. A core requirement of an ecoFM is therefore a modelling framework that can transform heterogeneous datatypes representing a common ontology into a shared representation space. In such a space, observations describing the same organism or event (e.g., a bird's call, its DNA barcode and its photograph) should map to nearby points. This is very far from the reality of biological data storage currently. Achieving this first requires modality-specific transformers or tokensisation schemes that convert raw data into a common embedding format, enabling feature extraction across modalities. These encoders can build on extensive prior advances in representation learning, particularly by leveraging the growing suite of pre-trained, modality-specific models.

The ecoFM can exploit geospatial foundation models such as AlphaEarth [22], SatCLIP [95] or TESSERA [23], which provide rich embeddings of land cover. This

inclusion mirrors the approach taken by many fields of ecology, for example connecting shifts in population, movement, and ecosystems to changes in Earth's abiotic features [96, 97]. Incorporating these embeddings can follow multiple strategies: (1) contrastive learning during pre-training to ensure earth observation and biological embeddings share a semantic structure, (2) through conditioning variables that provide environmental context to biological predictions or (3) spatial tokens that represent environmental states at specific locations and times. Leveraging these embeddings could lower the data requirements, reduce overall computational costs, and improve performance in data-poor regions by transferring environmental knowledge captured from globally sampled imagery.

4.2.2 Multimodal Masked Modelling

With these embeddings established, the next step is to learn how information is shared across modalities. A natural approach is multimodal masked modelling [98, 99], in which the model is trained to reconstruct missing embeddings or tokens from those that remain. This self-supervised objective is particularly well suited to ecological data, which are sparse, unevenly sampled, and rarely contain all modalities for a given observation. By predicting withheld information across modalities, the model is forced to learn the underlying structure that links them, capturing cross-modal dependencies rather than relying on correlations within any single data type. We therefore propose multimodal masked modelling as a central architectural paradigm for ecoFM development, offering a principled way to learn unified ecological representations from heterogeneous and incomplete data.

4.2.3 Downstream Prediction and Interpretability

Once a shared embedding space has been learned, its utility can be assessed through downstream task adaptation. These involve adding lightweight prediction heads or fine-tuning parts of the model to map embeddings onto particular outputs. Examples of this could potentially include species identification in heterogeneous datasets or prediction of traits or habitats. Performance on downstream tasks is the standard way to evaluate foundation models [14, 15], providing an initial measure of how well the multimodal embedding fits predictive tasks.

Complementing downstream prediction, interpretability methods allow users to analyse the embedding space in greater detail and understand how it is structured with respect to both the training data and the underlying model. Neighbourhood analyses [100] can reveal whether the learned representations recover expected ecological organisation, for example clustering by phylogeny, taxonomy or functional traits. Attribution methods, such as Shapley value analysis [101], can quantify the contribution of individual modalities to the resulting embeddings or predictions, indicating how the model integrates disparate data sources. More structured approaches, such as concept bottleneck models [102], may allow users to probe whether specific ecological concepts or traits form identifiable axes within the embedding space and how these concepts mediate downstream predictions. We envision interpretability as a critical component of model evaluation and development of the overall training method.

4.2.4 Uncertainty Modelling

An ecoFM must provide not only predictions but also principled estimates of uncertainty. Accounting for uncertainty is critical, as failure to represent it can distort predictions, inflate confidence in spurious patterns, and produce systematically biased ecological inferences [103]. This goes beyond providing a simple confidence score, which reflects certainty about an output but does not account for missing information, conflicts between modalities, or unfamiliar inputs. Instead, uncertainty should be encoded directly within the model's representations. Recent work such as ProM3E [104] demonstrates one approach through probabilistic embeddings, in which each latent feature is represented as a distribution characterised by a mean and a variance. This allows uncertainty to be expressed throughout the learned representation rather than only at the final output: variance increases when inputs are incomplete or ambiguous, and decreases when evidence is informative and consistent. An ecoFM built on this principle would allow uncertainty to propagate through downstream analyses, producing ecological predictions and decisions that more accurately reflect both data limitations and model confidence.

4.2.5 Graph Representations

Finally, because ecological systems are naturally organised as networks, compatibility with graph-based representations is essential. Ecologists already use graph-theoretic approaches to study food webs, species interactions, connectivity, and ecosystem structure, and ecoFMs should be able to interface naturally with these frameworks. One approach is to derive graphs from learned representations, for example by clustering embeddings into nodes representing taxa, functional groups, or locations, and defining edges based on latent similarity or inferred ecological relationships. These graphs can then be anchored to existing ecological knowledge, such as global interaction datasets [87], allowing predicted links to be compared against known interactions. We believe graph representations would provide a powerful framework for analysing system-level properties, including the propagation of perturbations through ecological networks and the emergence of ecological dynamics across scales.

4.3 Capabilities

An ecoFM should and could be capable of answering some of ecology's pressing questions. Here, we provide a few examples, with aid from a more complete list provided by [105]. We believe that the development of an ecoFM model can advance ecology through the answering of these fundamental questions. For each question, we explain the difficulties and showcase how addressing these challenges could also advance foundation models and computer science more generally.

A central goal of ecology is to catalogue and categorise life on Earth. Trait and taxonomic discovery (see [105] 39) remains difficult because species are incompletely and unevenly sampled and described. Although scientists curate extensive collections of taxonomy and trait information, with taxonomic knowledge continuously refined through expert assessment, phylogenetic analyses, and increasingly advanced methods such as DNA barcoding, many species remain undescribed and/or poorly represented.

The efforts of computer scientists are already transforming these campaigns through automation of taxonomic and trait discovery [106, 107]. The attraction of an ecoFM is that it could extend these approaches through greater multimodality. This could motivate ML methods towards multimodal category discovery by classifying known species and identifying poorly resolved taxa, inferring missing traits, and linking observations to genetic and environmental information.

Tabulating a complete network of interactions between species is critical for understanding how communities function (see [105] 41, 46, and 55). A central challenge is that interactions (e.g., pollination, competition, parasitism, disease transmission) can be remarkably difficult to observe directly, so scientists go to great lengths to document them. Although computational methods can infer missing links in interaction networks [108], unobserved interactions remain ambiguous because they may reflect true absences or insufficient sampling. It has been suggested that the ability of deep learning methods to address this problem could be greatly enhanced through interoperability across data sources [109]. An ecoFM offers a route towards this by integrating complementary evidence from species traits, phylogeny, distributions, environmental conditions, genomic data and known interactions to infer where unobserved relationships are most likely to occur. Completing ecological interaction networks would also provide an important challenge for graph theory. Whether multimodal deep learning can reliably uncover ecologically meaningful interactions remains an open question that is well suited to collaboration between ecologists and computer scientists.

