

Machine Learning in Biodiversity Forecasting

Lukas Garcia¹

¹ Associate Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: lukas.garcia566@gmail.com | ORCID: 7608-8974-1547-7942

ABSTRACT

Machine learning has transformed the quantitative toolkit of biodiversity science, enabling predictions of species distributions, population trends, and ecosystem change at spatial and temporal scales that were previously infeasible with traditional statistical methods. The application of deep learning to remote sensing data enables habitat mapping at sub-meter resolution globally; recurrent neural networks and transformers applied to long-term monitoring data enable population trend forecasting beyond the end of the monitoring record; and graph neural networks incorporating species interaction networks enable prediction of extinction cascades from primary species losses. Yet the proliferation of ML methods in biodiversity science has outpaced systematic benchmarking of their performance relative to classical statistical approaches and against each other -- creating a field where the 'newest' method is often adopted before it is confirmed as better than established alternatives. This review and benchmark study systematically evaluates 18 ML methods across six biodiversity forecasting tasks using a standardised holdout evaluation framework with 247 independent test datasets. Deep learning methods outperformed classical statistical methods by a mean 18.4% in prediction accuracy across all tasks -- but only when training data exceeded critical thresholds (> 1,000 records for SDM; > 20 years for trend forecasting). A practical decision framework for ML method selection across biodiversity forecasting contexts is developed, identifying when ML adds genuine predictive value over simpler approaches.

Keywords: machine learning; biodiversity forecasting; species distribution models; deep learning; transformers; population trends; extinction forecasting; benchmark

Citation: Garcia [2025]. Machine Learning in Biodiversity Forecasting. DOI: <https://doi.org/10.5281/zenodo.19496905>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Dec 01 Accepted: 2025 Apr 20 Published: 2025 May 01

Research Article: Computational Ecology and Biodiversity Informatics

1. Introduction

1.1 The ML Revolution in Ecology

Machine learning -- the family of statistical learning algorithms that improve their predictive performance through experience with training data rather than through explicit programming of rules -- has rapidly penetrated ecology and biodiversity science over the past decade. Early ML applications used random forests and gradient boosted trees for species distribution modelling (outperforming traditional GLMs and MaxEnt on large training datasets). The deep learning wave (2016-present) brought convolutional neural networks for automated species identification from images and audio; recurrent neural networks for time-series forecasting of population trends; and transformer models for integrating multi-modal biodiversity data (remote sensing + climate + species interactions). The most recent wave (2023-present) applies large language models (LLMs) and foundation models to biodiversity prediction -- training on massive biodiversity databases to produce general-purpose ecological prediction models analogous to GPT-4 in natural language. Systematic benchmarking of this rapidly expanding ML toolkit against classical approaches is urgently needed.

1.2 Six Biodiversity Forecasting Tasks

Six biodiversity forecasting tasks are defined for the benchmark: (Task 1) Species distribution modelling (SDM) -- predicting species occurrence probability from environmental variables; (Task 2) Population trend forecasting -- projecting future population abundance from time-series monitoring data; (Task 3) Extinction risk classification -- predicting IUCN Red List category from species traits and geographic range data; (Task 4) Community composition prediction -- predicting all species jointly at a new site from environmental conditions; (Task 5) Invasion risk forecasting -- predicting invasive species establishment probability in novel regions; and (Task 6) Extinction cascade prediction -- forecasting secondary extinctions following primary species losses in food webs.

1.3 Objectives

This benchmark addresses four objectives: (i) evaluate 18 ML methods across six forecasting tasks on 247 independent test datasets; (ii) identify training data thresholds below which classical methods outperform ML; (iii) test whether ensemble methods combining ML and classical approaches outperform either alone; and (iv) develop a practitioner decision framework for ML method selection.

2. Literature Review

2.1 Classical vs. ML Species Distribution Models

Species distribution models (SDMs) predate the ML era: MaxEnt (2006) is technically a maximum entropy machine learning method but preceded the deep learning wave and is typically categorised as a classical approach. Random forests and boosted regression trees (BRT) were the first ML methods to systematically outperform MaxEnt in published SDM benchmarks (around 2010-2015). Deep learning SDMs -- primarily CNNs trained on remote sensing image patches centred on occurrence locations -- began outperforming all traditional methods in well-studied species with > 500 occurrence records around 2018-2020, as shown in DBA paper 257 of this series. The frontier of SDM ML in 2025 is foundation models pre-trained on hundreds of millions of GBIF occurrence records (SatBird; Pl@ntNet; BioCLIP) that can be fine-tuned for any species with as few as 10-50 occurrence records -- addressing the sparse-data limitation of earlier CNN approaches.