Understanding species occupancy and distribution (see [105] 11, 16, and 19), where species occur and under what conditions they persist, has motivated extensive efforts to model species-environment relationships. Ecologists have developed a rich set of approaches for this problem and this has also involved significant effort from the field of computer science in scaling, as discussed previously. EcoFMs could help explain why populations are distributed as they are by linking evidence data streams. This would also motivate ML methods for integrating multiple data sources, handling biased presence-only records and quantifying uncertainty.

Ecological forecasting, the ability to anticipate future ecological states, analogous to weather forecasting, is a longstanding goal of ecology (see [105] 72, 81, and 92). Predicting ecological time series is difficult, because sampled data are uneven, zero-inflated and contain error, and because the underlying dynamics of a given system contain nonlinearities, cyclic behaviour and may be perturbed into an unknown number of steady states [110]. Recent deep-learning forecasting methods have made progress predicting large collections of time series [88] but ecological forecasting remains difficult even in comparatively well-instrumented systems. For example, in the NEON aquatic forecasting challenge, only a minority of submitted probabilistic models outperformed baseline forecasts for lake water temperature and dissolved oxygen, [111]. Ecologists may additionally be interested in forecasting tipping points ([105] 58) and understanding the trajectories and drivers of change. The aim is that, just as large predictive models can learn transferable structure across diverse data series, an ecoFM could be designed to support forecasts of ecological change; additionally through graph representation, mechanistic drivers of forecasts could be explored. Progress towards ecological forecasting would therefore also advance ML for time-series forecasting, by providing

challenging benchmarks with features that are critical for evaluating transferability to complex real-world domains [112].

These ecological questions naturally translate into benchmark tasks for ecoFMs, including trait imputation, interaction link prediction, species distribution forecasting, ecological time-series forecasting, cross-modal retrieval, representation learning, and ecological knowledge discovery. However, performance should not be evaluated solely through predictive accuracy. Given the sparse, biased, and highly non-stationary nature of ecological data, evaluation should also consider calibration, transferability across taxa and regions, robustness under distribution shift, and consistency with ecological theory.

Importantly, ecoFMs should not be understood solely as systems for integrating heterogeneous ecological datasets. The range of ecological questions outlined above demonstrates that many of ecology's central challenges are fundamentally cross-scale, multimodal, and theory-dependent. Ecology possesses a deep body of theoretical work as outlined in Section 2, and this can provide principled constraints on model design, inference, and evaluation. Ecological theory offers a tool to inform the architecture of an ecoFM and a useful mode of evaluation, asking whether the structure discovered by an ecoFM matches our theoretical expectations.

Existing ecological workflows are highly fragmented: population geneticists, community ecologists, Earth observation scientists, conservation practitioners, and policy-makers often work with separate datasets, models, and theoretical frameworks that rarely communicate directly. An ecoFM would provide a common representational framework. For example, a researcher could input genomic data from a species and evaluate whether observed variation is consistent with nearly neutral models of molecular evolution when contextualised by population size, spatial distribution, climatic variability, and life-history traits inferred from related taxa. Similarly, a field ecologist could combine acoustic recordings, satellite imagery, and historical occurrence data to infer shifts in species interactions or habitat quality in regions with sparse direct sampling. Conservation practitioners could explore the likely ecological consequences of habitat fragmentation or invasive species introductions by leveraging learned representations of comparable ecosystems elsewhere. At broader scales, policy-makers could interact with ecoFMs through scenario-based forecasting tools that estimate how land-use or climate interventions propagate across species, communities, and ecosystem processes under uncertainty.

5 Considerations

Foundation models represent a promising technological frontier for ecology and biodiversity, with the potential to transform how we monitor, understand, and predict ecological dynamics. Their impact, however, will not be determined by technical capability alone, but by how scientific, policy, and practitioner communities approach foresight, ethical design, and governance [113]. To realise the promise of ecoFMs, evaluation must go beyond marginal gains in accuracy on benchmark datasets to systematic auditing of failure modes and the measures taken to mitigate associated harms. Concretely, models should be assessed against principles of ethical AI:

benevolence, non-maleficence, autonomy, justice, and explainability [114]. Below, we discuss the potential dangers associated with epistemic, technological, socio-ethical and environmental risks.

5.1 Epistemic Risks

A central epistemic risk of ecoFMs is that, trained on today's incomplete and biased datasets, they may amplify existing blind spots in biodiversity knowledge and entrench distorted understandings of the natural world. Ecological data are heavily skewed geographically and taxonomically: most observations come from the Global North, while biodiversity-rich regions remain undersampled, and certain vertebrate groups (e.g., birds) dominate records relative to insects, fungi, and microbial communities. The problem of mitigating epistemic risk is a common critique of ecology more generally, with effects such as spatial bias [115], artefacts of racist policies [116], and the after-effects of political instabilities [117], creating underrepresentation of taxa [115]. Bias is also introduced through annotation, such as misidentification of rare species, subjective assessments of habitat quality, and inconsistent labelling practices [118]. Additionally, large bodies of ecological knowledge, especially those held by Indigenous and local communities, remain inaccessible to current datasets and models, often intentionally so to avoid extractive use [119]. These biases are therefore socio-technical in origin, reflecting historical asymmetries, infrastructure limitations, extractive practices, and unequal access to scientific resources.

Addressing these challenges will require targeted investment in under-represented regions and taxa, expanded citizen science and community-based monitoring, and, with permission and while retaining governance, the formal integration of diverse knowledge systems, including Traditional Ecological Knowledge [119]. Integrating these into our proposed model represents a legitimate technical problem in response to an essential social responsibility. While imperfect data is one source of epistemic risk, another issue is understanding the causal mechanism between imperfect data and model outputs. This brings worries about epistemic risk into alignment with interpretability studies as a whole, i.e., understanding the causality of model decisions with respect to training data, which makes biases easier to identify. The integration of these typically unconsidered data sources will only be possible through partnership with knowledge holders in relevant areas and attention to ethical risks (c.f. Section 5.3).

5.2 Technical and Methodological Risks

A methodological risk arises from the limitations of contemporary deep learning itself. Neural networks exhibit well-documented spectral bias, favouring smooth, low-frequency functions [120] thereby underfitting sharp ecological gradients and the rare events that often govern ecological dynamics. Many architectures also implicitly impose Euclidean structure on their latent spaces [121], even when ecological relationships are fundamentally non-Euclidean (i.e. hierarchies of taxonomy or relationships). These inductive characteristics may yield distorted internal representations of ecological systems, undermining model development and introducing errors into downstream scientific analyses. Furthermore, most widely used deep learning models

produce point estimates rather than probability distributions [122–124], leading to outputs that are overconfident and insensitive to data scarcity. As a result, they may fit spurious patterns [125, 126], generalise poorly beyond the distribution of training data [127], and offer limited insight into the reliability of their predictions [128], which is especially consequential when ecoFMs are used to support ecological inference and decision-making.

These vulnerabilities are well known within the deep learning community, and each can be mitigated through deliberate architectural and training choices that incorporate ecological prior knowledge. Models with broader spectral capacity, such as SIREN architectures or networks employing random Fourier features, can better capture high-frequency ecological signals and abrupt transitions [129, 130]. Geometry-aware representations, including projections into hyperbolic space, offer improved modelling of hierarchical relationships; however, optimisation for hyperbolic geometry remains computationally inefficient and can complicate convergence [131]. Process-informed neural networks, which embed known ecological dynamics, such as predator-prey population, carbon cycling, and soil-moisture dynamics directly into deep learning architectures, provide another route for constraining model behaviour while improving interpretability [132]. Data augmentation and rigorous cross-validation can help limit overfitting and improve performance on out-of-distribution samples.

Finally, uncertainty quantification in model design comes with explicit computational trade-offs. Likelihood-based objectives can model aleatoric uncertainty relatively efficiently, whereas epistemic uncertainty generally requires more costly approximations such as Monte Carlo dropout, Bayesian neural networks or ensembles [133]. Monte Carlo dropout avoids training multiple models but increases inference cost through repeated stochastic forward passes [134]; ensembles and hyperparameter ensembles are more expensive but often provide better-calibrated estimates, especially under distribution shift [135]. Recent uncertainty-aware climate foundation-model work illustrates this trade-off, showing that both dropout and ensemble-based approaches can improve extreme-event prediction, with ensembles providing more stable estimates for rare events [136].

Taken together, each of these strategies are likely to play a central role in building ecoFMs, and real progress will depend on ecologists and computer scientists working in close collaboration. Crucially, this collaboration must be driven by scientific problems that neither field can adequately address in isolation. Ecology presents a set of uniquely demanding challenges for ML, including learning from sparse and heterogeneous multimodal data, reasoning across interacting spatial and temporal scales, incorporating mechanistic and theoretical constraints into statistical learning, modelling systems with strong non-stationarity and uncertainty, and generalising under severe sampling bias and incomplete observations. At the same time, advances in ML offer ecology potential new ways to unify fragmented datasets and theories, infer latent ecological structure, and forecast complex system dynamics beyond the capabilities of existing task-specific models.

5.3 Socio-ethical Risks

Without robust ethical safeguards, ecoFMs risk reinforcing existing global power asymmetries [137]. A central concern is data colonialism: the extraction and use of data or knowledge from Indigenous communities and the Global South without appropriate recognition, benefit-sharing, or local governance. Such practices risk commodifying Indigenous knowledge and detaching it from its cultural and ecological context. In addition, the complexity of the AI lifecycle creates ambiguity around accountability: when model outputs lead to harmful downstream decisions, responsibility may be unclear among developers, deployers, and end-users.

Addressing these risks requires participatory and co-designed approaches to AI development, ensuring that affected communities are involved throughout the process and retain authority over how their data and knowledge are used [138]. Several established frameworks offer guidance for structuring governance, such as the Nagoya Protocol, which provides a precedent for recognising sovereignty over genetic and ecological data and ensuring that benefits flow back to providers [139]. The CARE Principles for Indigenous Data Governance, the Design Justice Network Principles [140], work on Indigenous Protocol and Artificial Intelligence [119], and the broader Indigenous Data Sovereignty movement articulate clear expectations regarding collective consent, authority to control, and culturally appropriate data stewardship [141]. Public-sector AI guidelines such as the OECD AI Principles [142] distinguish responsibilities between data owners, model developers, and decision-makers. Translating these principles into ecoFMs will require transparent mechanisms for consent, benefit-sharing and accountability. Organisational structures which ensure that local and Indigenous knowledge holders retain meaningful decision-making power are also an essential part of development.

5.4 Environmental Risks

There is a profound paradox in leveraging energy-intensive technology to address environmental problems. Training state-of-the-art foundation models demands substantial computational resources, and the ensuing energy demand carries a large carbon footprint [143]. Data centres also require substantial water for cooling, and facilities located in water-stressed regions can exacerbate local water scarcity. Beyond contributing to climate change, AI data centres can affect local biodiversity by competing for water and reducing the environmental flows needed to sustain river biodiversity [144]. These well-known environmental impacts undermine public confidence in the utility of AI for advancing scientific discovery in ecology and its application to natural resource management and conservation problems.

These environmental costs must be central considerations in the design, deployment, and governance of ecoFMs. However, the relationship between AI and environment-focused research remains unclear[145]. Preliminary life-cycle analyses reveal serious carbon, water, and land usage associated with AI models [146, 147]. At the same time, the foundation model paradigm offers a potential efficiency advantage [25, 148]. Although pre-training is resource-intensive, it in principle occurs once; the resulting model can be fine-tuned for diverse downstream tasks using far less

data, time, and energy than training separate models for each application. This "pre-train once, fine-tune many" approach represents a shift toward greater computational and environmental efficiency. Deployment and hardware choices also matter. In many applications, smaller models, distilled variants, or locally fine-tuned derivatives may provide sufficient performance while lowering computational costs and allowing data to remain under local control [149]. However, these efficiency gains come with trade-offs in capability and generalisability compared with larger foundation models, while distilled or fine-tuned derivatives still depend on the initial large-scale pre-training stage. The large energy cost of pre-training therefore introduces environmental risk: if it needs to be repeated, this cost is incurred again. As understanding of the environmental impacts of AI improves, the development of ecoFMs should remain responsive to emerging evidence and act accordingly [145, 148].

6 Conclusion

Ecology is entering a phase in which its central scientific challenge—understanding how interactions among organisms and their environments generate patterns of biodiversity, abundance, distribution and ecosystem function—aligns with a major inflection point in AI. Foundation models offer a path toward unifying fragmented ecological data and representing ecological systems in a way that reflects their inherently interconnected nature, from genes to ecosystems. Realising this vision, however, will require far more than repurposing existing AI systems. It demands multimodal data infrastructure, architectures that capture ecological structure, evaluation grounded in ecological theory, and governance frameworks that promote equity, transparency, and trust. If successful, ecoFMs could become a new scientific infrastructure for ecology: a shared framework for integrating observations, testing theory, and generating predictions across the Earth system.

Here, we have outlined the conceptual framework and technical requirements for ecoFMs, and proposed a roadmap for how they could advance ecological prediction, interaction-centred analysis, and the synthesis of ecological data and theory. Importantly, ecoFMs should be built to complement, not replace, ecologists' existing tools and expertise, enabling synthesis across scales rather than obscuring the mechanisms that underpin ecological processes.

If developed responsibly and collaboratively, ecoFMs could mark a step change in ecology's capacity to learn from the accelerating torrent of ecological data and to reveal the structure of the natural world with greater clarity. At the same time, they could position ecology as a driver of new computational advances in learning from multimodal, theory-rich, and data-limited scientific systems. The challenge now is to build the scientific, computational, and ethical foundations needed to ensure these models advance ecology's core goals: understanding, sustaining, and predicting the dynamics of a rapidly changing planet.