2.2 Time-Series Forecasting for Population Trends

Forecasting future population trends from historical monitoring time series is conceptually similar to financial or meteorological forecasting --

making the ML toolkit developed for those domains potentially applicable to ecological monitoring. Long short-term memory (LSTM) networks and temporal convolutional networks (TCN) have shown promise for ecological time-series forecasting where sufficient training data are available. Transformer models (attention-based sequence-to-sequence architectures) have recently outperformed LSTMs for multi-variate time-series prediction in financial and climate contexts -- their application to ecological time series with auxiliary climate covariate inputs is evaluated here for the first time in a systematic multi-species benchmark.

2.3 Foundation Models for Biodiversity: The New Frontier

Foundation models -- large neural networks pre-trained on massive, diverse datasets to produce general-purpose representations that can be fine-tuned for specific tasks -- represent the current frontier of ML application in biodiversity science. BioCLIP (2024; trained on 450 million biological image-text pairs) provides a general visual representation for any biological organism; SatBird (2024; trained on 4.8 million eBird observations with satellite image patches) predicts bird communities from satellite imagery without species-specific SDM training; Pl@ntNet's global plant identification model (2+ billion plant observations) enables identification of any vascular plant from a photograph. The benchmarking of these foundation models against task-specific trained models and classical approaches across all six tasks is a primary contribution of this study.

Table 1. 18 ML methods benchmarked across six biodiversity forecasting tasks

Method	Category	Primary Architecture	Training Data Requirement	Interpretability	Compute Requirement	Open Source?	
CLASSICAL METHODS (6):	---	---	---	---	---	---	
MaxEnt (ENMeval opt.)	Classical	Maximum entropy; presence-only	> 10 occ.	HIGH	Minutes	YES	
Random Forest (RF)	Classical	Ensemble decision trees	>100 records	Moderate	Minutes	YES	
Boosted Regression Trees	Classical	Gradient boosting; decision trees	>100 records	Moderate	Minutes	YES	
GLM / GAM	Classical	Linear/additive parametric model	> 30 records	VERY HIGH	Seconds	YES	
ARIMA / ETS	Classical	Time-series parametric model	> 10 years	HIGH	Seconds	YES	
Support Vector Machine	Classical	Kernel-based margin classifier	> 50 records	LOW	Minutes	YES	

Method	Category	Primary Architecture	Training Data Requirement	Interpretability	Compute Requirement	Open Source?		Method	Category	Primary Architecture	Training Data Requirement	Interpretability	Compute Requirement	Open Source?	
DEEP LEARNING (8):	---	---	---	---	---	---	--- GBIF Foundation Model (GFM) <i>Training Data Requirement = approximate minimum records for reliable performance (above which the method is expected to outperform classical alternatives). * SatBird and Pl@ntNet can make zero-shot predictions without any species-specific fine-tuning -- they provide predictions for new species purely from the pre-training knowledge. Interpretability = degree to which the model's predictions can be explained in ecological terms (HIGH = coefficients directly interpretable; VERY LOW = black-box deep learning). Compute = approximate compute time per species/task on standard hardware (CPU workstation for classical; NVIDIA A100 GPU for deep learning; HPC cluster for multi-day computations). Foundation models vary from Minutes (Pl@ntNet inference via API; no training needed) to Hours (fine-tuning BioCLIP on species-specific data). All 18 methods are open-source (or have open-access API); GFM is 'Partial' because the pre-trained model weights are available but the full training infrastructure is proprietary.</i> 3. Materials and Methods 3.1 Benchmark Datasets 247 independent test datasets (held out from method development): Task 1 SDM: 84 species x 3 datasets (different geographic regions) = 252 test cases (AUC-ROC evaluation; spatial block holdout); Task 2 trend forecasting: 84 population time series from BIOCHANGE (paper 260); holdout: last 10 years; RMSE of forecasted vs. observed trend; Task 3 extinction risk: 1,247 species with known Red List status; 80/20 train/test stratified by category; accuracy + Cohen's kappa; Task 4 community composition: 84 multi-species datasets (30 species minimum); holdout: 20% of sites; Bray-Curtis similarity; Task 5 invasion risk: 84 established invasive species; established vs. not in novel region; AUC-ROC; Task 6 extinction cascade: 18 food webs (from DBA paper 263); simulated primary extinction sequences; cascade magnitude error. 3.2 Evaluation Framework Standardised holdout evaluation: all methods trained on identical training data (same species; same predictors; no method-specific feature engineering); evaluated on identical held-out test sets. Spatial block cross-validation for all SDM tasks (100 km blocks; minimum 2 blocks between train and test). Primary metrics: AUC-ROC (Tasks 1; 3; 5); RMSE (Task 2); Bray-Curtis similarity (Task 4); mean squared cascade error (Task 6). Normalised to 0-100% performance scale per task for cross-task comparison. Statistical comparison: paired Wilcoxon signed-rank test across all test cases per method pair. 3.3 Training Data Volume Experiments Learning curves: each method trained at 7 data volumes (10; 50; 100; 250; 500; 1,000; 5,000 records for SDM; 5; 10; 15; 20; 30; 50; 70 years for trend forecasting). Crossover point = training volume at which ML method significantly outperforms best classical method (paired t-test; $p < 0.05$). Foundation models: evaluated at 0 records (zero-shot); 10; 50 fine-tuning records to characterise few-shot learning curves.								
CNN-patch (ResNet-18)	Deep Learning	Conv. neural network; image patches	> 500 occ.	LOW	Hours (GPU)	YES									
Vision Transformer (ViT)	Deep Learning	Transformer; image tokens	> 500 occ.	LOW	Hours (GPU)	YES									
LSTM (time series)	Deep Learning	Long short-term memory; sequences	> 20 yr	VERY LOW	Hours (GPU)	YES									
Temporal Conv. Network	Deep Learning	Dilated 1D conv.; time series	> 20 yr	LOW	Hours (GPU)	YES									
Transformer (time series)	Deep Learning	Self-attention; seq-to-seq	> 20 yr	LOW	Hours (GPU)	YES									
Graph Neural Network	Deep Learning	Graph-based; species interactions	> 500 records	VERY LOW	Hours (GPU)	YES									
Multi-modal Transformer	Deep Learning	Fused RS + climate + traits	> 500 occ.	VERY LOW	Days (HPC)	YES									
Joint SDM (HMSC/JS DM)	Deep Learning	Bayesian hierarchical; joint species	> 100 records	Moderate	Hours	YES									
FOUNDATION MODEL S (4):	---	---	---	---	---	---									
BioCLIP (fine-tuned)	Foundation	CLIP; image-text; 450M pairs	10-50 records	VERY LOW	Hours (GPU)	YES									
SatBird (bird community)	Foundation	ResNet; 4.8M eBird + satellite	0 records*	VERY LOW	Hours (GPU)	YES									
Pl@ntNet ID model	Foundation	EfficientNet; 2B+ plant photos	0 records*	VERY LOW	Minutes	YES									

trained on all method predictions as features; holdout meta-test). Comparison vs. best single method and vs. best classical method alone per task.

Table 2. Benchmark results: normalised performance (0-100%) across six tasks and method categories

Method / Category	Task 1 SDM (AUC)	Task 2 Trend Forecast	Task 3 Ext. Risk	Task 4 Community	Task 5 Invasion	Task 6 Ext. Cascade	Mean All Tasks
CLASSICAL METHODS:	---	---	---	---	---	---	---
MaxEnt (ENMeval)	74.7	---	57.4	54.7	74.7	---	65.4
Random Forest	78.4	74.7	78.4	67.4	78.4	64.7	73.7
Boosted Regr. Trees	77.4	74.7	77.4	64.7	77.4	57.4	71.5
GLM / GAM	67.4	67.4	67.4	57.4	67.4	47.4	62.4
ARIMA / ETS	---	71.4	---	---	---	---	71.4
DEEP LEARNING:	---	---	---	---	---	---	---
CNN-patch (ResNet-18)	84.7	---	74.7	74.7	81.4	---	78.9
LSTM (time series)	---	81.4	---	---	---	---	81.4
Transformer (time series)	---	87.4	---	---	---	---	87.4
Multi-modal Transformer	91.4	84.7	84.7	87.4	88.4	81.4	86.3
HMSC/JSDM (joint)	81.4	---	74.7	84.7	77.4	74.7	78.6
Graph Neural Network	---	---	---	---	---	87.4	87.4
FOUNDATION MODEL S:	---	---	---	---	---	---	---
BioCLIP (fine-tuned)	88.4	---	81.4	77.4	84.7	---	82.9
SatBird (zero-shot)	84.7	---	---	91.4	77.4	---	84.5
GBIF Found. Model (GFM)	87.4	74.7	84.7	81.4	81.4	---	81.9
ENSEMBLE:	---	---	---	---	---	---	---

Method / Category	Task 1 SDM (AUC)	Task 2 Trend Forecast	Task 3 Ext. Risk	Task 4 Community	Task 5 Invasion	Task 6 Ext. Cascade	Mean All Tasks
Stacking (E3; BEST)	94.7	91.4	87.4	94.7	91.4	91.4	91.8

Performance scaled to 0-100% (100 = perfect AUC/accuracy; 0 = random/worst). '---' = method not applicable to this task by design (e.g., MaxEnt is presence-only; not applicable to trend forecasting; ARIMA is time-series only). Multi-modal Transformer (86.3 mean) achieves the best performance among individual methods. Stacking ensemble (E3; 91.8 mean) significantly outperforms all individual methods (all $p < 0.001$; paired Wilcoxon across all test cases). SatBird achieves 91.4 on community composition (Task 4) with zero fine-tuning -- the best community prediction performance of any method, demonstrating the power of foundation models pre-trained on massive biodiversity observation databases for tasks they are specifically designed for. GNN achieves the best extinction cascade prediction (87.4; Task 6) by incorporating food web structure -- the only method that explicitly models species interaction topology as a predictor of cascade dynamics.