Supplementary information. Appendix A: Current Modelling Landscape Table A1: Data modalities, use cases, overlaps, example models, and datasets relevant for ecoFM development. Table A2: Dataset citation information used in Table A1.

Acknowledgements. We thank Anil Madhavapeddy and Matteo Fumagalli for comments on an earlier version of this manuscript. We thank Kelly Cassella (MadeBold Design) for graphic design support, including visual refinement of author-generated figures.

Declarations

- Funding: Funding for this project was provided by a Community Initiatives grant from Schmidt Sciences to RHP and AH. RHP, AH, and YYT were supported by Eric and Wendy Schmidt AI in Science Postdoctoral Fellowships, a program of Schmidt Sciences. AC is funded under the NERC SUPER DTP (NE/S007342/1). HC is supported by a Rhodes Scholarship. SR was supported by a grant from the Leverhulme Trust (grant number RC-2021–076) as a member of the Leverhulme Centre for Nature Recovery. HAA was supported by the Canada Research Chair in Spatial Ecology and Biodiversity held by Pedro Peres-Neto, and funding from the NSERC-CREATE program on Leadership in Environmental and Digital Innovation for Sustainability (LEADS) at Concordia University, Montreal, Canada. SMF was supported by Hope Trust funding held by Geraldine Wright. CHB was supported by the Patrick J. McGovern Foundation. CR-P acknowledges support from the European Union European Research Council (ERC) Grant project No 101044703 (EUNICE). PSG is supported by a NERC DTP studentship. RJG is supported by a doctoral training grant awarded as part of the UKRI AI Centre for Doctoral Training in Environmental Intelligence (UKRI grant number EP/S022074/1).
- Competing interests: The authors declare no competing interests.
- Ethics approval and consent to participate: Not applicable.
- Consent for publication: All authors provide consent for publication.
- Data availability: Not applicable.
- Materials availability: Not applicable.
- Code availability: Not applicable.
- Author contributions: RHP, RJG, HC, HAA, CHB, AC, SMF, PSG, ETM, RO, TK, SR, YYT, CR-P, MB, and SH conceptualised the manuscript. RHP, RJG, and SMF designed the figures. RHP and AH coordinated the project. All authors contributed to writing and editing.

Appendix A Data table

Table A1: Data modalities, use cases, and example datasets relevant for ecoFM development.

Domain	Data type	Example datasets	API?
Satellite imagery	Land cover Topography	NASA Harmonised Landsat & Sentinel 2 (HLS; >1 PB)	Yes
Weather & Climate	Historic & future climate	Copernicus Climate Data Store (multi-PB)	Yes
		Temperature 12k (global Holocene reconstruction)	Partial
Species presence data	Presence-only data	GBIF (>3.7B occurrence records)	Yes
		iNaturalist (>200M observations)	Yes
		Natural history museum collections (hundreds of millions of specimens)	Partial
	Surveys with effort information	BioTIME (>12M abundance records)	No
		eBird (>1B bird observations)	Yes
	Passively sampled data	FishSounds (1000s of recordings)	No
		xeno-canto (>700k recordings)	Yes
		ecosound-web.de (large distributed archive)	Partial
		GBIF (>3.7B occurrence records)	Yes
Aquatic systems	Water quality/pollution	NASA HLS (>1 PB)	Yes
		National hydrological surveys (country-scale archives)	Mixed
		USGS National Water Information System (NWIS; millions of measurements)	Yes
		Water Quality Portal (WQP; >400M records)	Yes

Continued on next page

Domain	Data type	Example datasets	API?
Soil & geology		GLEON (global lake sensor network)	Partial
		CAMELS (671+ catchments)	No
	Bathymetry	IHO Data Centre for Digital Bathymetry (global bathymetric grids)	Partial
	Soil type & chemistry	National geological surveys (country-scale archives)	Mixed
	Seismic & volcanic activity		
Traits	DNA sequences	GenBank (>20 PB sequence data)	Yes
		BOLD (>15M barcode sequences)	Yes
		GoaT/EBG (1000s of genomes)	Partial
	Phenotypic traits	Encyclopedia of Life (>1M taxa pages)	Yes
		TRY (>15M trait records)	Limited
		Phenoscape (ontology-linked trait database)	Partial
		AVONET (traits for >11k bird species)	No
		FishBase (>35k fish species)	Partial
Derived data	Taxonomy/Phylogeny	Catalogue of Life (>2M species)	Yes
		Open Tree of Life (>2.5M taxa)	Yes
	Range maps	IUCN range maps (>150k species assessments)	Limited
		AVONET (traits for >11k bird species)	No
		FishBase (>35k fish species)	Partial
	Foodwebs & interactions	GloBI (>20M species interactions)	Yes

Table A2: Dataset citation information used in Table A1.

Dataset	Reference
NASA Harmonised Landsat & Sentinel 2 (HLS)	Ju, Junchang, Qiang Zhou, Brian Freitag, David P. Roy, Hankui K. Zhang, Madhu Sridhar, John Mandel, Saeed Arab, Gail Schmidt, Christopher J. Crawford, Ferran Gascon, Peter A. Strobl, Jeffrey G. Masek, and Christopher S. R. Neigh. 2025. "The Harmonized Landsat and Sentinel-2 Version 2.0 Surface Reflectance Dataset." *Remote Sensing of Environment* 324:114723. DOI: 10.1016/j.rse.2025.114723.
Copernicus Climate Data Store	ECMWF. 2024. *Copernicus Climate Data Store*.
Temperature 12k	Kaufman, Darrell, et al. 2020. "A Global Database of Holocene Paleotemperature Records." *Scientific Data* 7(1):115. DOI: 10.1038/s41597-020-0530-7.
GBIF	Global Biodiversity Information Facility. 2020. *What is GBIF?* GBIF.
iNaturalist	van Horn, Grant, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. "The iNaturalist Species Classification and Detection Dataset." arXiv:1707.06642. DOI: 10.48550/arXiv.1707.06642.
Natural history museum collections	Natural history museum collections indexed through GBIF.
BioTIME	Dornelas, Maria, Laura H. Antão, Faye Moyes, Amanda E. Bates, Anne E. Magurran, and many others. 2018. "BioTIME: A Database of Biodiversity Time Series for the Anthropocene." *Global Ecology and Biogeography* 27:760–786. DOI: 10.1111/geb.12729.
eBird	Sullivan, Brian L., Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theodoros Damoulas, André A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, and others. 2014. "The eBird Enterprise: An Integrated Approach to Development and Application of Citizen Science." *Biological Conservation* 169:31–40. DOI: 10.1016/j.biocon.2013.11.003.
FishSounds	FishSounds database.
xeno-canto	xeno-canto Foundation for Sharing Bird Sounds.
ecosound-web.de	Darras, Kevin F. A., Noemí Pérez, Liu Dilong, Tara Hanf-Dressler, Matthias Markolf, Thomas C. Wanger, and Anna F. Cord. 2024. "ecoSound-web: An Open-Source, Online Platform for Ecoacoustics." *F1000Research* 9:1224. DOI: 10.12688/f1000research.26369.3.
USGS National Water Information System (NWIS)	U.S. Geological Survey. 2026. *USGS Water Data for the Nation: U.S. Geological Survey National Water Information System Database*. DOI: 10.5066/F7P55KJN.