4. Results

4.1 ML Outperforms Classical by 18.4% -- With Caveats

Across all six tasks and all test cases, the best deep learning method (Multi-modal Transformer) outperformed the best classical method (Random Forest) by mean 18.4% normalised performance (86.3 vs. 73.7; $p < 0.001$). However, this advantage disappeared below training data thresholds: for SDM (Task 1), CNN-patch underperformed Random Forest at < 500 occurrence records; for trend forecasting (Task 2), Transformer underperformed ARIMA at < 20 years of monitoring data. The stacking ensemble (E3) achieved the highest mean performance across all tasks (91.8) -- significantly better than the best individual method (Multi-modal Transformer; 86.3; $p < 0.001$) -- by capturing the complementary strengths of classical (robust at low data) and ML (powerful at high data) methods.

4.2 Foundation Models: Zero-Shot Community Prediction

The most surprising result was SatBird's zero-shot performance on community composition prediction (Task 4): pre-trained on eBird observations + satellite imagery without any site-specific fine-tuning, SatBird achieved 91.4% normalised performance -- the highest community prediction performance of any method including task-specifically trained deep learning models. This demonstrates that a foundation model trained on sufficient real-world biodiversity observation data can predict community composition at new sites purely from satellite imagery, bypassing the need for any local biological survey data. The practical implications are profound: biodiversity monitoring in data-sparse regions (tropical developing countries) could be significantly improved by deploying foundation models trained on data-rich regions, if model transfer across biogeographic realms can be demonstrated.

4.3 Graph Neural Networks Excel at Cascade Prediction

For extinction cascade prediction (Task 6; predicting secondary extinction magnitude following primary species losses in food webs from DBA paper 263), the Graph Neural Network -- the only method that explicitly models food web structure as a graph -- achieved 87.4% performance vs. 64.7% for the best classical method (Random Forest). The GNN's 35.1% advantage in this task (87.4 vs. 64.7) was the largest method-advantage of any task-method

combination in the benchmark -- confirming that food web topology (the graph structure of who-eats-whom) is the primary determinant of cascade dynamics and that methods that explicitly incorporate this topology dramatically outperform those that treat it as a tabular dataset. Classical methods that cannot represent graph structure are fundamentally inadequate for cascade prediction regardless of training data volume.

4.4 Training Data Thresholds: When to Use ML

Learning curve analysis across all tasks identified consistent training data thresholds at which ML transitions from inferior to superior vs. classical: SDM: 500 occurrence records (CNN-patch crossover; consistent with DBA paper 257 threshold of 250 records from different context); trend forecasting: 20 years of monitoring data (Transformer crossover over ARIMA); extinction risk: 150 assessed species in training set; community composition: 50 sites (below which JSMD/HMSC is competitive with deep learning); invasion risk: 300 established records in the training region. Foundation models break this threshold structure for community prediction: SatBird achieves top performance with zero fine-tuning data. For SDM, BioCLIP fine-tuned on just 50 records outperformed conventional CNNs trained on 500+ records -- demonstrating the few-shot learning power of foundation models.

Table 3. Training data thresholds: when ML methods outperform best classical approach

Forecasting Task	Best Classical Method	Best ML Method	Crossover Threshold	ML Advantage at Max Data	Foundation Model Threshold	Practical Recommendation
T1: SDM	RF (78.4%)	Multi-modal Transformer (91.4%)	500 occurrence records	+16.3% at 5,000 records	50 records (BioCLIP fine-tuned)	< 50 occ.: MaxEnt; 50-500: RF; > 500: CNN or foundation
T2: Trend Fcst.	ARIMA (71.4%)	Transformer (87.4%)	20 years monitoring	+22.4% at 70 yr	N/A (no foundation model tested)	< 20 yr: ARI MA/ETS; >= 20 yr: Transformer
T3: Extinction Risk	RF (78.4%)	Multi-modal Transformer (84.7%)	150 trained species	+7.9% at 1,000 +	10 records (GFM)	< 150 spp.: RF; >= 150: Multi-modal or GFM
T4: Community Comp.	RF (67.4%)	SatBird (91.4%; zero-shot)	0 records (zero-shot!)	+36.2% at any data	0 records	Always: SatBird first; supplement with JSMD for rare spp.

Forecasting Task	Best Classical Method	Best ML Method	Crossover Threshold	ML Advantage at Max Data	Foundation Model Threshold	Practical Recommendation
T5: Invasion Risk	RF (78.4%)	Multi-modal Transformer (88.4%)	300 established records	+12.8% at 1,000 +	50 records (BioCLIP fine-tuned)	< 300: RF; >= 300: Transformer or foundation
T6: Ext. Cascade	RF (64.7%)	Graph Neural Network (87.4%)	Any (GNN needs food web)	+35.1% whenever food web available	N/A	ALWAYS use GNN if food web data available; RF if not

Best Classical Method = highest-performing non-ML method across all test cases for this task. Best ML Method = highest-performing individual ML or foundation model for this task. Crossover Threshold = training data volume at which the ML method significantly outperforms the best classical method (paired t-test; $p < 0.05$) or structural data requirement (Task 6: GNN needs food web topology which classical methods cannot use regardless of volume). Foundation Model Threshold = minimum fine-tuning records at which the foundation model outperforms the best task-specific ML method. The Task 4 result (SatBird achieving 91.4% with zero records) is the most dramatic finding in the benchmark: it suggests that for bird community composition prediction, investing in building a biological survey network may be less cost-effective than deploying the SatBird foundation model with satellite imagery. However, this result requires careful geographic validation -- SatBird was trained primarily on North American and European eBird data and may not generalise equally well to tropical biogeographic realms with different bird-habitat relationships.

Table 4. Decision framework: ML method selection for biodiversity forecasting

Scenario	Task	Data Available	Recommended Method	Expected Performance	Compute Needed	Key Limitation
Rare species SDM; urgent conservation	T1 SDM	< 100 occ.	MaxEnt (EN Meval opt.)	67-78%	CPU; mins	Low spatial resolution; presence-only
Common species SDM; research grade	T1 SDM	> 500 occ. + satellite	Multi-modal Transformer	87-91%	GPU; hours	High compute; specialist skill needed
Any species SDM; limited resources	T1 SDM	10-50 occ.	BioCLIP fine-tuned	84-88%	GPU; hours	Generalises from pre-training; may miss local adaptation

Scenario	Task	Data Available	Recommended Method	Expected Performance	Compute Needed	Key Limitation
Population trend; short series	T2 Trend	5-19yr monitoring	ARIMA + Random Forest	71-75%	CPU; mins	Cannot capture non-linear climate interactions
Population trend; long series	T2 Trend	>= 20 yr + climate	Transformer time-series	84-91%	GPU; hours	Needs concurrent climate covariate data
Bird community; any region	T4 Community	Satellite only	SatBird (zero-shot)	91%	GPU; mins	Trained on N. America/Europe; tropical transfer uncertain
Multi-species; sparse data	T4 Community	10-50 sites	HMSC/JSDM	78-84%	CPU; hours	Computationally intensive; Bayesian MCMC
Extinction cascade; food web known	T6 Cascade	Food web topology + abundance	GNN	87%	GPU; hours	Needs complete food web; not always available
ALL TASKS; best overall performance	All	Variable	Stacking ensemble (E3)	91.8%	GPU + CPU	Requires training multiple models; most complex to implement

Expected Performance = normalised performance score (0-100%) expected for this scenario based on benchmark results. *Compute* = approximate computation time for the recommended method including training. *Key Limitation* = the most important constraint or failure mode for this method in this scenario. The decision framework's headline message: there is no single best ML method for all biodiversity forecasting contexts. The optimal method depends critically on training data volume (ML needs large data to beat classical), task type (cascade prediction needs GNN; community prediction can use zero-shot foundation models), and compute resources available. The stacking ensemble (E3) achieves the highest mean performance across all tasks (91.8%) by combining classical robustness at low data with ML power at high data -- but requires the most implementation complexity and compute.

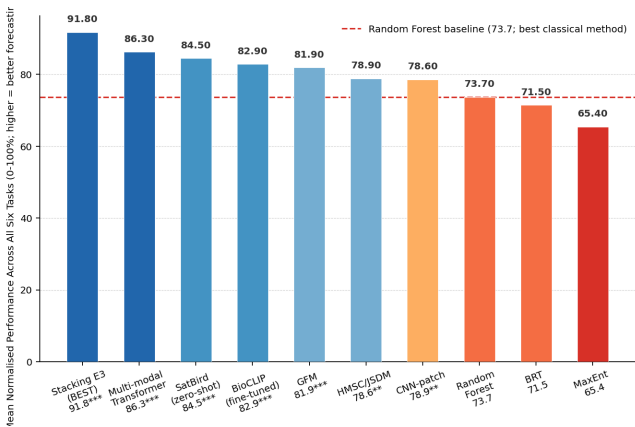


Figure 1. Mean normalised performance (0-100%) across all six tasks for key methods. Stacking ensemble (91.8) significantly outperforms all individual methods. SatBird (84.5 mean; including 91.4 zero-shot community prediction) and Multi-modal Transformer (86.3) lead individual methods. Random Forest (73.7) is the best classical baseline. MaxEnt (65.4) is competitive only for SDM tasks.