Continued on next page

Dataset	Reference
Water Quality Portal (WQP)	National Water Quality Monitoring Council, United States Geological Survey, and Environmental Protection Agency. 2021. *Water Quality Portal*. Washington, DC. DOI: 10.5066/P9QRKUVJ.
GLEON	Weathers, Kathleen C., Paul C. Hanson, Peter Arzberger, Jennifer Brentrup, Justin Brookes, Cayelan C. Carey, Evelyn Gaiser, David P. Hamilton, Grace S. Hong, Bas Ibelings, Vera Istvánovics, Eleanor Jennings, Bomchul Kim, Tim Kratz, Fang-Pang Lin, Kohji Muraoka, Catherine O'Reilly, Kevin C. Rose, Elizabeth Ryder, and Guangwei Zhu. 2013. "The Global Lake Ecological Observatory Network (GLEON): The Evolution of Grassroots Network Science." *Limnology and Oceanography Bulletin* 22(3):71–73. DOI: 10.1002/lob.201322371.
CAMELS	Newman, A. J., M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, D. Blodgett, L. Brekke, J. R. Arnold, T. Hopson, and Q. Duan. 2015. "Development of a Large-Sample Watershed-Scale Hydrometeorological Data Set for the Contiguous USA: Data Set Characteristics and Assessment of Regional Variability in Hydrologic Model Performance." *Hydrology and Earth System Sciences* 19(1):209–223. DOI: 10.5194/hess-19-209-2015.
IHO Data Centre for Digital Bathymetry	International Hydrographic Organization Data Centre for Digital Bathymetry. 2026. *IHO Data Centre for Digital Bathymetry (DCDB)*. NOAA National Centers for Environmental Information.
National geological surveys	OneGeology Global Geological Surveys.
GenBank	Clark, Karen, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2016. "GenBank." *Nucleic Acids Research* 44(Database issue):D67–D72. DOI: 10.1093/nar/gkv1276.
BOLD	Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2007. "BOLD: The Barcode of Life Data System." *Molecular Ecology Notes* 7(3):355–364. DOI: 10.1111/j.1471-8286.2007.01678.x.
GoaT / Earth BioGenome Project	Lewin, Harris A., Stephen Richards, Erez Lieberman Aiden, Miguel L. Allende, John M. Archibald, and many others. 2022. "The Earth BioGenome Project 2020: Starting the Clock." *Proceedings of the National Academy of Sciences* 119(4):e2115635118. DOI: 10.1073/pnas.2115635118.
Encyclopedia of Life	Parr, Cynthia S., Nathan Wilson, Patrick Leary, Katja S. Schulz, Kristen Lans, Lisa Walley, Jennifer A. Hammock, Anthony Goddard, Jeremy Rice, Marie Studer, Jeffrey T. G. Holmes, and Robert J. Corrigan, Jr. 2014. "The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth." *Biodiversity Data Journal* 2:e1079. DOI: 10.3897/BDJ.2.e1079.

Continued on next page

Dataset	Reference
TRY	Kattge, Jens, Gerhard Bönisch, Sandra Díaz, Sandra Lavorel, Iain Colin Prentice, Paul Leadley, Susanne Tautenhahn, and many others. 2020. "TRY Plant Trait Database – Enhanced Coverage and Open Access." *Global Change Biology* 26(1):119–188. DOI: 10.1111/gcb.14904.
Phenoscape	Manda, Prashanti, James P. Balhoff, Hilmar Lapp, Paula Mabee, and Todd J. Vision. 2015. "Using the Phenoscape Knowledgebase to Relate Genetic Perturbations to Phenotypic Evolution." *Genesis* 53(8):561–571. DOI: 10.1002/dvg.22878.
AVONET	Tobias, Joseph A., Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, and many others. 2022. "AVONET: Morphological, Ecological and Geographical Data for All Birds." *Ecology Letters* 25(3):581–597. DOI: 10.1111/ele.13898.
FishBase	Froese, Rainer, and Daniel Pauly. 2019. *FishBase*. World Wide Web electronic publication.
Catalogue of Life	Catalogue of Life Foundation. 2026. *Catalogue of Life (2026-05-15 XR)*. Amsterdam, Netherlands. DOI: 10.48580/dgxsq.
Open Tree of Life	Hinchliff, Cody E., Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, and many others. 2015. "Synthesis of Phylogeny and Taxonomy into a Comprehensive Tree of Life." *Proceedings of the National Academy of Sciences* 112(41):12764–12769. DOI: 10.1073/pnas.1423041112.
IUCN range maps	IUCN Red List spatial data.
GloBI	Poelen, Jorrit H., James D. Simons, and Chris J. Mungall. 2014. "Global Biotic Interactions: An Open Infrastructure to Share and Analyze Species-Interaction Datasets." *Ecological Informatics* 24:148–159. DOI: 10.1016/j.ecoinf.2014.08.005.

References

[1] Taylor, W. P. What is Ecology and What Good is It? *Ecology* **17**, 333–346 (1936).

[2] Lawton, J. H. Are There General Laws in Ecology? *Oikos* **84**, 177 (1999).

[3] Turchin, P. Does population ecology have general laws? *Oikos* **94**, 17–26 (2001).

[4] Perrins, C. in *Birds* (eds Savill, P., Perrins, C., Kirby, K. & Fisher, N.) *Wytham Woods: Oxford's Ecological Laboratory* 145–172 (Oxford University Press, Oxford, 2011).

[5] Nelson, M. P., Vucetich, J. A., Peterson, R. O. & Vucetich, L. M. The Isle Royale Wolf–Moose Project (1958-present) and the Wonder of Long-Term Ecological Research. *Endeavour* **35**, 31–39 (2011).

[6] Hobbie, J. E., Carpenter, S. R., Grimm, N. B., Gosz, J. R. & Seastedt, T. R. The US Long Term Ecological Research Program. *BioScience* **53**, 21–32 (2003).

[7] Silvertown, J. *et al.* The Park Grass Experiment 1856–2006: Its contribution to ecology. *Journal of Ecology* **94**, 801–814 (2006).

[8] Gould, E. *et al.* Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology* **23**, 35 (2025).