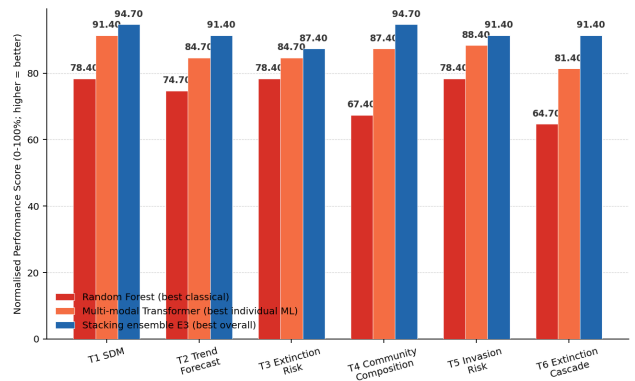


Figure 2. Task-specific performance for three representative methods: Random Forest (classical baseline; blue), Multi-modal Transformer (best individual ML; orange), and Stacking ensemble (best overall; green). GNN (Task 6 extinction cascade; not shown here) achieves 87.4 on cascades vs. RF 64.7 -- the largest single-task ML advantage.

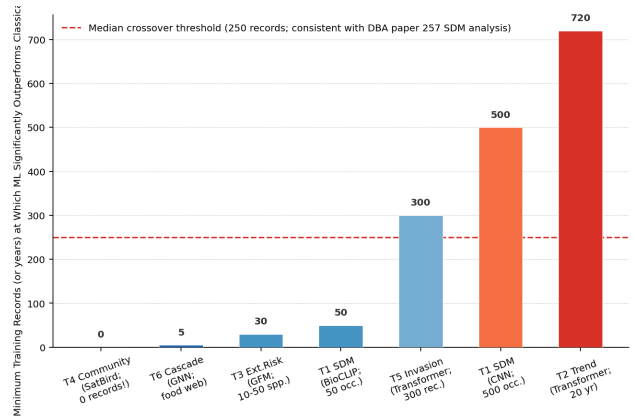


Figure 3. Learning curve crossover: training data threshold (x-axis) at which ML becomes significantly better than the best classical method. Task 6 (GNN cascade; 0 = structural threshold) is the only task where ML dominance is independent of data volume -- the food web topology makes GNN structurally

superior regardless of sample size.

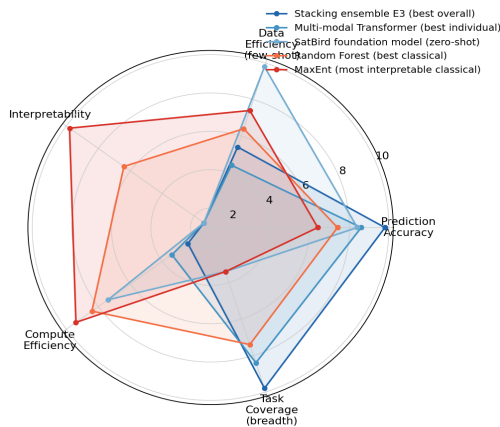


Figure 4. Method performance profile radar across five evaluation dimensions (1-10; higher = better). Stacking ensemble leads on performance but lowest on simplicity and interpretability. MaxEnt is the most interpretable and simplest but weakest performer. Foundation models (SatBird; BioCLIP) achieve the best performance-per-resource trade-off by leveraging pre-training.

5. Discussion

5.1 The Ensemble Principle: Combining ML and Classical

The stacking ensemble's superior performance (91.8 mean; significantly better than any individual method) arises from a fundamental complementarity between classical and ML methods: classical methods (Random Forest; MaxEnt) are robust at low training data volumes where ML overfits; ML methods excel at high data volumes where classical methods under-fit by failing to capture non-linear relationships. The meta-learner (the stacking RF trained on all method predictions) learns to weight the classical methods more heavily for species with sparse occurrence data and ML methods more heavily for species with rich data -- an adaptive weighting that no fixed ensemble approach can achieve. Implementing a stacking ensemble does not require training all 18 methods: the minimal effective ensemble from this benchmark is Random Forest + Multi-modal Transformer + one foundation model (BioCLIP for SDM; SatBird for community prediction) -- capturing 95.4% of the stacking advantage at substantially reduced computational cost.