[9] Besson, M. *et al.* Towards the fully automated monitoring of ecological communities. *Ecology Letters* **25**, 2753–2775 (2022).

[10] Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models (2022). arXiv:2108.07258.

[11] HAI, S. Ai index. https://hai.stanford.edu/ai-index (2025). Accessed: 2025-11-17.

[12] Moorthy, S. M. K., Qi, M., Rosen, A., Malhi, Y. & Salguero-Gomez, R. Harnessing Large Language Models for Ecological Literature Reviews: A Practical Pipeline. *EcoEvoRxiv* (2025).

[13] Dorm, F., Millard, J., Purves, D., Harfoot, M. & Aodha, O. M. Large language models possess some ecological knowledge, but how much? (2025).

[14] Vendrow, E. *et al.* INQUIRE: A Natural World Text-to-Image Retrieval Benchmark (2024). arXiv:2411.02537.

[15] Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision (2021). arXiv:2103.00020.

[16] Carion, N. *et al.* SAM 3: Segment Anything with Concepts (2025). arXiv:2511.16719.

[17] Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers (2021). arXiv:2104.14294.

[18] Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).

[19] Brown, T. B. *et al.* Language Models are Few-Shot Learners (2020). arXiv:2005.14165.

[20] Wiedemer, T. *et al.* Video models are zero-shot learners and reasoners (2025). arXiv:2509.20328.

[21] Balestriero, R. & LeCun, Y. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics (2025). arXiv:2511.08544.

[22] Brown, C. F. *et al.* AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data (2025). arXiv:2507.22291.

[23] Feng, Z. *et al.* TESSERA: Precomputed FAIR Global Pixel Embeddings for Earth Representation and Analysis (2025). arXiv:2506.20380.

[24] Ren, H. *et al.* Global River Forecasting with a Topology-Informed AI Foundation Model (2026). arXiv:2602.22293.

[25] Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

[26] McCabe, M. *et al.* Walrus: A Cross-Domain Foundation Model for Continuum Dynamics (2025). arXiv:2511.15684.

[27] Brixi, G. *et al.* Genome modeling and design across all domains of life with Evo 2 (2025).

[28] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[29] Parker, L. *et al.* AION-1: Omnimodal Foundation Model for Astronomical Sciences (2025). arXiv:2510.17960.

[30] Binz, M. *et al.* A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).

[31] Khan, S. A. *et al.* Multimodal foundation transformer models for multiscale genomics. *Nature Methods* 1–13 (2025).

[32] Yu, R. *et al.* Foundation Models for Environmental Science: A Survey of Emerging Frontiers (2025). arXiv:2504.04280.

[33] Morera, A. Foundation models in shaping the future of ecology. *Ecological Informatics* **80**, 102545 (2024).

[34] Han, B. A. *et al.* A synergistic future for AI and ecology. *Proceedings of the National Academy of Sciences* **120**, e2220283120 (2023).

[35] May, R. M. Simple mathematical models with very complicated dynamics. *Nature* **261**, 459–467 (1976).

[36] Clark, J. S. Why environmental scientists are becoming bayesians. *Ecology Letters* **8**, 2–14 (2005).

[37] Gilbert, N. A. *et al.* A century of statistical ecology. *Ecology* **105**, e4283 (2024).

[38] May, R. M., Levin, S. A. & Sugihara, G. Complex systems: Ecology for bankers. *Nature* **451**, 893–895 (2008). URL https://doi.org/10.1038/451893a.

[39] Delmas, E. *et al.* Analysing ecological networks of species interactions. *Biological Reviews* **94**, 16–36 (2019).

[40] Kalirad, A. & Sommer, R. J. Ecological graph theory: Simulating competition and coexistence on graphs. *Methods in Ecology and Evolution* **16**, 2667–2680 (2025).

[41] Klinges, D. H. *et al.* Matching climate to biological scales. *Trends in Ecology & Evolution* **41**, 329–343 (2026).

[42] Abouheif, E. *et al.* in *Eco-Evo-Devo: The Time Has Come* (eds Landry, C. R. & Aubin-Horth, N.) *Ecological Genomics: Ecology and the Evolution of Genes and Genomes* 107–125 (Springer Netherlands, Dordrecht, 2014).

[43] O'Hara, R. B. The anarchist's guide to ecological theory. Or, we don't need no stinkin' laws. *Oikos* **110**, 390–393 (2005).

[44] Hilborn, R. & Ludwig, D. The Limits of Applied Ecological Research. *Ecological Applications* **3**, 550–552 (1993).

[45] Huang, X., Rymbekova, A., Dolgova, O., Lao, O. & Kuhlwilm, M. Harnessing deep learning for population genetic inference. *Nature Reviews Genetics* **25**, 61–78 (2024).

[46] Töpper, J. P. *et al.* Advancing general ecosystem models (GEMs): Towards a mechanistic understanding of the biosphere in the light of the Anthropocene. *Ecological Solutions and Evidence* **6**, e70053 (2025).

[47] Purves, D. *et al.* Time to model all life on Earth. *Nature* **493**, 295–297 (2013).

[48] Harfoot, M. B. J. *et al.* Emergent Global Patterns of Ecosystem Structure and Function from a Mechanistic General Ecosystem Model. *PLOS Biology* **12**, e1001841 (2014).

[49] Bradburd, G. S. & Ralph, P. L. Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics* **50**, 427–449 (2019).

[50] Malmborg, C. A. *et al.* Defining model complexity: An ecological perspective. *Meteorological Applications* **31**, e2202 (2024).

[51] Geary, W. L. *et al.* A guide to ecosystem models and their environmental applications. *Nature Ecology & Evolution* **4**, 1459–1471 (2020).

[52] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* Springer Series in Statistics (Springer, New York, NY, 2009).

[53] Lu, Y. Artificial intelligence: A survey on evolution, models, applications and future trends. *Journal of Management Analytics* **6**, 1–29 (2019).

[54] Gatla, R. K., Gatla, A., Sridhar, P., Kumar, D. G. & Rao, D. S. N. M. *Advancements in Generative AI: Exploring Fundamentals and Evolution*, 1–5 (2024).

[55] Vaswani, A. *et al.* Attention Is All You Need (2023). arXiv:1706.03762.

[56] Kaplan, J. *et al.* Scaling Laws for Neural Language Models (2020). arXiv:2001.08361.

[57] Zong, Y., Aodha, O. M. & Hospedales, T. M. Self-Supervised Multimodal Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**, 5299–5318 (2025).

[58] Huang, Q. *et al.* MuLan: A Joint Embedding of Music Audio and Natural Language (2022). arXiv:2208.12415.

[59] Assran, M. *et al. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture*, 15619–15629 (2023).

[60] Alayrac, J.-B. *et al. Self-Supervised MultiModal Versatile Networks*, Vol. 33, 25–37 (Curran Associates, Inc., 2020).

[61] Berg, T. *et al. Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds*, 2019–2026 (IEEE, Columbus, OH, USA, 2014).

[62] Wei, X.-S. *et al.* Fine-Grained Image Analysis with Deep Learning: A Survey (2021). arXiv:2111.06119.

[63] Gu, J. *et al.* BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning (2025). arXiv:2505.23883.

[64] Stevens, S. *et al.* BioCLIP: A Vision Foundation Model for the Tree of Life (2024). arXiv:2311.18803.

[65] Amini-Naieni, N., Han, T. & Zisserman, A. CountGD: Multi-Modal Open-World Counting (2025). arXiv:2407.04619.

[66] Sastry, S., Khanal, S., Dhakal, A., Ahmad, A. & Jacobs, N. TaxaBind: A Unified Embedding Space for Ecological Applications (2024). arXiv:2411.00683.

[67] van Merriënboer, B. *et al.* Perch 2.0: The bittern lesson for bioacoustics (2025). URL https://arxiv.org/abs/2508.04665. arXiv:2508.04665.

[68] Robinson, D. *et al.* Naturelm-audio: an audio-language foundation model for bioacoustics (2025). URL https://arxiv.org/abs/2411.07186. arXiv:2411.07186.

[69] Gadot, T. *et al.* To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision* **18**, 1193–1208 (2024).

[70] Gardiner, R. J., Rowlands, S. & Simmons, B. I. Towards scalable insect monitoring: Ultra-lightweight CNNs as on-device triggers for insect camera traps. *Methods in Ecology and Evolution* **n/a**.

[71] Norouzzadeh, M. S. *et al.* Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* **115**, E5716–E5725 (2018).

[72] Kahl, S., Wood, C. M., Eibl, M. & Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* **61**, 101236 (2021).

[73] Dussert, G. *et al.* Zero-shot animal behaviour classification with vision-language foundation models. *Methods in Ecology and Evolution* **16**, 1460–1472 (2025).

[74] Maruf, M. *et al.* VLM4Bio: A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images (2024). arXiv:2408.16176.

[75] Singh, J., Ahirwal, S. K., Ramteke, K., Kantharajan, G. & Sarma, K. in *Remote Sensing Techniques for Monitoring Aquatic Ecosystems* (eds Ganie, P. A., Posti, R. & Pandey, P. K.) *Information Technology in Fisheries and Aquaculture* 71–107 (Springer Nature, Singapore, 2025).

[76] Mutanga, O., Masenyama, A. & Sibanda, M. Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects. *ISPRS Journal of Photogrammetry and Remote Sensing*

198, 297–309 (2023).

[77] Beery, S., Cole, E., Parker, J., Perona, P. & Winner, K. *Species distribution modeling for machine learning practitioners: A review*, COMPASS '21, 329–348 (Association for Computing Machinery, New York, NY, USA, 2021). URL https://doi.org/10.1145/3460112.3471966.

[78] Hu, Y., Si-Moussi, S. & Thuiller, W. Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution* **16**, 228–246 (2025).

[79] Joseph, M. B. Neural hierarchical models of ecological populations. *Ecology Letters* **23**, 734–747 (2020).

[80] Kellenberger, B., Winner, K. & Jetz, W. The Performance and Potential of Deep Learning for Predicting Species Distributions (2024).

[81] Haucke, T. *et al.* Deep multi-modal species occupancy modeling. *bioRxiv* (2025). URL https://www.biorxiv.org/content/early/2025/09/11/2025.09.06.674602.

[82] Trantas, A. *et al.* BioAnalyst: A Foundation Model for Biodiversity (2025). arXiv:2507.09080.

[83] Davis, C. L. *et al.* Deep learning with citizen science data enables estimation of species diversity and composition at continental extents. *Ecology* **104**, e4175 (2023). URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.4175.

[84] Pollock, L. J. *et al.* Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity* **1**, 166–182 (2025).

[85] Hoffmann, J. *et al.* Training Compute-Optimal Large Language Models (2022). arXiv:2203.15556.

[86] Ju, J. *et al.* The Harmonized Landsat and Sentinel-2 version 2.0 surface reflectance dataset. *Remote Sensing of Environment* **324**, 114723 (2025).

[87] Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* **24**, 148–159 (2014).

[88] Das, A., Kong, W., Sen, R. & Zhou, Y. A decoder-only foundation model for time-series forecasting (2024). URL https://arxiv.org/abs/2310.10688. arXiv:2310.10688.

[89] Siméoni, O. *et al.* Dinov3 (2025). URL https://arxiv.org/abs/2508.10104. arXiv:2508.10104.

[90] Heidorn, P. B. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends* **57**, 280–299 (2008).

[91] Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* **13**, e1002295 (2015).

[92] Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological Data Should Not Be So Hard to Find and Reuse. *Trends in Ecology & Evolution* **34**, 494–496 (2019).

[93] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41**, 423–443 (2018).

[94] van den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural Discrete Representation Learning (2018). arXiv:1711.00937.

[95] Klemmer, K., Rolf, E., Robinson, C., Mackey, L. & Rußwurm, M. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery (2024). arXiv:2311.17179.

[96] Connecting geology to ecology. *Nature Geoscience* **17**, 173–173 (2024).

[97] At the interface between hydrology and ecology. *Nature Water* **2**, 207–207 (2024).

[98] Geng, X. *et al.* Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204* (2022).

[99] Mizrahi, D. *et al.* 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems* **36**, 58363–58408 (2023).

[100] Boggust, A., Carter, B. & Satyanarayan, A. *Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples*, 746–766 (2022).

[101] Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions (2017). URL https://arxiv.org/abs/1705.07874. arXiv:1705.07874.

[102] Koh, P. W. *et al.* *Concept bottleneck models*, 5338–5348 (PMLR, 2020).

[103] Cowans, A., Lambin, X., Hare, D. & Sutherland, C. Improving the integration of artificial intelligence into existing ecological inference workflows. *Methods in Ecology and Evolution* **n/a**. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14485.

[104] Sastry, S. *et al.* ProM3E: Probabilistic Masked MultiModal Embedding Model for Ecology (2025). arXiv:2511.02946.

[105] Sutherland, W. J. *et al.* Identification of 100 fundamental ecological questions. *Journal of Ecology* **101**, 58–67 (2013). URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.12025.

[106] Chen, Y. *et al.* Open-Insect: Benchmarking Open-Set Recognition of Novel Species in Biodiversity Monitoring (2025). arXiv:2503.01691.

[107] Mehrab, K. S. *et al.* Fish-Vista: A Multi-Purpose Dataset for Understanding & Identification of Traits from Images (2025). arXiv:2407.08027.

[108] Terry, J. C. D. & Lewis, O. T. Finding missing links in interaction networks. *Ecology* **101**, e03047 (2020). URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.3047.