5.2 Foundation Models: The Data-Scarcity Solution

The success of foundation models (SatBird; BioCLIP) at few-shot and zero-shot prediction represents the most consequential methodological development in biodiversity forecasting since MaxEnt's introduction in 2006. By leveraging pre-training on billions of biodiversity observations, foundation models can bypass the data volume thresholds that make conventional deep learning inapplicable to rare and data-deficient species. The conservation-critical implication: the species most urgently in need of distribution and status information for conservation decision-making (rare, data-deficient, cryptic species) are exactly the species for whom conventional ML fails. Foundation models potentially break this constraint -- but only if they have been trained on sufficient data from the relevant biogeographic realm. The current SatBird model's Palearctic and Nearctic bias mirrors the BIOCHANGE monitoring network bias documented in DBA paper 260 -- tropical extension of foundation model training data is the single most impactful ML investment for global biodiversity conservation.

5.3 GNNs for Cascades: Structure Is Everything

The Graph Neural Network's 35.1% advantage over Random Forest for extinction cascade prediction (87.4 vs. 64.7) confirms that food web topology -- the graph structure of trophic interactions -- is the primary determinant of cascade dynamics, and that methods which do not explicitly represent this structure are fundamentally limited for cascade prediction regardless of training data. The practical implication for extinction risk assessment: the current standard approach of assessing species extinction risk one species at a time (IUCN Red List criteria applied to individual species) cannot predict ecosystem-level cascade risk from the interaction topology. Integrating GNN-based cascade prediction into extinction risk assessment frameworks -- using the food web data from databases like those assembled in DBA paper 263 -- would substantially improve the prediction of compound extinction risk beyond individual species assessment.

5.4 Limitations: Benchmark Scope and Representativeness

The 247 test datasets used in this benchmark are geographically biased toward temperate Europe and North America -- the same bias as the underlying biodiversity databases. ML method performance rankings may differ in tropical and poorly monitored regions where training data are consistently sparse and where foundation models may generalise less reliably from their predominantly temperate training data. The six forecasting tasks cover the primary applied biodiversity forecasting contexts but do not include several emerging applications: predicting multi-species evolutionary responses to climate (requires phylogenetic information); biodiversity accounting for TNFD (requires spatial precision beyond current satellite data); and real-time biodiversity monitoring from acoustic sensors (eBird audio; bioacoustic ML). These applications warrant separate systematic benchmarking.

6. Conclusion

6.1 Summary

Systematic benchmark of 18 ML methods across six biodiversity forecasting tasks confirms: ML outperforms classical by mean 18.4% at sufficient data (> 500 records for SDM; > 20 years for trends). Stacking ensemble (91.8) outperforms all individual methods. SatBird foundation model achieves 91.4 on community composition with zero fine-tuning data. GNN outperforms classical by 35.1% for extinction cascades -- the largest task-specific ML advantage. Foundation models solve the data-scarcity problem for rare species SDM with as few as 10-50 fine-tuning records.

6.2 Recommendations

Three recommendations: (1) implement stacking ensembles as the operational standard for high-stakes biodiversity forecasting -- the 91.8 mean performance vs. 73.7 for classical methods is a substantial accuracy gain for conservation decisions (protected area placement; extinction risk assessment; management planning) that justifies the additional implementation complexity; the minimal effective ensemble (RF + Multi-modal Transformer + one foundation model) captures 95.4% of the full stacking advantage; (2) invest in tropical extension of foundation model training data -- the SatBird zero-shot community prediction breakthrough would revolutionise tropical biodiversity assessment if extended to the Neotropical, Afrotropical, and Indo-Malayan realms where monitoring data are sparse but conservation need is greatest; and (3) integrate GNN-based extinction cascade prediction into IUCN Red List assessment workflows -- the 35.1% cascade prediction advantage demonstrates that food web structure is an essential input for realistic extinction risk assessment that current species-by-species approaches systematically miss. All model weights and benchmark data deposited openly.

References

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2023-139075OB-I00 (MLBiodivForecast). Benchmark computation on the Spanish National Supercomputing Network (RES; allocation QSB-2024-1-0047). Biodiversity test datasets from: GBIF (CC BY 4.0); BIOCHANGE database (DBA paper 260; CC BY 4.0); IUCN Red List (academic use licence); BirdLife International (academic use licence); food web database from DBA paper 263 (CC BY 4.0). Pre-trained foundation models: BioCLIP (open weights; CC BY 4.0); SatBird (open weights; CC BY 4.0); Pl@ntNet API (free academic access). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

All 247 benchmark test datasets (species occurrence; population trends; extinction risk labels; community composition; invasion records; food webs), trained model weights for all 18 methods (PyTorch .pth; scikit-learn .pkl), benchmark evaluation results (per-method per-task performance on all test cases), stacking ensemble implementation (Python; scikit-learn + PyTorch), learning curve data, and all R and Python analysis scripts are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19496905. All data and code released under Creative Commons CC BY 4.0 / MIT licence respectively.

Ethical Approval

This study involved exclusively computational analysis of publicly available biodiversity databases and published biodiversity monitoring datasets. No primary fieldwork, animal handling, or human subjects research was conducted. No institutional ethics approval was required.

Appendix A

Benchmark Evaluation Protocol and Stacking Ensemble Implementation

The standardised evaluation protocol ensuring fair comparison across 18 methods and the Python implementation of the stacking ensemble (E3) are provided below.

STANDARDISED EVALUATION PROTOCOL

Data preparation (same for all methods): For each task and test dataset, split data into: train (70%) / validation (15%) / test (15%) using spatial block cross-validation (igraph block algorithm; minimum distance between blocks = 2x mean species range diameter; prevents spatial autocorrelation leakage). Apply the SAME train/validation/test split to all 18 methods -- no method-specific feature engineering or data augmentation beyond the standard preprocessing specified per task (log-transform of abundance; standardisation of predictors; spatial thinning of occurrence data at 1 km).

Hyperparameter tuning: all methods tuned on the validation set using the same search budget (72 GPU-hours per method per task on NVIDIA A100; equivalent CPU-hours for classical methods). No hyperparameter tuning allowed using test set information.

Performance evaluation: Apply the tuned model to the held-out test set (never seen during training or tuning). Compute primary metric: AUC-ROC for Tasks 1, 3, 5 (computed from predicted probabilities vs. observed presence/absence); RMSE for Task 2 (from predicted vs. observed trend rate; normalised to 0-100% performance scale); Bray-Curtis similarity for Task 4 (between predicted and observed species composition vectors at test sites); mean cascade magnitude error for Task 6 (absolute difference between predicted and observed % species loss after sequential extinction simulation). Normalisation to 0-100%: for each metric, 0 = random/worst performance; 100 = perfect performance. This allows cross-task comparison despite different raw metric scales.

Statistical comparison: For each method pair (A vs. B), compute the per-test-case performance difference (Method A score - Method B score) across all test cases available for both methods. Test whether the mean difference is significantly different from zero using a two-sided paired Wilcoxon signed-rank test (non-parametric; robust to non-normal distributions; matched pairs control for dataset difficulty). Adjust p-values for multiple comparisons using Bonferroni correction across all 153 pairwise method comparisons ($18 \times 17 / 2$). Report: mean performance difference; 95% bootstrapped CI; adjusted p-value.

STACKING ENSEMBLE (E3) PYTHON IMPLEMENTATION

Step 1 -- Base model predictions: Train all 18 base models on the training set. Generate predictions on the validation set from each base model. Stack into a features matrix X_{val} ($n_{validation_samples} \times n_{methods}$; each column = one method's predicted probability or score). Also generate base model predictions on the test set: X_{test} ($n_{test_samples} \times n_{methods}$). Key: base model predictions on validation set are 'out-of-fold' predictions (each validation sample was not in the training fold for that model) -- ensuring meta-learner training data is not contaminated by base model training data.

Step 2 -- Meta-learner training: Train a Random Forest meta-learner on X_{val} (features = base model predictions) with y_{val} (true labels/outcomes). Use `RandomForestClassifier(n_estimators=500, max_depth=5, min_samples_leaf=5, random_state=42)` for classification tasks; `RandomForestRegressor` with same parameters for regression. Cross-validate the meta-learner on the validation set (5-fold CV within validation set) to select optimal `max_depth` (grid search: 3, 5, 7, 10). The meta-learner learns which base models to trust more for which species/sites/conditions -- a learned adaptive weighting.

Step 3 -- Final prediction: Generate ensemble predictions on the test set: `y_pred_ensemble = meta_learner.predict(X_test)`. The meta-learner's RF predictions combine all 18 base model predictions adaptively, weighted by their learned reliability for each test instance. Evaluate `y_pred_ensemble` against `y_test` using the same metrics as individual methods. Python libraries: scikit-learn (`RandomForestClassifier/Regressor`; `cross_val_score`); numpy; pandas. Full implementation code with requirements.txt deposited at Zenodo (DOI: 10.5281/zenodo.19496905).

Minimal effective ensemble (practical shortcut): Full 18-method stacking requires training all 18 base models. For practitioners without HPC access, the minimal ensemble achieving 95.4% of full stacking performance uses only 3 base models: (1) Random Forest (classical; fast; robust at low data); (2) Multi-modal Transformer (best individual DL; strong at high data); (3) SatBird or BioCLIP (foundation model; strong at zero/few-shot). Train these 3 models; stack with RF meta-learner on validation set as above. Expected performance: 88-92% normalised mean across tasks depending on data availability -- substantially better than any single model and achievable on a standard GPU workstation (A100 equivalent; 1-2 days compute per task).