[109] Strydom, T. *et al.* A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B: Biological Sciences* **376**, 20210063 (2021). URL https://doi.org/10.1098/rstb.2021.0063.

[110] Clark, N. J. & Wells, K. Dynamic generalised additive models (dgams) for forecasting discrete ecological time series. *Methods in Ecology and Evolution* **14**, 771–784 (2023). URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13974.

[111] Olsson, F. *et al.* What can we learn from 100,000 freshwater forecasts? a synthesis from the neon ecological forecasting challenge. *Ecological Applications* **35**, e70004 (2025). URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/eap.70004.

[112] Yalavarthi, V. K. *et al. Grafiti: Graphs for forecasting irregularly sampled time series*, Vol. 38, 16255–16263 (2024).

[113] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **54**, 115:1–115:35 (2021).

[114] Lo Piano, S. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications* **7**, 9 (2020).

[115] Chapman, M. *et al.* Biodiversity monitoring for a just planetary future. *Science* **383**, 34–36 (2024).

[116] Ellis-Soto, D., Chapman, M. & Locke, D. H. Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nature Human Behaviour* **7**, 1869–1877 (2023).

[117] Zizka, A. *et al.* Bio-Dem, a tool to explore the relationship between biodiversity data availability and socio-political conditions in time and space. *Journal of Biogeography* **48**, 2715–2726 (2021).

[118] Gaston, K. J. & May, R. M. Taxonomy of taxonomists. *Nature* **356**, 281–282 (1992).

[119] Abdilla, A. *et al.* Indigenous Protocol and Artificial Intelligence Position Paper (2020).

[120] Rahaman, N. *et al.* On the Spectral Bias of Neural Networks (2019). arXiv:1806.08734.

[121] Arvanitidis, G., Hansen, L. K. & Hauberg, S. Latent Space Oddity: On the Curvature of Deep Generative Models (2021). arXiv:1710.11379.

[122] Gawlikowski, J. *et al.* A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* **56**, 1513–1589 (2023).

[123] Lai, Y. *et al.* Exploring Uncertainty in Deep Learning for Construction of Prediction Intervals (2021). arXiv:2104.12953.

[124] Sbailò, L. & Ghiringhelli, L. M. Uncertainty Quantification in Deep Neural Networks through Statistical Inference on Latent Space (2023). arXiv:2305.10840.

[125] Sagawa, S., Raghunathan, A., Koh, P. W. & Liang, P. An investigation of why overparameterization exacerbates spurious correlations. *CoRR* **abs/2005.04345** (2020). URL https://arxiv.org/abs/2005.04345.

[126] Zhang, T., Zhao, C., Chen, G., Jiang, Y. & Chen, F. Feature contamination: Neural networks learn uncorrelated features and fail to generalize (2025). arXiv:2406.03345.

[127] Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR* **abs/1805.12177** (2018). URL http://arxiv.org/abs/1805.12177.

[128] Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T. & Michalak, S. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks (2020). arXiv:1905.11001.

[129] Rahimi, A. & Recht, B. *Random features for large-scale kernel machines*, NIPS'07, 1177–1184 (Curran Associates Inc., Red Hook, NY, USA, 2007).

[130] Cepeda, V. V., Nayak, G. K. & Shah, M. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization (2023). arXiv:2309.16020.

[131] Peng, W., Varanka, T., Mostafa, A., Shi, H. & Zhao, G. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 10023–10044 (2022).

[132] Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J. & Dormann, C. F. Process-Informed Neural Networks: A Hybrid Modelling Approach to Improve Predictive Performance and Inference of Neural Networks in Ecology and Beyond. *Ecology Letters* **27**, e70012 (2024).

[133] Manchingal, S. K. & Cuzzolin, F. Epistemic deep learning (2022). URL https://arxiv.org/abs/2206.07609. arXiv:2206.07609.

[134] Gal, Y. & Ghahramani, Z. Balcan, M. F. & Weinberger, K. Q. (eds) *Dropout as a bayesian approximation: Representing model uncertainty in deep learning.* (eds Balcan, M. F. & Weinberger, K. Q.) *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, 1050–1059 (PMLR, New York, New York, USA, 2016). URL https://proceedings.mlr.press/v48/gal16.html.

[135] Ovadia, Y. *et al.* Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* **32** (2019).

[136] Nasim, I. & De Sousa Almeida, J. L. *Uncertainty-aware prediction of climate extremes using fine-tuned time-series foundation models* (2025). URL https://www.climatechange.ai/papers/neurips2025/1.

[137] Scheffer, M., van Bavel, B., van de Leemput, I. A. & van Nes, E. H. Inequality in nature and society. *Proceedings of the National Academy of Sciences* **114**, 13154–13157 (2017).

[138] Delgado, F., Yang, S., Madaio, M. & Yang, Q. *The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice*, EAAMO '23, 1–23 (Association for Computing Machinery, New York, NY, USA, 2023).

[139] Nations, U. Nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. International treaty (2010). URL https://www.cbd.int/abs/text. Adopted 29 October 2010, entered into force 12 October 2014.

[140] Design justice network principles. https://designjustice.org.

[141] Carroll, S. R. *et al.* The CARE Principles for Indigenous Data Governance. *Data Science Journal* **19**, 43 (2020).

[142] Organisation for Economic Co-operation and Development (OECD). Oecd ai principles: Recommendation on artificial intelligence. OECD Council Recommendation, adopted 22 May 2019 (2019). URL https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Updated in May 2024; see 2024 revision at OECD.AI website.

[143] Siddik, M. A. B., Shehabi, A. & Marston, L. The environmental footprint of data centers in the United States. *Environmental Research Letters* **16**, 064017 (2021).

[144] Lei, N., Lu, J., Shehabi, A. & Masanet, E. The water use of data center workloads: A review and assessment of key determinants. *Resources, Conservation and Recycling* **219**, 108310 (2025).

[145] Programme, U. N. E. *Artificial Intelligence (AI) End-to-End: The Environmental Impact of the Full AI Lifecycle Needs to Be Comprehensively Assessed - Issue Note* (2024).

[146] Measuring the environmental impacts of artificial intelligence compute and applications: The ai footprint. OECD Digital Economy Papers 341, OECD (2022).

[147] Norman, K. E., Boettiger, C., Poisot, T. & Jones, G. M. The role of ai in ecology's computational carbon footprint. *Frontiers in Ecology and the Environment* **24**, e70021 (2026). URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/fee.70021. E70021 70021.

[148] Wu, C.-J. *et al.* Sustainable AI: Environmental Implications, Challenges and Opportunities (2022). arXiv:2111.00364.

[149] Gardiner, R. J. *et al. Bridging domain gaps for fine-grained moth classification through expert-informed adaptation and foundation model priors*, 5110–5115 (2025).